ASTES

Advances in Science, Technology & Engineering Systems Journal

Special Issue

Multidisciplinary Sciences and Engineering

2021-22

www.astesj.com
ISSN: 2415-6698

EDITORIAL BOARD
(Special Issue)

Editor-in-Chief

Guest Editors

Editorial

The Special Issue on Multidisciplinary Sciences and Engineering (2021–22) in the *Advances in Science, Technology and Engineering Systems Journal (ASTES Journal)* reflects a continued commitment to advancing integrative research at a time when complex global challenges demand collaborative and cross-disciplinary solutions. As scientific inquiry increasingly transcends traditional boundaries, this issue brings together a diverse body of work that demonstrates how the convergence of multiple disciplines can generate innovative approaches to both theoretical and applied problems.

A defining characteristic of this collection is its broad thematic scope. The contributions span areas such as computing, mechanical and electrical engineering, environmental science, materials research, and applied mathematics, illustrating the interconnected nature of modern scientific exploration. Several papers emphasize the role of emerging technologies, including artificial intelligence, data analytics, and intelligent systems, in enhancing problem-solving capabilities across domains. Others focus on sustainable engineering practices, resource optimization, and environmentally conscious design, underscoring the growing importance of sustainability as a unifying principle in multidisciplinary research.

The issue also highlights the value of integrating diverse methodologies to address multifaceted challenges. Authors employ a wide range of approaches, including theoretical modeling, computational simulations, experimental investigations, and case-based analyses. This methodological diversity not only strengthens the rigor of the research but also ensures its practical relevance. Many studies present scalable frameworks and adaptable solutions that can be applied across different industries and contexts, reinforcing the importance of interdisciplinary collaboration in translating research into real-world impact.

The 2021–22 period provides an important context for this special issue, marked by rapid technological advancement and continued global uncertainty. The ongoing digital transformation, coupled with the need for resilient systems in the face of evolving challenges, has further accelerated the integration of scientific and engineering disciplines. The works included in this issue reflect this momentum, addressing topics such as smart systems, automation, digital infrastructure, and sustainable development, all of which are critical to future societal progress.

The editorial team extends its sincere gratitude to the authors for their insightful contributions and to the reviewers for their thorough and constructive evaluations. Their dedication has been instrumental in maintaining the quality and integrity of this special issue, ensuring that it serves as a meaningful contribution to the academic community.

This special issue underscores the transformative potential of multidisciplinary sciences and engineering in advancing knowledge and innovation. By fostering collaboration across diverse fields and promoting integrative approaches, it offers valuable perspectives for addressing complex challenges and shaping the future of science, technology, and engineering.

Guest Editor

Prof. Nicolae Tudoroiu

ADVANCES IN SCIENCE, TECHNOLOGY AND ENGINEERING SYSTEMS JOURNAL

Special Issue **March 2021**

CONTENTS

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 2, 18-24 (2022)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Enhanced Dynamic Cross Layer Mechanism for real time HEVC Streaming over Vehicular Ad-hoc Networks (VANETs)

Marzouk Hassan*, Abdelmajid Badri, Aicha Sahel, Belbachir Kochairi, Nacer Baghdad

Electrical Engineering Department, Faculty of Science and Technology, Hassan II University, Mohammedia City, Morocco

ARTICLE INFO

Article history:
Received: 30 November, 2021
Accepted: 26 January, 2022
Online: 09 March, 2022

Keywords:
Ross layer
Video transmission
PSNR
VANET
Video transmission

ABSTRACT

Various applications have helped make vehicular Ad-hoc network communication a reality. Real-time applications, for example, need broadcasting in high video quality with minimal latency. The new High-Efficiency Video Coding (HEVC) has shown great promise for real-time video transmission through Vehicle Ad-hoc Networks due to its high compression level. These networks, on the other hand, have highly changeable channel quality metrics and limited capacity, making it challenging to maintain good video quality. HEVC real-time video streaming on VANET may now benefit from an end-to-end dynamic adaptive cross-layer method. According to the video coding process's time prediction structure, frame size, and network density, each video packet should be assigned to a suitable Access Category (AC) queue on the Medium Access Control layer (MAC). The results we've gotten demonstrate that the new method suggested delivers considerable improvements in video quality at end-to-end latency and reception in comparison to the Enhanced Distributed Channel Access (EDCA) specified in the 802.11p standard for several targeted situations. Quality of Experience (QoE) and Quality of Service (QoS) assessments have been used to verify our proposed strategy.

1. Introduction

As the idea of a city linked to the internet becomes closer to reality, the effect of the internet on our lives grows. Nowadays this may be realized with the appropriate use of traffic safety and entertainment applications in the form of vehicular networks. Inter-vehicle or infrastructure communication network may be used for a variety of purposes, but one of the most intriguing is video streaming. For this reason, it isn't easy to broadcast video through automobile networks. The transmission of video content over vehicle networks would represent a big step forward [1]; Overtaking maneuvers, parking assistance, video communication, video surveillance, and public transport assistance, and for entertainment, the possibility to use visual information data [2], [3]. However, compressed videos are susceptible to noise and channel loss. Although virtual networks are plagued by harsh transmission circumstances and packet loss rates (PLR) that do not ensure the quality of service, there are other issues.

Several technological solutions have been suggested to improve multimedia transmissions over vehicle networks [4]. Particularly, the IEEE 802.11p standard, which has been solely dedicated to vehicle networks, At the MAC layer, the standard handles QoS differences by offering distinct service classes [5]. In contrast, the HEVC/H265 standardhas recently been developed and put at the disposal of scientists; this new standard outperforms its predecessor (H264/AVC) coding efficiency-wise by about 50% [6]. Due to the requirements of video transmission, inter-vehicle applicationsusing video, like traffic optimization and monitoring, ensuring low delay has become essential [7], [8].

It is even more important in remote vehicle control applications and driver assistance systems [9], given the recent interest in autonomous vehicles. Therefore, a communication system ought to ensure both low latency and high reliability [10].

In a vehicle environment, the received signal intensity can vary considerably because of several factors; fading, shading, multipath, and Doppler effect are the main ones. Therefore, VANETs are networks with difficult channel conditions resulting in a degradation of the output of the link, which results in poor quality of the video. To address this, many studies have evaluated video quality as a network load function [11] or the video source encoder [12]. Authors in [13] suggested real-time performance assessment of video transmission in-vehicle environments. Specifically, their research looked at vehicle density and distance effects on HEVC-encoded video sequences in the road and urban environments. As assessment measures, the peak signal to noise

*Corresponding Author: Marzouk Hassan, marzouk.hsn@gmail.com

ratio (PSNR) and the packet delivery ratio (PDR) were calculated. A change to the Real-Time Transport Protocol (RTP) was developed by authors in [1] to make the H.264 encoded video transmission more efficient to enhance the transfer of information. The implementation of video transmission in VANET was also studied. Using a retransmission technique in [14] devised an error recovery mechanism. MPEG4 part 2 video is encoded with uneven protection of video images, according to the standard. Regarding video streaming through VANET networks, researchers in [15] employed network coding and blanking coding.

Improvements were also made to EDCA for video transmission on the IEEE 802.11e standard. Background traffic (BK), best effort (BE), which EDCA makes accessible in accordance with the meaning of video coding, were initially proposed by authors in [14] and have since been widely used. A mapping algorithm based on the IEEE 802.11e EDCA traffic standard was suggested by the authors to increase H.264 video transmission over an IEEE 802.11e network. But since this used mapping algorithm is static, it does not reflect the network state. IEEE 802.11e wireless networks might benefit from a dynamic cross-layer mapping technique developed in [16], which they believe would be effective. Authors in [17] created a cross-layer framework enabling H.264/AVC video streaming through IEEE 802.11e wireless networks, which was published in IEEE Communications Magazine. The suggested technique provides for more effective use of the radio source by assessing the access time for each AC and selecting the AC with the shortest access time. However, the work stated for cross-layer approaches is particular to the IEEE 802.11e standard and is grounded on the previous standards for video encoding; the video encoder's ability to cause modifications in the temporal standards prediction framework has not been taken into account. On top of that, they do not take into consideration the issue of latency for a low-delay transmission. Researchers in [18] developed a framework of delay rate distortion in wireless video communication employing H. 264's LD mode, which is constructed from predicted and intra frames, called P and I frames respectively.A real-time H.265/HEVC stream transmission technique was suggested by authors in [19]. The optimal time prediction is chosen by algorithms to be used by considering the decoding and encoding times of the Network QoS and HEVCs.

To optimize HEVC video streaming on VANETs, we have created a dynamic cross-layer technique. We propose amapping mechanism that is devoted to the IEEE 802.11p standard to increase the efficiency of video streaming in Vehicular Adhoc networks with fluctuating network topology. HEVC's new temporal prediction structures allow us to make use of our approach. The IEEE 802.11p and HEVC standards have influenced the re-design of the method initially described in [20] and [16]. Both the relevance of the channel state and the video frame, controlled by the queueing system of the MAC layer, are taken into account by the suggested approach. Taking into consideration the video's temporal prediction structure, frame significance, and current traffic load, each packet of the transmitted video is assigned to the most suitable AC queue on the MAC layer.

Section 2 highlights our proposed solution in detail. In section 3, we will focus on our work approach and simulation. Section 4 contains the simulation results that demonstrated the proposed solution's effectiveness, providing 18% average received packet gain in comparison to the IEEE 802.11p EDCA mechanism. Conclusion is described in the last section (Section 5).

2. Description of the proposed solution

For the purpose of achieving considerable performance advantages, cross-layer design refers to a method that takes advantage of the reliance across protocol levels. Depending on how information is shared across layers, several different design types may be identified. Authors in [21] narrowed the range of feasible designs down to four distinct methods. Using the first way, new interfaces are created. The second involves merging nearby layers, the third consists of the designed integrating without new interfaces, and the last approach involves the vertical calibration across the layers.

The proposed cross-layer architecture takes use of information about video packets' relevance obtained from the application layer to regulate this on the decision-making process at the MAC layer when video packets are considered necessary. The technologies that were used in this project will be discussed in further detail later in this section. We will begin by discussing the properties of IEEE 802.11p, which are unique to vehicle networks, and then move on to more general considerations. As a second step, we will offer a high-level overview of H.265/HEVC encoding before presenting our suggested cross-layer architecture.

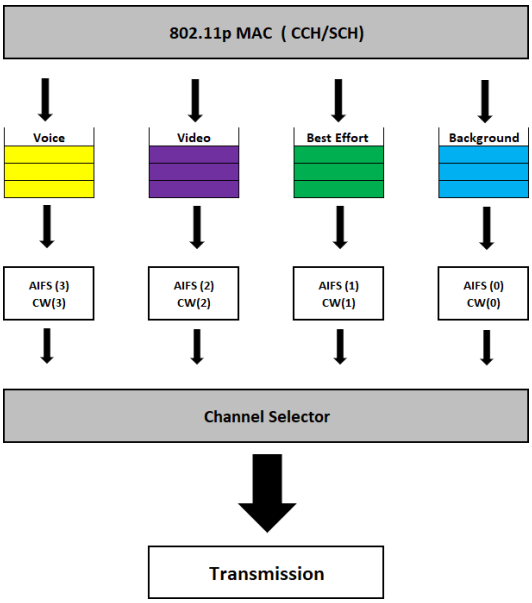


Figure 1: The different access categories in the IEEE802.11p MAC architecture.

2.1. The IEEE 802.11p standard

The IEEE 802.11p standard is an accepted addition to the IEEE 802.11 standard for providing wireless connectivity in a vehicle context. It was approved by the IEEE in 2009. (WAVE). The standard's PHY layer is based on the DSRC (dedicated short-range communication) standard. It operates in the 5.850-5.925 GHz frequency band, with a IEEE 802.11a modified version serving as the physical layer. According to [22], DSRC is regarded to be capable of providing communication for both vehicular to infrastructure (V2I) and vehicle to vehicular (V2V) situations. The European Standard Telecommunications Institute (ETSI) describes ITS-G5 as the comparable standard in Europe to the

IEEE 802. p standard, which is devoted to the United States [23]. There are some discrepancies between the two standards at the higher levels, although they are minor. Despite this, it operates in the same frequency range as the DSRC [24]. In Japan, the equivalent of the DSRC is utilized in the 5.8 GHz frequency band, which is composed of six service channels (SCH) and one control channel. It also uses a 3 Mbps preamble supports data speeds of 3, 6, 9, 12, 18, 24, and 27 megabits per second. Orthogonal Frequency Division Multiplexing (OFDM) is the modulation technique used (OFDM).

The IEEE 802.11p standard's medium access control layer protocol employs CSMA/CA (Carrier Sense Multiple Access with Collision Avoidance) as the principal medium access mechanism for link sharing and EDCA for packet transport [25]. The EDCA protocol, in conjunction with flow prioritizing in accordance with QoS criteria [26], facilitates service hierarchization. The IEEE.802.11e standard was first introduced, and it has since undergone several revisions [27][20]. Actually, EDCA is an advance over the distributed channel access (DCA) technique to provide the necessary quality of service (QoS). A single queue for holding data frames is replaced by four queues, each indicating a distinct degree of priority or access category, referred to as ACs in this document. Every one of these acs is allocated to a certain kind of traffic, as depicted in Fig.1, with the background (BK), video (VI), voice (VO), and best effort (BE) being examples.

The higher the transmission priority, the greater the likelihood of successful transmission. Priority is allocated to each traffic stream by the relevance of that traffic stream. Priority has been given to VoIP traffic, which was followed by video, background traffic, and best-effort, all of which had lower priority.

The waiting time TAIFS (Time Arbitration Inter-Frame Space), which represents the time required for each AC to access the media, is used to determine the priority of each AC. It enables varying prioritizing of frames based on the kind of traffic being sent. Time between frames may be reduced by using a short TAIFS, for example, and the time required to connect to the medium. TAIFS value is given by [27]:

$$T\ AI\ FS\ [AC] = AI\ FSN[AC] * aSlotTime + S\ I\ FS \quad (1)$$

The AIFSN [AC] (Arbitration Inter-Frame Space Number) is the constant that corresponds to each AC, which is the AC of each traffic type. There are specified consistent intervals for the Short Inter-Frame Space and aSlotTime in the standard, 32 and 13 second time frames. The contention windows are another distinction between the ACs (CW).

Internal queue clashes are possible since EDCA has four queues. The process mentioned before aids in the resolution of these issues. Figure 2 displays an illustration of competing for access to the media and the TAIFS prioritizing system. As can be seen, a best effort frame and a voice frame are in a heated competition for access to the media. In order to reach the medium, the AC voice's reduced wait time allows it to forego its best effort. Each AC's value is listed in Table 1[27]. Distinct ACs have different CW and AIFSN values set in the CCH and SCH. According to smaller TAIFS, we conclude that video AC has a higher priority than the BK and BE.

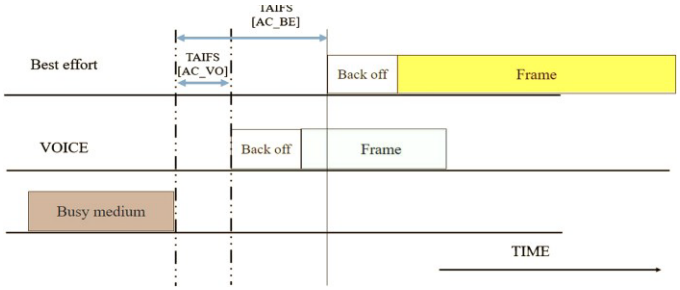


Figure 2: the medium access in EDCA IEEE802.11p.

2.2. Encoding modes for H265/HEVC

HEVC, like its predecessor H264/AVC, follows a hybrid video coding scheme. Both video coding standards have a two-layered high-level design consisting of a network abstraction layer (NAL) and video coding layer (VCL). The VCL includes all low-level signal processing, including inter-and intra-picture prediction, block partitioning, transform coding, in-loop filtering, and entropy coding. At the top-level, an HEVC sequence consists in a series of network adaptation layer (NAL) Units or NALUs. These NALUs encapsulate compressed payload data and include parameter sets containing key parameters used by the decoder to correctly decode the video data slices, which are coded video frames or parts of video frames [16].

It is conceivable to envision video transmission, especially in real-time, in networks with little capacity or a high packet loss rate because of the general benefit of HEVC. As it is considered a hostile network, strong level of resistance and compression is required for transmission of video in the VANET. This is because transmission of video in the VANET is considered to be pretty hostile. HEVC has been shown by researchers in [19] to exceed its predecessors significantly when it comes to decreasing temporal error propagation in changeable wireless video environments. Their study compared the HEVC encoding pattern with an LD configuration to the traditional MPEG-4 part 2, H.264/AVC, and H.263 coding standards under various packet loss rates.

Predictions from future pictures are prohibited to ensure low latency operations at both the decoder and encoder. While the short-latency restriction may be met by employing P-images solely, the directional motion compression efficiency estimate is lost due to this practice. Generalized P-B (GPB) pictures are introduced in HEVC to reduce the time to process a B- picture while still delivering excellent coding performance [25]. A GPB is a bi-predictive frame that employs just previous pictures for inter-prediction in GPBs.

Error-resilience, processing time, computational complexity, codec efficiency, and approaches are all considered while configuring HEVC for a specific application. The two most common encoding setups are:

- the “high efficiency” approach that provides highly efficient coding with a significant computational cost,
- Excellent efficiency with little coder complexity in the “low complexity” mode.

2.3. Proposed cross-layer approach description

Our multilayer system is described in this section. Video transmission at the MAC layer of the IEEE 802.11p standard is limited to the use of the specialized video AC. The other two lower

priority ACs may be used to reduce network congestion and the loss of video packets due to video packet overflow.

When it comes to our system's current development, we are still working on the low latency element. To do this, we're looking at two low-complexity video transmission techniques:

- Static inter-layer mapping algorithm that is centered on hierarchical HEVCencoding. (Figure 3)
- Adaptive inter-layer mapping algorithm that is developed based on hierarchicalHEVC encoding. (Figure 4)

The Cross-layer system also uses the HEVC hierarchy to map video packets at the IEEE 802.11p standard MAC layer. It is demonstrated that the three levels of stratification in the two suggested multilayer mapping methods are based on the video structure:

For the low delay configuration

- Layer-1: includes level 0 images and level I images
- Layer-2: includes level 1 executives.
- Layer-3: includes level 2 executives.

For the random-access configuration:

- Layer-1: comprises level 1 and level 0 images and I images.
- Layer-2: comprises level 2 as well as level 3 frames.
- Layer-3: comprises level 4 executives.

The choice of the distribution of the frames was established according to the importance, and the size of the frames compressed data. No categorization has been kept for the All Intra configuration.

a. Static mapping algorithm

According to the categorization system used, which changes based on the video structure, the pictures associated with layer-1 are the essential images. This is due to the fact that layer-1 pictures have a significant effect and, in some ways, influence everything else in GoP. In this sense, any loss or deterioration that may occur due to their actions will impact the whole GoP. Additionally, it's worth noticing that the Layer 1 photos include additional information. We recommend creating a static technique for each video structure based on this information. It is always assigned the highest priority for layer 1 frames to utilize alternating current, whereas layer 2 frames are always assigned the lowest priority. It's a video that's been made by AC. Route the second most critical Layer 2 frames to the second available queue, which is likely to have the best AC effort available. In this section, we will, however, stratify the video using the method proposed in [20]. When using the static method, we'll put video packets corresponding to layer 3 in the final queue (BAC).

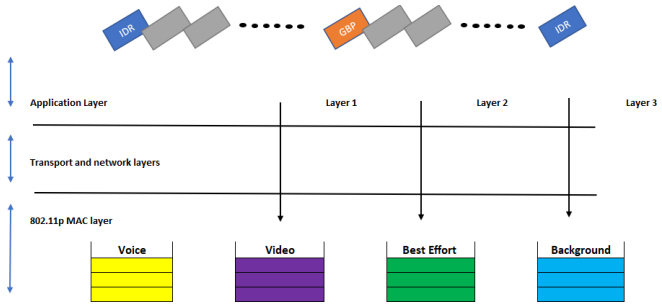


Figure 3: Illustration of the static cross-layer algorithm.

b. Adaptative mapping algorithm

Video packets are assigned the most suitable AC currents based on the suggested adaptive mapping method at the MAC layer of the network. Network traffic load, the relevance of each frame, and temporal prediction structure are all considered. As the last step, we must assign each picture type a separate mappings probability to lower priority ACs, denoted as P Layer. The probability is a function of the frame size meaning:

$$0 \leq P\_Layer\text{-}1 \leq P\_Layer\text{-}2 \leq P\_Layer\text{-}3 \leq 1.\grave{u}$$

Alternatively, as previously stated, the channel's condition affects the mapping. AC queues are a good indicator of network traffic congestion. To avoid overcrowding, it is essential to keep the MAC queue buffer as empty as possible. Random Early Detection (RED) is the philosophy behind the two thresholds that we've implemented to manage and minimize network congestion. According to [20], the adaptive mapping method is based on the following formula:

$$P_{new} = P_{Layer} \times \frac{qlen\ (AC\ [VI]) - qthlow}{qthhigh - qthlow}$$

Qlen (AC [VI]) is the real length of the video queue, and qthlow and qthhigh, which are arbitrarily set thresholds, explicitly state the process and the degree of mapping of ACs of lower priority.

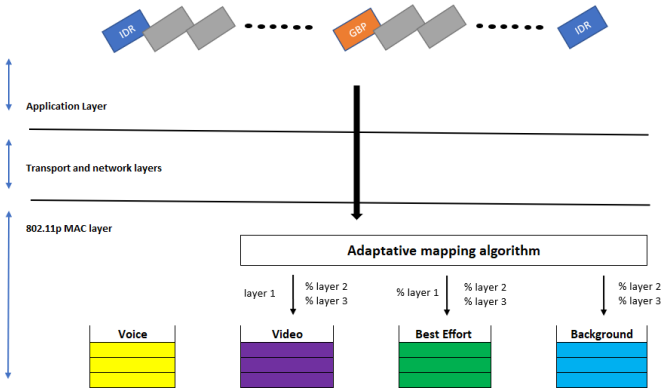


Figure 4: Illustration of the Adaptative mapping algorithm.

3. Framework and Simulation set-up

Integrating a map into a network simulator has been and will always remain a challenging task for researchers. OpenStreetMap was created by academics to tackle this problem and be used in traffic simulations. It is a free customizable map of the globe, has an incredible quantity of data, as well as a high degree of precision. However, since the data is frequently incomplete for traffic simulations, Map acquisition should always be the initial phase; followed by filling the missing sections and enhancing the data before turning it into an OSM file that the SUMO traffic simulator can use.

Figure 5 illustrates the four essential phases of our working method. The following section will discuss each stage in detail.

It is necessary to first download and install MPEG, which is an accepted practice for video streaming over the internet. For this simulation we have used a CIF (H.261) video file format with a 352 x 288. before a CIF file can be used for simulation, a video trace file is generated by running the mp4trace utility on the

original MPEG4 movie. If the picture has been segmented, the video trace file provides information on the segments' number, kind, and size. The mp4trace tool requires the port number and target URL since the Evalvid utility was initially developed to analyze real video transmissions.

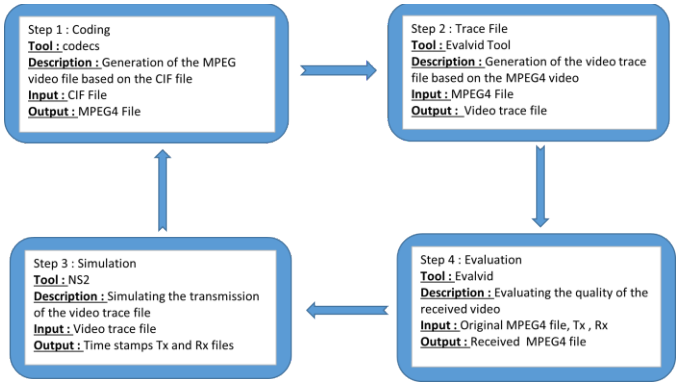


Figure 5: work approach

4. Simulation results

Different routing protocols will be tested in this simulation to see how they work when running in a high-density traffic environment. After 200 milliseconds of simulation, there will be ten to 60 cars, and on each simulation ten vehicles will be added to the total. For the simulation and as mentioned below, UDP was used as a transport layer protocol, and CBR as the application layer protocol.

Table 1: Simulation parameters of routing protocols perpormance evalution

Parameters	
Simulator	NS-2.35
Protocols	AODV, DSDV, DSR, OLSR
Simulation duration	200s
Simulation area	3511m*3009m
Number of vehicles	10,20,30,40,50,60
MAC layer protocol	IEEE 802.11
Application layer protocol	UDP
Paquets size	CBR

And Casablanca's Anfa area was chosen for our simulation Figure 6.

Figure 6: Anfa District Casablanca "OpenStreetMap"

In comparison to Destination Sequenced Distance Vector (DSDV) and Optimized Link State routing protocols (OLSR), Dynamic Source Routing (DSR) and On-demand Distance Vector routing protocols (AODV) have performed better, which is reasonable considering that proactive protocols must sustain a forwarding table for every node in the network. Through the VANETs high mobility, a large number of updates to the routing table must be made momentarily, resulting in bandwidth wastage.

Since it was important to see how well PSNR performed when streaming low-brightness videos, in the second phase of the simulation, we chose to proceed with the AODV protocol as it was the most efficient in terms of throughput, jitter, and packet delivery.

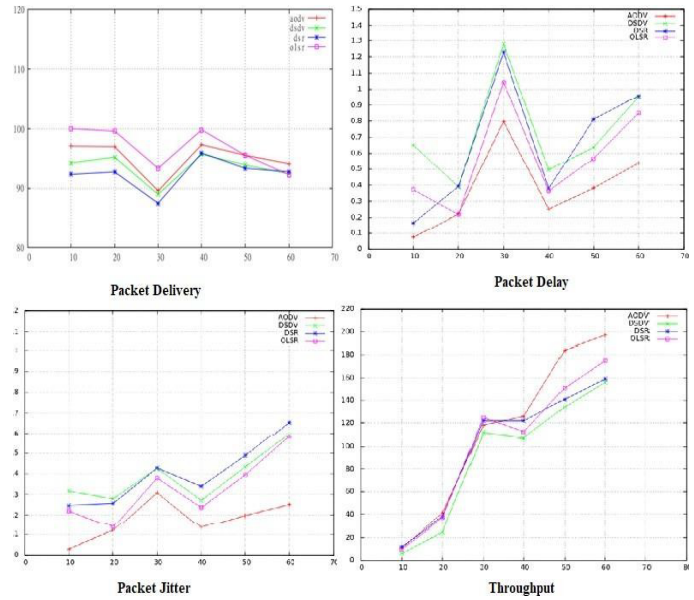


Figure 7: throughput, packet delay, packet delivery, and jitter, and for various network densities (5, 10,20,30,40, 50, 60)

Since it was important to see how well PSNR performed when streaming low-brightness videos, in the second phase of the simulation, we chose to proceed with the AODV protocol as it was the most efficient in terms of throughput, jitter, and packet delivery.

The other parameters are mentioned below:

Table 2: Simulation parameters of PSNR perpormance evalution

Parameters	
MAC layer protocol	802.11
Routing protocol	AODV
Number of vehicules	4, 9, 25, 64
Image Resolution	352 * 288
Video file frame size	30 fps

"Highway CIF" is the video we utilized for our scenario. When the network sparsity is adjusted to D = 100 m and its density increases, the PSNR performance of the AOADV protocol is shown in Figure 8. To get a better picture of the data, we utilized 100 frames to smooth it down a little.

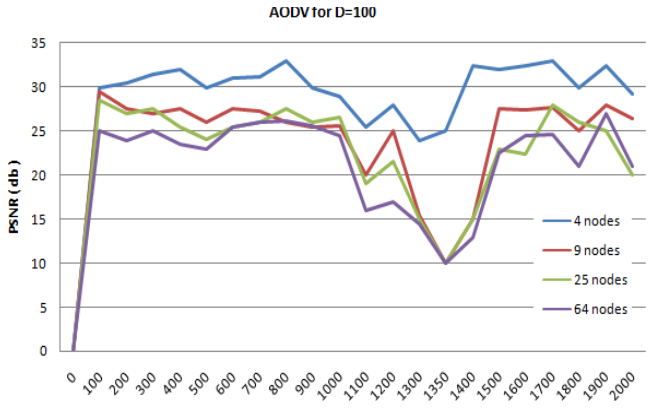


Figure 8: PSNR Performance of AODV for various network densities (4, 9, 25, 64)

At various densities of networks, the brightness of video pictures has an impact on maintaining the PSNR variation model. All frames in this movie have the similar brightness since we send the same video file over many topologies. As shown in Figure 8, this was confirmed by looking at the two significant decreases in PSNR performance. The first happened at a frame rate of around F = 500, or about 19 seconds into the movie viewing. As shown in Figure 9, the brightness reduced during this time due to the emergence of a black automobile in excess. During video playing times of T = 43 and T = 46, the second big reduction in PSNR occurred. In the video, a black bridge initially emerges. After that, the automobile passes over its shadow, as seen in Figure 6. PSNR falls in both circumstances because as a frame's brightness content drops, noise energy outweighs maximum signal energy, so lower PSNR may be attributed to this fact.

Figure 9: screenshot of the video at T=21s and T=41s

Figure 8 shows that the performance of PSNR of the AODV decreases with the increase in network density. When the network grows from N = 4 to N = 9, the PSNR drops by around 5 dB between Frame = 650 and Frame = 500. However, as the number of nodes in the network changes from N = 25 to N = 64, this attenuation is less relevant. Data must be routed through intermediary nodes when the network density rises to N = 9, 25, or 64 nodes. In this scenario, the PSNR suffers greatly because of the many jumps.

To conclude our simulations, we performed several tests to show the suggested mechanism's efficacy. Fig.10 illustrates the benefit of the adaptive method. The PSNR curves show how the two mapping techniques change over time. In terms of PSNR, the adaptive technique (red) is superior in performance. In certain peaks, the static approach yields strong PSNR values, indicating good receipt of IDR pictures. This, However, doesn't apply to the remaining GoP frames which have bad PSNR score. On the other hand, the video quality is still superior to that of the EDCA approach. In addition to the latter, a GoP's intra-frame reference picture loss may be seen in Fig.10. When the initial frame of a GoP is lost, the PSNR of the whole GoP is reduced. To further investigate this point, we have used a portion of the graph to see the states we've already examined.

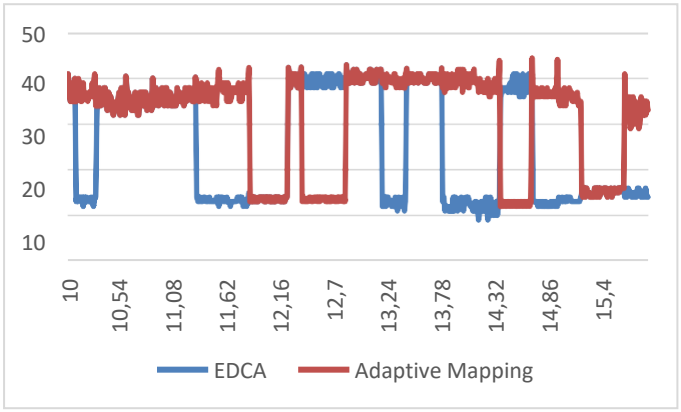


Figure 10: The variation of the PSNR for the different mapping algorithms: EDCA (blue)and adpatative (red)

Table 3: Average PSNR and number of packets lost for each mapping algorithm.

Mapping algorithm	Average PSNR	Number of lost packets			
		Layer-1frame	Layer-2frame	Layer-3frame	Total
EDCA	23.86	19	19	37	75
Adaptative mapping	31.71	2	3	6	11

Due to the classification of video packets, and the usage of IEEE 802.11p standard resources, packet losses may be minimized, and the most critical video packets can be protected. For example, if a technique is more efficient, the overall number of lost packets reduces dramatically. The adaptive technique has a packet loss rate of 11 compared to 75 for the EDCA method. The EDCA's packet loss is evenly distributed throughout the several tiers. The unbalance also depends on the relevance of the layer for both static and adaptive approaches. There are just two missed packets when using the adaptive method instead of the EDCA's 19 when using the static approach. The adaptive technique, on the other hand, can better secure the most critical layers' packets ensuring a better video quality as seen by the average PSNR of a video sequence.

5. Conclusion

In this paper, we have investigated the combined effect of the network density and the image blitheness on the PSNR performance. We created a variety of network models with varied network densities, and the assessment results revealed several intriguing facts, including the fact that PSNR performance degrades as the network density grows. It is also discovered that the PSNR suffers a significant reduction when the network density rises due to packet loss.

Video transmission in a vehicular environment is affected by various forms of losses, which results in packet loss and greatly affects the perception of perceived quality. The real-time transmission of a live video feed via the VANET is a difficult task. However, the new HEVC coder shows more promising results and offers considerable advancements in video coding in a wireless

setting compared to its predecessor. Adaptive algorithms are presented in this study.

Low-latency HEVC streaming over IEEE 802.11p vehicular networks may now be improved using a new cross-layer map-ping approach. MAC layer application layer information is used in a cross-layer manner in the suggested enhancement. Indeed, the method can optimally transport video packets based on information about the MAC layer buffer filling status, frame type, and temporal prediction video structure.

Simulation findings reveal that the suggested alternatives outperform the typical EDCA in many distinct situations and scenarios. In addition, a comparison of the suggested adaptive algorithm's QoS and QoE results showed that it gives the best outcomes for the various HEVC temporal forecast structures.

The present AI encoding setup does not include any kind of categorization. As a result, our next step would be to look into a more efficient video packet classification algorithm for this kind of transmission. Also, packets that aren't received in the allocated time aren't included in the calculation. As a result, sending them through the network is a waste of time and bandwidth, therefore they can be eliminated at the transmitter. Hence an algorithm capable of doing so should be considered, an algorithm connects the queue buffering time, delay constraints at application level and end to end delay.

References

[1] M.G. W.L. Junior, D. Rosário, E. Cerqueira, L.A. Villas, "A game theory approach for platoon- based driving for multimedia transmission in VANETs," Wirel. Commun. Mob. Comput., **2414658**, 1– 11, 2018, doi:https://doi .org /10 .1155 /2018 /2414658.

[2] A.V.V. M. Jiau, S. Huang, J. Hwang, "Multimedia services in cloud-based vehicular networks," IEEE Intell. Transport. Syst. Mag., **7**(3), 62–79, 2015, doi:https://doi .org /10 .1109 /MITS .2015 .2417974.

[3] X.Z. M. Gerla, C. Wu, G. Pau, "Content distribution in VANETs, Veh.," (Veh. Commun. **1**(1), 3–12, 2014.

[4] R.S. C. Campolo, A. Molinaro, "From today's VANETs to tomorrow's planning and the bets for the day after," (Veh. Commun. **2**(3), 158–171, 2015, doi:https://doi .org /10 .1016 /j .vehcom .2015 .06 .002.

[5] R.F.S. D. Perdana, "Performance comparison of IEEE 1609.4/802.11p and 802.11e with EDCA implementation in MAC sublayer," in International Conference on Information Technology and Electrical Engineering (ICITEE), 285–290, 2013.

[6] T.W. G.J. Sullivan, J.R. Ohm, W.J. Han, "Overview of the high efficiency video coding (HEVC) standard," IEEE Trans. Circuits Syst. Video Technol. 22(12), 1649–1668, 2012, doi:https://doi .org /10 .1109/TCSVT.2012 .2221191.

[7] H.D. I. Parvez, A. Rahmati, I. Guvenc, A.I. Sarwat, "A survey on low latency to-wards 5G: RAN RAN, core network and caching solutions," IEEE Commun. Surv. Tutor, (1), 2018, doi:https://doi .org /10.1109 /COMST.2018 .2841349.

[8] E.C. C. Quadros, A. Santos, M. Gerla, "QoE-driven dissemination of real-time videos over vehicular networks," Computer Communications, **91**–**92**, 91–92, 2016.

[9] M.F. P. Gomes, C. Olaverri-Monreal, "Making vehicles transparent through V2V video streaming," IEEE Trans. Intell. Transp. Syst, **13**(2), 930–938, 2012, doi:https://doi .org /10 .1109 /TITS.2012 .2188289.

[10] T.H. R. Alieiev, A. Kwoczek, "Automotive requirements for future mobile networks," IEEE MTT-S International Conference on Microwaves for In-Telligent Mobility (ICMIM), 1– 4, 2015.

[11] J.I.A. M. Oche, R.M. Noor, "Network centric QoS performance evaluation of IPTV transmission quality over vanets," Comput. Commun, **61**, 34–47, 2015, doi:https://doi .org /10 .1016 /j.comcom .2014 .12 .001.

[12] M.P.M. P. Pinol, A. Torres, O. Lopez, M. Martinez, "Evaluating HEVC video delivery in VANET scenarios," IFIP Wireless Days (WD), (Nov. 2013), 2013.

[13] Y.J. A. Torres, C.T. Calafate, J.-C. Cano, P. Manzoni, "Evaluation of flooding schemes for real- time video transmission in VANETs."

[14] D.A. I. Zaimi, Z.S. Houssaini, A. Boushaba, M. Oumsis, "An evaluation of routing protocols for vehicular ad-hoc network considering the video stream," Wirel. Pers. Commun., **98**(1), 945–981, 2018, doi:https://doi.org/10.1007/s11277- 017-4903-y.

[15] G.J.S. C. Rosewarne, B. Bross, M. Naccari, K. Sharman, "High efficiency video coding (HEVC) test model 16 (hm 16) improved encoder description update 9," in document: Jctvc-ab1002, joint collaborative team on video coding (jct-vc) of itu-t sg16 wp3 and iso/iec jtc1/sc29/wg11 28th meeting, 15–21, 2017.

[16] M. Wien, "High Efficiency Video Coding: Coding Tools and Specification," Springer-Verlag, 2015.

[17] H.Y.W. C.H. Mai, Y.C. Huang, "Cross-layer adaptive H.264/AVC streaming over IEEE 802.11e experimental testbed," in IEEE 71st Vehicular Technology Conference, 1–5, 2010.

[18] D.W. Q. Chen, "Delay-rate-distortion model for real-time video communication," IEEE Trans. Circuits Syst. Video Technol., **22**(12), 1376–1394, 2015, doi:https://doi .org /10 .1109/TCSVT.2015 .2389391.

[19] Y.I. G. Kokkonis, K.E. Psannis, M. Roumeliotis, "Efficient algorithm for transferring a real-time HEVC stream with haptic data through the internet," J. Real-Time Image Process., 343–355, 2016, doi:https://doi .org /10 .1007 /s11554 -015 -0505 -7.

[20] X.S. C. Han, M. Dianati, R. Tafazolli, R. Kernchen, "Analytical study of the IEEE 802.11p MAC sublayer in vehicular networks," IEEE Trans. Intell. Transp. Syst., **13**(2), 873–886, doi:https://doi.org/10.1109/TITS.2012.2183366.

[21] V. Srivastava, M. Motani, "Cross-layer design: a survey and the road ahead," IEEE Commun. Mag, **43**(12), 112–119, 2005, doi:https://doi .org /10 .1109 /MCOM .2005 .1561928.

[22] G.J. van R. M.J. Booysen, S. Zeadally, "Survey of media access control protocols for vehicular ad hoc networks," IET Commun., **5**(11), 1619–1631, 2011, doi:https://doi .org /10 .1049/iet -com .2011.0085.

[23] Intelligent transport systems (ITS); access layer specification for intelligent transport systems operating in the 5 GHz frequency band, ETSI EN 302 663, 1.2.1, 1–24, 2013.

[24] A. Festag, "Standards for vehicular communication—from IEEE 802.11p to 5G," E&I, Elektrotech. Inf.Tech, **132**(7), 409–416, 2015.

[25] IEEE standard for wireless access in vehicular environments (WAVE)–multi-channel operation, IEEE Std, **1609.4**, 1–89, 2010, doi:https://doi.org /10 .1109 /IEEESTD .2011.5712769.

[26] R. Zhang, L. Cai, J. Pan, "Resource Management for Multimedia Services in High Data Rate Wireless Networks," Springer-Verlag, 2017.

[27] IEEE standard for wireless access in vehicular environments (WAVE)–multichannel operation, IEEE Std, **1609.4**–**201**, 1–89, 2011, doi:https://doi.org/10.1109/IEEESTD.2011.5712769.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 3, 19-31 (2022)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Towards a Model-based and Variant-oriented Development of a System of Systems

Sylvia Melzer[*1,2], Stefan Thiemann[3], Hagen Peukert[3], Ralf Möller[1]

[1] *Universität zu Lübeck, Institute of Information Systems, Lübeck, 23562, Germany*

[2] *Universität Hamburg, Centre for the Study of Manuscript Cultures, Warburgstraße 26, 20354 Hamburg, Germany*

[3] *Universität Hamburg, Center for Sustainable Research Data Management, Monetastraße 4, 20146 Hamburg, Germany*

ARTICLE INFO

Article history:
Received: 15 February, 2022
Accepted: 14 May, 2022
Online: 25 May, 2022

Keywords:
System of Systems
SysML
Information System
Variants

ABSTRACT

The development of an aggregated system consisting of autonomously developed components is usually implemented as a self-contained unit. If such an aggregation is understood as a system of systems (SoS) that communicates via interfaces with its autonomous subsystems and components, the interfaces and communication exchange should play a central role in the architectural design. In fact, complete and exact interface specifications simplify loose coupling of independent systems into an aggregation. Since an SoS consists of variant and non-variant subsystems, the main challenge in SoS development is the identification of all true variants and its deviating attributes within an SoS. If the system variants are identified at an early stage of the development process, redundant work in the interface design can be substantially reduced. This paper presents an efficient method to identify SoS variants with regard to life cycle management and it shows how to configure a variant-oriented SoS with a standardized communication interface. For the development, the forward-looking model-based systems engineering approach is recommended to create executable specification parts and to detect errors early on through simulations.

1 Introduction

The *Centre for the Study of Manuscript Cultures* (CSMC) at Universität Hamburg hosts a steadily growing number of autonomously developed database systems, e.g., for the projects *Epigraphic database of ancient Asia Minor* (EDAK) (https://www.epigraphik.uni-hamburg.de), *Going From Hand to Hand: Networks of Intellectual Exchange in the Tamil Learned Traditions* (NETamil) (https://www.manuscript-cultures.uni-hamburg.de/netamil/), and *Thesaurus Defixionum* (TheDefix) (www.thedefix.uni-hamburg.de). As a first priority requirement, the database schema reflect the high data variety of these research projects while maintaining the same overall structure. If these databases are now combined into an aggregated information system, new functionalities that are designed to the overall structure and not to the peculiarities of each schema can be defined as it is the case for federated searches. In addition, one can very well imagine that new database systems would like to connect to the aggregated information system later on as long as the structure remains clear.

An illustration for the usefulness of an aggregated information system are trove discoveries, whose associated parts, for some reason, are scattered at different places in the world as it often happens for old manuscripts. Such script fragments are administered in different information systems. As an example, one fragment AO 29196 [1] is located at the Louvre and the counterpart of this fragment, KUG 15 [2], is located in Germany. Indeed, both fragments were discovered without using federated search queries, but for analyzing data from different databases it would be desirable to find related data in an aggregated information system. This example highlights the need to combine, analyze, and query data from different database systems.

The requirement for the development of an aggregated system is, on the one hand, to develop autonomous systems in such a way that the variant parts are not implemented redundantly and, on the other hand, that external databases can be added to the aggregated system without much effort.

A systematic approach to model variant parts was developed at the *Institute of Product Development and Mechanical Engineering Design* (PKT) at the Hamburg University of Technology. The variant-oriented approach is called *integrated PKT* approach (see

[*]Corresponding Author: Sylvia Melzer, Universität Hamburg, Centre for the Study of Manuscript Cultures, Warburgstraße 26, 20354 Hamburg, Germany, sylvia.melzer@uni-hamburg.de

[3]). The integrated PKT approach aims at satisfying a wide variety of customer requirements while developing a component. Ideally, the approach applies to a product family that is developed within one organization, for which the marketed products are supposed to have as few variants as possible. Also, adding an external sub-product to the product family at a later date must be considered very early on in the planning phase. Early consideration of coupling systems or products can lead to modularization of systems to support the approach. The integrated PKT approach also supports modularization. Nevertheless, it is not always possible to extend the product family by a new variant. Another core idea of the integrated PKT approach, besides the variant-oriented development of products, is the combination of the database into an SoS and keeping the main focus on the development of a communication interface. While there are some good approaches to SoS development already, an approach that considers system variants during development of an SoS is to the best of our knowledge not yet available.

For the development of an SoS, model-based methods using the *Systems Modeling Language* (SysML) are increasingly used. Users also benefit from the general *Model-Based Systems Engineering* (MBSE) advantages, such as making complexity manageable. The first model-based SoS developing methods are evolving, cf. [4] as well as [5].

In this paper, we present a provident, model-based, and variant-oriented approach to develop new functions for an aggregated information system so that all functions can be used simultaneously to the benefit of all database systems.

The paper is structured as follows. In Section 2 and Section 3 we give an overview on related work and preliminaries for developing a model-based and variant-oriented SoS. In Section 4 we describe how to develop variant-oriented and sustainable information systems such that as many customer requirements as possible are considered while increasing the number of variants and reducing the necessity of redundant information system development.

In Section 5 we present, first, how variant and SoS relevant requirements are elicited during the requirements engineering process, and second, how to design the structure of an SoS. In Section 6 we describe modeling and simulating the systems' behavior to execute a federated search as an example. The application of our new approach and its results are presented in Section 7. We like to close in Section 8 with a summary and a preview of future work.

2 Related Work

Despite its frequent usage, there is little agreement nowadays on a concise and general definition of the term *system of systems* (SoS). Some approaches of SoS distinguish between SoS and traditional systems. These approaches elaborate specifically on the heuristics of SoS development. They also emphasize the differences to traditional systems. More particularly, it has been noted that the architecture of an SoS aims at optimal communication for the vast majority of all SoS (see e.g. [4, 6, 7]).

The application in [4] illustrates the development of a *Trusted Forwarder System* (TFS) for a secured air cargo transport chain as an SoS using a set of standards that enable useful communication between existing and newly developed components. For the TFS, a communication standard is used that satisfies the new requirements. Thus, after SoS integration, the systems and components are enabled to follow their original tasks without diversion.

In [7], the authors show how, from autonomously developed database systems, an aggregated information system enables relatively simple data exchange through a standardized communication interface. If the individual systems are combined into a federated system via a communication interface, new functions can be added easily, such as the federated search functionality. With the new search function, the individual systems can perform searches in all databases (as opposed to only one database) provided that access rights are correctly assigned. The new search feature can be considered as a new service relevant for a broader range of users.

Both aggregated systems, the TFS and the aggregated information system, were initially developed as single systems, cf. [8, 9]. The single system and the SoS development of the TFS were compared in [4]. It was observed that the SoS development approach mainly is advantageous for traceability beyond the system perspective to the service. The advantages of re-usability of a system, which was developed as SoS, are plain to see: Communication interfaces give more flexibility to add new functionalities or remove subsystems from the overall system.

In [6], the author recommends a stable architectural design for SoS. Such stability can be achieved by admitting independently, i.e. autonomously, developed systems in the architectural design together with a communication interface.

Model-based approaches, such as the *Variant Modeling with SysML* (VAMOS) presented in [10], pure::variant (`https://www.pure-systems.com/purevariants`), the *Variety Allocation Model* (VAM) (variant-oriented developing process of the integrated PKT approach) with SysML [11], exist to identify possible variants in the early phase of system development. The methods VAMOS and VAM with SysML can be represented in the SysML modeling tool *Cameo Systems Modeler* by extending the language elements. Pure::variant can be used as a stand-alone entity for variant modeling. In this paper we decided for VAMOS, since it was already applied in [7] for the development of a cross-domain information system.

Cameo Systems Modeler and the broker-based SysML Toolbox have been successfully used for simple modeling of communication networks in several projects such as SiLuFra [12], ConCabInO [13], KomKab [14], and KMUDigital [15].

3 Preliminaries

This section describes the languages, methods, and tools proposed for a model-based and variant-oriented development of an SoS.

3.1 Modeling Languages

According to the recommendations of MBSE, systems are described or documented using semi-formal modeling languages such as the *Unified Modeling Language* (UML) or the SysML.

Systems Modeling Language SysML was specified by the *Object Management Group* (OMG) to support the model-based devel-

opment of complex systems during the system development process. SysML is a subset of the standardized language UML 2 including some additional extensions. A SysML model can be used to describe the structure as well as the behavior of a system, and can be used to simulate the behavior of systems. In this paper, the focus of variant modeling is on structure. For variant behavior modeling, further challenges are to be expected (cf. [16]), which should be addressed separately due to the complexity of SoS development.

Variant Modeling with SysML A variant is characterized by a base model and differentiating parts, where the base model represents the core of the system and the differentiating parts represent the distinctions of the system components (see [17]). In [10], the author specified VAMOS to model variants with SysML. To use this, the existing model elements in SysML are extended (see Figure 1). The two model elements *Package* and *NamedElement* are extended with the stereotypes *Variant*, *Variation*, *VariationPoint*, and *VariationElement*. The *Variation* is a stereotype of the *Metaclass Package*, which contains all the elements of an option of *Variation*. A *Variation* is also a stereotype of the *Metaclass Package*, which contains multiple variation packages. A *VariationPoint* is a stereotype of the *Metaclass NamedElement*.

VAMOS is suitable for systems with some variability. Furthermore, VAMOS is applicable for the use of structural elements. For the description of variant system behavior, VAMOS can be applied conceptually. Practical applications usually exploit extensions of further SysML language elements related to the behavior necessitating the variant behavior description (see [16]).

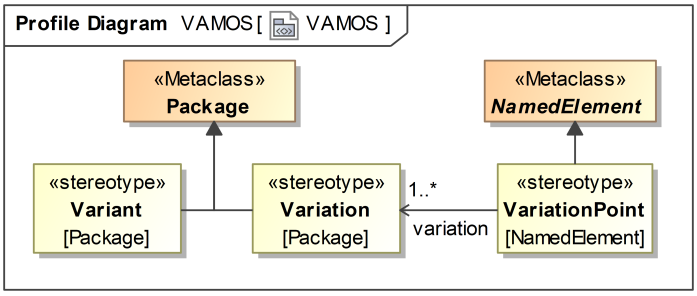


Figure 1: Profile diagram: Variant Modeling with SysML (VAMOS) for the development of a cross-domain information system

3.2 Methods

Context-Based Requirements Engineering The identification of variants should be done as early as possible, so it is necessary that views of all stakeholders involved are considered during the requirements engineering process. The associated systems have to be identified during the SoS context description process. In [18], the authors have defined the *Approach for Context-Based Requirements Engineering* (ACRE) ontology with the goal to capture the requirements of all stakeholders and manage them during the entire system development process. A fundamental approach of the ACRE ontology is the context, since it emphasizes and defines the view of use cases and requirements. Depending on the use cases, some new model-based systems engineering approaches add specific contexts defined for system or product development, e.g. the life phase modularization context in [8], the variety context in [11], and the system of systems context in [19]. In [20, 21], a supplemented ACRE ontology with the SoS aspect are presented and also introduce new terms for system development. Adding the variety context to the SoS approach results in a new variant-oriented approach, which is described in this paper.

V-model For all IT projects in the federal public administration, the V-model is a mandatory procedural standard. The processes of the V-model, inspired by the V-model XT (see www.v-modell-xt.de), can be described as follows: analysis of requirements, functional analysis, high-level design, low-level design, implementation, component test, system test, integration test, and acceptance test. Verification and validation also belongs to the processes.

We argue that the V-model is a good basis for system development, so we have used this approach to develop information systems. For a variant-oriented development and implementation of a communication interface, it is important to consider in the individual process steps like the high-level design, that, among other properties, combined approaches are used to develop the SoS efficiently and correctly. Approaches for the development of an SoS are described in Subsection 5.2.

Broker Federation The brokerage network enables the creation of message routing networks, in which messages in one broker are automatically routed to another broker. These routes may be defined, e.g., between exchanges in the source and destination brokers, or from a message queue in the source broker to an exchange in the destination broker [22]. The principle of coupling systems via a broker federation is a practically proven approach that is used in many applications. In this paper, broker federation is used to create a communication interface between the systems to develop the SoS.

3.3 Tools

Communication Tool The open-source message broker RabbitMQ (https://www.rabbitmq.com/) can be used to create communication networks. RabbitMQ uses the *Advanced Messaging Queuing Protocol* (AMQP) as a standardized communication technology. AMQP defines three components which are essential to implement a message-based architecture. 1) The *message queue* stores messages which can be consumed by client applications. 2) The *exchange* receives messages from publisher applications and routes these to message queues. 3) The *binding* defines a relationship between a message queue and an exchange. Using these components, classic communication paradigms can be implemented and used such as 1) send and receive, 2) work queues, 3) publish and subscribe, 4) routing, 5) topics, and 6) request and reply.

In [4, 7, 9, 15], the authors show that the developed communication interfaces with RabbitMQ can be used for implementing real software or hardware in the model with little effort. For this reason, we choose RabbitMQ to support a communication interface for the individual systems that become part of the SoS.

Modeling and Simulation Tool Cameo Systems Modeler (version 2021x) is a modeling and simulation tool that was originally

developed specifically for the development of systems using the SysML. To simulate behavioral diagrams, Cameo Systems Modeler uses a subset of the UML elements on the OMG *Foundation Subset for Executable Models* (fUML) and *W3C State Chart XML* (SCXML) standards. The broker-based SysML Toolbox is an extension of the Cameo Systems Modeler and provides the integration of real software and hardware [15]. The Toolbox also offers predefined SysML elements that can be used to create database interactions. The SysML Toolbox contains an implementation of these six messaging paradigms. These communication paradigms are implemented via the SysML element *opaque action* and the usage of the Java-like scripting language *BeanShell*.

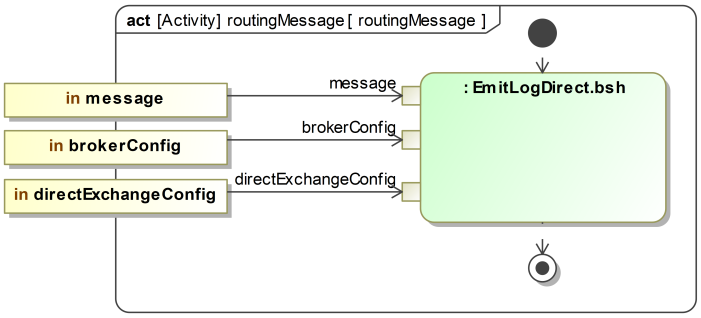


Figure 2: The opaque action *EmitLogDirect.bsh* for sending a routing message to the RabbitMQ server

An implementation of the routing communication paradigm as an opaque action element and the respective BeanShell code are presented in Figure 2 and in Figure 3. The opaque action is called *EmitLogDirect.bsh*. The other paradigms are also available as opaque actions. These opaque actions are implemented as drag-and-drop communication elements, with the aim to increase efficiency and to avoid coding effort.

Language:

BeanShell

Body:

EmitLogDirect.bsh(message, brokerConfig, directExchangeConfig)

```
18 try
19 {
20     ConnectionFactory factory = new ConnectionFactory();
21     factory.setHost(host);
22     factory.setVirtualHost(virtualHost);
23     factory.setPort(port);
24     factory.setUsername(userName);
25     factory.setPassword(pw);
26     Connection connection = factory.newConnection();
27     Channel channel = connection.createChannel();
28
29     try {
30        channel.exchangeDeclare(EXCHANGE_NAME, "direct", true);
31     } catch (IOException e1) {
32        // TODO Auto-generated catch block
33        e1.printStackTrace();
34     }
35
36     String severity = routingKey;
37
38     channel.basicPublish(EXCHANGE_NAME, severity, null, message.getBytes("UTF-8"));
39     print(" [x] Sent :'" + message + "'");
40
41     channel.close();
42     connection.close();
```

Figure 3: The source code for sending a routing message to the RabbitMQ server

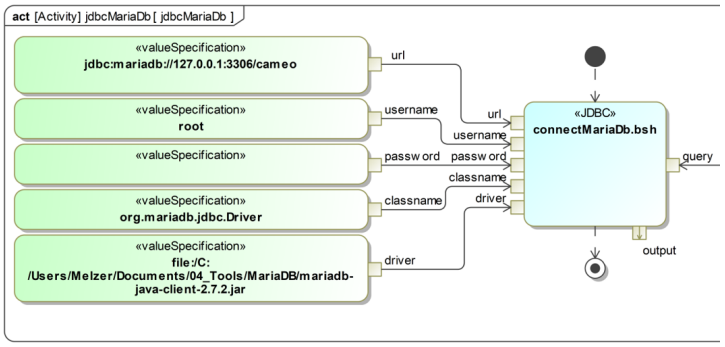


Figure 4: The opaque action *connectMariaDb.bsh* for sending a message to the database MariaDB, source: [23]

The following SysML blocks are used to define input and output parameters: *MessageBroker*, *MessageQueue*, and *MessageExchange*. The *MessageBroker* contains the properties: *host*, *virtualHost*, *port*, *username*, and *password*, which are input parameters for the opaque behavior *EmitLogDirect.bsh*. The properties of the MessageExchange are *exchangeName* and *routing key*. In order to set individual configurations, it is possible to create instances of the SysML blocks. An instance of the *MessageBroker* is *brokerConfig*. An instance of the *MessageExchange* is *directExchangeConfig* (see Figure 2). More details of the broker-based SysML Toolbox are given in [15].

For modeling database expressions, the extension of the broker-based SysML Toolbox can be used or replicated. The predefined database expressions for creating, manipulating, and querying databases are implemented as opaque actions. These predefined actions can also be used as drag-and-drop elements (cf. [23]).

Language:

BeanShell

Body:

connectMariaDb.bsh(url, username, password, classname, driver, query, output)

```
20 import org.mariadb.jdbc.Driver;
21
22 ConnectionMariaDB c = new ConnectionMariaDB();
23 conn = c.connectDb(classname, url, username, password, driver);
24
25 if(conn!=null) {
26     // create the beanshell statement
27     Statement st = null;
28     try {
29         st = conn.createStatement();
30     } catch (SQLException e) {
31         // TODO Auto-generated catch block
32         e.printStackTrace();
33     }
34     // execute the query, and get a beanshell resultset
35     ResultSet rs = null;
36     try {
37         rs = st.executeQuery(query);
38     } catch (SQLException e) {
39         // TODO Auto-generated catch block
40        e.printStackTrace();
41     }
```

Figure 5: The source code for sending a message to the database MariaDB

Figure 5 shows that the opaque action element *connectMariaDB.bsh* has the input values *classname*, *url*, *username*, *password*,

and *driver* to create a database connection. In addition, the BeanShell code is required to send a request to the database (Figure 5, line 29) and get a response (Figure 5, line 37).

The source code in the opaque behaviors is tested by running the simulation. The code can therefore be adopted when implementing, e.g., systems or SoS.

Database Management System The open-source web-based database management system Heurist was specially developed for the Humanities. Heurist allows researchers without prior IT knowledge to develop databases, store and search their data, and publish it on an automatically-generated website.

4 Information System Development

Data projects in the Humanities depict a perfect test scenario for information system development for two reasons. First, the projects tend to be comparatively small and, second, both data and usage show high degrees of heterogeneity. This phenomenon, known as the *long-tail* problem of the Humanities, is due to an institutional decision of most universities to subsume all kinds of subjects under one departmental unit called Humanities. Left the reasons aside, software architects and researchers alike find themselves in the situation to cope with the high variability of requirements, software quality attributes, and missing standards. From a point of view of information system development, one can think of several solutions for data heterogeneity. In fact, they can also all be viewed as an SoS. An analysis of the current situation reveals three strategic strains of data management. Firstly, isolated applications fully independent and maintained by decentralized units such as a single chair. Second, single data applications implemented with a set framework such as *My Content Repository* (MyCoRe) managed centrally. And third, a globally maintained platform with limited but extensive data curation functionality for archiving, publishing, and analyzing data such as Heurist. Since sooner or later, the isolated applications are transferred to one of the centralized data solutions, we will take a closer look at the two later approaches.

MyCoRe The framework MyCoRe (https://www.mycore.de/en/) contains all the functionality of a data repository. Some public institutions such as libraries and universities implement instances of MyCoRe to administer publication inventories and research data. As a typical client-server application, it can be used to host any kind of data. Among the main configurable components of the MyCoRe system are a Solr (https://solr.apache.org/) search engine, a data base access handler via Hibernate, a system management for user rights and access as well as a content store. Interfaces to external systems are restricted to library formats Z39.50, but also comprise REST, OAI and SWORD. Generally all documents and metadata are saved as XML, however, some information is stored in relational database tables for reasons of performance and modifiability. Other interfaces include information exchange to the application layer, that is, a layout engine rendering XSL stylesheets and some functionality to configure the data model as well as other system variables.

Although the structure of a MyCoRe database is known and could be used for automatic retrieval, the data models of a MyCoRe application are very flexible and represented without a standard as a XML schema definition. Its retrieval and analysis depend on how different data models are related to each other and which structural information on how to process the data is *hidden* in the application. Generally it is possible to parse the data model schema definitions and based on this information automate the data retrieval. Yet, for MyCoRe applications that make use of several data models whose interaction and processing became part of the business logic of the program, a semi-automated retrieval process seems to be the only doable solution.

It is a valid data management strategy to have these projects set up as independent MyCoRe instances if larger amounts of data need to be handled or if many users with many different tasks and views on the data require clear and comprehensible workflows. It ensures more flexibility while keeping data maintenance and server administration on an acceptable workload. Although the structure of the data, its formats and processing, is the same for all instances and it therefore has a lot of technical scalability potential, the operation of many MyCoRe instances still leads in the long run into maintenance problems if new versions have to be adjusted to the specific needs and the changing requirements of the project stakeholders. Thus, if specific needs such as a federated search are desired, this cannot be easily added. The implementation of a new function would have to be done for all instances. And if there are variant instances, a new function would have to be developed separately for each instance.

An elegant way around the growing maintainability dilemma is to find a new optimum between usability and scalability. More specifically, it means trading off the flexibility of front end layouts and some cut back on performance to integrate projects into one platform. Indeed, the tendency to focus on services rather than entire system development plays a role in the design decisions of SoS. A practical solution is to devise a system that allows for just so much adjustability as necessary for requirements satisfaction (variant-oriented system development), but leave the components responsible for all other quality requirements untouched. Heurist can be seen as such a way in the middle. Within the approach of SoS, one could push it a step further and classify data projects according to their requirements or one could also embrace all smaller projects into a new platform solution, such as Heurist, and leave the few projects with a large data inventory on MyCoRe instances to keep performance on an acceptable level.

Heurist The data management system Heurist is suitable for variant-oriented system development such as presented in [7, 9]. Even if the development of the systems, here database instances of Heurist, hold the same functions, these can be used to create a project-specific database autonomously. If the individual database instances were to be combined into an aggregated system, it would be possible to develop the complete system as a single system, as a product family or as an SoS. However, the system development of a single system has little flexibility to make extensions. Single systems cannot be used for different purposes as variants as effortless and cost-saving than SoSs.

With Heurist, for each project a project-specific web page can be constructed as a variant with the same functional range. In order

to create a website, the search area, the display of a result set, the display of the contents as well as the integration of a map can be straightforwardly arranged. The view can either be programmed with PHP or implemented via an editor interface.

To sum up, Heurist makes it possible to create a database instance as a variant and supports further development with individual properties.

4.1 Information Systems

The three information systems EDAK, TheDefix, and NETamil are autonomously developed information systems at the CSMC using the tool Heurist while the *Collection of Greek Ritual Norms* (CGRN) is an external application that was not modeled in Heurist.

The first three information systems mentioned above represent how systems can be developed in a variant-oriented manner. The CGRN system represents a system which becomes part of the SoS without being an instance of Heurist.

In practice, other systems are often developed as different instances, but they should also have the possibility to use the same functionality if required. Then, it is desirable that these systems can also be integrated into the existing overall structure without having a complete redesign of the SoS.

In what follows it is shown that both variant and non-variant systems can be part of the SoS and thus all these systems can use the federated search function.

NETamil During the project NETamil at the Universität Hamburg a repository was created containing digital images of classical Tamil manuscripts on palm leaves and on paper from Indian and European libraries along with a descriptive catalog, e-texts along with critical editions and annotated translations. The data was originally stored in a Word document.

In general, the database schemes can either be created individually or they can also be converted into well established XML standards such as *Text Encoding Initiative* (TEI) (`https://tei-c.org`). The TEI format is more common used in the humanities for data storage and exchange.

The automatic transformation of XML-encoded formats into a Heurist database instance has the feature of transferring a large data set into a new database instance in short time. In this paper, the created database systems have been automatically created using a word to TEI transformation process [24].

EDAK During the project EDAK the Department of History at the Universität Hamburg created an epigraphic database of ancient Asia Minor. This database contains a collection of Greek and Latin inscriptions in the area of modern-day Turkey. The data are stored in the format EpiDoc to enable easier data exchange between machines. EpiDoc is a widely used scheme for encoding scholarly and educational editions of ancient documents. It uses a subset of the TEI's standard for the representation of texts in digital form [25].

The EDAK Information system has been automatically created using an EpiDoc to Heurist transformation process.

TheDefix The database TheDefix contains curse inscriptions of the ancient world. The data are represented in a project-specific scheme. In Figure 6 the information system for the *TheDefix* project is presented: the search area is located at the left, in the middle is the result set, and on the very right the project specific data representation (text and map representation) are displayed.

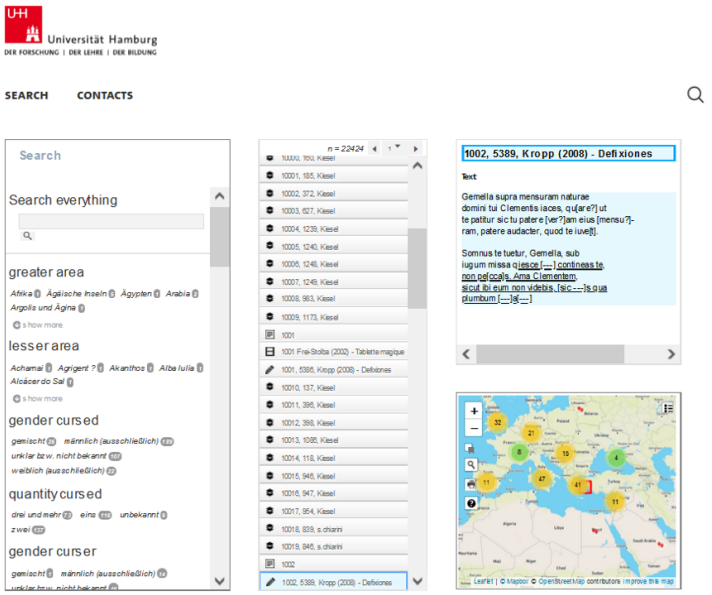


Figure 6: TheDefix Information System

CGRN The CGRN presents epigraphical data on a website. Its primary goal is to gather epigraphical material for the study of Greek rituals and to make these sources widely available [26]. The data are additionally stored in the EpiDoc format.

Merging information systems into an aggregated system, in general, requires addressing the complex issue of information integration. "Information integration is the merging of information from heterogeneous sources with differing conceptual, contextual, and typographical representations" (see [27]). For computers it is difficult to merge information without the knowledge of the syntax, semantics, model, and access of the data representations (see [28]). The approaches therefore require a standardized framework for representing data that, while supporting autonomy to some degree, can make heterogeneity manageable.

In the next chapters, we will reveal how to integrate different autonomously developed information systems, which can also be physically located in different places, as one SoS, taking into account that variant parts are not developed redundantly.

Product family The integrated PKT approach includes the VAM which, in a hierarchical approach of four levels, is used to develop a variant component for each custom-relevant differentiating property, whenever possible in a 1:1 relationship. This approach is very suitable if as many different customer requirements as possible have to be satisfied while still remaining competitive. The VAM approach was also transformed into a model-based approach, using VAMOS to represent the variants. It was observed that the structured package overview in the SysML model avoids redundancies

and improved transparency and traceability for large and complex projects (see [11]).

The integrated PKT approach aims at developing a product family comprising variants. The idea of developing an SoS in a standardized manner is obvious. However, it must be ensured that the development of product families also involves the development of variant hardware components and not just a communication interface. In addition, one has an influence on all systems with the development of product families. If, however, an information system were designed as an SoS consisting of both internal and external systems, it would be recommendable for a clear focus on the communication interface. This recommendation must be taken into account when the SoS is actually modeled on variants.

5 Variant-oriented SoS Development

In this section we present how context information relevant to the SoS and the variants are identified as part of the requirements engineering process ACRE. Additionally, it is described how to design the structure of an aggregated information system. Finally, this chapter depicts how to create a communication network as well as how to simulate the common function *federated search* using Cameo Systems Modeler and the broker-based SysML Toolbox.

5.1 Variant-Based Requirements Engineering

For successful system development it is essential that the needs of all stakeholders are sufficiently satisfied. Therefore, it is necessary to have identified all persons and institutions that have requirements or interest in the system. The respective requirements of all identified stakeholders are collected, documented, and structured according to the ACRE ontology, presented in [18], with the goal to identify the requirements of all stakeholders and to be able to manage them throughout the system development process.

For the variant-oriented development of an information system as an SoS, a lean version of the ACRE ontology was specified and used for modeling an aggregated information system. The ACRE ontology with SoS and variety contexts is presented in Figure 7.

An (abstract) requirement has the specializations business, functional, and non-functional requirements. A requirement description explains requirements, where some rules are applied. The rules could be that requirements have to be formulated in accordance with the ISO 29148:2011 [29] and RFC2119 [30] to use the linguistic syntax profitably. A requirement description is elicited from one or more *Source Element*(s). Source elements can be standards, conversations, requirement lists, and specifications among others.

The contexts are defined as:

- A system of systems context defines views on aggregated systems.
- A system context contains views of system, subsystem, assembly, and component.
- A stakeholder context defines views on different stakeholders.
- A variety context defines views on system variants.

Use cases are validated by one or more *Scenario*(s). A *Scenario* describes the “what ifs” in a semi-formal or formal way. SysML activity and sequence diagrams can present *Semi-formal Scenario*s to describe communication processes and interactions between elements in the system. SysML parametric diagrams present the “what ifs” formally. Both scenario types support the analysis of “what ifs” to validate the use cases. Through simulation, the modeled scenarios can test the interactions between all participants within the communication network.

The developed systems have a satisfying relationship with the requirements they meet. A *System of System* element is a generalization and has two or more systems as parts.

It should be noted that there are a number of other approaches to the requirements engineering process. However, it is crucial that the variety context will be considered during the requirements engineering process. Contextual information has to be added in all other approaches as well. The ACRE ontology has already been successfully applied in many projects over several years using the SysML [11, 13, 14]. Due to the well-known and proven approach of applying ACRE with a variety context in a model-based way during system development, the ACRE approach was chosen for the development of an SoS.

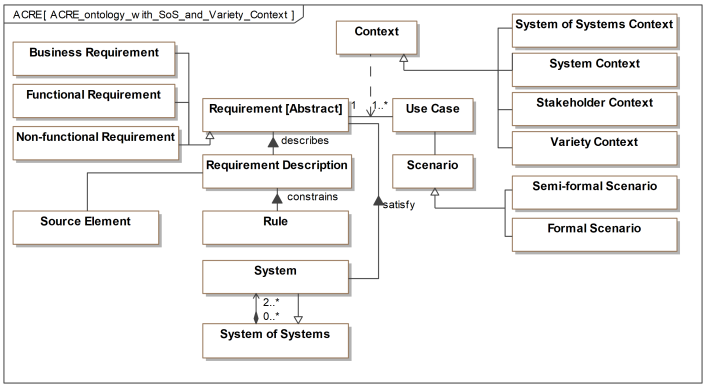


Figure 7: ACRE ontology with systems of system and variety contexts

5.2 SoS Development

Heurist can be used to create variant database instances and is realized as a client-server architecture. Although established design patterns are missing in the still evolving software, further development tends in the direction to have Heurist fully operational as an *Model-View-Controller (MVC)* application. The MVC separates an application into three main logical components: the model, the view, and the controller. Each of these components are built to handle specific development aspects of an application. In the context of creating an information system, the model represents a data scheme, the view a graphical user interface, and the controller accepts user inputs and converts it to commands for the model or view.

The new planned architectural approach is important when it requires adding another layer, the SoS layer. The development is currently still in the conceptual phase. As of now, Heurist is initially used for variant system development and the SoS layer is first tested out through simulations and prototype implementations.

In addition, the Universität Hamburg operates Heurist as a pub-

lic institution, which recommends the use of an adjusted version of the V-model. It follows that further adjustments will be made to Heurist in the area of verification and testing.

6 Modeling and Simulation of an SoS

Modeling and Simulating of an SoS using the SysML and the tool Cameo Systems Modeler has the advantage to test the system's behavior before implementation because the specification is executable. "This quality of executable specifications promises to remedy the most serious problem of software – its lack of correctness and reliability." [31]

In the following we present how to develop an executable specification for an SoS during the requirements engineering process.

6.1 Requirements Profile

For the special requirements (business, functional, and non-functional), new stereotypes were defined as an extension of the *Extended Requirement* stereotype. An *Extended Requirement* is a standard requirement extension that adds some properties to a requirement element. The requirements are devised in accordance with the ISO 29148:2011 and RFC2119.

6.2 SoS Profile in SysML

For representing Systems and SoS the new stereotypes *System* and *System-of-Systems* are defined as extensions of the *Metaclass Class*, see Figure 8.

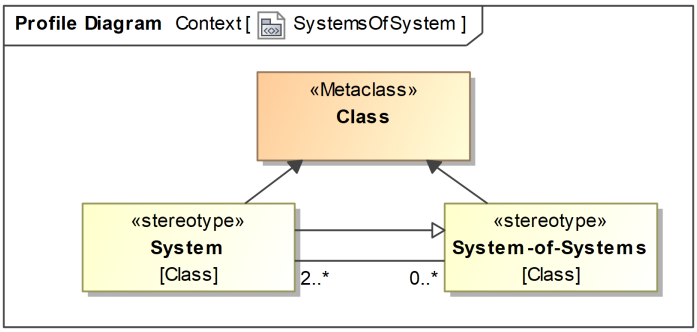


Figure 8: Profile Diagram: new stereotypes *Systems of System* is a specialization of a system and has an association to the stereotype *System*

6.3 Variety Profile

The VAMOS profile which is presented in Figure 1 is used for variant modeling. Figure 9 presents a concrete application of VAMOS.

The package *Variation 1* has the stereotype *Variation* and contains the System *Heurist*. The packages V1, V2, and V3 are variants of *Variation 1*. The variant *V1* contains the System EDAK, the variant *V2* contains the system TheDefix, and the variant *V3* contains the system NETamil, respectively.

One way to introduce a redundancy-reduced communication interface for all variants is to add a SysML port element to the Heurist system. All variants inherit the port via the specialization. However, if an external system were added to the aggregated system at a later point in time, a separate communication interface would have to be implemented for this external system. This is precisely the crux of the matter. If a communication interface is to be offered for internal variant systems as well as for external systems, the communication interface should be inherited by the systems via a specialization using an SoS element.

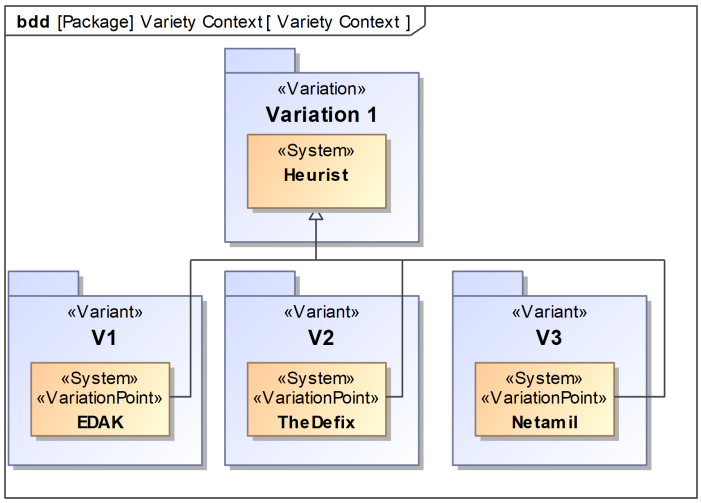


Figure 9: Representation of three variants using VAMOS

6.4 Use Cases

Figure 10 shows the representative use cases for different search functionalities while considering the variety and SoS contexts. The main actor is a CSMC user. The CGRN, EDAK, NETamil, and TheDefix users are specializations of the CSMC user.

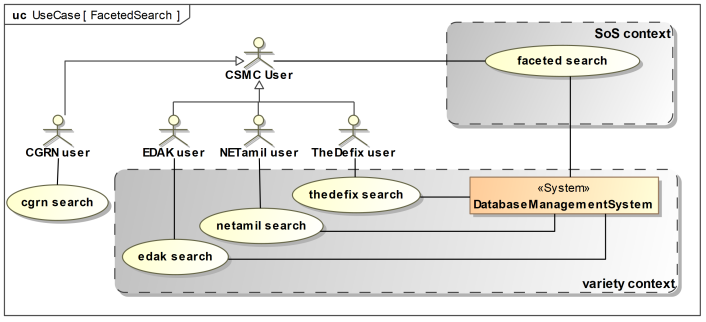


Figure 10: Use case diagram with variety and SoS context

One can see that CGRN users do not belong to the variant context as the other three users but all users also have an association to the use case faceted search. As described in Section 4, the three systems should be developed as variant systems using Heurist, while the development of the CGRN system was done externally. Nevertheless, all systems should be networked so that each system can use the federated search function.

A CSMC user can execute a faceted search. The specialized users can also execute this search while all users have (project-)specific search functionalities, e.g., EDAK users search for specific names mentioned in editions or for object types of inscriptions (use

case: *edak search*), NETamil users look up which word occurs in which poem and in which line (use case: *netamil search*), and TheDefix users want to know the curse id of curses (use case: *netamil search*). The different contexts are presented in SysML use case diagrams.

6.5 Scenarios

Each use case can be validated by one or more scenarios. The scenarios can be represented in behavioral diagrams such as activity or sequence diagrams. We use activity diagrams for modeling and simulating, e.g., the federated search functionality.

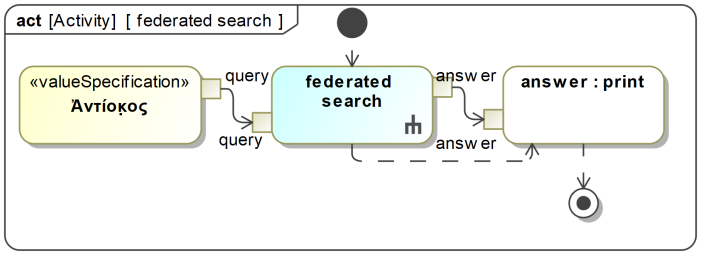


Figure 11: Search for the word “Antiochus” (engl.)

Figure 11 depicts the word “Antiochus” (engl.). It is the input value (=query) for the federated search activity. Behind the federated search activity is a more detailed federation process. As a result, the responses of all databases are printed.

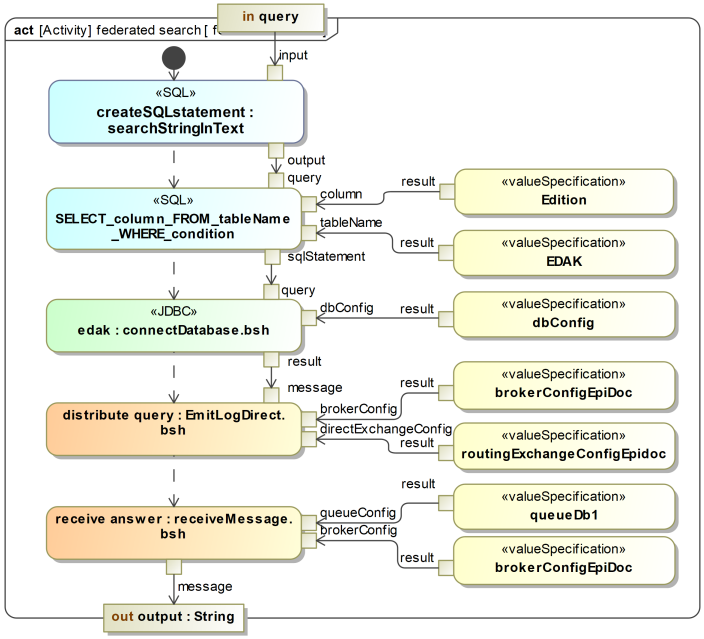


Figure 12: Faceted search actions

Figure 12 illustrates the faceted search process in more detail. The query is the input value. The first action *createSQLstatement:searchStringInText* creates the SQL statement *“Text LIKE ’%Antiochus%’;”*. The action *SELECT_column_FROM_tableName_WHERE_condition* creates the SQL expression *SELECT Edition FROM EDAK WHERE ’%Antiochus%’;*, which is the input value for the next action. The action *distribute query:EmitLogDirect.bsh* sends the SQL expression to a server. The action *receive answer:receiveMessage.bsh* sends a response from a server. The opaque action *EmitLogDirect.bsh* can also contain a forwarding process to another database. To realize a federated search, a script was implemented and must be active on the server side. In fact, the script calls the requests from the server (query queue), processes the schema mapping, and passes the response to the server (response queue). We implemented the server side scripts in Java. The source code is very similar to that of Beanshell (see https://www.rabbitmq.com/tutorials/tutorial-three-java.html). However, other programming and script languages can also be used such as Python, PHP, C#, or JavaScript (see https://www.rabbitmq.com/getstarted.html).

It should be noted here that the scenario at hand already incorporates decoupling of the systems using a communication interface. In a very early phase of system development, communication could take place directly with the database. And yet, communication interfaces are to be used in the development of SoS. Briefly put, this has already been taken into account in the scenarios. As intended by ACRE, the use cases were validated by the scenarios during the requirements engineering process.

6.6 Communication Interface

Communication interfaces ensure the coverage of the need for information and are used for data exchange. For creating communication networks, RabbitMQ is used as an *Application Programming Interface* (API) for SoSs. RabbitMQ offers broker federation and therefore allows the exchange between source and destination brokers, or from a message queue in the source broker to an exchange in the destination broker (see [15]). To model these communication interfaces the stereotype *interfaceBlock* is used. One port of the SoS has at least this communication interface to establish a communication network between the systems which are part of the SoS.

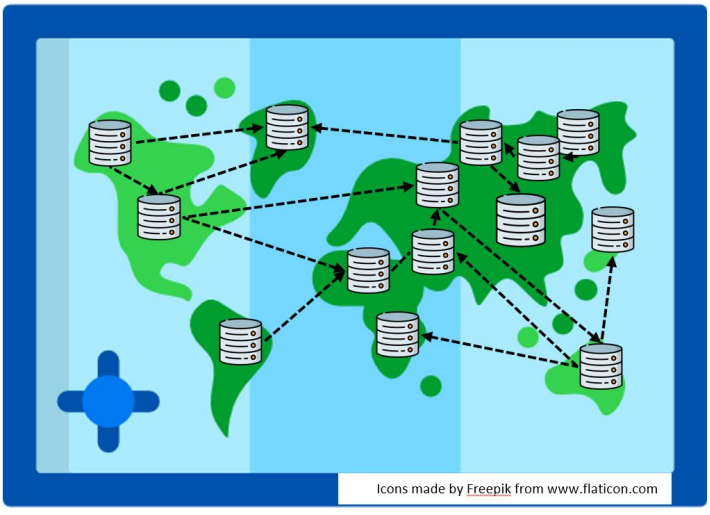


Figure 13: Federated Search Network

Figure 13 illustrates a communication network between loosely-coupled systems which are to be transferred to an aggregated system including a communication interface. By coupling the systems, e.g., federated searches can be realized. The idea is to provide each participant with its own RabbitMQ message broker to easily realize this communication network.

6.7 *Structure of the CSMC Information System*

Figure 14 illustrates the structure of the SoS named *CSMC Information System*. The SoS has the parts of systems *EDAK*, *TheDefix*, *NETamil*, and *CGRN*. These systems are also specializations of the SoS and inherit all activities of the SoS. In this case, federated search is part of each system. The SoS has a communication interface which is modeled as a port. The systems *EDAK*, *TheDefix*, and *NETamil* also inherit all elements of the system *Heurist*. The system *CGRN* is not a specialization of the system *Heurist* and thus does not inherit all activities of the Heurist system, but only those of the SoS.

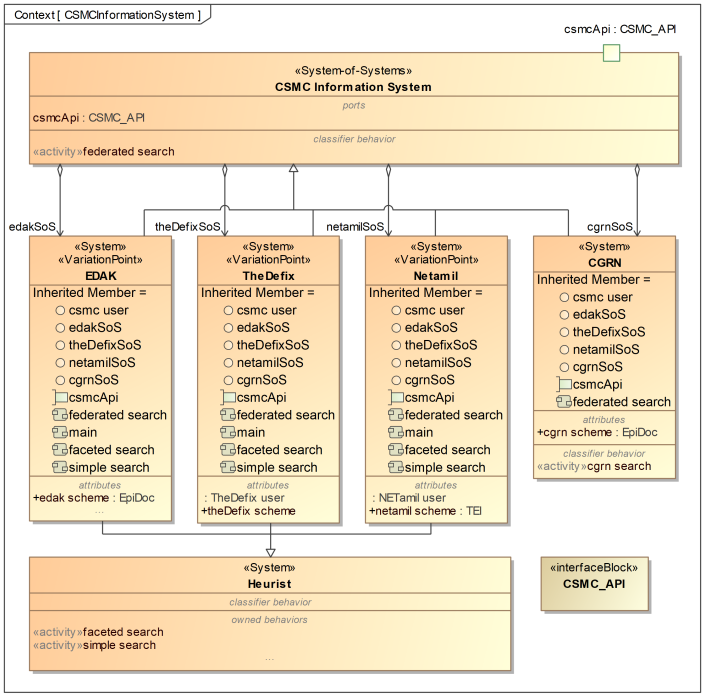


Figure 14: CSMC Information System

Figure 15 gives another overview about the dependencies *generalization* and *inherit* members between the SoS and the systems. In the allocation diagram it can be seen at a glance which systems inherit which activities or which do not. When adding more activities to an SoS or when adding more external systems, this overview can be used to quickly determine which elements will be added to a system when it becomes part of an SoS.

In the development of interfaces, the allocation diagram is an excellent way to illustrate the dependencies of all the systems involved. In the diagram, the separation between the interfaces of the SoS or other interface dependencies can be clearly highlighted.

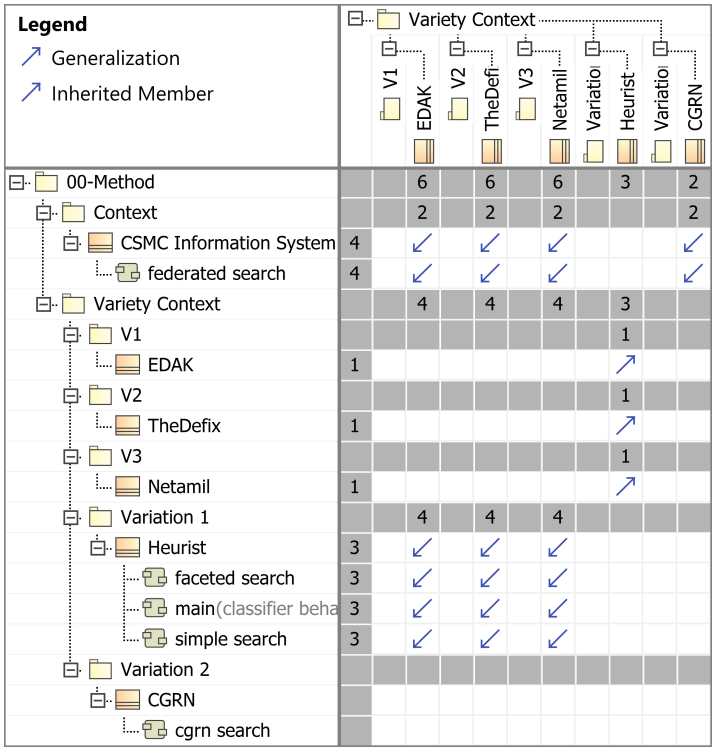


Figure 15: Allocation diagram which represents the dependencies *generalization* and *inherit* members

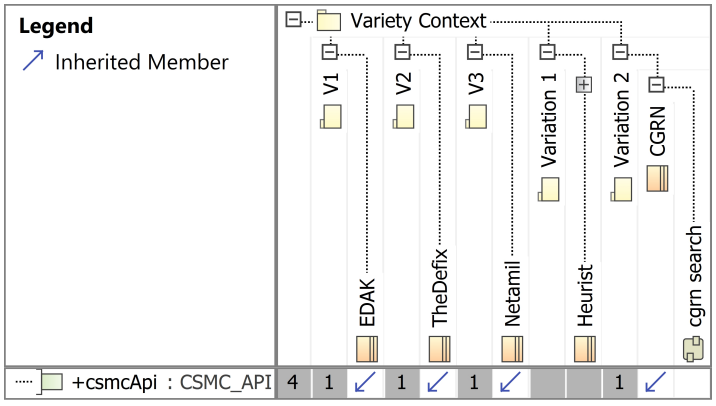


Figure 16: Allocation diagram which represents the systems which have the SoS communication interface

7 Application and Results

We evaluate our approach by a feasibility study. For this purpose, we use a notebook where the tool Cameo Systems Modeler (version 2021x) and the broker-based SysML Toolbox, a RabbitMQ server (version 3.8.9), and MariaDB (10.5.6) are installed. We emulate the databases EDAK, and NETamil on the database MariaDB which represents the Heurist database instances. On a Raspberry Pi 4 we also installed a RabbitMQ server and MariaDB where the database CGRN is simulated. Both RabbitMQ servers are configured with particular message queues, exchanges, and bindings as follows.

The message queues *queueDb1* for EDAK, *queueDb2* for NETamil, and *queueDb3* for CGRN are defined. They are all in the

same virtual host *dbFederatation* (see Figure 17).

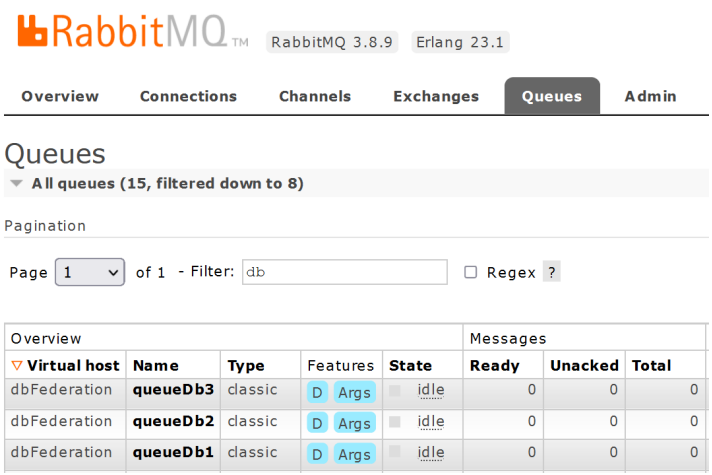


Figure 17: Defined queues on a RabbitMQ server

The exchange is called *db.direct*. The bindings with the particular routing key are: queueDb1 → epiDoc, object, query; queueDb2 → object, queueDb3 → epiDoc, object.

The EDAK data model is represented by the entity type "description." A description has the attributes "identifier", "description_id", "edition", "category", "region", "location", "find spot", "text", and "date." Each "description" has the unique identifier "description_id."

The CGRN data model is represented by the entity type "description." A description has the attributes "idno", "date", "provenance", "support", "layout", "bibliography", "text", "translation", "traduction", "commentary", "publication", "authors", and "project director."

The NETamil data model is represented by the entity types "poem", "commentary", and "dictionary." A poem has the attributes "edition", "transliteration", "word_by_word_translation_into_english", "translation_into_english", and "source."

We simulate federated searches, such as presented in Figure 12. During the simulation the query *SELECT Edition FROM EDAK* is sent to the EDAK database via the opaque action *edak:connectDatabase.bsh*. The SQL expression is published via the opaque action *distribute query EmitLogDirect.bsh*. The databases EDAK receives this expression via the opaque action *receive answer:receiveMessage.bsh*. The database CGRN is queried with the same SQL expression because of the defined routes in the RabbitMQ servers. For the search query "Antiochus", written in Greek language, (*SELECT Edition FROM EDAK WHERE '%Antiochus%';*) we received 1 answer from EDAK and 0 answers from CGRN. For the search query "Zeus", written in Greek language, (*SELECT Edition FROM EDAK WHERE '%Zeus%';*) we received 5 answers from EDAK and 1 answer from CGRN (see Figure 18).

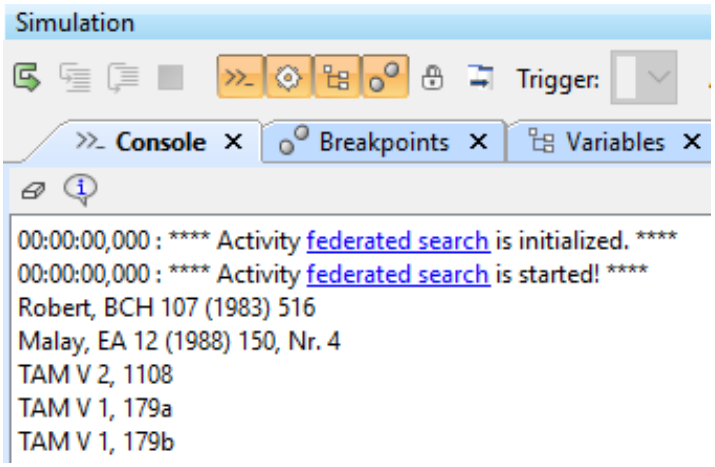


Figure 18: Database results for the search query "Zeus" written in Greek language

For the search query *SELECT COUNT (e.Date) AS Number, e.Date AS Date FROM EDAK e GROUP BY e.Date* we receive the following results (excerpt):

```
+--------+----------------+----------+
| Number | Date           | Database |
+--------+----------------+----------+
|     31 | 4. Jh. v. Chr. | EDAK     |
|    200 | 1. Jh. n. Chr. | EDAK     |
|    818 | 2. Jh. n. Chr. | EDAK     |
|    642 | 3. Jh. n. Chr. | EDAK     |
|    156 | 4. Jh. n. Chr. | EDAK     |
+--------+----------------+----------+
|      1 | ca. 250-200 BC | CGRN     |
|      2 | ca. 350-300 BC | CGRN     |
+--------+----------------+----------+
```

The responses returned by EDAK and CGRN show that the date is differently represented in both databases. The date differs in language and representation (indication as century or year). When the query is filtered by year, one of the two databases returns an empty result set as response. A translation of the date representations can lead to a complete answer. A mapping between the representation of the date is required to ensure *correct* query results. Schema mapping is generally required when defining federated search queries.

This simulation example also presents that queries from EDAK are answered using both the EDAK and the CGRN databases. NETamil is not involved in this specific query process because of the missing routing key in the RabbitMQ configurations. At this point it makes no sense in terms of content. If one wants to compare another repository with Tamil poems, a route can be defined via the RabbitMQ configuration that supports the sensible federated search. In this example, the Cameo System Modeler's console represents the CSMC information system, which receives all responses from the various databases from the federated search. If either a Heurist database instance or an externally developed system is to be aggregated to the SoS, this can be realized by installing a RabbitMQ broker, programming a script for publishing and receiving messages from the broker, and setting the broker configurations.

In this way, new databases can be added so that our principle "*bring your own database* is supported. Then all new systems of the

SoS will also benefit from the federated search. Consider that the challenges of information integration must be resolved for mapping to use federated searches successful.

In the humanities, as well as other fields, it is important that existing functions such as federated searches can be used without large expenditure of resources. After all, resources are limited. Existing data can thus be enriched with further information in a short time and users can focus more on editing content. Using the allocation diagram to keep an overview of all systems to be aggregated also helps to keep track of the growing number of systems (cf., Figure 15 and Figure 16). All systems, whether variants or not, can be specifically developed and integrated into the overall SoS. Without considering this overview, existing systems could be produced mistakenly from scratch simply because they are unknown and as such a variant is regarded as an external (unregistered) system. As a positive effect of the present approach, one has the advantage of being able to cooperate with external parties whether their system can also be developed as a variant of one's own system, so that the same communication interface (*csmcInterface*) can be used.

At the Universität Hamburg Heurist has already been set up and it is being used for the autonomous development of information systems. More than ten information systems have already been created. When changing functions, such as the web page design, the allocation diagram can be used to see which systems are affected.

The special feature of the model-based documentation of SoSs and the variants with VAMOS make it convenient for the developers to get an overview between the different diagram types and the filter properties for displaying data. Developers also see which systems are relevant and whether changes have an impact on the system properties or not. It is self-explanatory to switch between a display of specific SoS or variant elements, or view everything together in one diagram, as shown in Figure 15. The model-based and variant-oriented approach also ensures that the interfaces of the systems are not developed redundantly. The realization of a communication interface using RabbitMQ supports the aggregation of decoupled systems by implementing message scripts that are publicly available.

It is planned to transfer the prototypically implemented CSMC Information System with the presented communication interface into a product. Regarding the product development, the variant-oriented approach points in the right direction as to even more relevant parameters, such as performance and security attributes, that should be tested in advance.

The approach at hand also fits in other areas of system development, e.g., in the aviation industry. There, it has already been shown that the networking of systems after the digitization of business processes can be helpful to automate ordering processes between supply chain tiers [32]. It would be conceivable to design the ordering process as part of the SoS and thereby identify the variants in the ordering process as well as the external systems that want to become part of the SoS. As an assumption, there will be more individual solutions that will be aggregated into an SoS. Here, too, the approach offers the advantage of keeping an overview of all systems and working towards a common interface in a targeted manner so that the connection to the SoS can be made with little resource effort.

8 Conclusion and Outlook

In this paper, we presented how to develop an aggregated system, which, understood as an SoS, was put into practice in a model-based and variant-oriented way. The aim was to identify the number of variants easily at an early stage of the requirements engineering process so that the development of elements has neither to be done holistically nor redundantly. For this purpose, we used the ACRE ontology with the extended contexts on SoS and variety, and applied this approach with the SysML tool Cameo Systems Modeler as well as the VAMOS profile. In addition, we defined an SoS profile which helped us to distinguish between developing variants and merging variants as well as non-variants into an SoS. For the implementation of a communication interface, RabbitMQ was used as a message broker, which allows loosely coupled systems to be brought together in a simple way. The variant database systems were developed with Heurist which on the one hand supports the development of automated database systems and on the other hand keeps the heterogeneity under control, often resulting from the many requirements. The prototype implementation showed us that this path is promising and should be further pursued.

The advantage of merging multiple database systems is that functions such as a federated search can be implemented, however, the problem of data integration between all the database instances must be solved beforehand so that a search query does not lead to a faulty response. Therefore it must be ensured that the data or their representations have the same syntax, semantics, model, and access. At the CSMC, a feasible study is currently in progress.

Conflict of Interest The authors declare no conflict of interest.

References

[1] E. Werner, "Clay Tablet (AO 29196) from the Louvre in Paris (3D model)," 2020, doi:http://doi.org/10.25592/uhhfdm.918.

[2] E. Werner, "Clay Tablet (KUG 15) from the University Library Giessen (3D model)," 2020, doi:https://doi.org/10.25592/uhhfdm.766.

[3] D. Krause, G. Beckmann, S. Eilmus, N. Gebhardt, H. Jonas, , R. Rettberg, "Integrated Development of Modular Product Families - a Methods Toolkit," in In T.W. Simpson, J. Jiao, Z. Siddique, K. Hölttä-Otto (Eds.): Advances in product family and product platform design: Methods & applications, 245–269, Berlin Springer, 2014.

[4] O. C. Eichmann, S. Melzer, R. God, "Model-based Development of a System of Systems Using Unified Architecture Framework (UAF): A Case Study," in Proceedings of 2019 IEEE International Systems Conference, IEEE, 2019.

[5] C. B. Nielsen, P. G. Larsen, J. Fitzgerald, J. Woodcock, J. Peleska, "Systems of Systems Engineering: Basic Concepts, Model-Based Techniques, and Research Directions," 2015.

[6] M. W. Maier, "Architecting principles for systems-of-systems," Systems Engineering, **1**, 267–284, 1998.

[7] S. Melzer, H. Peukert, H. Wang, S. Thiemann, "Model-based Development of a Federated Database Infrastructure to support the Usability of Cross-Domain Information Systems," in Proceedings of 2022 IEEE International Systems Conference, IEEE, 2022.

[8] O. C. Eichmann, S. Melzer, M. Hanna, R. God, D. Krause, "A Model-Based Approach for the Development of Modular Product Families Considering Different Life Phases," in Proceeding EMEA Systems Engineering Conference, EMEASEC 2018 / TdSE 2018, 2018.

[9] S. Melzer, S. Thiemann, R. Möller, "Modeling and Simulating Federated Databases for early Validation of Federated Searches using the Broker-based SysML Toolbox," in Proceedings of 2021 IEEE International Systems Conference, IEEE, 2021.

[10] T. Weilkiens, Variant Modeling with SysML, MBSE4U Booklet Series, 2016.

[11] T. Bahns, S. Melzer, R. God, D. Krause, "Ein modellbasiertes Vorgehen zur variantengerechten Entwicklung modularer Produktfamilien," in Tagungsband zum Tag des Systems Engineering (Eds.: Chr. Muggeo, S.O. Schulze), Carl Hanser Verlag GmbH & Co. KG, 2015.

[12] R. God, S. Melzer, U. Wittke, "SiLuFra Schlussbericht - Sichere Luftfracht-Transportkette: Konzepte, Strategien und Technologien für sichere und effiziente Luftfracht-Transportketten; Teilvorhaben: Modellbasierte Architektur- und Lösungsspezifikation; Laufzeit des Vorhabens: 01.07.2013 - 31.08.2016," 2016.

[13] R. God, S. Melzer, "Teilvorhaben: Spezifikation und Integration cyber-physischer Betriebs- und Geschäftsprozesse : Schlussbericht : Laufzeit des Vorhabens: 01.05.2016-30.09.2019 : Berichtzeitraum: 01.05.2016-30.09.2019, Spezifikation und Integration cyber-physischer Betriebs- und Geschäftsprozesse," Technical report, Technische Universität Hamburg, Institut für Flugzeug-Kabinensysteme, Hamburg, 2020, doi:10.2314/KXP:1726105857.

[14] R. God, U. Wittke, S. Melzer, C. Witte, "KomKab Schlussbericht - Kommunizierende Kabine; Teilvorhaben: Digitaler Ramp-Agent; Laufzeit des Vorhabens: 01.01.2016 - 31.03.2019," 2019.

[15] S. Melzer, J. P. Speichert, O. C. Eichmann, R. God, "Simulating cyber-physical systems using a broker-based SysML Toolbox," in AST 2019 - 7th International Workshop on Aircraft System Technologies, Hamburg University of Technology, 2019.

[16] D. Arndt, S. Melzer, R. God, M. Sieber, "Konzept zur Verhaltensmodellierung mit der Systems Modeling Language (SysML) zur Simulation varianten Systemverhaltens," in Tagungsband zum Tag des Systems Engineering (Eds.: S.O. Schulze, C. Tschirner, R. Kaffenberger, S. Ackva), Carl Hanser Verlag, 2017.

[17] S. Melzer, R. God, T. Kiehl, R. Möller, M. Wessel, "Identifikation von Varianten durch Berechnung der semantischen Differenz von Modellen," in Tagungsband zum Tag des Systems Engineering (Eds.: M. Maurer, S. O. Schulze), Carl Hanser Verlag GmbH & Co. KG, 2014.

[18] J. Holt, S. A. Perry, M. Brownsword, Model-Based Requirements Engineering, volume 9 of *Professional Applications of Computing Series*, Institution of Engineering and Technology, Stevenage, 2012.

[19] O. C. Eichmann, S. Melzer, F. Giertzsch, R. God, "Stakeholder Needs and Requirements Definition During Service Development in a System of Systems," in Proceedings of 2020 IEEE International Systems Conference, IEEE, 2020.

[20] J. Holt, S. Perry, R. Payne, J. Bryans, S. Hallerstede, F. Hansen, "A Model-Based Approach for Requirements Engineering for Systems of Systems," IEEE Systems Journal, **9**(1), 252–262, 2015, doi:10.1109/JSYST.2014.2312051.

[21] J. Holt, S. Perry, SysML for Systems Engineering: A Model-Based Approach, Computing, Institution of Engineering and Technology, 2018.

[22] A. Qpid, "Messaging built on AMQP," https://qpid.apache.org/releases/qpid-cpp-master/cpp-broker/book/chap-Messaging_User_Guide-Broker_Federation.html, 2015, accessed January 22, 2022.

[23] S. Melzer, O. C. Eichmann, H. Wang, R. God, "Modeling and Simulation of Database Interactions," Tag des Systems Engineering 2021 (TdSE2021), 2021, doi:10.25592/uhhfdm.9696.

[24] S. Schiff, S. Melzer, E. Wilden, R. Möller, "TEI-based Interactive Critical Editions," in 15th IAPR International Workshop on Document Analysis Systems, Lecture Notes in Computer Science (LNCS), Springer, 2022.

[25] T. Elliott, G. Bodard, E. Mylonas, S. Stoyanova, C. Tupman, S. Vanderbilt, et al., "EpiDoc Guidelines: Ancient documents in TEI XML (Version 9)." Available: https://epidoc.stoa.org/gl/latest/., (2007-2022), accessed January 22, 2022.

[26] J.-M. Carbon, S. Peels, V. Pirenne-Delforge, "A Collection of Greek Ritual Norms (CGRN)," Liège, http://cgrn.ulg.ac.be, consulted in 2020, 2016–2020, online; accessed 10 December 2021.

[27] W. Hao, S. De-wen, F. Xujian, X. Haitao, "Application of information fusion technologies for multi-source data," Journal of chemical and pharmaceutical research, **5**, 2013.

[28] U. Leser, F. Naumann, Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen, dpunkt.verlag GmbH, 2007.

[29] ISO/IEC/IEEE 29148:2011(E), "Systems and software engineering - Life cycle processes - Requirements engineering," https://wiki.unix7.org/_media/ict/lib/iso-iec-ieee-29148-2011.pdf, accessed January 22, 2022.

[30] S. Bradner, "Key words for use in RFCs to Indicate Requirement Levels," Harvard University, https://datatracker.ietf.org/doc/html/rfc2119, 2017, accessed January 22, 2022.

[31] N. E. Fuchs, "Specifications are (preferably) executable," Softw. Eng. J., **7**, 323–334, 1992.

[32] H. Wang, S. Melzer, "Simulation of Ordering Processes across different Supply Chain Tiers in the Aviation Industry," in 2022 IEEE International Systems Conference (SysCon) (IEEE SysCon 2022), Montreal, Canada, 2022.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 1, 27-46 (2022)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Physics behind the Concept of a Sodium-Potassium-Cesium-Cooled Martian Nuclear Reactor

Okunev Viacheslav Sergeevich*

Bauman Moscow State Technical University, Department of Physics, The Faculty of Fundamental Science, Moscow, 2-ya Baumanskaya ul. 5, 105005, Russia

ARTICLE INFO

Article history:
Received: 22 November, 2021
Accepted: 25 December, 2021
Online: 22 January, 2022

Keywords:
Martian Nuclear Power Plant
Fast reactor
Eutectic alloy of sodium, potassium and cesium
Safety
Cermets fuel

ABSTRACT

The main goal of the work is to determine the basic conceptual solutions of a nuclear reactor operating on the surface of another planet. The problem was solved using the example of the Martian nuclear power plant. The article uses calculation and optimization research methods and corresponding program codes (well-known and author's). The results of the study made it possible to formulate the basic requirements for the Martian nuclear power plant and select the type of reactor. This is a new type of reactor: pressurized liquid metal fast reactor. It is proposed to use an innovative cermets nuclear fuel based on mixed mononitride and uranium metal nanopowder, which was previously considered by the author for new generation BN and BREST ground-based reactors. It is proposed to use a eutectic (or near-eutectic) NaKCs alloy as a coolant. Optimization of the alloy composition has been carried out. The fuel and coolant of the reactor contains long-lived radioactive waste to be transmuted. NaKCs alloy is less reactive than pure alkali metals including Na, K and Cs. With an electric power of 600 MW, it is possible to ensure the internal self-protection of the reactor. All emergency modes of the ATWS type (anticipated transient without scram) are not hazardous. This means that with a decrease in power to values characteristic of the initial stages of the colonization of Mars, the safety of the reactor is easily ensured. The relatively low chemical activity of the coolant makes it possible to use a two-circuit energy conversion scheme. The second circuit can use water or carbon dioxide. Carbon dioxide is preferred because of its presence in the atmosphere of Mars (95% CO_2). The significance of the research lies in the possibility of constructing a Martian nuclear power plant within the framework of existing technologies.

1. Introduction

The exploration of outer space, the planets closest and distant from us, requires the creation of compact, economical, reliable, low-maintenance (practically autonomous) and safe energy sources with a high energy density. Technologically mastered energy sources using the fission reaction of heavy atomic nuclei are attractive.

There are two directions of using space nuclear reactors. The first is related to the need to travel long distances. This is a nuclear electric propulsion system that combines the concepts of "nuclear space tug", "interorbital tug", "and transport and power module". The second direction is associated with the supply of energy to the spacecraft, space or alien station and crew. It can be a low-power alien nuclear power plant (NPP) that provides power to the station, or a space NPP that provides power to a spacecraft. Its functions are not related to the movement of the spacecraft. This work is devoted to solving the second problem and is a continuation of the research published by the author earlier [1]. At present, the development of nuclear electric propulsion systems is being actively pursued, capable of operating in the mode of obtaining thrust (engine) and in the mode of generating electrical energy (NPP) [2, 3]. If it is necessary to supply power to stations located on other planets, it will be necessary to build a NPP (still of low power, that is, electric power up to 100 ... 300 MW). The present work is devoted to this problem.

All reactors that have ever experienced severe accidents were considered safe. They met all the rules and regulations that were in force at the time of the accident on the territory of the country where these reactors were operated. Serious accidents occurred at the plutonium production reactor of the Windscale nuclear complex (Great Britain, 1957); Unit 2 of the Three Mile Island NPP (USA,

*Corresponding Author: Okunev Viacheslav Sergeevich, okunevvs@bmstu.ru

1979); 4th block of the Chernobyl NPP (USSR, 1986); four units of the Fukushima-1 NPP (Japan, 2011).

After a series of accidents at NPPs, the philosophy of ensuring the safety of nuclear power facilities was radically rethought. The first conceptual designs of nuclear reactors appeared, in which it is possible to exclude all accidents leading to unacceptable releases of radioactive substances outside the NPP.

A comprehensive solution to the following tasks was proposed:

- Provision of energy on the required scale;
- Economic efficiency;
- Safety;
- Provision (self-sufficiency) with fuel.

Safety is understood not only to exclude severe accidents at nuclear power facilities, but also to safely handle waste and create physical barriers to the theft of nuclear materials.

Nuclear power in the 20th century was not limited to terrestrial reactors. Nuclear power has entered low earth orbit. The first developments of NPPs were started simultaneously in the USSR and the USA in 1954 [4]. Space nuclear power was ahead of its time. Modern interplanetary spacecraft use radioisotope energy sources (less than 1 kW) [2]. Russia has experience in the development and use of nuclear electric propulsion systems and NPPs with electric power up to 6 kW in outer space. But such installations have not yet been in demand [2]. Until now, energy sources with a capacity of about 1 MW and more have not been in demand. The provision of spacecraft and orbital stations with energy is achieved through the use of solar batteries (the power of which is estimated at several kilowatts, but can reach ~ 10^5 W), chemical sources of electric current (galvanic cells, batteries, fuel cells), radioisotope sources, nuclear reactors, etc. [5].

Despite the relatively low efficiency (up to 13% [5]), the preference is still given to solar panels. Their specific weight is 4.6 kg / kW; in the future it can be reduced by almost 2 times [5]. Flights to Mars, the creation of stations on the Moon and Mars will require much more energy (from 1 MW and more). Space nuclear power can easily solve this problem.

2. Background

2.1. Specific Requirements for a Space Reactor

Unlike ground-based nuclear power reactors for space reactors, it is necessary to provide a high density of energy release in the core with much lower safety requirements, and the possibility of long-term autonomous operation. For reactors and NPPs built on the surface of the planet, minimization of mass and size is not so urgent.

Relatively low safety requirements are associated with the fact that space reactors are designed to operate with an extremely low population density (crew of a ship or an orbital station; personnel of a station located on the planet under study). When placing a NPP on the surface of a planet (the Moon or Mars), it is advisable to bury the reactor under the soil layer. This is necessary to protect against meteorites in a highly rarefied atmosphere or in the absence of an atmosphere. With this arrangement of the reactor, a severe accident would not destroy or damage the entire space station. In addition, long-term dust storms are characteristic of Mars. Martian dust is iron oxide particles about 1.5 microns in size [6].

The situation is aggravated if the reactor is located on a spacecraft or orbital station. A severe accident of a space reactor will lead to huge economic losses. According to various estimates, the cost of the International Space Station is $ 150 billion. The cost of the Mir station is $ 4.2 billion (in 2001 prices). The commercial value of the Soyuz-2.1 disposable spacecraft (without additional modules) is about $ 35 million ($ 20-30 thousand per 1 kg of payload) [7]. The commercial value of the heavy rocket Proton (Russia) and Falcon 9 (USA, SpaceX) is $ 60-65 and $ 62 million, respectively [7]. The Soyuz-2.1 spacecraft is capable of putting up to 8.25 tons of cargo into low orbit, the Proton spacecraft - about 22 tons, the Falcon 9 - 22.8 tons [7]. NASA estimates the cost of a manned flight to Mars at between $ 400 billion and $ 1 trillion [8]. For comparison, the cost of building a land-based NPP with four VVER-1200 reactors (Turkish nuclear power plant "Akkuyu") is $ 22 billion [9]. The cost of building a fast reactor BN-800 (4th power unit of the Beloyarsk NPP, Zarechny, Russia, electric capacity 880 MW) is estimated at $ 2 billion [10]. The cost of building BREST-OD-300 (Seversk, Russia) is estimated at $ 1.4 billion. [11] The high cost of the project is explained by the rise in prices for materials [11].

When designing space reactors, compactness is important: minimum weight and dimensions. This is achievable for liquid metal cooled reactors (LMR) operating on intermediate neutrons (in the resonance spectrum of neurons). One of the advantages of intermediate neutron LMRs is its relatively small size and lighter weight compared to light water reactors (LWRs). This is attractive for the use of LMRs as transport reactors (on ships of surface and submarine fleets, spaceships and space stations).

2.2. Existing Devices and Developments

The first space reactor on intermediate neutrons (SNAP-10A, thermal power 40 kW, electric power 500 ... 650 W, thermoelectric power converter) was developed by Boeing by order of the Air Force and the US Atomic Energy Commission [12]. Highly enriched fuel was used in the core, the moderator was zirconium hydride, and the coolant was a eutectic alloy of sodium and potassium [12]. In the period from 1970 to 1988, 31 spacecraft with the fast reactor "Buk" (thermoelectric energy conversion, electric power 3 kW, service life up to 4400 hours) were launched into low-earth orbits in the USSR. Two experimental Soviet reactors on intermediate neutrons "TOPAZ" (thermionic energy conversion, electric power 5 kW) were launched into space in 1987 and 1988.

Modern concepts of space reactors, performing the functions of a tug, being developed from Russia and the United States, are guided by an electrical power of ~ 1 MW. In Russia, since 2010, research has been carried out within the framework of the project of the state corporations "Roscosmos" and "Rosatom" ("Creation of a transport and energy module based on a megawatt-class NPP") [13]. The installations are designed for flights into deep space, to the Moon and Mars. For the same purposes, the USA is developing a nuclear electric motor plant "Kilopower" with an electric power of 1 kW (in the future, with an increase in power up to 2 MW) [14]. The service life of such installations is designed for 15 years [14].

At present, the development of nuclear electric propulsion systems is carried out in three directions: three types of installations are being developed [2, 3]:

- The first type is based on the concept of a reactor with direct conversion of thermal energy into electrical energy (thermal emission conversion) and an electric rocket motor;
- The installation of the second type is a reactor with machine energy conversion and an electric rocket engine;
- The installation of the third type should operate in two modes: obtaining traction in engine mode and generating electrical energy in a closed loop.

Several conceptual designs of new generation space NPPs have been developed in Russia [4, 15].

- Four projects with electrical capacity from 0.1 to 150 MW for operation as part of a transport an energy module for the study of distant planets of the solar system;
- Four projects of the Martian NPP with electric power from 25 to 500 kW;
- Four projects of the lunar NPP with electric power from 25 to 500 kW.

It is proposed to use a high-temperature fast reactor in all projects.

NASA and the US Department of Energy have selected three nuclear thermal propulsion concepts. The reactors are intended for deep space exploration (see [16] with reference to "World Nuclear News"). The concepts were developed by the following companies.

- "BWX Technologies" и "Lockheed Martin";
- "General Atomics Electromagnetic Systems", "X-energy" and "Aerojet Rocketdyne";
- "Ultra Safe Nuclear Technologies", "Ultra Safe Nuclear Corporation", "GE Hitachi Nuclear Energy", "GE Research", "Framatome" и "Materion".

Companies are encouraged to develop "different design strategies for specific performance requirements that could help deep space exploration" [16].

The plans of the Russian state corporation "Roscosmos" (within the framework of the "Nucleon" project) include the construction of a NPP on Mars to supply the future Russian Martian base with electrical energy [17]. The NPP is planned to be delivered to Mars orbit using the "Zeus" nuclear tug (nuclear rocket engine, 1 MW) [17]. The NPP will descend to the surface of the planet by parachute. Immediately after landing, the NPP will be ready for operation [17]. The developer of the project is the "Arsenal" design bureau (St. Petersburg). The "Arsenal" design bureau includes a pilot production plant. The plant is capable of performing a full cycle of work from the development of the corresponding (design, technological) documentation to the manufacture of finished products. The United States is planning to build a small NPP (with an electrical power of 10 kW, based on the "Kilopower" concept) on the Moon and Mars.

2.3. Operating Conditions

In the absence of a dense layer of the atmosphere in order to minimize the risk of damage to an alien NPP by meteorites, it should be buried in the soil of the planet [1]. In addition, it is necessary to take into account the daily and seasonal temperature drops on the planet's surface [1]. According to [18, 19], daily temperature drops on the lunar surface average 300 ºC: from –173 ºC (at night) to +127 ºC (during the day). During eclipses (lasting about one and a half hours) the temperature of the lunar surface during the day decreases by 250 ... 300 ºC. In lunar craters near the pole, a temperature of –249 ºC was recorded. On the other hand, at a depth of 1 m, the temperature of the lunar regolith is stable and amounts to –35 ºC [18, 19]. This is attractive for burying NPPs into the ground.

The seasonal change in the temperature of the atmosphere near the surface of Mars is 105 ºC: from –80… 125 ºC to + 20… 25 ºC. The maximum temperature spread is 178 ºC: from –143 ºC (in winter at the pole) to +35 ºC [18, 19]. In mid-latitudes, temperatures range from -50 °C in winter nights to 0 °C in summer days [18]. The average atmospheric pressure on the average radius of Mars is 636 Pa (data from NASA, 2004) [18]. Depending on the season, the pressure varies from 400 to 870 Pa [18]. Due to the large difference in altitude on Mars, the pressure at the surface varies greatly. In deep depressions (Hellas plain), atmospheric pressure is 1.24 kPa [20]. Atmospheric pressure on the lunar surface is about 10 nPa [21]. The lower layer of the Mars atmosphere contains carbon dioxide (95.32%), nitrogen (2.7%), argon (1.6%), other gases and traces of organic compounds [22]. The atmosphere of the Moon is composed mainly of hydrogen, helium, neon and argon [21].

One of the features of space NPPs operating in space far from the planet (relatively large gravitational mass) is associated with the absence of natural circulation of the coolant when the pumps fail. The high level of natural circulation in land-based reactors plays an important role (with rare exceptions) in ensuring their safety. An exception to this pattern is a fast reactor with a heavy coolant (lead or lead-bismuth alloy). For such reactors (for example, BREST-OD-300 [23]), a two-circuit energy conversion scheme is adopted, and water is used as a working fluid (in the second circuit). In emergency modes with depressurization of the steam generator tubes, a high level of natural circulation promotes the entrainment of water vapor bubbles into the core, which leads to an increase in reactivity and can lead to an accident.

In reactors located on Mars and the Moon, the level of natural circulation is noticeably lower than in terrestrial reactors. The reason lies in the lower value of the acceleration g of gravity. If on the surface of the Earth the average value of g is 9.81 m / s^2, then on the surface of Mars $g \approx 3.86$ m / s^2, on the surface of the Moon $g \approx 1.61$ m / s^2. The flow rate G of the coolant through the core in the natural circulation mode is determined by the formula [1]:

$$G = [Hg\Delta\gamma/\Delta p]^{1/1,75}, \quad (1)$$

where H is the height of the natural circulation contour (thrust section), equal to the difference in heights between the average levels of the core and the intermediate heat exchanger for

emergency cooling; $\Delta\gamma$ is the difference in the density of the coolant at the inlet and outlet to the core; Δp is hydraulic pressure loss in the primary circuit.

In fig. 1 shows the qualitative dependence of the coolant flow rate in the natural circulation mode on the acceleration of the combined fall, all other things being equal (i.e., at $H\Delta\gamma / \Delta p = \text{const} = B$).

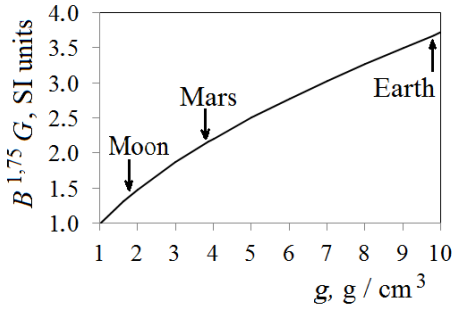


Figure 1: Dependence of the coolant flow rate in the natural circulation mode on the acceleration of the consolidated fall

As noted in [1], "to compensate for the decrease in G, all other things being equal (in comparison with a ground-based reactor), it is necessary to increase H and, consequently, the total height of the NPPt". For reactors located on Mars or the Moon, high values of the height of the natural circulation loop should be chosen. This is possible due to the deepening of the reactor into the soil of the planet. Table 1 shows the characteristic values of H for the lunar and Martian NPPs in relation to H_0 of a terrestrial-based reactor (according to [1]).

Based on the data in table 1, it can be concluded that for a space reactor located at a shallow depth under the surface of Mars or the Moon, acceptable values of the level of natural circulation are achievable to ensure safety in emergency modes with violation of forced circulation of the coolant. Circulation pumps can be eliminated in the design of low-power reactors located on the surface of planets with relatively high gravity.

Table 1: Characteristic values of the height of the natural circulation contour of terrestrial, Martian and lunar-based reactors

The planet on which the reactor is located	H/H_0	Typical values of H, m
Earth	1	5…10
Mars	2.54	up to 10
Moon	6.09	24…25

2.4. Possible Borrowing of Technologies of Ground-Based New Generation of Fast Power Reactors

The design of NPPs (installations capable of operating only in the mode of generating electrical energy) on other planets can be based on borrowing technologies from new generation of ground-based power reactors, taking into account the specifics of the operating conditions of the reactor.

Liquid metal cooled fast reactors (fast reactors or LMFRs) have a higher core energy density than light water reactors (LWRs) and have good potential for improving internal security. Currently, two directions of high-power, ground-based fast power reactors are being developed. The first direction (traditional) involves the use of a liquid sodium coolant. The second direction (innovative) involves the use of a liquid lead coolant. In 2021, Russia began construction of a pilot demonstration reactor BREST-OD-300 with a lead coolant and mixed mononitride fuel (UN-PuN) based on waste uranium and plutonium recovered from spent LWR fuel. Another advantage of fast reactors is the high value of the average number of neutrons ν_f produced during the fission of a heavy nucleus (Fig. 2). This is due to the relatively high kinetic energy of neutrons in the fast reactor.

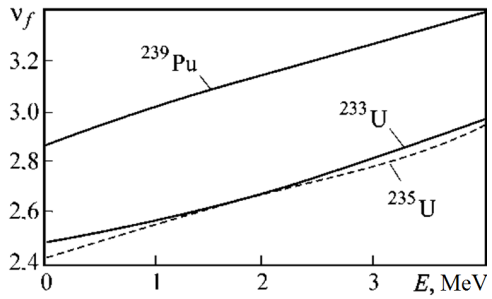


Figure 2: Dependence of the average number of neutrons produced during nuclear fission on the kinetic energy of the neutron that caused the fission [24]

In space reactors, fuel based on highly enriched uranium (without plutonium in the launch load) is used. This is due to some of the properties of plutonium. When using plutonium-239 as the main fissile nuclide, the following advantages are achieved (in comparison with the fissile isotopes of uranium ^{233}U and ^{235}U):

- Maximum energy release;
- Maximum value of parameter ν_f;
- when using mixed fuel (based on ^{238}U and ^{239}Pu), the reactor can operate for a long time (up to 20 years) in a self-fueling mode, i.e., the number of fissile nuclides in the core is constant and practically does not change over time (the BRC - breeding ratio in the core is approximately equal to 1).

These advantages are attractive for the use of mixed fuel (based on ^{238}U and ^{239}Pu) in terrestrial power engineering.

There are two significant disadvantages of plutonium fuel.

- A small fraction of β delayed neutrons produced during the fission of heavy nuclei (during the fission of uranium-235 β = 0.0065, plutonium-239 β = 0.0021 [25]) complicates reactor control and leads to a potential accident hazard (reactor is prompt critical) [24];
- Unpredictable behavior of the crystal lattice (when using metallic plutonium) and a spontaneous change in its shape during operation of the reactor leads to a noticeable change in the volume of fuel [26].

These disadvantages impede the use of plutonium-239-based fuel in transport reactors (space, ship, etc.).

One of the topical and fundamental tasks is the development of space NPPs, combining the advantages of new generation fast power reactors and existing or under development space reactors.

2.5. Tasks solved by a space NPP (based on data [2, 3])

In the near future, space NPPs with an electric power of up to 250 kW are designed to solve the problems of a global communication system, remote sensing of the Earth, environmental monitoring, warning of natural disasters, etc .; comprehensive study of the Earth and solar-terrestrial relations; detailed studies of planets and their satellites; explore the possibilities of industrial production in space; the task of cleaning from space debris; warnings about the approach of dangerous space objects. In the future, space nuclear power (the electric power of the reactor is from 5 to 40 MW) is designed to provide manned flights to the Moon and Mars; sending probes outside the solar system; delivery to Earth and processing of asteroids; combating natural disasters.

3. Materials and Methods

In addition to the general requirements for the choice of materials for space reactors, specific requirements must be taken into account, taking into account the extreme operating conditions. Extreme conditions include high temperatures (usually oriented at 3000 K), intense radioactive irradiation (neutron flux density 10^{14} s^{-1} cm^{-2}), chemically active media (hydrogen in a nuclear rocket engine) [27, 28]. Most of the "traditional" materials for ground-based nuclear technology do not withstand these operating conditions.

In the manufacture of structural materials, preference is given to refractory metals (tungsten, molybdenum, chromium) [29]. Preferred fuels are carbides and nitrides of transition metals, ceramics [29]. To expand the temperature range of operation of fuel and structural materials, introduction phases are used [29-31]: ordered structures that are formed during the interaction of some metals with carbon, nitrogen, hydrogen and oxygen atoms. The disadvantage of interstitial phases is brittle fracture at relatively low temperatures (up to 1000 K) [29-31]. This drawback is eliminated by a special design technology. Materials must be in a compressive stress field. Tensile stresses should be minimized [29, 31].

3.1. "Traditional" promising fuel

Uranium dioxide is rightfully considered the most widespread nuclear fuel. The melting point of UO_2 is 2750 °C [32]. Such fuel is characterized by low density and thermal conductivity. Uranium mononitride melts congruently at 2850 °C if the nitrogen pressure in the system exceeds 250 kPa [32]. Low nitrogen pressure leads to the decomposition of UN to form a liquid uranium phase. The melting point of uranium monocarbide is 2275 °C [32]. Mononitride and monocarbide are characterized by high density and thermal conductivity. This is attractive for the use of such fuels in reactors with a high bulk heat density.

A mixture of uranium and zirconium hydride was used as fuel for the SNAP-10A space reactor. An alloy of uranium with molybdenum was used as fuel for the "Buk" reactor, a uranium dioxide was used for the "TOPAZ" reactor.

Fuel materials (introduction phases) are produced by interaction of heavy metals (Th, U, Pu) with non-metals (oxygen, carbon and nitrogen). The simplest ceramic compounds are UO_2, UC and UN. The advantages and disadvantages of these fuels are well understood. They were studied at the dawn of the development of fast reactors. In the BOR-60 reactor, fuel based on UC, UPuC and UNC was studied [33]. The fuel (UPu) $C_{1-x}N_x$ (where $x < 0.8$), PuO_2-MgO, PuN-CrN, (UZr) N [18, 33], UZrCN (uranium carbonitride) [34] was studied. It is proposed to use highly enriched uranium (up to 90% ^{235}U) [34]. For a nuclear rocket engine, high-temperature fuel based on solid solutions of carbides is considered: UC-ZrC, UC-NbC and ZrC-NbC-UC [35]. The density of ZrC-NbC-UC is relatively low: 7.6 g / cm^3, melting point 3520 K [35]. The carbonitride fuel UC-ZrC-ZrN is being investigated [35]. Fuel based on uranium-zirconium carbonitrides has the highest density among high-temperature fuels.

A space NPP is being developed in the United States. The NPP was developed on the basis of the "Kilopower" project. Uranium metal, UMo alloy, or uranium dioxide is considered as fuel [4]. The installation is designed to generate electricity (up to 10 kW) on the surface of Mars. The service life is 10 years.

So, in Russian space reactor designs, UZrCN carbonitride fuel is considered the most promising. The multi component composition opens up possibilities for optimizing the composition and properties of such a fuel. In the USSR, high-temperature and high-density carbonitride fuel based on highly enriched (96%) and low-enriched (less than 20%) uranium was developed for various types of reactors [36].

Scientists from Belarus, Russia and the United States are conducting joint computational and experimental studies of 0.9U-0.1Zr-0.5C-0.5N fuel (with 19.75% enrichment in ^{235}U) with a density of 12 ... 12.5 g / cm^3 and thermal conductivity

$$\lambda = 11,4 \text{ W m}^{-1} \text{ K}^{-1}$$

for the Russian high-flux research reactor SM-3 [36, 37].

3.2. New Promising Fuel

When designing Martian and lunar NPPs (in fact, space nuclear power plants), borrowing technical solutions suggests that the cermets fuel previously considered by the author for ground-based BN and BREST power reactors can be used in space NPPs [1]. This fuel is sintered pellets of mixed mononitride fuel micro grains (UN-PuN with additives of ^{237}Np, ^{241}Am, ^{243}Am long-lived waste) and uranium metal nanopowder. The composition of UN uses natural (^{235}U content about 0.72%) or depleted (^{235}U content about 0.4%) uranium. PuN uses plutonium recovered from LWR spent fuel and purified from the ^{238}Pu isotope. Plutonium isotopic composition: 0.6 ^{239}Pu - 0.25 ^{240}Pu - 0.109 ^{241}Pu - 0.041 ^{242}Pu. The Pu / (Pu + U) ratio does not exceed 0.2. The mass content of radioactive waste (^{237}Np, ^{241}Am and ^{243}Am) in fresh fuel is about 5%.The total content of fissile nuclides (^{235}U, ^{239}Pu and ^{241}Pu) in the fuel is about 16%.

The nanopowder occupies the voids between the micro grains of the ceramic. As a result, the porosity of the fuel is reduced to 5%. "The ratio of the volumes of micro grains of ceramics and metal nanopowder 20/75 corresponds to a mass ratio of 0.39. This is the maximum permissible mass content of metallic nanopowder in the fuel, associated with the limitation of the maximum permissible temperature" [1]. When using enriched uranium

nanopowder, the reactor power at fixed dimensions can be increased by 40%.

The average density of such a fuel is about 16 g / cm^3, the thermal conductivity is about 27.8 W m^{-1} K^{-1}, which significantly exceeds the corresponding values for mononitride and carbonitride. This contributes to the safe end of emergency situations, including those accompanied by a failure of the emergency protection (ATWS - anticipated transient without scram).

When the fuel temperature rises in emergency modes, melting of uranium nanopowder is possible. However, it will not go beyond the fuel element cladding. At the end of the emergency mode, the nanopowder will again go into the solid phase.

Uranium metal nanopowder acts as a getter of free nitrogen released from micro grains during reactor operation. Nitrogen migrates to the fuel element cladding. In the presence of free nitrogen, the corrosion rate of the inner surface of the shell increases. Getter binds free nitrogen, helping to minimize the rate of corrosion of the inner surface of the shell. After a long-term operation of the reactor, which "survived" emergency situations, uranium metal gradually loses its functions of an effective getter. Its functions in ensuring the safety of the reactor (due to its high thermal conductivity and density) are preserved.

Optimization of fuel properties is possible by changing the ratio of micro grains of mononitride and uranium nanopowder, as well as by changing the enrichment of uranium nanopowder.

From the point of view of ensuring safety in emergency situations (including the most dangerous of them - ATWS), it is sufficient to ensure the mass content of uranium nanopowder in the fuel at the level of 15 ... 20%. From the point of view of maximizing the reactor power, it is advisable to increase the nanopowder content to the limiting value (39%). This value corresponds to the filling of the pores between the micro grains of the ceramic. When melting a nanopowder in emergency situations, liquid droplets should not completely surround the ceramic micro grains.

Particular attention is paid to solving the problem of non-proliferation of materials that can be used in the creation of weapons of mass destruction. As noted, when using a mixed mononitride fuel, a self-fueling regime is achieved, i.e., the breeding ratio in the core is approximately equal to 1. On the one hand, at all stages of the fuel cycle, the need to separate isotopes of heavy nuclei is completely eliminated. The burnout of the fissile isotopes of uranium (a small amount of ^{235}U) and plutonium (^{239}Pu, ^{241}Pu) is compensated by the production of ^{239}Pu from ^{238}U. On the other hand, NPP fuel contains material for direct use. This is attractive to the terrorist kidnappers. But terrorists are unlikely to be able to reach the Moon or Mars. If they can, then these will only mean that they are representatives of states with high technologies and (most likely) possessing nuclear weapons. If necessary, they will be able to produce the required amount of materials for direct use on Earth. It is extremely difficult to steal irradiated fuel characterized by a high background radiation, especially in space.

3.3. Traditional Coolants

The eutectic alloy NaK was used as a coolant for the first space reactors SNAP-10A, "Buk" and "TOPAZ".

In fast space reactors of the second generation, it was proposed to use the eutectic alloy NaK (electric power 70 ... 150 kW) and lithium (two reactors with electric power 80 ... 150 and 160 ... 400 kW) as a coolant [38]. The eutectic NaK and ^7Li (not used in space) can be considered traditional coolants for space reactors with liquid metal cooling [39].

Alkali metals are characterized by the highest chemical activity. The activity towards oxygen increases from lithium to cesium. In the series of alkali metals from lithium to cesium (Li, Na, K, Rb, and Cs), the melting point decreases from Li to Cs, the chemical activity increases, and the hardness weakens. Of all alkali metals, only lithium interacts with water without exploding and has the highest melting point. There is practically no oxygen in the lunar and Martian atmosphere. Carbon dioxide (the main constituent of the Mars atmosphere [22]) oxidizes all alkali metals.

Table 2 shows the melting and boiling points of some liquid metal coolants with a relatively low fast neutron absorption cross section (according to [40–42]).

Table 2: Operating temperatures of some liquid metal coolants

Coolant	Melting point, ºC	Boiling point, ºC
Lithium	180.54	1339.85
Sodium	97.86	883.15
Potassium	63.7	758.85
Mercury	-36	357
Bismuth	271	1490
Lead	327.4	1740
77,2 % K - 22,8 % Na	−12.6	785
55,5 % Bi - 44,4 % Pb	123.5	1670

A NaK alloys are chemically active, low-melting, and flammable in air. With a mass fraction of potassium of 40 ... 90%, the alloy remains liquid under normal conditions [41]. An alloy of 56% K and 44% Na is also a eutectic.

3.4. Alternative Coolants. Sodium, Potassium and Cesium Alloys

Is there an alternative to lithium-7 and NaK alloy used for cooling the core of space NPPs?

In the reactors of nuclear submarines, a coolant based on a eutectic alloy of blue bream with bismuth was used [43]. (Eutectic is a liquid solution that crystallizes at the lowest temperature for the alloys of this system.) The first such reactor (BM-40A) of the nuclear submarine "Lira" was lighter than traditional reactors of the LWR type by 300 tons at the same power (155 MW) [43]. An eutectic alloy of lead and bismuth provides a high level of inherent safety in relation to accidents [44]. The operating temperature range of such a coolant is wide: from 123.5 °C to 1670 °C [40, 41]. The technology of using such an alloy as a coolant for a nuclear reactor has been developed [43]. The PbBi eutectic alloy cooled reactor is intrinsically safe. In such a reactor, severe accidents can be excluded. The use of the PbBi coolant will require the reliable operation of the heating system of the idle reactor circuit.

Gallium has a maximum operating temperature range. The freezing point is 30 °C; the boiling point is 2230 °C [42]. Gallium is classified as a rare element. It is highly dispersed in the earth's

crust (on average, its mass content in the earth's crust is about 19 ppm) [42]. Gallium is extracted from zinc and aluminum ores. Gallium is attractive as a high-temperature heat carrier. But gallium is a very expensive coolant. In addition, gallium is highly corrosive to traditional structural materials for fast reactors (AISI 316 SS, Russian stainless steel "12X18H9") [44]. Gallium has never been used as a coolant in nuclear reactors [44]. Gallium is characterized by a relatively high fast neutron absorption cross section. When using mononitride fuel or cermets based on micro grains of mononitride and uranium metal nanopowder, it is possible to ensure a good balance of neutrons in the reactor core. The fuel self-sufficiency mode is easily achievable, i.e., the fuel breeding ratio in the core is slightly more than 1. However, the void reactivity effect is positive (when bubbles are drawn into the central part of the core) even in low-power reactors (electric power of 150 MW and more).

Among alkali metals, NaKCs alloys with different concentrations of components are attracted. Sodium, potassium and cesium are highly reactive. The NaKCs alloy forms a eutectic with a sodium content of 12% (at.), Potassium 47%, and cesium 41%. The properties of such an alloy are well studied [44–46]. The melting temperature of such an alloy ("Soviet alloy") is minimal for all metals and is –78 ºC [47]. This eliminates the need for heating the coolant in the idle reactor. It is possible to exclude freezing of the coolant in emergency modes such as OVC WS (overcooling accident without scram). Alloy NaKCs with a certain content of components are characterized by a much lower chemical activity than their individual components. It was found experimentally that NaKCs alloys with a certain content of components "are capable of A NaK alloys are chemically active, low-melting, and flammable in air. With a mass fraction of potassium of 40 ... 90%, the alloy remains liquid under normal conditions [41]. An alloy of 56% K and 44% Na is also a eutectic.

"NaKCs alloys with a certain content of components are capable of oxidizing in air at 300-1000 K without intensive release of aerosols and a noticeable increase in temperature" [48]. Eutectic alloy NaKCs does not ignite in the temperature range 293 ... 973 K [45]. Thus, by changing the concentration of Na, K and Cs in the alloy, it is possible to control the properties of this alloy. The manufacturing technology of NaKCs alloys is presented in [45].

The melting point, density, electrical conductivity, electric potential, reactivity, plasticity, and flammability of the NaKCs alloy depend on the concentration of the components [45]. These differences are a consequence of differences in the interactions between the components of the alloy during its formation [45]. In addition to the eutectic, alloys 20% (at.) Na - 30% K - 50% Cs and 40% (at.) Na - 30% K - 30% Cs have been well studied. Melting point of 20% Na - 30% K - 50% Cs alloy is equal minus 37.90 ºC [45]. Melting point of 40% (at.) Na - 30% K - 30% Cs alloy is equal 0.00 ºC [45].

Ignition of liquid metal occurs due to a sharp violation of thermal equilibrium. This leads to a self-accelerating temperature rise [45]. The ability of a metal to ignite is largely determined by the characteristics of the oxide film formed on the surface [45].

When analyzing three-component systems (liquid alloys), appropriately delineated triangles are used (they are called Gibbs-Rosebohm triangles or Rosebohm triangles). On such a triangle for the NaKCs alloy, it is possible to distinguish the concentration range corresponding to the non-flammability of the alloy [48]. The boundary of the region is determined by the equation of an ellipse with the ratio of the semi axes 2.4 m center at the point corresponding to an alloy close to eutectic [48].

The center of the ellipse is determined by the content of sodium 12% (at.), potassium 44%, cesium 44% (or the mass content of 3.52% Na - 21.88% K - 74.60% Cs) [48]. The boundaries of the non-flammable range (% at.): 2 ... 24 Na - 19 ... 69 K - 19 ... 69 Cs or (% wt.) 0.5 ... 7.8 Na - 7.6 ... 43.7 K - 46.5 ... 89.2 Cs [48]. The region of non-flammability is determined by the inequality [48]:

$$(6C_{Na} - 1)^2 - 8C_K C_{Cs} + 1 \le 0, \qquad (2)$$

where C_{Na}, C_K and C_{Cs} are the atomic concentrations of sodium, potassium and cesium, respectively.

Sodium and potassium are considered the most common in the earth's crust (the mass content of sodium in the earth's crust is 2.3%; potassium is 2.1% [49]). Cesium is a rare element (the mass content in the earth's crust is $3 \cdot 10^{-4}$% [42]). The volume of world production of cesium is small: 20 tons / year [49]. Increase in cesium production is possible. It can be extracted from spent nuclear fuel (~ 10 t / year).

In [50], with reference to [51], the total yield of relatively long-lived cesium isotopes during the fission of uranium and plutonium nuclei is given (Table 3). This yield is quite large: from 6.2 to 7.5%.

Table 3: Half-life and total (cumulative) yield of some isotopes of cesium during fission of heavy nuclei

Parameter	Cesium isotope		
	^{133}Cs	^{135}Cs	^{137}Cs
Half-life, years	Stable	$2.3 \cdot 10^6$	30.0
The yield upon fission of ^{235}U nuclei by a thermal neutron, %	6.6149	6.5803	6.2444
The yield upon fission of ^{235}U nuclei by a fast neutron, %	6.6100	6.3696	5.9962
The yield upon fission of ^{239}Pu nuclei by a thermal neutron, %	6.9034	7.2411	6.5090
The yield upon fission of ^{239}Pu nuclei by a fast neutron, %	7.0466	7.5384	6.3534

The distribution of the yield w of fission fragments over the mass numbers A is determined by the presence of about magic, magic, and doubly magic nuclei. According to the shell structure (model) of the atomic nucleus, asymmetric fission is a consequence of the predominant formation of fission fragments with filled neutron and proton shells.

The w (A) curve corresponding to the fission of uranium or plutonium contains two maxima. This reflects the well-known fact that fission into two fragments with a mass ratio of 1.6 is most

likely. The first maximum corresponds to the values $A \approx \approx 90$ ... 100, the second corresponds to the values $A \approx 130$ ... 140 (Fig. 3). The yield of cesium isotopes illustrates the second (right) maximum.

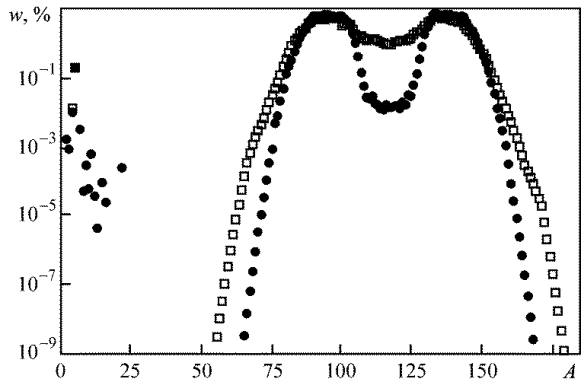


Figure 3: Distribution of the yield of fission fragments of uranium ^{235}U by atomic masses A [24]: □ - fission of nuclei by neutrons with an energy of 14.06 MeV; ● - fission of nuclei by neutrons with an energy of 0.0253 eV

Cesium is strongly activated in the core of a nuclear reactor [40, 44]). This is a serious problem that prevents the use of such alloys as coolants for ground-based power reactors. The use of cesium as a component of the NaKCs alloy, if it can be justified, is only by the non-flammability of such a coolant. However, the specific operating conditions of the reactor on the surface of Mars or the Moon (extremely low population density - the personnel of the alien station) make it possible to use radioactive waste from terrestrial nuclear power engineering as reactor materials (fuel and coolant).

Due to the wide range of operating temperatures at an extremely low freezing point and a relatively high boiling point (721 ºC [47]), the coolant based on the NaKCs eutectic is of great interest for space reactors. You can eliminate the need for heating the coolant in the idle reactor. Freezing of the coolant in OVC WS is excluded.

The use of such a coolant will make it possible to exclude the combustion of the coolant (as a result of which the void reactivity effect is realized), and the boiling problem can be solved due to the optimal choice of the reactor layout parameters. For medium to high power LMFRs, there is a problem with the positive void reactivity effect. However, in low-power reactors, this effect is negative. It can be expected that the use of such a coolant will make it possible to create a safe space reactor.

3.5. The Research Methods

The author uses calculation and optimization methods and codes to substantiate the possibility of creating a space reactor for an alien NPP. The author's codes [1] and well-known codes [52, 53] are used:

- FRISS-2D for simulation of emergency modes (including ATWS) [1];
- Multiple modernized calculation and optimization code Dragon-M [1];
- MCU code for precision neutron-physical calculation [52];
- WIMS code for neutron-physical calculation of a fuel element cell [53].

The FRISS-2D mathematical model is based on the solution of the following equations:

- Systems of equations of point neutron kinetics with a six-group description of delayed neutrons (for the relationship between power and reactivity);
- Non-stationary heat conduction equation, which is the heat balance equation, and the coolant energy equation (to determine the relationship between the reactor power and changes in the fuel and coolant temperatures);
- balance equation of reactivity (linking temperature changes with reactivity).

One of the versions of the FRISS-2D code can work as part of the Dragon-M computational and optimization code.

Dragon-M code is intended for solving mathematical programming problems in a deterministic setting and under conditions of uncertainty in the initial data. To solve optimization problems, the method of sequential linearization is used [54]. The code allows you to solve problems with a given criterion of optimality. The restrictions for the functionals characterizing the economic efficiency of the NPP, reliability and safety are considered. The emergency modes of the ATWS type of reactor operation are taken into account. Comprehensive optimization is carried out: the code contains modules for neutron-physical calculation (in the multi-group diffusion approximation), calculation of changes in the nuclide composition of fuel, thermal-hydraulic calculation, strength calculation, and assessment of the economic characteristics of NPPs. A homogeneous cylindrical reactor consisting of radial and axial zones differing in composition is considered. The control parameters include the geometric characteristics of the fuel element array, the fuel composition, the coolant flow rate, the dimensions of the reactor zones, etc.

The MCU code allows obtaining the solution of the stationary gas-kinetic equation of neutron transport (Boltzmann equation) by the Monte Carlo method.

In the calculations using the WIMS code (WIMS-D5B), the S4-approximation of the discrete ordinate method was chosen to solve the integral neutron transport equation (Peierls equation).

The author has developed the FRISS-2D, Dracon-M codes and a number of auxiliary codes. The author for the first time included in the optimization problem the constraints for functionals that simulate the safe termination of emergency processes (including ATWS). Limitations are considered for extreme temperatures of fuel, coolant, fuel-element cladding, power, pressure in the cavity for collecting gaseous fission products, etc.

As an optimality criterion, it is customary to consider the functional characterizing the economic efficiency of a NPP. The rest of the functionals can be taken into account among the limitations of the mathematical programming problem. However, when safety functionals are taken into account, the area of safe layouts in the space of control parameters is sometimes so small

that the choice of the economic criterion of optimality does not matter. The criterion related to the safety of the reactor can be selected as the target functional (for example, the void reactivity effect). The control vector includes the characteristics of the core.

The FRISS-2D and Dracon-M codes assume that the reactor consists of several radial and axial zones of homogenized composition. The Dracon-M code has no analogues yet.

With the help of the developed software, the author for the first time obtained the results of optimizing the safe layouts of ground-based and alien-based LMFRs, taking into account safe operation. In such reactors, severe accidents can be ruled out.

The software developed by the author allows solving computational and optimization problems in conditions of uncertainty, including the uncertainty of emergency scenarios. This is important at the initial stages of nuclear reactor design (when its characteristics are not finally determined).

4. Results

4.1. Initial Provisions

The relatively low chemical activity of the NaKCs alloy in comparison with its components makes it possible to consider a two-circuit energy conversion scheme. The first circuit is filled with liquid NaKCs alloy, the second - with water. The transition to a two-circuit scheme will greatly simplify the design and reduce the cost of NPPs.

The most dangerous are emergency modes, accompanied by the failure of emergency protection (ATWS). As you know, all emergency modes of LMFR operation can be triggered by the following events.

- Draining the core or part of it (including entrainment of bubbles into the core) - modes of the LOCA WS (loss of coolant without scram) type;
- Input of limited value of positive reactivity - modes of TOP WS (transient overpower without scram) type;
- Violation of the forced circulation of the coolant in the primary circuit - modes of the LOF WS (loss of flow without scram) type;
- Violation of heat sinks from the first circuit to the second - modes of the LOHS WS type (loss of heat sink without scram);
- Overcooling of the primary coolant - modes of the OVC WS type.

The most dangerous are the combinations of these modes.

It is assumed that the reactor of the Martian NPP has an electrical capacity of 600 MW. (This is an abnormally high power for such a NPP.)

The traditional layout of the core is considered. It is assumed that there are two zones of profiling along the radius of the reactor. The initial loading of the reactor assumes the presence of enriched uranium and does not contain plutonium isotopes. Plutonium builds up as the reactor runs.

Both zones contain uranium mononitride fuel of different enrichment (11 and 15% for the central and peripheral zones, respectively) with the addition of long-lived transuranic elements (5% by weight ^{237}Np, ^{241}Am and ^{243}Am). A nanopowder of metallic uranium (39% by weight) with an enrichment of 16% in ^{235}U is placed in the fuel pellet between the micro grains of the mononitride.

In order to increase the flow rate of the coolant in the natural circulation mode (when the main circulation pumps are de-energized), it is necessary to increase the effective height of the natural circulation circuit in comparison with ground-based NPPs. To ensure the required level of natural circulation up to 14 ... 15% (which is required for the trouble-free completion of the LOF WS mode), it is necessary that the height difference between the averages levels of the core and the emergency cool-down heat exchanger be 23 ... 25 m. This can be easily achieved when the reactor is buried under soil of the planet.

Ground-based fast reactors, cooled with sodium or NaK alloy, operate and operated at a pressure in the core close to the Earth's atmospheric pressure (0.1 MPa). The average atmospheric pressure at the surface of Mars is about 636 Pa [18], i.e., almost 160 times less than at the surface of the Earth. The boiling point of the coolant decreases with decreasing operating pressure. For this reason, it is necessary to provide pressure in the core at the level of the earth's atmosphere, i.e., 0.1 MPa. The result is a new type of reactor called the LMFR under pressure (PLMFR).

Reactors of the LMFR type with increased pressure in the core are being developed in Russia. An example is the lead-cooled BREST-OD-300 reactor under construction [23]. Under a thick and heavy layer of lead, the pressure in the core is about 1 MPa or 10 atm. This does not preclude the safety of the reactor.

4.2. Optimal Layout of the Martian Reactor

The calculations were performed using the Dragon-M code, which includes the FRISS-2D modules. The void reactivity effect was chosen as an optimality criterion. When the entire reactor is drained, it is negative. The WIMS code is used to calculate some of the coefficients of reactivity. The MCU code is used to refine the meaning of the void reactivity effect. (The diffusion approximation used in the Dragon-M code can lead to large errors in estimating this effect.)

It is assumed that the reactor core operates in a self-fueling mode. The breeding ratio in the core is slightly more than 1 (which is necessary to compensate for the production of "slag").

Several layouts of the Martian reactor core were obtained. Two types of fuels based on cermets were considered. The first is a mixture of micro grains MN and nanopowder U, the second is a mixture of micro grains MOX and nanopowder U. The nanopowder fills the pores between micro grains of the ceramic. In both cases, the condition BRC ≈ 1 is satisfied. This limitation takes into account the minimum reactivity margin for fuel burnup.

The void reactivity effect was considered as an optimality criterion, realized when draining the reactor.

The core contains two radial zones surrounded by a reflector. The zones differ in the value of the fuel enrichment. The initial charge is formed from enriched uranium and does not contain

plutonium. The fuel rods are located at the nodes of the triangular lattice. Perforated covers for fuel assemblies are used.

The reflector consists of two zones, assembled from assemblies containing steel tubes with long-lived radioactive waste (^{14}C, ^{99}Tc). The assemblies are washed with a coolant. The first rows of assemblies (closest to the core) contain ^{99}Tc, while the far rows contain a mixture of ^{14}C and ^{99}Tc powders. Carbon-14 slows down neutrons, which contributes to the creation of favorable conditions for the transmutation of ^{14}C and ^{99}Tc. In addition, the presence of such nuclides in the reflector allows the thickness of the reflector to be reduced. The end reflectors are technetium-99 tablets. The height of the upper and lower reflectors is 20 cm each. The lower reflector may be absent.

Two optimal versions (A and B) of the MN-U-fuel core have been proposed.

Option A corresponds to the highest inlet temperature value. The total radius of the core and side shield is limited to 2.6 m. The thicknesses of the side and end reflectors are specified and did not change during the optimization process.

For both variants the diameter of the fuel pellet is 5 mm. The fuel rods are located at the nodes of the triangular lattice.

The grating spacing is 5.75 and 6.2 mm for two zones, respectively (Option A). The height of the fuel column is 1 m. The outer radius of the core is 2.15 m. The temperature at the entrance to the core is 198 ºC. Fuel enrichment with respect to ^{235}U is 10.9 and 14.5% for the central and peripheral zones, respectively. The radius of the core is 129 cm. The radial dimensions of the two zones (with fuel of different enrichment) are 86.7 and 42.3 cm, respectively. When the reactor is operating at rated power, the maximum fuel temperature does not exceed 641 ºC, the coolant temperature does not exceed 516 ºC, and the fuel element cladding does not exceed 524 ºC. The presence of long-lived radioactive cesium-135 in the coolant (decaying into stable barium-135) and short-lived cesium-137 (decaying into metastable barium-135 and with the subsequent emission of γ-quanta) require taking into account the possibility of contamination of the coolant with barium. The melting point of barium is 727 ºC. There is no data on the formation of intermetallic compounds in the Ba-Na, Ba-K, Ba-Cs systems.

Thus, the presence of barium in the coolant can cause increased erosion of structural materials (in particular, fuel-element cladding). This is a payment for the use of "free" radioactive cesium-137 as part of the coolant. For this reason, the speed of the coolant in the core should be limited. In the considered options A and B, the average speed of the coolant does not exceed 3.2 and 3.5 m / s, respectively. (In sodium reactors, the average speed does not exceed 5 m / s.)

There are three ways to solve the problem:

- limiting the speed of the coolant in comparison with a sodium reactor;
- refusal to use cesium-137 as part of the coolant;
- use of cladding of fuel elements with tungsten sputtering (the technology is based on low-temperature plasma spraying of tungsten powder).

Options A and B are taught using the first path. In the options under consideration, the average speed of the coolant does not exceed 3.2 and 3.5 m / s, respectively. (In sodium reactors, the average speed does not exceed 5 m / s.)

Option B was obtained with the same lattice of fuel elements in two radial zones. This is a less compact core. The fuel element lattice spacing is 6.3 mm. Fuel enrichment in 235U in zones is 9.9 and 13.5%. The radius of the core is 144.5 cm. The radial dimensions of the two zones (with fuel of different enrichment) are 111.0 and 33.5 cm, respectively. The height of the fuel column is 0.92 m. The temperature at the inlet to the core is 198 ºC. When the reactor is operating at rated power, the maximum fuel temperature does not exceed 684 ºC, the coolant temperature does not exceed 502 ºC, and the fuel cladding does not exceed 527 ºC.

4.3. On the role of Doppler void and density coefficients and reactivity effects in ATWS

Doppler broadening of resonances in the energy dependence of the neutron-matter interaction cross sections plays an important role in ensuring the self-protection of reactors. This broadening is characterized by the Doppler effect and the coefficient of reactivity. If resonances are present in the energy dependence of the cross sections, then with increasing temperature they broaden (the width increases and the amplitude decreases). With a significant increase in temperature, narrow isolated resonances can overlap. If the kinetic energy of a neutron after collision (scattering) with a nucleus becomes comparable to the energy of some resonance, a nuclear reaction, in the energy dependence of which this resonance is present, is realized with a high probability (a neutron is captured by a nucleus, a compound nucleus is formed, decaying along some channel). If the kinetic energy of the neutron corresponds to the energy between resonances, where the cross section is close to zero, the neutron continues to slow down until its energy coincides with the energy of another resonance (up to the width of the level) and the neutron enters into a nuclear reaction.

If a neutron is captured by the uranium-238 nucleus, then one neutron decreases in the reactor. If a neutron is captured by a uranium-235 nucleus, then the fission of this nucleus with the production of new ν_f (as a rule, $\nu_f > 1$) neutrons is possible, and the number of neutrons in the system increases. (In the resonance region, ^{238}U nuclei practically do not fission.) Thus, broadening of resonances in the energy dependence of the cross sections for ^{238}U leads to a decrease in the number of neutrons in the system, i.e., to a decrease in the effective neutron multiplication factor and reactivity. Broadening of resonances in the energy dependence of the fission cross section for fissile fuel nuclei (^{235}U, ^{239}Pu, etc.) promotes an increase in the number of fissions and, consequently, the number of neutrons, and hence, an increase in the effective neutron multiplication factor and reactivity. The first case corresponds to a negative Doppler reactivity effect, the second to a positive one. Thus, the sign and significance of the Doppler reactivity effect are mainly determined by the nuclei on which the broadening of resonances predominates.

If the reactor uses fuel with a relatively low enrichment in the fissile nuclide, then with an increase in the fuel temperature, the effect of broadening of resonances on the raw material nuclides (^{238}U, etc.) dominates. This effect is negative. It has a beneficial effect on the safety of the reactor. In reactors with high enrichment

(some research, transport), the dominant effect is associated with the broadening of resonances on the dependence of the fission cross section for fissile nuclides, which leads to an increase in the number of neutrons and a positive Doppler reactivity effect. For this reason, it is proposed to use fuel with a relatively low enrichment for LMFR (about 16%) in our space reactor.

It should be noted that in LMFRs, the negative Doppler reactivity effect can compensate for the positive void reactivity effect, since when the core is drained, the heat removal from the fuel elements deteriorates and the fuel temperature increases. Depending on the design of the LMFR and the type of emergency, the role of the Doppler coefficient of reactivity is different. The generally desirable negative Doppler coefficient of reactivity can in some cases exacerbate an accident. This remark applies primarily to LMFRs with medium to high power oxide fuels. This is well known. In such reactors, an increase in the absolute value of the negative Doppler coefficient leads to a favorable termination of emergencies caused by an unauthorized increase in the reactor power (for example, due to the introduction of limited positive reactivity or overcooling of the coolant), and the aggravation of emergency processes initiated by a decrease in the coolant flow rate (for example, as a result of blackout main circulation pumps).

It is known that the negative Doppler coefficient of reactivity helps to minimize the deviation of power and maximum fuel temperature from the nominal values. Fig. 4 illustrates the role of the Doppler coefficient of reactivity with decreasing and increasing maximum power W and maximum fuel temperature T_f. (These parameters are considered among the functionals of the optimization problem.) The dashed line corresponds to the values of these functionals in the nominal operating mode of the reactor. The arrows indicate the trend in the change in the functionals (ΔW or ΔT_f) with an increase in modulus of the negative Doppler coefficient of reactivity. It can be seen that the negative Doppler coefficient of reactivity, which is large in absolute value, plays a favorable role. For this reason, a negative Doppler factor contributes to the safe termination of the TOP WS emergency mode (initiated by injection of limited positive reactivity). Moreover, the higher this coefficient in absolute value, the smaller the deviation of the temperature and power of the reactor in emergency mode.

The nature of the change in the maximum fuel temperature over time during the LOF WS process depends on the type of fuel used (the relative amount of light and heavy atoms). As a consequence, the nature of the change in the maximum fuel temperature depends on the temperature difference between the fuel and the coolant in the nominal mode (before the start of the emergency). If this difference is large (for example, when using oxide fuel), the following scenario is realized. As the flow rate decreases, the coolant temperature rises and the temperature difference between the fuel and the coolant decreases. Consequently, the heat flux from the fuel to the coolant decreases. Under the influence of negative feedback on reactivity due to the thermal expansion of the core, the thermal power of the reactor decreases. The power decreases with time faster than the heat flux from the fuel to the coolant, i.e. more heat is removed from the fuel than is generated in it. As a result, the maximum fuel temperature decreases over time, which leads to a positive contribution to reactivity from the Doppler reactivity effect. To neutralize the latter, a higher temperature of the coolant and structural materials of the core is required. A negative Doppler coefficient of reactivity always prevents the power from deviating from the nominal value (in this case, it prevents it from decreasing) and leads to a stronger heating of the coolant. Thus, in order to increase the self-protection of a fast reactor with oxide fuel from accidents of the LOF WS type, it is necessary to decrease the negative Doppler coefficient of reactivity in absolute value. This fact was known at the dawn of the development of LMFR technologies and was described by many authors. However, the ATWS modes were not considered.

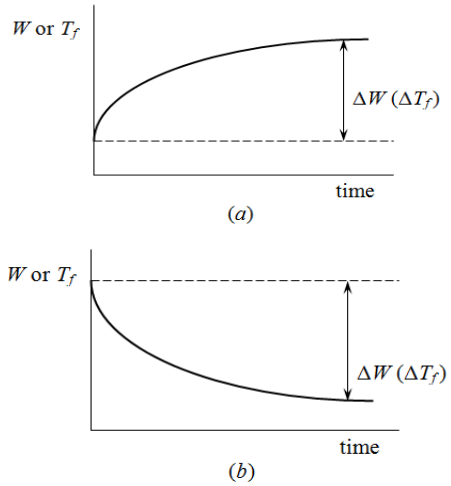


Figure 4: Influence of the Doppler coefficient of reactivity with decreasing (*a*) and increasing (*b*) functional W or T_f

When using metallic fuel, the temperature difference between the fuel and the coolant in the nominal mode is small and the temperatures of the fuel and the coolant in the LOF WS process increase. In this case, negative Doppler reactivity coefficient and the effect of thermal expansion of the core prevent an increase in temperatures, and in order to increase self-protection from accidents, it is necessary to increase the negative Doppler coefficient in absolute value, i.e., the Doppler reactivity effect plays a favorable role. This is also known.

In reactors with nitride or carbide fuel, one of these scenarios can be realized or an intermediate case, when the fuel temperature increases, reaches a maximum, and then decreases to a value below the nominal value. Different scenarios of changing the maximum fuel temperature can be realized in different grading zones.

When using the fuel considered in this work (UN-PuN-U or UN-U) with a relatively high density and thermal conductivity, the role of the Doppler reactivity coefficient in the emergency modes LOF WS and TOP WS is the same. An increase in the absolute value of the negative Doppler coefficient of reactivity contributes to an increase in the self-protection of the reactor from such accidents.

The results of the analysis of the Doppler, void and density coefficients and reactivity effects are presented in [1]. In fig. 5 shows the dependence of the Doppler reactivity effect on the deviation of the fuel temperature from the nominal. Fig. 6 illustrates the dependence of the spectral component of the density

effect on the density of the coolant. Zero density in Fig. 6 corresponds to the void reactivity effect.

Dependence *1* in Fig. 5 and 6 correspond to traditional mixed mononitride fuel (porosity 25%). Dependences *2*, *3* and *4* correspond to fuel (5% porosity) based on a mixture of micro grains UN-PuN and metallic uranium nanopowder (dependence *2* corresponds to waste uranium in the composition of a nanopowder; *3* - enriched to 20% in ^{235}U; *4* - enriched to 80% in ^{235}U). Obviously, the point of intersection of all straight lines corresponds to the reactor operating mode at rated power.

4.4. ATWS Analysis

The LOF WS emergency mode is initiated by de-energizing all main circulation pumps. For variant A, with a pump run-on time of 30 s, the maximum temperature of the heating medium reaches 694 ºC, which corresponds to a 27-degree boiling margin. For variant B it is 6 ºC higher than the boiling point. With an increase in the run-on time of the pumps to 40 s, the maximum temperature of the heat carrier is 709 ºC. In the steady-state mode of natural circulation, the maximum temperature of the heat carrier decreases by 68 and 58 ºC for options A and B, respectively. The maximum fuel temperature in the LOF WS transition mode (with a pump run-on time of 30 s) reaches 714 and 762 ºC for options A and B, respectively. The reactor power is reduced in LOF WS mode.

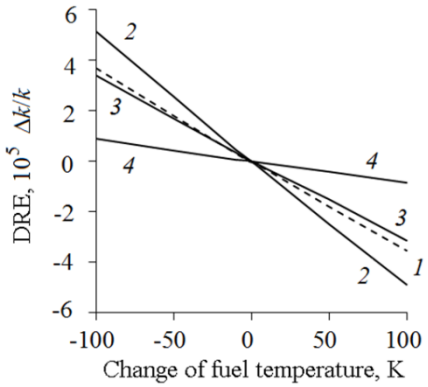


Figure 5: Dependence of the Doppler reactivity effect (DRE) on the deviation of the fuel temperature from the nominal [1]

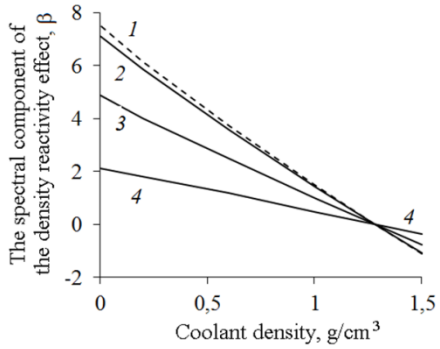


Figure 6: Dependence of the spectral component of the density and void reactivity effects on the density of the coolant [1]

The TOP WS emergency mode is initiated by the input of reactivity 0.9 β (where β is the effective fraction of delayed neutrons) for 10 s according to the linear law. This corresponds to the erroneous (unauthorized) removal from the core of both absorbing rods - reactivity compensators. The maximum fuel temperature increases in the transient mode to 777 and 789 ºC for options A and B, respectively. The maximum coolant temperature is 577 and 562 ºC for options A and B, respectively. The steady state power (under the action of reactivity feedbacks) is approximately 24% higher than the nominal power.

In OVC WS mode, the heating medium temperature decreases over time. A significant temperature margin is maintained until the coolant freezes. When the main circulation pumps are converted to increased performance (with an increase in flow rate by 20%), the maximum fuel temperature reaches 733 and 767 ºC for options A and B, respectively. The power increases by 58 and 24% (respectively) of the nominal. This is a hypothetical emergency that practically cannot be realized. When a “cold” backup loop is connected (if it is provided), the maximum fuel temperature is 746 and 787 ºC for options A and B, respectively.

LOHS WS emergency mode is initiated by de-energization (or failure) of all secondary circulation pumps. Even if a coolant with a temperature equal to the outlet temperature is supplied to the entrance to the core, the emergency mode is not dangerous. Fuel temperature decreases over time. In the LOHS WS emergency mode, the maximum temperature of the coolant is 535 ºC. The maximum fuel temperature decreases over time.

When the LOF WS and OVC WS, LOHS WS and OVC WS modes are superimposed, their mutual neutralization is observed. The most dangerous is the overlap of the LOF WS, LOHS WS and TOP WS modes and the non-simultaneous overlap of LOF WS, TOP WS and OVC WS processes (with OVC lag). In this case, boiling of the coolant is possible.

Optimal configurations of the reactor with fuel based on oxide micro grains and uranium nanopowder have been obtained. The problem statement (optimality criterion, control parameters, constraints) is the same. The same scenarios for the development of emergency modes are postulated. It is a more heat-resistant fuel. Let us consider one of such layouts with the most compact core. Fuel enrichment in two zones is 12.5 and 19.8%, respectively. The diameter of the fuel pellet in both zones is the same and equal to 5.21 mm. The spacing of the fuel element lattice in the zones is 5.84 and 6.40 mm, respectively. The radial dimensions of the zones are 86.7 and 42.3 cm, respectively. The radius of the core is 129 cm. The height of the core is 1 m. The temperature of the coolant at the entrance to the core is 198 ºC. The maximum temperatures of fuel, coolants and fuel element cladding in the nominal mode do not exceed 938, 523 and 527 ºC, respectively. The void effect of reactivity is positive and dangerous. In the most dangerous combination of emergency modes, the coolant boils. The behavior of the reactor in ATWS is similar to that of using MN fuel without uranium nanopowder.

Thus, MN-U fuels are of the greatest safety interest.

4.5. Optimization of the Composition of the NaKCs Alloy

Multi component alloys of liquid metals have the greatest potential for regulating the properties of the LMFR coolant. If the problem of optimizing the ratio of the concentrations of the components of a binary alloy (for example, NaK used in space

reactors) from the point of view of one or another quality criterion is quite trivial, then when switching to coolants based on three-component systems (for example, NaKCs), the solution of such a problem is laborious and can be associated with serious problems.

The complexity of solving such problems increases many times, if individual components of the alloy are capable of forming a eutectic, and if, at certain ratios of the concentrations of the constituent elements, a qualitative change in the chemical properties of the alloy is possible. First of all, these remarks refer to three-component alloys of alkali metals based on sodium, potassium and cesium, which, like individual components, including those forming binary eutectic alloys, can be considered as potential coolants for space reactors. Among the components of the NaKCs alloy, the eutectic is formed by the binary alloys NaK, NaCs, KCs [45, 55]. The melting point of the eutectic 21% (at.) Na - 79% Cs is equal to minus 31.70 ºC; the melting temperature of the eutectic 30% (at.) Na - 70% Cs is equal to minus 7.85 ºC [45]. The melting point of the eutectic 50% (at.) K - 50% Cs is equal to minus 37.90 ºC [45].

Optimization of the composition of the NaKCs alloy in terms of minimizing the cost of the coolant, increasing the internal self-protection against accidents LOHS WS, OVC WS and LOCA WS is of a conflict nature.

An increase in the concentration of cesium in the alloy is hindered by its high activability in the reactor core and its high cost. The problem is not relevant when using radioactive waste in the coolant of an alien reactor. An increase in the concentration of cesium in the NaKCs alloy leads to a decrease in the void reactivity effect (hence, to an increase in self-protection against a LOCA WS accident) and a deterioration in the self-protection against accidents of LOF WS and LOHS WS. Heavier elements (in comparison with sodium) slow down neutrons worse due to elastic scattering. The deceleration due to inelastic scattering is characterized by a high energy threshold. As a result, when the coolant is lost, the neutron spectrum in the reactor does not change significantly. As a result, the void reactivity effect is relatively small, and the danger of a LOCA WS emergency is less. The boiling points of the NaKCs alloy are lower than that of pure sodium, potassium (but higher than that of cesium).

The content of sodium and potassium in the alloy is higher than the content of cesium. As a result, the danger of accidents LOF WS and LOHS WS is higher. At the same time, the freezing point of the NaKCs eutectic is so low that this advantage can only be used in space. (For a Martian NPP buried in the ground, it is enough to ensure the coolant is non-freezing.) A slight change in the concentration of cesium in the NaKCs alloy will lead to an insignificant increase in the freezing point of the alloy (in comparison with the eutectic) from the point of view of ensuring self-protection from accidents with coolant cooling.

Thus, it makes sense to solve the problem of minimizing the concentration of cesium in the NaKCs alloy, provided that the main advantages of using such a coolant are retained: non-flammability and the elimination of freezing in the reactor. The first condition is determined by expression (1), the second - by the non-freezing of the alloy. The condition of non-flammability is more important when choosing a heat carrier. According to inequality (1), the alloy non-freezing is provided for a fairly wide range of cesium concentration in the alloy: from 19 to 69% (at.), Or from 46.6 to 89.2% (wt.). Non-freezing of the alloy is ensured at and near the eutectic point, i.e., with a slight deviation of the cesium concentration in the alloy from 73.8%, as well as with a small (ideally zero) cesium content in the NaKCs alloy. As noted earlier, in the absence of cesium, sodium and potassium can form eutectic alloys.

The optimization problem can be formulated as a continuous multi criteria problem with constraints for several functionals. These functionals include the maximum and minimum coolant temperatures, the maximum temperatures of the fuel and cladding of fuel elements, and the maximum power of the reactor in emergency modes of the ATWS type. The control parameters of the problem are the concentration of alloy components, the parameters of the reactor and core layout, the parameters of the lattice of the fuel elements, the flow rate of the coolant, the enrichment of the fuel, etc.

The solution to a multi criteria problem is a set of informal procedures. For example, the formation of a system of quality criteria (at the stage of setting the problem) and the choice of the most preferred solution (at the final stage of the solution) are not formalized and are subjective. To solve the problem, multiple correlation coefficients are introduced. This coefficient is equal to zero when the void reactivity effect and the cost of the coolant are selected as criteria. For the criteria related to the cost of the coolant and internal self-protection against accidents LOF WS and LOHS WS, the correlation coefficient is maximal and is equal to 1.

Let us analyze the preferred solutions, the number of which can be large. Let us choose nine solutions U_1, U_2, U_9, each of which is obtained under the condition that one or another quality criterion is preferable.

- The U_1 solution is obtained if the cost of the heat carrier is selected as the most preferable criterion: U_1 = {0% Cs, 0% K, 100% Na} (hereinafter, the mass content is indicated as a percentage);
- The U_2 solution meets the criteria of "cost of the coolant" and "internal self-protection against accidents OVC WS": U_2 = {0% Cs, 78% K, 22% Na}, i.e. eutectic Na-K;
- The U_3 solution corresponds to the minimum void reactivity effect: U_3 = {100% Cs, 0% K, 0% Na};
- The U_4 solution corresponds to internal self-protection against accidents LOF WS and LOHS WS: U_4 = {0% Cs, 0% K, 100% Na};
- The U_5 solution corresponds to ensuring the non-flammability of the coolant: a set of optimal concentrations that satisfy condition (1), i.e., the optimality area, centered at the point 3.5% Na - 21.9% K - 74.6% Cs;
- The U_6 solution corresponds to the internal self-protection against accidents OVC WS: U_6 = {73.8% Cs, 22.0% K, 4.2% Na}, i.e. eutectic;
- The U_7 solution corresponds to internal self-protection against accidents OVC WS and non-flammability: U_7 = {73.8% Cs, 22.0% K, 4.2% Na}, i.e. eutectic;

- The U_8 solution corresponds to the non-flammability of the coolant and the minimum void reactivity effect: determined by condition (1) at the maximum cesium content - 89.2% Cs;
- The U_9 solution corresponds to the low cost of the coolant, the non-flammability of the coolant and the minimum void reactivity effect: determined by condition (1) with the minimum cesium content U_9 = {46.6% Cs, 46.3% K, 7.1% Na}, i.e. eutectic.

The next task is to determine the most preferable solution from options U_1, U_2, U_9.

Optimum corresponding to the low cost of the coolant; negative void reactivity effect; the best self-protection against accidents OVC WS, LOF WS, LOHS WS, LOCA WS (loss of coolant as a result of its combustion); and the minimum cost of the coolant can be determined unambiguously only when taking into account the weight coefficients reflecting the significance of all the listed criteria. (This is not a feature of the Martian reactors. This is a feature of the materials used in the core and the secondary circuit (mainly a non-flammable coolant with a relatively high boiling point and the lowest freezing point, with a small absorption and deceleration cross section for fast neurons) and the optimal layout of the core.) In other words, the determination of optimal control requires a preliminary ranking of the above criteria and emergency situations according to the degree of significance and danger. Obviously, this optimum belongs to the range of cesium concentration at which the NaKCs alloy does not ignite, i.e. 46.6 ... 89.2% Cs. Internal self-protection against accidents such as ATWS and the negative void reactivity effect can be achieved by the optimal choice of the parameters of the layout, the fuel element array, and other design solutions. For this reason, the criterion of non-flammability, other things approximately equal, should be given preference when choosing a coolant for an alien reactor.

Using the well-known ideas and methods for solving discrete multi criteria problems, let us rank the options U_1, U_2, ..., U_9 according to the degree of significance. The set of analyzed objects of the problem (U_1, U_2, ..., U_9) is finite. Of these nine objects of the discrete optimization problem, a set of the most preferable solutions (objects) can be distinguished. This set includes options U_5, U_6, U_7, U_8 and U_9.

We will give preference to economic efficiency (in the problem under consideration, the goal is to minimize the cost of the coolant) and safety. In single power units (in the absence of massive construction of reactors on Mars), improving the economic characteristics is not so urgent. Giving preference to three criteria of the problem: non-flammability, the smallest void reactivity effect and the best self-protection against accidents LOF WS and LOHS WS, we can single out effective objects: U_5, U_6, U_7, U_8 and U_9. They are preferred. Thus, in the problem of finding the most preferred object, options U_1, U_2, U_3 and U_4 can be ignored.

Let F_1, F_2, and F_3 be dimensionless optimality criteria that determine non-flammability, the smallest PER and the best self-protection against accidents with a violation of heat removal (LOF WS and LOHS WS), respectively. Let's define them. The area of effective objects is limited by three planes A, B and C (Fig. 7), corresponding to $F_i \geq L_i$ (where i = 1, 2, 3), where L_i is the minimum allowable value of the corresponding functional.

The lines of intersection of planes A and B; B and C; A and C, are parallel to the axes [0 F_1), [0 F_3) and [0 F_2), respectively. Point 0_2 meets the conditions: $F_1 = L_1$, $F_2 = L_2$ and $F_3 = L_3$. The goal of the problem is to maximize F_i under the condition $F_i \geq L_i$ (in the case under consideration, i = 1, 2, 3).

Let's sort options U_5, U_6, U_7, U_8 and U_9 according to preference. Note that the solutions U_6 and U_7 are coincide (correspond to the eutectic). One of them (for example, U_6) can be excluded from consideration. Let's write down the remaining objects in order of preference.

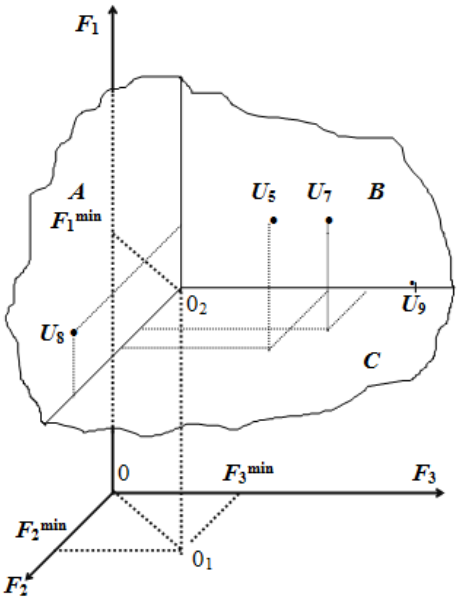


Figure 7: General view of the area of effective objects for choosing the preferred composition of the NaKCs alloy

- From the point of view of maximizing the functional F_1: $U_5 > U_6 > U_7 > U_8 > U_9$ (here the symbol ">" means "more preferable");
- From the point of view of maximizing the functional F_2: $U_8 > U_5 > U_7 > U_9$;
- From the point of view of maximizing the functional F_3: $U_9 > U_7 > U_5 > U_8$.

Object U_8 belongs to plane A (Fig. 7), i.e., for it $F_3 = L_3$. Object U_9 belongs to the F_3 axis, that is, to the planes B and C. For it, $F_1 = L_1$ and $F_2 = L_2$. Thus, object U_8 can be considered more preferable than U_9. Variants U_5 and U_7 lie inside the area of dominant objects at some distance from the border of this area. They satisfy the strict inequalities $F_i > L_i$ (i = 1, 2, 3); moreover, U_5 dominates twice, i.e., according to two criteria: when maximizing F_1, and when maximizing F_2, and the U_7 variant dominates once: when the functional F_3 is maximized.

Thus, variant U_5 can be considered the most preferable object. If we consider all alloys based on sodium, potassium, and cesium with combinations of concentrations of components from region (1) "equally non-flammable", then approaching the center of the

non-flammable region inside (1) makes no sense, in other words, it makes no sense to maximize the functional F_1 inside the region (1). In this case, it is enough to satisfy the condition $F_i \geq L_i$ and the most preferable solution should be considered U_7 (U_7 is equivalent to the object U_6), corresponding to the eutectic. Strictly speaking, the choice of the best solution from U_5 and U_7 should be carried out taking into account the coefficients of significance of the criteria F_1, F_2 and F_3.

Solution U_9 belongs to the line of intersection of planes B and C, that is, it is determined by the condition $F_3 = L_3$. A further increase in the economic efficiency of the reactor is possible with a comprehensive optimization of its parameters, which presupposes the search for the optimal control u, the components of which are not only the concentrations of sodium, potassium and cesium in the alloy, but also the parameters of the layout, the lattice of the fuel elements, the flow rate of the coolant, the enrichment of the fuel, etc.

So, the optimal composition of the NaKCs alloy can be obtained by solving the problem of multi criteria optimization. Moreover, the most preferable solutions may be different, and for alloys forming a eutectic, the optimum does not always coincide with the point of the eutectic.

5. Discussion

5.1. Further Prospects in Fuel Use

A two-circuit scheme of energy conversion with water in the second circuit was considered. Due to the relatively low chemical activity of the primary coolant (eutectic alloy NaKCs) during depressurization of the heat exchanger tubes, the interaction of NaKCs with water is not explosive. Water is the simplest and cheapest heat carrier. The technology of using water for cooling is well established. This applies to land-based reactors. The construction of a NPP on Mars will require large volumes of water. Water will have to be transported from Earth. This will increase the cost of the Martian NPP.

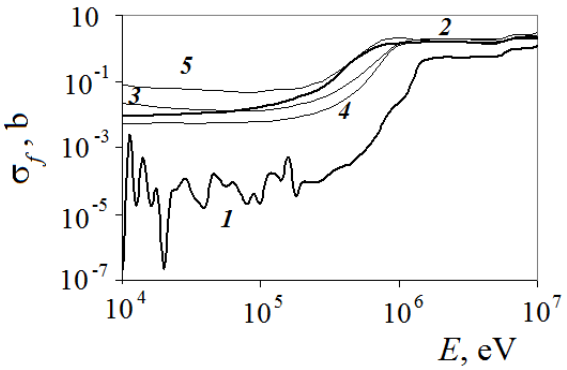


Figure 8: Dependence of the fission cross section on the kinetic energy of a neutron for ^{238}U (curve 1) and some minor actinides: ^{237}Np (*2*), ^{241}Am (*3*), ^{243}Am (*4*), ^{244}Cm (*5*)

Among the possible innovations, it is possible to propose an increase in the proportion of long-lived radioactive waste (isotopes of neptunium and americium with an even number of neutrons) in the fuel composition by reducing the proportion of waste uranium. The fission cross section for ^{237}Np, ^{241}Am and ^{243}Am in the neutron energy range of more than 1.4 MeV is 2 - 3 times higher than that of ^{238}U [51].

In fig. 8, according to [51], the dependences of the microscopic fission cross section σ_f on the kinetic energy E of the neutron for uranium-238 and some minor actinides are given.

When water is used in the second loop, the ingress of a small portion of the moderator into the core will lead to a decrease in reactivity and shutdown of the reactor. (Initiating event: emergency situation with depressurization of the tubes of the heat exchanger "liquid metal - water".)

5.2. NPP with a Closed Gas Turbine Energy Conversion Cycle

What can be considered as an alternative to water (in the second circuit)? The atmosphere of Mars is 95.32% carbon dioxide [22]. Carbon dioxide was used as a coolant for the British nuclear reactors "Magnox", AGR and the French reactor UNGG [56]. It is possible to offer a NPP with a closed gas turbine energy conversion cycle. In this case, carbon dioxide or unpurified air of Mars can serve as a coolant. A closed gas turbine cycle is proposed to be used in high-temperature gas-cooled reactors [57].

It is often proposed to use helium as a working medium for reactors of NPPs with a gas turbine energy conversion cycle. As is known, helium is the second most abundant in the Universe (after hydrogen) and makes up about 23% of the mass of the Universe. Natural helium consists of two stable isotopes with mass numbers 3 and 4. The fraction of ^{4}He in the natural mixture of isotopes is 99.99986%) and can vary considerably within wide limits [58].

Helium (consisting mainly of the doubly magical isotope ^{4}He) is not activated in the reactor core. Its practical use is hindered by its high cost and high fluidity [59]. For this reason, instead of helium, it is advisable to use expensive helium – xenon mixture [59]. Heat exchange equipment using a He-Xe mixture is the most compact [59].

In three-circuit NPPs with sodium reactors, the void reactivity effect can be realized in the following cases.

- Boiling of the coolant;
- Combustion of the coolant;
- Coolant leak (loss of coolant);
- Entrainment of gas from the argon bed into the core;
- Depressurization of fuel element cladding and release of gaseous fission products into the coolant.

In two-circuit NPPs with a gas turbine energy conversion cycle, another additional scenario for the implementation of the void reactivity effect appears: gas entrainment into the core when the gas heater tubes (liquid metal - gas heat exchanger) are depressurized.

Potential coolants of fast gas-cooled reactors (superheated water vapor, helium, carbon dioxide, dissociating gas, for example, $N_2O_4 \leftrightarrow 2\ NO_2$, helium - xenon mixtures, etc), as well as working fluids of gas turbine plants, can be considered as the working fluid (coolant of the secondary circuit).

One problem with medium to high power LMFRs is the positive void reactivity effect. The most dangerous is the draining of the central part of the core (for example, when the working fluid is involved in the event of depressurization of the heat exchanger tubes). For the considered reactor, the void reactivity effect in the most dangerous scenario of its implementation reaches 6.1 β (slightly less than when using a sodium coolant).

When the steam generator pipes are depressurized, steam bubbles (the working fluid of the second circuit) emerging from them can be entrained by the coolant flow into the core. From the point of view of possible depressurization of the steam generator pipes, the most dangerous mode is LOF WS (or its overlap with TOP WS and LOHS WS), accompanied by the greatest increase in the temperature of the primary coolant.

When the main circulation pumps of the primary circuit are de-energized, the mode of natural circulation of the coolant is established. The higher the flow rates of the coolant in this mode, the greater the likelihood of the entrainment of bubbles in the core. Thus, the normally desired high level of natural circulation can worsen a steam generator pipe leakage accident.

The mechanism of the effect of bubbles (in the core) on reactivity is as follows. If water is used in the second loop, then, entering the core, water vapor (hydrogen) effectively slows down neutrons. Hydrogen also absorbs neutrons, but this effect is relatively small at LMFR energies. Deceleration of neutrons leads to a decrease in the average neutron energy in the core. In this case, the microscopic fission cross sections for all fissile nuclides (^{235}U, ^{239}Pu, etc.) increase significantly. As a result, the effective neutron multiplication factor and reactivity increase. This could lead to an accident.

If, as a working fluid (in the second circuit), a substance is used that strongly absorbs fast neutrons and weakly slows them down, then the result is exactly the opposite. Additional absorption of neutrons in the core leads to the shutdown of the reactor: the void effect is negative.

Let us assume that a substance that weakly interacts with neutrons is used as a working fluid. In this case, the void effect of reactivity is realized in “pure form”. Any coolant one way or another absorbs and slows down neutrons. The NaKCs alloy is no exception. When entering the core in the form of bubbles, the efficiency of absorption and slowing down of neutrons deteriorates. As a result of the worst absorption in the core, additional neutrons appear. This leads to increased reactivity and the possibility of an accident. As a result of the worst moderation of neutrons in the core, the neutron energy increases. This leads to intensification of the fission of raw material nuclides (^{238}U) by neutrons with energies above 1.4 MeV (the fission threshold for most raw material nuclides). As a result, the reactivity and the risk of an accident increase.

Thus, it is necessary either to exclude the ingress of bubbles into the core, or to carefully select the materials of the reactor. It is preferable to use substances that efficiently absorb fast neutrons as a working fluid (secondary circuit). In this case, the void reactivity effect is negative.

Suppose, as a result of depressurization of the heat exchanger tubes, the material of the second circuit (working fluid) is drawn into the central part of the core (the most dangerous scenario). In fig. 9 shows the dependences of the neutron infinite multiplication factor k on the volumetric content ε of the working fluid of the second circuit in the coolant of the core. (Calculations were performed using the WIMS code [53]). As a working fluid, the following were considered: water vapor (fig. 9, curve *1*), ^{4}He (*2*), ^{3}He (*3*), CO_2 (*4*), N_2 (*5*), N_2O_4 (*6*), air of the earth's atmosphere (*7*) and Xe. Curve *8* corresponds to vacuum and is close to xenon. The dependence of k on the volume fraction f of gas when steam, He or CO_2 enters the core is non-monotonic, and it is possible to increase and decrease k, therefore, the scenario of the development of an emergency is less predictable.

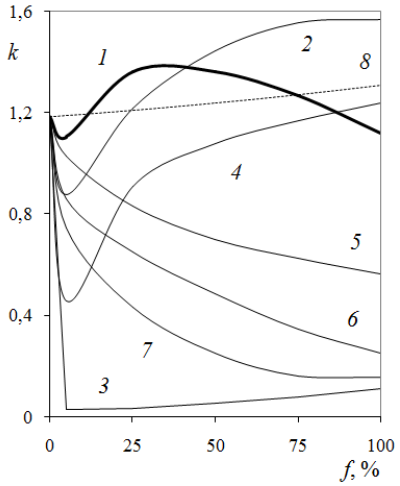


Figure 9: Dependence of k on the volume fraction f of gas in the primary coolant during depressurization of the heat exchanger tubes and the entrainment of gas bubbles into the core

As follows from the data in Fig. 9, with a small fraction ε of water vapor in the coolant, deceleration beyond the fission threshold of ^{238}U, ^{237}Np, ^{241}Am and ^{243}Am prevails (k decreases). With an increase in the vapor fraction, a slowdown towards resonances is observed in the fission cross sections of ^{235}U, ^{239}Pu and ^{241}Pu (k increases). The key role in these processes is played by elastic deceleration on hydrogen nuclei in the composition of water. With a further increase in ε, an intense absorption of neutrons slowed down to resonances in the radiation capture cross section (on these nuclei and on hydrogen) is observed. When ^{3}He (exotic), CO_2, N_2, N_2O_4 and air are used as a working medium, the deterioration of the conditions for heat removal from fuel elements can be compensated for by negative reactivity from the ingress of gas into the core.

The ingress of carbon dioxide (Martian atmospheric air) into the central part of the core and the replacement of the liquid metal coolant with gas by 7 ... 8% by volume, leads to a sharp decrease in the criticality of the reactor (k decreases by 2.4 times), i.e., to the shutdown of the reactor.

The disadvantage of a NPP with a gas turbine energy conversion cycle is its large size. However, if a reactor of relatively small (according to earthly estimates) electrical power (500 ... 600 MW) is built on the surface of another planet, this is not a serious problem.

5.3. Plans

The Russian design bureau "Arsenal" plans to implement the "Nucleon" project to build a NPP on Mars. It is assumed that the NPP will be fully built in terrestrial conditions, and not on Mars. The nuclear tug "Zeus" will deliver it to Mars [17]. The NPPs will start operating immediately after landing on Mars. Certain problems can arise here. The atmosphere of Mars is so rarefied that NPPs will be attacked by meteorites. Without special protection (containment), this can lead to damage or complete destruction of the NPP and reactor. (NASA's MRO robotic interplanetary station, which orbited 60,000 times around Mars from 2006 to 2019, "received images of an unusual trail of recent meteorite impacts on the surface of Mars, similar to the traces of shots from a giant space shotgun" [60].)

It is advisable to use the Martian NPP "Nucleon" with an electrical capacity of about 1 MW at the initial stage of Mars exploration. It can be used, for example, to generate energy for the construction of NPPs of greater capacity, buried under the ground. Larger NPPs can be assembled from blocks built on Earth and delivered to Mars by a nuclear tug. If the concept of a NPP with a gas turbine energy conversion cycle is chosen as the basis, then carbon dioxide for the second circuit can be pumped from the Martian atmosphere.

There are at least three reasons for burying the Martian NPP under the ground.

- Burying the NPP under the ground is necessary to protect against meteorites and dust storms;
- It is necessary to ensure the required height of the natural circulation circuit (to ensure the natural circulation of the primary coolant in emergency LOF WS modes);
- It is also necessary to maintain a stable ambient temperature during NPP operation.

An erosion hazard for NPPs can be posed by Martian dust storms and tornadoes (particles consisting of iron oxide) [6]. Their speed reaches 100 m / s, duration from 50 to 100 days [6].

When building a NPP on Mars, the following factors should be taken into account.

- The soil of Mars is a rocky rock (basalt) with an admixture of iron oxide (that is, it does not differ much from the earth's soil), covered with a layer of dust (consisting mainly of iron oxide);
- Mars is characterized by large daily and seasonal temperature drops;
- Mars is characterized by dust storms, meteorite attacks;
- In some regions of Mars (covering about 25% of the planet's surface) volcanic and tectonic activity is pronounced [61].

It is possible to single out specific tasks that are not typical for ground-based NPPs. One of them is the production of oxygen from the atmosphere of Mars (from carbon dioxide) [62, 63]. Oxygen is needed to support the life of researchers working on Mars and as an oxidizer for traditional rocket (non-nuclear) fuel, if used. Another challenge is to provide energy to greenhouses for growing food crops. At the University of Dartmouth, by order of NASA, a project for a Martian greenhouse was developed [64]. The greenhouse will provide plant food for four astronauts daily for 600 days on Mars. [64] Another challenge is to provide energy for vehicles and machines for laying mines and tunnels.

Economic assessments of the construction of the considered rector (electric power of 600 MW) on Mars have not yet been carried out. It is obvious that the proposed concept of the core (and the reactor as a whole) in terms of construction costs significantly exceeds a NPP with a low-power reactor (~ 10 ... 1000 kW). The diameter of the BN-600 reactor vessel is 12.8 m, the height is 12.6 m. (For comparison, the Hubble telescope is 13.3 m long and 4.3 m in diameter.) It is not yet possible to transport such a "product" to Mars. But at first glance, unsolvable problems have always stimulated and intensified scientific research in related fields. The coolant should be transported separately by elements in solid form (in the form of small briquettes). It is proposed to use long-lived cesium isotopes. This requires the use of special containers for the transport of cesium or full automation of loading and unloading a space tug. By the time a NPP of such a large capacity is in demand, a certain infrastructure and sources of electricity will already exist on Mars (for example, a low-power NPP). Robots will play an important role. The process of building a Martian NPP, filling it with coolant and other procedures can be fully robotized. It is advisable to make the reactor without a vessel. Instead of a body, a concrete shaft lined with a thin layer of steel (5 ... 6 mm) can be used. The technology of creating such a mine (for the BREST-OD-300 reactor) exists in terrestrial conditions [23]. The lining of a concrete shaft can be made with steel or with a composite material. There is a technology for obtaining such a material. It is lightweight, durable, heat-resistant, radiation-resistant material. The material is reinforced with nanofibers of another polymer composite material with a high capture cross section for neutron and γ-radiation [65].

Concrete is required for the construction of the mine. It can be produced from Martian soil. There are many types of cement. As in terrestrial conditions, the technological process for the production of Martian cement consists of three main stages: extraction and processing of raw materials (CaO, Si_2O_3, Al_2O_3, Fe_2O_3); roasting of crushed and mixed raw materials to obtain clinker; *and* grinding of clinker, gypsum and additives. Gypsum can exist as minerals on Mars. Gypsum or its analog can be made artificially from Martian rock. Cement can be produced from Martian soil. It is not necessary to use earthy materials as additives. Mars has significant and accessible deposits of water ice [66]. In 2008, the Phoenix space probe (NASA) obtained water from soil samples collected from the surface of Mars [66].

6. Conclusion

The proposed innovations associated with the use of new nuclear fuel based on UN-PuN micro grains and non-powder of metallic uranium and a coolant based on the NaKCs alloy will make it possible to create an alien NPP within the framework of existing technologies. Computational and optimization studies demonstrate the possibility of creating a safe space reactor with an electric power of up to 600 MW. A reactor of such a relatively large capacity may be in demand only in the distant future. It is important to note that if it is possible to ensure the internal safety

of a 600 MW reactor, then there will be no problems in ensuring the safety of a smaller reactor.

The following physical foundations of the concept of a Martian NPP can be noted:

- Fast reactor with electric power up to 600 MW (at the initial stage it is possible to restrict oneself to a much lower power, for example, 50 ... 100 MW);
- Low-maintenance or maintenance-free (fully autonomous) NPP, operating without fuel replacement for 10 ... 20 years;
- Coolant based on eutectic (or near eutectic) NaKCs alloy;
- Double-circuit power conversion scheme;
- As an option, a NPP with a gas turbine cycle (CO_2) energy conversion;
- Burying the NPP under the ground;
- Land-based factory fabrication of NPP modules and their transportation to Mars;
- Use of well-mastered technologies (it makes sense for the construction of a ground-based prototype of the Martian NPP).

LMFR type reactors are characterized by high safety potential. They are compact. The neutron spectrum (the absence of a neutron moderator and the fuel used) allows such a reactor to operate in the mode of self-supply of the core with fuel (the breeding ratio in the core is about 1), i.e., for a long time. Essentially, the Martian NPP will use a new type of nuclear reactor: a pressurized fast reactor (PLMFR).

The proposed new cermets fuel will allow not only to ensure accident-free completion of ATWS modes, but also to increase the content of fissile materials in the fuel elements of the core (hence, to increase the reactor power).

The power of the NPP will be determined by the range of tasks to be solved, including those unconventional for the terrestrial energy sector. For example, with the help of electricity generated by a NPP, the problem of producing oxygen from carbon dioxide will be solved. As Mars exploration progresses, the relevance of charging batteries for an increasing number of electric vehicles capable of moving across the surface of Mars will increase. There will be a need to provide power to electric machines for walking tunnels.

One should not be afraid of plutonium theft (material of direct use in weapons of mass destruction) from "fresh" fuel before the launch of the reactor. States whose representatives will visit Mars on their spacecraft (especially with a nuclear rocket engine) already possess nuclear weapons. In addition, the fuel cycle is protected from proliferation.

The coolant based on the NaKCs alloy is characterized by a much lower chemical activity compared to pure alkali metals. This allows us to consider a two-circuit energy conversion scheme (even with water in the second circuit). For Mars, a NPP with a gas turbine cycle (based on the use of carbon dioxide in the second loop) energy conversion with a low reactor power is more preferable.

The double-circuit scheme helps to simplify and reduce the cost of the NPP design. With such an energy conversion scheme, an additional scenario for the implementation of the void reactivity effect is possible. It is associated with the possible involvement of the working fluid in the reactor core. When carbon dioxide is used in the second circulation loop, the involvement of the working fluid into the core helps to reduce the reactivity and shutdown the reactor.

It is known that it is much easier to ensure the safety of a low-power reactor than a high-power reactor. The electric capacity of the 600 MW power units is unlikely to be in demand for the exploration of Mars in the near future. However, if it is possible to ensure the safety of such a reactor, then the safety of a reactor with a power of 10 times less does not pose a problem. The results are focused on the distant future.

Deepening under the ground, ground-based factory production of all modules and equipment, the use of well-mastered technologies are the general tasks of creating infrastructure facilities for the future Martian station.

It is obvious that the maximum use of the natural resources of the developed planet can significantly reduce research costs.

Obviously, for the implementation of the construction of a relatively large Martian NPP, international cooperation will be required. Before the construction of such a NPP on Mars, it is advisable to build a terrestrial prototype for testing technologies. Of course, the construction of a NPP on Mars with a reactor of electrical power of 600 MW may remain science fiction. But work on such an idea stimulates the development of new technologies that may be in demand in the future when exploring other planets.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The author is grateful to the head of the Department of Physics at the Bauman Moscow State Technical University, Corresponding Member of the Russian Academy of Sciences A.N. Morozov for comprehensive support.

References

[1] V.S. Okunev, "Analysis of Reactivity Effects and Coefficients of a Space Nuclear Reactor Cooled by a Sodium-Potassium-Cesium Alloy", AIP Conference Proceedings, **2318**, 040007, 2021, https://doi.org/10.1063/5.0036182

[2] A.S. Koroteev, "Analysis of Advanced Space Problems and the Place of Nuclear Power Propulsion Plants in their solution ", in 2005 International Conference Nuclear Power Engineering in Space-2005 (NPS-2005), Moscow, Podolsk, 2005, Plenary report 01.

[3] A.S. Koroteev, "Actual problems of modern rocket and space technology", Izvestiya RAN: Energetika, **5**, 9-18, 2004.

[4] Yu.G. Dragunov, B.A. Gabaraev, E.L. Romadova, "Space nuclear power: past, present, future", in 2018 5th Int. Scientific and Technical Conf. Innovative Design and Technologies of Nuclear Power (ISTC NIKIET-2018), Moscow, 689-705.

[5] V.N. Gushchin, "Fundamentals of Spacecraft Construction", Moscow: Mashinostroenie, 2003.

[6] M.T. Lemmon et. al., "Atmospheric Imaging Results from the Mars Exploration Rovers: Spirit and Opportunity", Science, 2004, **306**(5702), 1753-1756, DOI:10.1126/science.1104474.

[7] "The cost of the commercial launch of Soyuz with the Fregat block has become known", RIA Novosti: Science (Moscow), 02 October 2018, https://ria.ru/20181002/1529826936.html

[8] S. Kalashnikov, "The cost of a flight to Mars", Ferra.ru (Moscow), May 10, 2017, https://www.ferra.ru/news/techlife/mars-cost-10-05-2017.htm

[9] "Akkuyu NPP (Turkey)", RIA Novosti (Moscow), December 10, 2017, https://ria.ru/20171210/1510584872.html

[10] "The cost of building a fast breeder reactor BN-800 is estimated at 145.6 billion rubles (Russian)", Moscow: TASS, January 19, 2018, https://tass.ru/ural-news/2597666

[11] "The cost of building a fast reactor BREST is estimated at 100 billion rubles", Moscow: TASS, June 8, 2021, https://tass.ru/ekonomika/11594071

[12] R.F. Wilson, "SNAP 10A - A Status Report. / Book: Space Power Systems Engineering", Progress in Astronautics and Rocketry, 1966, **16**, 581-593, https://doi.org/10.1016/B978-1-4832-3056-6.50029-3

[13] N. Yachmennikova. "Roskosmos will continue to create a space tug with a nuclear engine", Rossiyskaya Gazeta RGRU, 04/29/2020, https://rg.ru/2020/04/29/roskosmos-prodolzhit-sozdanie-kosmicheskogo-buksira-s-iadernym-dvigatelem.html

[14] "Nuclear Reactor for Mars Outpost Could Be Ready to Fly by 2022", SPACE.com, 12 August 2019, https://www.youtube.com/watch?v=vQLStYhIR7Q

[15] Yu.G. Dragunov, E.L. Romadova, L.A. Sleptsov, V.V. Kudinov, "Space Reactor Installation of a New Generation",in 2012 Int. Scientific and Technical Conf. Innovative Design and Technologies of Nuclear Power (ISTC NIKIET-2012), Moscow.

[16] "NASA and US Department of Energy Selected Three Reactor Concepts for Thermal Nuclear Engines", Atominfo.Ru, 15 July 2021, http://www.atominfo.ru/newsz03/a0869.htm

[17] A. Demidov, "Roscosmos proposed to build a nuclear power plant on Mars", Gazeta.ru (Moscow), July 10, 2021, https://www.gazeta.ru/science/news/2021/07/10/n_16224692.shtml

[18] R.D. Williams, "Mars Fact Sheet", National Space Science Data Center, NASA, Sept. 1, 2004, https://nssdc.gsfc.nasa.gov/planetary/factsheet/marsfact.html.

[19] "Extreme Planet Takes Its Toll. Mars Exploration Rover Mission: Spotlight", NASA, California Institute of Technology, Jet Propulsion Lab., June 12, 2007, https://mars.nasa.gov/mer/spotlight/20070612.html.

[20] "Making a Splash on Mars", NASA Science, June 29, 2000, https://science.nasa.gov/science-news/science-at-nasa/2000/ast29jun_1m/

[21] S.A. Stern, "The Lunar atmosphere: history, status, current problems, and context", Rev. Geophys, 1999, **37** (4), 453-491, DOI: 10.1029/1999RG900005.

[22] "Mars Pathfinder - Science Results", Atmospheric and Meteorological Properties, https://mars.nasa.gov/MPF/science/atmospheric.html (дата обращения 7 ноября 2021)

[23] V.V. Lemekhov, A.V. Moiseev, M.K. Sarkulov, V.S. Smirnov, O.A. Yarmolenko, Yu.V. Lemekhov, Yu.S. Cherepnin, V.P. Vasyukhno, D.A. Afremov, "Present-Day Status and Development Prospects of Fast-Neutron Lead-Cooled Reactors", in. 2018 5th Int. Scientific and Technical Conf. Innovative Design and Technologies of Nuclear Power (ISTC NIKIET-2018), Moscow, 35-37.

[24] V.S. Okunev, "Fundamentals of Applied Nuclear Physics and an Introduction to the Physics of Nuclear Reactors", Moscow: Publishing house of Bauman Moscow State Technical University, 2015.

[25] DOE Fundamental Handbook: "Nuclear physics and Reactor Theory", DOE-HDBK-1019/1-93, U.S. Department of Energy, Washington, D.C. 20585, FSC-6910, **1**, 32, 1993.

[26] O.J. Wick (Ed), "Plutonium Handbook, A Guide to the Technology", American Nuclear Society, 1980, **2**.

[27] N.M. Vlasov, I.I. Fedik, "Fuel elements of nuclear rocket engines", Moscow: TsNIIatominform, 2001.

[28] L.H. Caveny (ed.), "Orbit-Raising and Maneuvering Propulsion: Research Status and Needs", Progress in Astronautics and Aeronautics. **89**, New York: American Institute of Aeronautics and Astronautics, 1984, https://doi.org/10.2514/4.865633

[29] I.I. Fedik, N.M. Vlasov, "New Materials in Space Nuclear Power Engineering", Perspective Materials, 2001, **6**, 24-30.

[30] L.E. Von Toth, "Transition Metal Carbides and Nitrides", New York – London: Academic Press, 1971.

[31] M.L. Taubin, "Some Radiation Effects in the Interstitial Phases of ZrC, NbC, ZrN", Atomic Energy 1990, **69** (3), 176-177.

[32] A.E. Waltar, A.B. Reynolds, "Fast Breeder Reactors", New York, Oxford, Toronto, Sydney, Paris, Frankfurt: PERGAMON PRESS, 1981.

[33] IAEA-TECDOC-1083, "Status of liquid metal cooled fast reactor technology", Vienna: IAEA, 1999.

[34] N.N. Ponomarev-Stepnoy, V.S. Rachuk, V.P. Smetannikov, I.I. Fedik, "Space nuclear power and propulsion systems based on a reactor with external heat conversion of a solid-phase core", in 2005 International Conference Nuclear Power Engineering in Space-2005 (NPS-2005), Moscow, Podolsk, 2005, Plenary report 07.

[35] A.G. Lanin and I.I. Fedik, "Selecting and using materials for a nuclear rocket engine reactor", Phys.-Usp, 2011, **54** 305, DOI: 10.3367/UFNr.0181.201103f.0319

[36] S.N.Sikorin, A.V.Kuzmin, S.G. Mandzik, S.A. Polozov, T.K. Grigorovich, Sh.T. Tukhvatulin, I.E. Galev, A.N. Bakhin, A.L. Izhutov, V.E. Alekseev, D. Kaiser, I. Bolshinsky, "Low Enriched Nuclear Fuel on the Basis Uranium Zirconium Carbonitride: Preparation to Reactor Test and Research on the Critical Assemblies", in 2018 5th Int. Scientific and Technical Conf. Innovative Design and Technologies of Nuclear Power (ISTC NIKIET-2018), Moscow, 522-532.

[37] A.N. Bakhin, I.E. Galev, D.S. Kiselev, D.M. Soldatkin, A.A. Urusov, D.A. Chesnokov, A.F. Ginevsky, Ya.S. Volgin, V.S. Volgin, "Study of Thermomechanical Strength of Uranium-Zirconium Carbonitride Fuel", in 2018 5th Int. Scientific and Technical Conf. Innovative Design and Technologies of Nuclear Power (ISTC NIKIET-2018), Moscow, 533-543.

[38] V.S. Vasilkovsky, P.V. Andreev, G.A. Zaritsky, N.N. Ponomarev-Stepnoy, G.V. Kompaniets, V.A. Usov, V.V. Vasilenko, A.Yu. Danilyuk, V.N. Zubrev, K.A. Pavlov, L.A. Rachev, V.P. Chupakhin, G.A. Shevtsov, V.V. Viter, M.V. Arakin, E.A. Ksenofontov, L.N. Tararin, B.I. Poletaev, E.G. Lyannoy, A.Yu. Pavlov, A.V. Romanov, "Problems of space energy and the role of nuclear power plants in their solution", in 2005 International Conference Nuclear Power Engineering in Space-2005 (NPS-2005), Moscow, Podolsk, 2005, Plenary report 05.

[39] A.A. Gafarov, A.B. Prishletsov, N.M. Rozhdestvensky, "Comparative analysis of transport and energy modules based on nuclear power plants with direct and dynamic energy conversion systems", in 2005 International Conference Nuclear Power Engineering in Space-2005 (NPS-2005), Moscow, Podolsk, 2005, Section 1, report 01.

[40] V.M. Borishansky, S.S. Kutateladze, I. I. Novikov, S. S. Fedynsky, "Liquid metal coolants", Moscow: Atomizdat, 1976.

[41] N.S. Akhmetov, "General and Inorganic Chemistry", Moscow: Higher School, 2001.

[42] L.M. Sulimenko, "Gallium. Popular library of chemical elements", T. 1. M .: Nauka, 1983.

[43] Yu. V. Apalkov, "Submarines of the Soviet fleet 1945-1991", Moscow: Morkniga, 2012, **3**, 229-233.

[44] V.S. Okunev, "Some Issues of Using Liquid Metals and Their Alloys for Cooling Fast Reactors", Moscow: Bauman Moscow State Technical University - Research Institute of Power Engineering, 2004.

[45] E.V. Sulim, N.G. Bogdanovich, O.V. Starkov, E.A. Kochetkova, V.E. Levchenko, "Investigation of the properties of the ternary system of alkali metals sodium-potassium-cesium in the temperature range 293-973 K" , Obninsk (Russia): Institute of Physics and Power Engineering, 2007, FEI-3087.

[46] A.G. Mozgovoy, V.V. Roshchupkin, S.N. Skovorodko, M.A. Pokrasin, A.I. Chernov, "Density and pressure of saturated vapors of liquid sodium-potassium-cesium eutectic", High Temperature, **41(**4), 560-562, 2004. http://mi.mathnet.ru/tvt1699

[47] P.I. Bystrov, D.N. Kagan, G.A. Krechetova, E.E. Spielrain, "Liquid metal coolants for heat pipes and power plants", Moscow: Nauka, 1988.

[48] O.D. Kazachkovsky, A.V. Starkov, E.A. Kochetkova, N.G. Bogdanovich, E.V. Sulim, V.E. Levchenko, N.F. Lyubchenko, V.I. Zharinov, "Some features of the alloys of the sodium-potassium-cesium system", Atomic Energy, 1992, **73** (6), 500-502.

[49] J. Emsley, "The Elements", Oxford: Clarendon Press, 1991.

[50] V.S. Okunev, **"**Nuclear Power in Space: a Look into the Future**",** AIP Conference Proceedings, **2171**, 040001, 2019.

[51] A. Koning, R. Forrest, M. Kellett, R. Mills, H. Henriksson, Y. Rugama (Ed.), "The JEFF Nuclear Data Library", OECD 2006 NEA No. 6190, Nuclear energy agency organisation for economic co-operation and development, 2011, https://inis.iaea.org/collection/NCLCollectionStore/_Public/45/026/45026295.pdf

[52] E.A. Gomin, "MCU-4 status", VANT: Physics of nuclear reactors, 2006, **1**, 6-32.

[53] "WIMSD-IAEA Library", IAEA, Nuclear Data Services, 2014, https://www.iaea.org/resources/databases/wimsd-iaea-library

[54] Geraskin N.I., Kuzmin A.M., Kashutin A.A, A.V. Kobiak, D.V. Morin, A.E. Novikov, V.S. Okunev, M.O. Shvedov, V.V. Khromov, "Complex Optimization of the Fast Reactors", in 1990 proc. Of International Conference on the Physics of Reactors (PHYSOR-90), Marseille, Fr., **4,** 1990.
[55] I.S. Grigorieva, E.Z. Meilikhova (Ed.), "Physical quantities: Handbook", Moscow: Energoatomizdat, 1991.
[56] IAEA-TECDOC-1521, "Characterization, Treatment and Conditioning of Radioactive Graphite from Decommissioning of Nuclear Reactors", Vienna: IAEA, 2006.
[57] IAEA-TECDOC-1198, "Current Status and Future Developments of Modular High Temperature Gas C"ooled Reactor Technology", Vienna: IAEA, 1999.
[58] V.G. Fastovsky, A.E. Rovinsky, Yu.V. Petrovsky, "Inert gases", Moscow: Atomizdat, 1972.
[59] E.A. Manushin, "Gas turbines problems and prospects", Moscow: Energoatomizdat, 1986.
[60] A.Bina, E. Harrington, E. Pilles, L. Tornabelle, "A recent Cluster of Impacts", 2017 (A HiRISE. Lunar & Planetary Laboratory. The University of Arizona), https://hirise.lpl.arizona.edu/ESP_047768_1995
[61] S.C. Solomon, J.W. Head, "Evolution of the Tharsis Province of Mars: The Importance of Heterogeneous Lithospheric Thickness and Volcanic Construction", J. Geophys. Res., 1982, **87**(B12), 9755-9774, DOI:10.1029/JB087iB12p09755.
[62] "A plant that extracts oxygen from carbon dioxide was created in the USA, Pasadena, USA", Federal News Agency Regnum (Russia), May 30, 2019, https://regnum.ru/news/innovatio/2638707.html
[63] A. Deryabin, "Decomposed CO2: Nizhny Novgorod residents figured out how to effectively decompose carbon dioxide", Russian newspaper RGRU - Week - Volga region, 160, 22 July 2020, https://rg.ru/2020/07/22/reg-pfo/nizhegorodcy-pridumali-kak-effektivno-razlozhit-uglekislyj-gaz.html
[64] K. Damadeo (Ed.), "NASA's 2019 BIG Idea Challenge Winner Designs Best Planetary Greenhouse", NASA, Space Tech, Apr 24, 2019, https://www.nasa.gov/feature/langley/nasa-s-2019-big-idea-challenge-winner-designs-best-planetary-greenhouse
[65] P.V. Matyukhin. Heat-resistant polymeric composites for neutron and gamma protection. International research journal. **9**(28), 39-40. DOI: 10.18454/IRJ.2227-6017.
[66] Phoenix managed to get water from Martian soil. Lenta.ru. August 1, 2008. https://lenta.ru/news/2008/08/01/water/

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 3, 32-43 (2022)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Encompassing Chaos in Brain-inspired Neural Network Models for Substance Identification and Breast Cancer Detection

Hanae Naoum[1,*], Sidi Mohamed Benslimane[1], Mounir Boukadoum[2]

[1]*LabRI-SBA Lab., Ecole Superieure en Informatique, Sidi Bel Abbes, 22000, Algeria*

[2]*Design and fabrication of Microsystems Research Laboratory, Department of Computer Science, University of Quebec At Montreal, Montreal, H3C 3P8, Canada*

ARTICLE INFO

Article history:
Received: 28 January, 2022
Accepted: 06 May, 2022
Online: 25 May, 2022

Keywords:
Pattern recognition
Artificial Neural Networks
Chaos
Medical diagnosis aid systems
Breast cancer detection
Substance identification
Noise resilience

ABSTRACT

The main purpose in this work is to explore the fact that chaos, as a biological characteristic in the brain, should be used in an Artificial Neural Network (ANN) system. In fact, as long as chaos is present in brain functionalities, its properties need empirical investigations to show their potential to enhance accuracies in artificial neural network models. In this paper, we present brain-inspired neural network models applied as pattern recognition techniques first as an intelligent data processing module for an optoelectronic multi-wavelength biosensor, and second for breast cancer identification. To this purpose, the simultaneous use of three different neural network behaviors in the present work allows a performance differentiation between the pioneer classifier such as the multilayer perceptron employing the Resilient back Propagation (RProp) algorithm as a learning rule, a heteroassociative Bidirectional Associative Memory (BAM), and a Chaotic-BAM (CBAM). It is to be noted that this would be in two different multidimensional space problems. The later model is experimented on a set of different chaotic output maps before converging to the ANN model that remarkably leads to a perfect recognition for both real-life domains. Empirical exploration of chaotic properties on the memory-based models and their performances shows the ability of a specific modelisation of the whole system that totally satisfies the exigencies of a perfect pattern recognition performance. Accordingly, the experimental results revealed that, beyond chaos' biological plausibility, the perfect accuracy obtained stems from the potential of chaos in the model: (1) the model offers the ability to learn categories by developing prototype representations from exposition to a limited set of exemplars because of its interesting capacity of generalization, and (2) it can generate perfect outputs from incomplete and noisy data since chaos makes the ANN system capable of being resilient to noise.

1. Introduction

This paper is an extension of a work originally presented in the First International Conference on Cyber Management and Engineering (CyMaEn`21) [1].

During more than 300 years, there were only two kinds of movements known in simple dynamical systems: the uniform and the accelerated movements. Maxwell and Poincare were among the minority of scientists who disagreed with those facts. It was only in the last quarter of the 20th century that the third kind of movement appeared: chaos [2].

The existence of dynamics and nonlinearity in the brain has been the topic of numerous research investigations since the 1980s. In [3], it was revealed in neurosciences that the activity of the olfactory bulb of rabbits is chaotic and, at any time, it may switch to any perceptual state (or attractor). In fact, the experimentations assessed that when rabbits inhale an odorant, their Electroencephalograms (EEGs) display gamma oscillations, signals in a high-frequency range [4, 5]. The odor information represents then an aperiodic pattern of neural activity that could be recognized whenever there was a new odor in the environment of after a session of training.

*Corresponding Author: Hanae Naoum, Email: h.naoum@esi-sba.dz

Furthermore, during the same period of Freeman's research, other works figured out the existence of chaos in the temporal structure of the firing patterns of squid axons, of invertebrate pacemaker cells, and in temporal patterns of some brain disorders such as schizophrenia and human epileptic EEGs [4-7]. Moreover, in red blood cells, chaotic dynamics of sinusoidal flow were determined by 0-1 test. In fact, numerous simulations identified the existence of chaotic dynamics and complexity in the sinusoidal blood flow [8]. In addition, the exploration of dreaming through the application of concepts from chaos theory to human brain activity during Rapid-Eye-Movement state (REM-state) sleep/dreaming proved that chaos is on the flow of thoughts and imagery in the human mind [8-10]. Finally, chaos is ubiquitous in the brain operations and cognitions according to cognitive sciences, linguistics, psychology, philosophy, medical sciences, and human development [11-15]. The later issues are still addressed in depth in the context of research on the complex systems, to which the brain obviously belongs. This is at this level that the ultimate goal of AI has to be considered. Indeed, creating a machine exhibiting human-like behavior or intelligence, cannot be, with keeping the chaos properties aside.

Moreover, being an offshoot of Artificial Intelligence (AI) paradigms, pattern recognition techniques focus on the identification of regularities in data in an automated process [16]. It is worth noting the fact that, pattern recognition is a cognitive functionality in the brain. In fact, in real life, human beings are capable of recognizing and recalling patterns of different natures and forms (not necessarily perfect patterns) and in different conditions, naturally without significant effort. An intelligent pattern recognition system must thus include brain properties, such as, the presence of chaos.

Furthermore, ANNs represent a discipline of AI that has successfully been applied on different nature of pattern recognition problems. In fact, ANNs models were employed for data compression, data classification, data clustering, feature extraction, etc. Data classification is particularly one of the most active search and application fields in connectionism [16-18]. Consequently, ANN approaches encompasses potential techniques to face pattern recognition problems.

Several works can be noticed in the literature, that focus on the construction of ANN models that implement NDS properties [19-23]. Those proposed models are challenging the classical kinds of ANNs in terms of biological plausibility [24-28], and in some cases, even in terms of computational efficacy of the model [29-32]. Except that, most of the proposed models were developed including the stabilities of attractors with no attention to the ongoing instabilities. The present work shows the chaos potentials in a recurrent ANN model in comparison with two other conventional ANN models. The potentials are such as a perfect pattern recognition accuracy and an excellent resilience to noise. In addition, the model proposed in this paper faces two aspects in the biological plausibility formerly mentioned; the resilience to noise, and as a matter of fact that chaotic properties are actually present in the brain.

In the present work, three different ANN models are investigated as pattern recognition systems in two different real-life classification domains: substance identification and breast cancer identification.

1.1. Substance identification

Sensing technology encloses various instrumentation techniques for variable characterization in diverse aspects of human life [33]. From a hybridization of chemical and physical measurement devices, results the construction of biosensors. Those devices are capable of converting a chemical or physical characteristic of a particular analyte into a measurable signal [34]. Those devices offer a great potential for several integrated applications for rapid and low-cost measurement and were widely used in contrastive scientific practice, certainly owing to their remarkable outcome [33-39]. Biosensors were developed throughout various applications and principally put an accent on the construction of sensing components and transducers. Those applications fall under a multiple substance analysis for diagnosis [40-44], and estimations [45-48].

The technology of sensor devices led to remarkable achievements that are undeniable. Except that, it remains important in any sensing process to create a steady and precise pattern recognition model admitted to the sensory system for substance detection [33]. The raw data that are collected from the sensor need to be analyzed. For that purpose, and to offer a complete integrated instrument, the classical method has suggested incorporating optical filters to the basic sensor, also, the use of statistical and threshold-value based techniques for data processing. Recent researches offer a potential and more efficient alternative: the use of AI paradigms. Plenty of applications can be found in the literature that use these pattern recognition methods for substance detection such like, Decision Trees [38], random forest [34, 43], K Nearest Neighbor [34, 40, 42], Support Vector Machine (SVM) [40-43], and ANNs [36-47].

The authors use in [47] a biochemical sensor to acquire fluorescence measurements from a variety of substances at different concentrations. The sensor prototypes utilized Light Emitting Diodes (LEDs) as excitation sources, as detailed in [47], and LEDs and/or photodiodes as photodetectors.

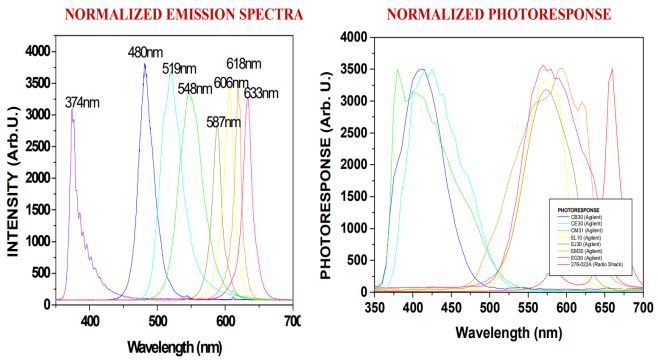


Figure 1: Sample mission and photo response spectra of the color LEDs [47]

Basically, the excitation light procured an interaction with the tested analyte of various aspects such as fluorescence, reflection, absorption, and scattering in a synchronous manner. The resulting light was quantified by the photodetectors having distinct spectral sensitivities. The LEDs were excited synchronously, one at a time, detecting the resultant fluorescence with the remaining LEDs. That was the process followed to collect an amount of data for each analyte at a specific concentration .This process generated a data collection that characterizes a singular spectral signature for a compound at a specific concentration, as illustrated in Figure 1. That process was repeated for all components at different

concentrations. Then, after detecting and amplifying the data with a multi-wavelength sensor front end, they are used to train the ANN and, upon satisfactory training, the network is assigned to the identification of other data collected by the biosensor. The ability to determine very low substance concentration levels using the ANN dramatically increases the specificity of the biochemical sensor.

The focus of classification techniques is, for given input patterns, to detect target classes, determined to define a particular substance concentration pairs. In this context, the authors in [47] developed a MultiLayered neural network, that was trained with the collected data from the biosensor, to process the classification phase on data reserved to test the network. The topology of the network model is basically a Multi-Layered Neural Network (MLNN) which consists of two hidden layers with 56 processing neurons for each layer, and a single output neuron. The RProp algorithm was employed as a learning rule on the network to process the training phase. The learning algorithm in multilayered network models consists basically of two phases. At the beginning, an input pattern is randomly selected from the training dataset and is assigned to the input layer of the ANN. Then, the network propagates that pattern from layer to layer until a corresponding output pattern is computed by the output layer. In case there is a difference between the resulting pattern and the desired output, the error is estimated and then propagated in the opposite direction through the network, from the output to the input layers. In the meanwhile, the weights' values are readjusted, as the error value is propagated backward [16].

In [47], the authors developed the MLNN to detect four fluorescent organic compounds at different concentrations, as one can notice on Table 1. The resulting performance attests a good classification capacity of the network, reaching more than 94% of perfect analyte detection. The error curves for both the training and the recall phases are plotted bellow in Figure 2.

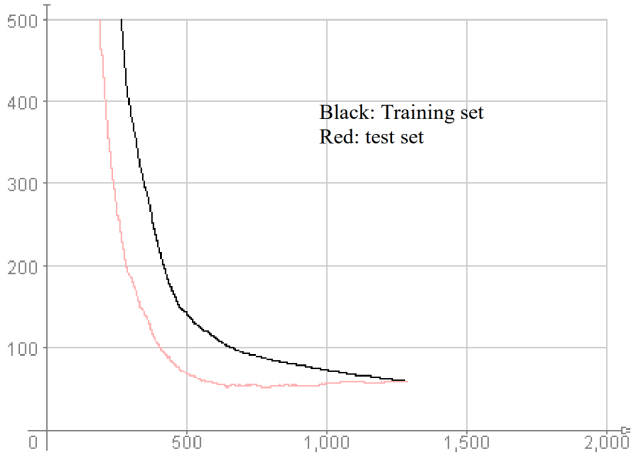


Figure 2: Sum-of-Squares classification error curves versus the number of cycles for the 56-56-56-1 MLNN topology.

Dealing with the same substance identification problem, the authors in [48] developed an evolutionary AI approach, based on Particle Swarm Optimization (PSO), viewing the detection problem as an optimized search. Indeed, the same datasets used in [47] were employed in the experimental set-up. The obtained results with the PSO model enhanced the recognition accuracy reaching 98% of exactitude.

The results of the latter two works and the same data collection were reserved in the present work, with the aim of trying to enhance even more the recognition rates obtained, by incorporating the properties of chaos in ANN models. We investigate a BAM and a CBAM recurrent models to process fluorescence data for the substance identification task.

1.2. Breast cancer detection

Cancer is a disease that might attack numerous human organs. A scourge that continues spreading all over the world with alarming new statistics each year. Statistics report that, breast cancer is the 2nd dangerous disease all over the world, in fact, the rate of mortality from this disease is overwhelming.

The WHO (*World Health Organization*), states that breast cancer affects more than 2 million women every one year across-the-board [49, 50]. In 2020, 2.3 million women were diagnosed with breast cancer and globally 15% of all death among women from cancer was from breast cancer disease. An early diagnosis of the disease, increases the chances of survival for the patient.

The main purpose of medical diagnosis aid systems for breast cancer disease is the detection of non-cancerous and cancerous tumors [49-53]. The only valid prevention approach for breast cancer disease remains the early diagnosis [54-56]. In the 1980s, in developed countries, with the establishment of early detection protocols and a set of treatment processes generate enhancements in survival rates. For a prevention purposes, the National Breast Cancer Foundation (NBCF) prescribe a mammogram once a year for women that are over 40 years old.

Technologies based on AI paradigms are getting more accurate and reliable results than conventional ones. AI tools such as pattern recognition techniques [49, 51, 53, 54, 55, 56], are estimated for being of great help in the medical diagnosis aid field. In fact, as part of breast cancer detection, doctors need to be able to differentiate between categories of tumors through a reliable procedure of examination. Specialists assess the fact that tumors' diagnosis is a task that is considered to be very hard to accomplish. It is thus crucial to diagnose breast cancers in an automated manner to overcome that difficulty.

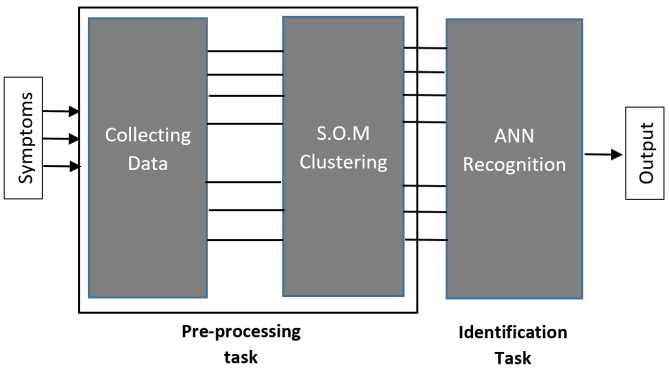


Figure 3: ANN for medical diagnosis aid system

In the field of breast cancer detection, numerous paradigms were employed to construct medical diagnosis aid systems. Those techniques use pattern recognition tools as Random forest [49-55], K-Nearest-Neighbor [52, 54], Logistic regression [51, 53], Naïve Bayes [49, 52, 53, 56], Decision Tree [49, 51-55]; not only but also ANN models [49-55], in particular, Support Vector Machines (SVM), Multi-Layer Neural Networks, Convolutional Neural

Networks, or BAM neural network model. All those techniques have the same aim: the automation of breast cancer identification to assist medical diagnosis protocols.

As for the first pattern recognition task, the substance identification problem, the same three different ANN models are used in this work to face a breast cancer identification problem, such as, a MLNN, a BAM and a CBAM model. Indeed, a different real-life pattern recognition problem is investigated with a different dataset to highlight the potential of chaos in an ANN model's performance. The proposed ANN models are operating as breast cancer diagnosis aid systems according to the process shown in Figure 3. First, symptoms are collected from the diagnosed patients, thereby creating a raw database. Subsequently, raw data undergoes a preprocessing phase before being assigned to the ANN model for the identification phase. The preprocessing phase is detailed in the *Data Acquisition* section.

The rest of this paper is organized as follows. Section 2 proposes briefly an overview about models theory. Section 3 defines the properties and the pre-processing methods used on the two different datasets employed in the experimental set up, such as, the fluorescence based measurements and the breast cancer dataset. Section 4 presents the different parameters' details and implementations' descriptions of the developed ANN models with their respective results. The obtained results are discussed in the last section, the conclusion.

2. Theory of Models

Among the various AI techniques applied in pattern recognition problems, connectionist approaches proved their good ability to process classification, which represents the most active field in the research on ANNs [16, 18, 57]. We present in the rest of this section an overview of research works employing multilayered neural network models and memory-based ones. The kind of ANN models that we are about to present in this paper.

2.1. Multi-Layered Neural Network (MLNN)

The MLNN is built in a multilayer Perceptron fashion. The architecture of such a model is composed of an input layer, one or more hidden layers of computing neurons, and an output layer generating an output pattern corresponding to the input assigned formerly to the network [18]. The MLNN performs a supervised learning. In addition, most of MLNNs processes their learning phase according to the basis of the pioneer backpropagation (Backprop) algorithm or one of its variants. In fact, an error is calculated according to the difference between an actual and a desired output patterns, which is propagated backward again through the network layers and the values of the weights are then updated to reduce it [16].

Basically, the backpropagation algorithm represents a chain rule to estimate the effect of each weight value in the network according to an arbitrary error-function [58]. Once all the weights of the connections are computed, the purpose is to make the error value as small as possible through an error-function. The commonly used error-function is the simple gradient descent including a learning rate parameter. The choice of that parameter has a considerable impact on the number of learning epochs needed for the ANN to converge. On the one hand, the smaller is the learning parameter the greater is the number of learning cycles. On the other, under a large value of the learning rate, the network risks to generate oscillation, which makes difficult to diminish the error value.

Given the limits of the classical Backpropagation rule, it has gone through several improved versions [16, 18, 58, 59], among which, introducing a momentum-term. A parameter that was supposed to make the learning algorithm more stable and the learning convergence quicker. However, it turns out experimentally that the optimal value of the momentum parameter is equally problem dependent as the learning rate, and finally, no general improvement can be carried out.

Later, numerous algorithms were proposed to face the problem of appropriate update of the weight values by using an adaptation parameter in the training epochs. That parameter is used actually to estimate the weight-step. The adaptation algorithms are approximately grouped into two classes, local and global rules [58]. On the one hand, local adaptation rules employ the partial derivative to adjust weight-specific parameters. On the other hand, global ones employ the information concerning the state of the whole ANN model, meaning, it uses the orientation of the previous weight-step, to update global parameters. Basically, the local rules fit better the conception of learning in ANNs. Again, the adaptive enhanced version of the Backpropagation algorithm has certain limitations. Concretely, the impact of the chosen value of the adapted learning parameter is very sensitive to the partial derivative [16, 18, 58].

Finally, the weaknesses of all the aforementioned backpropagation variants took over the conception of the Rprop. The fact that this algorithm updates the size of the weight-update directly and without taking into account the partial derivative's size, keeps the system away from the '*blurred adaptivity'* phenomenon. All the modified verisons of the backpropagation algorithm have the aim to accelerate the neural network convergence, and through various experimentations, the Rprop have proven to be more useful than the others. The Rprop learning scheme offers a great efficacy compared to the classical backpropagation algorithm and its above-mentioned modifications [58]. Basically, that learning rule processes the weight-step adaptation according to a local gradient information. In addition, the contribution made by the Rprop rule is that, the introduction of an individual update-value for each weight avoids the effort of adaptation to be blurred by the gradient behavior. Basically, the individual update-value determines the size of the weight-update. The value of the update parameter is estimated while the training is processed on the network, and that estimation is based on its error-function local information.

Through a set of experimentations on several MLNN models [47], the global error performance of the MLNN employing the RProp algorithm was the smallest. In addition, the speed of convergence of the model was the quickest one. Consequently, the resilient backpropagation is implemented in the learning phase for the MLNN model in the present work. Apart from that, the sigmoid function [59] is employed as the neuron output function of the multilayered model.

2.2. Bidirectional Associative Memory (BAM)

The first ANN model offering a learning process operating in a heteroassociative scheme, was proposed in the 1970s [19]. That ANN was designed with a focus on constructing a formal system that demonstrates the way that brain associates different patterns.

In fact, when training the brain perceive something (input pattern), another one is recalled (output pattern, or a category). The pioneer memory-based ANN model is linear [19], whether through its training rule or its the unit activation function. That in fact limits the recall capacity of the network, especially in case correlation occurs in input patterns. That weakness of the model has pushed the research on the field to evolve towards the construction of recurrent auto-associative and interpolative models of memory that includes nonlinearity [20]. The later models generate dynamic functionalities through a nonlinear output feedback function. That function nonlinearity offers to the system the possibility to converge to stable fixed-points. Consequently, if the networks learned training patterns in correspondence with given fixed-point attractors, it would be capable of recalling them despite the presence of noise in data. The author in [60], used those characteristics to incorporate the nonlinear feedback of the Hopfield model to a hetero-associative memory model. Finally, the BAM came up, a new kind of brain-inspired connectionist models.

A Multilayered ANN model has different functionalities from a memory-based ANN model. Instead of propagating the signal in a layer-by-layer fashion from the input layer to output layer, the BAM consists of generating feedback loops from its output to its inputs; in fact, this model has a recurrent topology [18]. In addition, the learning process in memory-based ANN models are a brain-inspired artificial approach. Indeed, that learning process allows to develop attractors for each pattern since the recurrent architecture offers the ability of feedback connections [21]. Furthermore, learning with BAMs demonstrates a remarkable stability and adaptability against noise and a great capacity of generalization. The memory-based model has also exhibited a great potential for pattern recognition especially given its capacity to be trained under a supervised or an unsupervised scheme. The Hebbian rule is the common learning algorithm used for the unsupervised trainings in the BAM models [16]. If two units in the network are activated in a simultaneous way on either side of a connection, the corresponding weight is then increased. Otherwise, if two units in the BAM are activated in an asynchronous way on either side of a connection, the corresponding weight is thus decreased. Concretely that is the Hebb's law basis. Being the fact that the Hebbian learning consists of an unsupervised learning, the process is local to the network and is performed independently of any external interaction.

The BAM training is originally performed with a classical Hebb's law [16, 20]. Because of its multiple limitations, among which, pattern-correlation, there were numerous enhanced versions of the hebbian learning principle. The first memory-based model employed a nonlinear activation function (the Signum function) in the recall phase. Again, the latter learning rule had some limitations such as it is accomplished offline and the network is limited to bipolar/binary input patterns. In addition to, the BAM generates numerous inaccurate attractors and its memorization aptitude is limited.

To confront those limits, the learning algorithm was modified with the use of a projection matrix following the principle based on least mean squared error minimization. Other alternatives were put forward in the literature tempting to enhance the learning algorithm behaviors. In fact, the proposed models tried to overcome the classical learning rule by increasing the model's storage capacity and his performance, but also by reducing the number of inaccurate states. Unfortunately, most of the proposed processes increases the neural network complexity [20].

In the present work, the authors use a BAM model that allows either an offline or an online learning of the patterns, and most of all, a model that is not limited to memorize binary or bipolar patterns. Indeed, the memory-based ANN model has to be capable of learning real-valued to deal with both of our real-life problems, substance identification and breast cancer detection. In the present work, the learning rule used in the BAM model is derived from the Hebbian/anti-Hebbian rule detailed in [43, 44].

$$W_{[k+1]} = W_{[k]} + \eta(y_{[0]}x_{[0]}^T + y_{[0]}x_{[t]}^T - y_{[t]}x_{[0]}^T - y_{[t]}x_{[t]}^T \quad (1)$$

$$V_{[k+1]} = V_{[k]} + \eta(x_{[0]}y_{[0]}^T + x_{[0]}y_{[t]}^T - x_{[t]}y_{[0]}^T - x_{[t]}y_{[t]}^T \quad (2)$$

W and V in equations (1) and (2) are the weight matrices for both network directions, $x_{[0]}$ and $y_{[0]}$ are the initial inputs to be associated. The variable η represents the training parameter, whole k represents the number of learning cycles. Through $x_{[t]}$ and $y_{[t]}$, a feedback from a nonlinear activation function is included in the learning algorithm; which offers to the network the ability to learn online and then contributes to the convergence of the BAM's behavior. Given those particularities, we opt to develop this learning function on the BAM model in the present work.

It is worth noting that, the cubic map detailed in [20, 22], is used as the unit output function of the memory-based model. The cubic map is employed for the BAM model under a non-chaotic mode, as detailed in section IV.

The training process of the BAM was performed under the basis of the following algorithm:

1) *Selecting randomly a pattern pair from the learning dataset;*
2) *Computing X_t and Y_t according to the output function employed (Cubic-map);*
3) *Computing the adjusted values of the weight matrix according to (1) and (2);*
4) *Reiteration of steps 1) to 3) until the weight marix converges;*

This same learning rule is used further in the third ANN developed model, the C-BAM.

2.3. Chaotic Bidirectional Associative Memory (C-BAM)

Since chaotic patterns were found in the brain [11, 12, 13, 14, 24, 25], numerous research works were proposed in the literature, tempting to include dynamic properties in ANN models. It must be noted that, time and change are the two properties that particularly defines the impressive properties of the NDS approach. As a result, those proposed NDS-based models are confronting the classical doctrines on brain functionalities and most of the theories that were assessed since the inception of neuroscience and cognitive sciences. This comes up with, challenging also the disciplines that focus on the construction of brain-inspired models such like artificial intelligence, and specifically ANNs paradigms.

Furthermore, most of proposed models are of a computational nature and leave dynamic principles aside. Moreover, concerning the models that encompasses NDS characteristics, only fixed points are taken into account to store and retrieve information. As a result, characteristics of the NDS approach are kept aside [46, 50]. Basically, perceived as a nuisance, chaos is, most of the time, excluded from the models.

The main purpose in the construction of brain-inspired AI models, particularly ANNs; cannot however ignore principles of nonlinear dynamical systems.

Moreover, in spite the various existing models in connectionism, not all those models are applicable to include NDS properties. Indeed, memory-based models such as BAMs offer the ability to develop nonlinear and dynamic behaviors. In fact, their recurrent architecture offers characteristics that allow an ANN model to generate oscillations [20, 21, 22, 23, 24, 25].

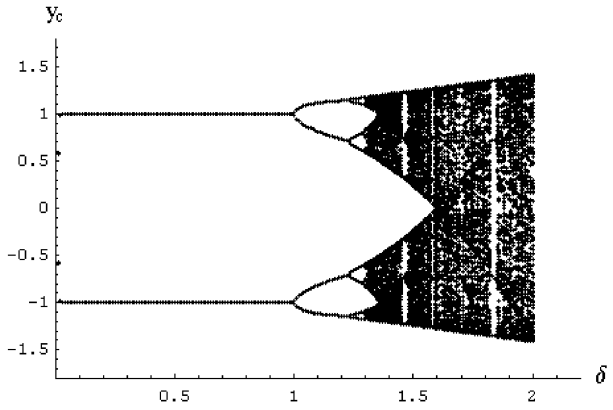


Figure 4: Bifurcation diagram of the cubic map [20].

The parameters employed in the CBAM model are as follows: the learning rule employed in the former BAM model, derived from the Hebbian/antiHebbian algorithm; also, the BAM's topology is kept the same; and concerning the output function, 23 chaotic maps (including the former cubic-map) operating in a chaotic mode are used after training, at the recall phase.

Chaotic maps characteristics

We test a set of 22 chaotic function defined in detail in [2, 61, 62, 63, 64, 65], among which, the Spikin maps family, the Tent maps, the Mira group, the Bernoulli map, and the Henon map. The last map is the one that completes the CBAM model's performance to perfect, at performing both the substance identification task and the breast cancer identification problem, as detailed further in this paper in the Experimentation section.

The 23rd function is the same cubic map used in the former BAM model, except that we set its parameters this time so that it can work in a chaotic mode. The function parameters are set according to its bifurcation diagram in the Figure 4, as detailed further in section IV. As one can notice, the process is capable of leading the system to stable attractors for value $\delta < 1$, although an aperiodic behavior can occur when the parameter value exceeds 1, and then, switching the system into a chaotic phase (black areas in the bifurcation diagram). The 23 output functions tested on the CBAM model are listed in Table 2.

The Henon map

This function is detailed in [2] as follows:

$$f(x, y) = (\alpha - x^2 + \beta.y, x)$$

The Henon map has to inputs, such as x and y, and two outputs, the new values of x and y [2]. The use of the control parameters' values α = 1.28 and β = -0.3, we notice that the orbit converges to a 2-period attractor (as shown in Figure 5), when at the value α=1.4 the attractor becomes fractal [2].

Table 1: The performances of the CBAM with the 23 maps, the BAM and the MLNN model.

ANN Model	Breast Cancer Recognition	Substance Identification	Map ID.
CBAM-Henon	100%	100%	1
CBAM-Bernoulli	100%	75.35%	2
CBAM-Logistic3	98.59%	77.32%	3
CBAM-Mira1	98.59%	72.16%	4
CBAM-Spikin Map3	96.48%	76.83%	5
CBAM-Spikin Map	95.43%	76.66%	6
CBAM-Logistic2	94.38%	74.89%	7
CBAM-Tent	91.92%	75.78%	8
CBAM-Logistic	89.47%	68.77%	9
CBAM-Spikin Map2	84.90%	76.78%	10
CBAM-PWAM2	77.89%	55.93%	11
CBAM-TailedTent1	75.78%	53.52%	12
CBAM-Logistic1	72.62%	51.95%	13
CBAM-PWAM4	70.17%	49.23%	14
CBAM-PWAM3	65.96%	46.34%	15
CBAM-Tent1	62.45%	44.87%	16
CBAM-PWAM1	53.32%	40.21%	17
CBAM-Logistic-Cubic	44.21%	72.98%	18
CBAM-Mira2	22.45%	31.44%	19
CBAM-Spikin Map1	18.94%	45.76%	20
CBAM-Ideka	18.94%	45.89%	21
CBAM-Mira-Gumolski	12.27%	23.91%	22
CBAM-Tent2	4.56%	18.67%	23
BAM	96.56%	90.17%	---
MLNN	89.32%	94.79%	---

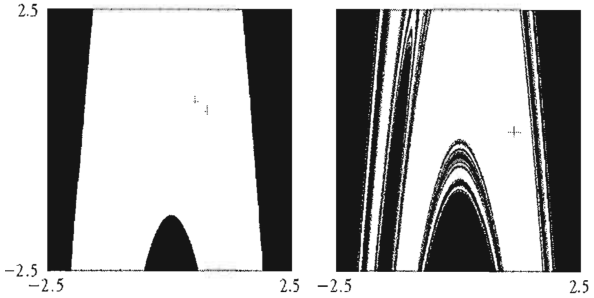


Figure 5: The attraction bassin of the Henon map, β = -0.3 [65].

On the one hand, the black zones in the Figure 5 represent the initial values whose trajectories diverge towards infinity. On the other hand, the white zones represent the initial values that are pulled by the 2-periods attractor. The basin limit is a curve that moves from the inside to the outside of initial values' space.

The Bernoulli map

This chaotic function is defined in [31] as follows:

$x[n] = y[n-1]$,

$y[n] = mod(\delta.x[n-1], 1)$

Where δ is a control parameter set to 1.99 so the map can operate in a chaotic mode [64]. *x* and *y* represent the input and the output of the system respectively.

The Mira maps group

The original Mira map is defined as follows [61]:

$x_{[n]} = y_{[n-l]}$,

$y_{[n]} = y_{[n-l]} - ax_{[n-1]}$ if $x_{[n-1]} < 6$,

$y_{[n]} = y_{[n-l]} + bx_{[n-1]} - 6\ (a + b)$, otherwise

where *x* and *y* represent the initial conditions for the trajectory of the map. After numerous trials on the values of the control parameters of the Mira map, we set them during the final experimentation to: a=1.05; and b=2. The modified Mira map is used to imitate the spiking phenomenon of the biological neurons [32], this map is defined as follows:

$f(x) = y$,

$f(y) = a\,x + b\,x^2 + y^2$

We use this map with the following control parameters' values:

- Mira_map1, a = 0.8, b = 1, where the map has one breast cancer fixed-point with two positive eigenvalues.
- Mira_map2, a = -0.8, b = 0.2, where the map has a stable set.

The dynamic characteristics of the Mira map and its versions are detailed in [47, 48].

The Spiking maps group

The original Spiking_map is defined in [63] as follows:

$$X_{[n]} = f\left(X_{[n-1]}, Y_{[n-1]}\right),$$

$$Y_{[n]} = Y_{[n-1]} - \mu\left(X_{[n-1]} + 1\right) + \mu\sigma,$$

$$f(x,y) = \begin{cases} \dfrac{\alpha}{(1-x)} + y, & if\ x \le 0 \\ \alpha + y, & if\ 0 < x < \alpha + y, \\ -1, & if\ x \ge \alpha + y \end{cases}$$

Where, $x_{[n]}$ is the fast dynamical variable, μ is a constant value set at 0.001, $y_{[n]}$ is the slow dynamical variable, its moderate evolution is on account of to the small value of the parameter μ. The map's control parameters are the variables α and σ. We use also the Spiking Map with three different modifications. We experiment the parameters' values used in [63] so that the map operates in a chaotic mode:

- Spiking_Map1: $\mu = 0.001$; $\alpha = 5.6$ and $\sigma = 0.322$
- Spiking_Map2: $\mu = 0.001$; $\alpha = 4.6$ and $\sigma = 0.16$
- Spiking_Map3: $\mu = 0.001$; $\alpha = 4.6$ and $\sigma = 0.225$

In the present work, we aim at experimenting 23 chaotic maps with the CBAM model, among which we have presented few ones. These selected functions are among the ones that have laid the best accuracy rates.

Furthermore, we have relied on useful mathematical tools to set the different chaotic maps' parameters, among which, the Lyapunov exponent. It represents a logarithmic estimation for the mean expansion rate per cycle of the existing distance between two infinitesimally close trajectories [2]. The interest in the present work concerns particularly the case where that value is positive. It must be noted that, a NDS with a positive Lyapunov exponent characterizes the fact that the system is chaotic. That system is in particular sensible to initial conditions. Figure 6 shows a bifurcation diagram of the Henon map.

For each vertical slice shows the projection onto the *x*-axis of an attractor for the map for a fixed value of the parameter α.

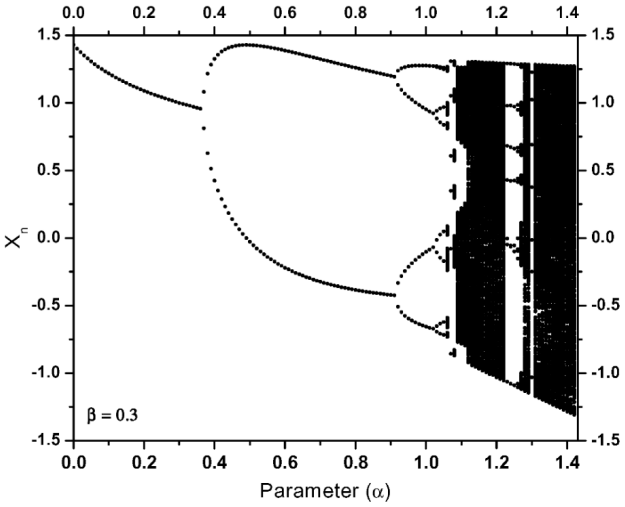


Figure 6: The bifurcation diagram for the Henon map [65].

Whereas, a stable movement has a negative Lyapunov exponent. An example is illustrated in the Figure 7, which concerns the Henon map Lyapunov Exponent.

Both of those mathematical tools, the bifurcation diagram and the Lyapunov exponent, are useful in our experimental set up to have control over the dynamical behaviors of the performed output unit functions.

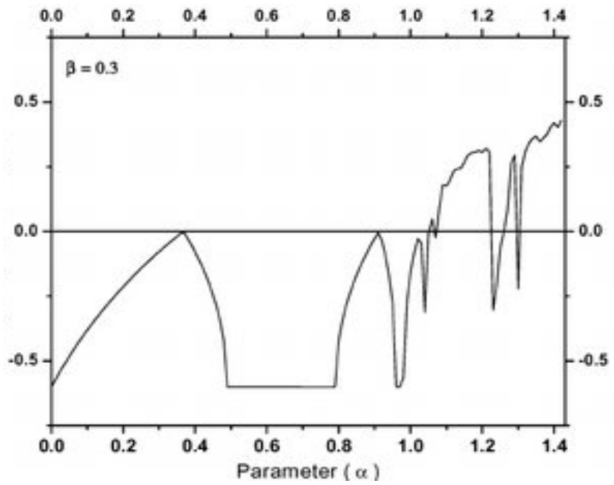


Figure 7: The Lyapunov exponent [65].

3. Data Acquisition

It is a common routine to prepare data before performing the learning and the recall phases on an ANN model. Numerous preprocessing techniques exist in the literature varying according to the nature of the data. We detail in the following subsections the preprocessing procedure applied on both the set of fluorescence-based measurements, and the breast cancer data collection. It is worth noting here that, noise and missing data are two characteristics that are kept in the datasets.

3.1. Fluorescence based measurements

A set of fluorescence-based measurements was collected from the optoelectronic biosensor for each analyte at a specific concentration through the procedure detailed in [47]. The first pattern recognition problem to aboard in the present work concerns the identification of different analytes at different concentrations. Table 1 illustrates the test compounds under their different concentrations.

Table 2: Test substances with their concentrations

Compound	Concentration	Category
Chlorophyll	10-4 M	1
	10-5 M	2
	10-6 M	3
	10-7 M	4
	10-8 M	5
Coumarin	10-3 M	6
	10-4 M	7
	10-5 M	8
	10-6 M	9
	10-7 M	10
Rhomadine B	10-4 M	11
	10-5 M	12
	10-6 M	13
	10-7 M	14
Erythrosin B	10-4 M	15
	10-5 M	16
	10-6 M	17
	10-7 M	18
	10-8 M	19

We have 19 classes among which each class represents one compound at a specific concentration.

Concretely, each measurement in the dataset is composed of 64 values forming an 8x8 matrix. The row in the matrix provides the outputs of 7 photodetectors LEDs [47], and one more output corresponding to the excitation LED which is fixed to 0, and then removed leading to the resulting matrix 8x7. Consequently, the size of each vector in the data collection is 56. Furthermore, a random division of the dataset was employed to get two distinct ones, the first one is dedicated to the learning process (two thirds of the entire collection), while the second one is reserved for the testing phase (the remaining third). The resulting datasets contains 2103, and 1051 vectors respectively.

3.2. Breast Cancer database

Concerning the second pattern recognition problem investigated in the present work, the authors employ the *Yougoslavia Breast Cancer* dataset in the experimentations. Clinical data have been collected by Matjaz Zwitter & Milan Soklic at the oncological institute of the university medical center of Ljubljana in Yougoslavia. The entire data collection contain 286 tumor cases. Each tumor case is represented as a vector of 10 attributes: the tumor frequency, the patient age, the type of the menopause, the tumor-size, inv-node, node-caps, deg-malig, breast position, breast-quad, the irradiation value. It must be noted here that few attributes are missing in the breast cancer dataset. The authors apply scaling on the data collection with dividing the values by 10. The digit 0 was excluded to prevent the system from instable fixe points. As a result, the normalized dataset is represented by digits in the interval [1, 13].

Finally, the data collection is composed of 286 tumor cases coded in a set of rows of dimension 10.

Moreover, the different tumor cases in the breast cancer dataset are not identified. A categorization phase must be accomplished to determine them. As a consequent, the authors developed a Self-Organized Map (SOM), an ANN model detailed in [66]. The designed SOM model consists of a map that is composed of 400 neurons (20*20 cells). Computing units are represented through the matrix cells. As one can notice in Figure 8, the numbers displayed in the 73 cells consist of the resulting categories' identifiers processed by the SOM network.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	16	28	94		159	172	147	47	148	56	9	100	78	97		91			89	115
2	19			104		98	93		107			27	80	1		149	87	156		90
3	12		163					20	22		69	133				143	92	162		166
4	7	38		18	74		55	129	23		135	110			175	112				105
5		70	68		102		17			76	32		36			101	154	127		145
6	41		165			146		15	117											103
7	174		26	108		122	88		111		21	150	120				167	86		153
8								24	161	168	72		79		71	109				126
9	54				119								134	4		53		171		164
10		57	4		66			77									62			158
11	106			65		29		10	114	3	81		2	67						169
12	142	152	125	113			116	40			43	42		128		45	96			140
13	46				49	31	130				75		51	59			50			141
14						64			25					11					137	
15							84	37	60		35	96		5	33	39	85	34		
16							13						48					82		131
17			52					8	121	61	30	63						58		
18								44										83		
19							6		139		73	173	123		160	99				155
20								132	136	138	170		144	118	157	124		151		

Figure 8: The SOM breast cancers' categorization

The obtained map indicates that each neuron in the network has the medical specificities of the category that it represents. As a result, the 175 classes (plotted in Figure 8) generated by the SOM are used for breast cancer identification.

For the experimentation needs, and according to the procedure followed in the substance identification task, the breast cancer dataset was separated into two subsets. The first set contains the equivalent of two-thirds (191 vectors), while the second set contains the remaining third (95 vectors). The first dataset is reserved for the training phase, and the second for the recall phase.

The cross-validation method is used to determine the accuracy rates for the different ANN models developed, and this, whether in substance identification task or in the breast cancer detection problem.

4. Experimentations and Results

It is worth noting that, the topology, the activation function and the learning rule are the main parameters that characterizes an ANN model. We detail in the following subsections those parameters in each of the MLNN, the BAM and the C-BAM models.

4.1. MLNN model

The authors in [47] investigate a set of experimentations on several MLNN architectures and different parameters and converged on the multilayered model employed for the substance identification task. The Stuttgart Neural Networks Simulator (SNNS) of the Stuttgart University in Germany was used for the different MLNN experimentations. The best network performance obtained was with the MLNN topology consisting of an input layer of the size 56 according to the size of the input vectors, two hidden layers with the same size as the input layer, and one output unit generating the output pattern corresponding to input assigned to the network. The aforementioned learning algorithm was used as the learning rule, the RProp. The sigmoid was the output function of the MLNN model.

We kept the same conception principle for the breast cancer detection task. The topology of the resulting MLNN model consists of an input layer containing 10 units according to the size of the tumor-case vector, two hidden layers with 10 computing neurons for each, and one output. The MLNN topology for the breast cancer detection task is plotted in the Figure 9 bellow.

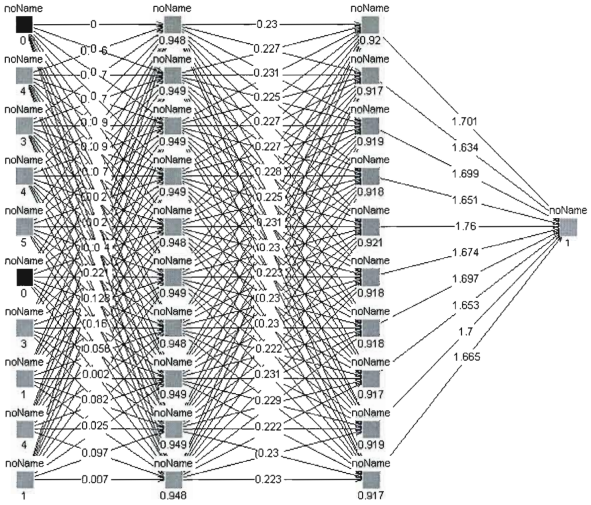


Figure 9: The MLNN topology 10-10-10-1.

The implemented ANN models are tested in the recall phase with the patterns that were not used during the learning process. Meaning that those patterns were not affected to the network during its learning phase, but during the recall only. In addition, the cross validation technique is used to determine the different ANNs' recognition accuracies. Finally, once the tests are achieved, the average of the three experimentations on each of the ANN models is considered as its overall classification exactitude. Empirically, the MLNN model reached 94,79% of good overall recognition for the substance identification task, 89.32% of exactitude for the breast cancer identification problem.

4.2. BAM model

The memory-based network architecture is the second ANN model developed in the present work and it is composed of two Hopfield-like neural networks interconnected in head-to-tail fashion, as one can notice in Figure 10. We employ in the BAM model the parameters experimented in [20, 21].

The network topology describes an interconnection that allows a recurrent flow of information that is processed bidirectionally. In that way, the vectors composing the pairs to be learned do not have to be specifically of same dimensions and that, contrary to the conventional BAM designs, the weight matrix from one side is not necessarily the transpose of that from the other side.

The unit activation function employed in the BAM model is the cubic map described in [21]. Figure 4 illustrates the bifurcation diagram of that function according to δ, the parameter that dictates the dynamic behavior of the outputs.

Fundamentally, this cubic function has three fixed points, -1, 0, and 1, of which both the values-1 and 1 are stable fixed points. They offer to the memory the possibility to develop two attractors at these values. The cubic output function takes several time steps to converge. First, the given stimulus is projected from the network space to the stimuli space. Second, in the following time steps, the stimulus is progressively pushed toward one of the stimuli space corners.

Furthermore, according to Figure 4, one can notice that the learning rule leads to stable attractors for value $\delta < 1$, when it gets an aperiodic behaviour when exceeding 1 before leading the system into a chaotic phase (black areas in the bifurcation diagram).

The experimentations on the BAM were realized with the cubic map operating in a fixed-points mode. The Hebbian/antiHebbian learning rule includes a feedback from the nonlinear output function via the couple of patterns to be associated; which allows the BAM to learn online, thus contributing to the convergence of the weight connections.

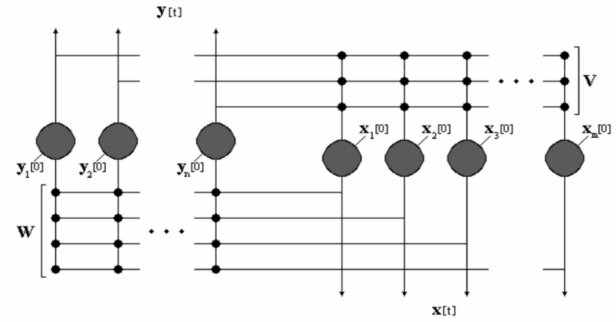


Figure 10: The BAM architecture.

The breast cancer recognition accuracy of the BAM model is plotted in Figure 11.

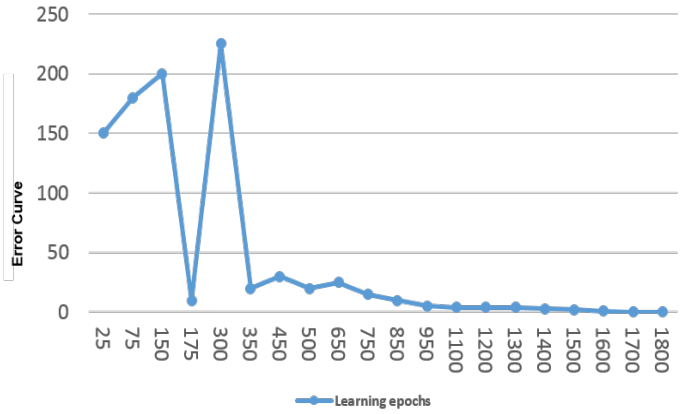


Figure 11: Error curve relative to the number of learning epochs.

The network could reach a steady state after a reasonable number of learning cycles. The error was less than 0,0005 after 1700 epochs. The BAM could correctly categorize 85 tumors cases from the testing dataset. That means that the accuracy rate of the network reached 90,17% at the recall phase. It is worth here that, that recognition rate was accomplished inspite the missing values in the data, which proves a good capacity of generalization on the one hand and a good resilience to noise on the other.

The accuracy increased with the substance identification task to 96,56 % of exactitude. That is certainly due to the better conditions of the data collection despite the larger problem space.

4.3. CBAM model

The third model developed in the present work is the chaotic BAM model. The same main BAM's parameters are kept in the C-BAM network, consisting of the same topology and learning rule as the ones of the former BAM, detailed in [20], except that, the neuron activation function was replaced by chaotic functions. Distinction between the two memory-based ANN models throughout that parameter offers the possibility to concretely estimate the network pattern recognition performance with, and without chaos. The first chaotic output function tested on the C-BAM is the same cubic map employed in the BAM model, operating in a chaotic mode during recall. For that purpose, the value of δ was set to 0.1 during training and to 1.5 during recall.As one can notice in Figure 4, those values correspond to a fixed-point and chaotic behavior respectively for that output map. Subsequently, the experimentation of 22 other chaotic maps on the CBAM model were investigated. Indeed, after reaching a steady state in the learning process, the parameters of each function were fixed so it can operate in a chaotic mode, according to the bifurcation diagram of each map [2, 61, 62, 63, 64].

The recognition accuracy rates are illustrated in Table 2 for both substance identification and the breast cancer identification tasks. For the purpose of results comparison, the MLNN's and the classical BAM's performances are also mentioned.

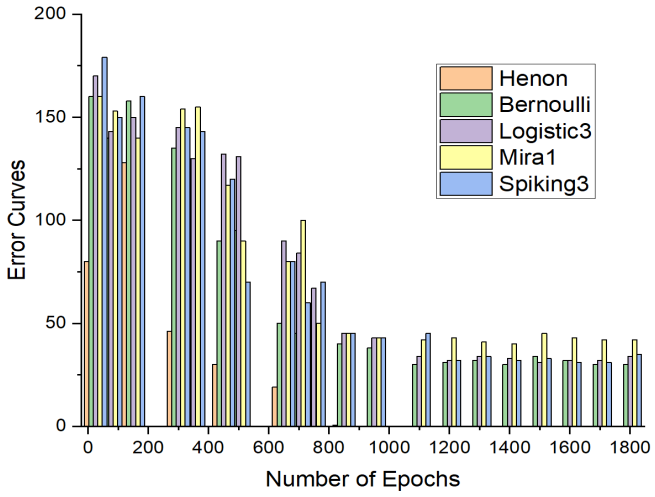


Figure 12: The CBAM error curves for the substance identification's task with the best 5 performing chaotic maps.

The Figure 12 and the Figure 13 show the best recognition accuracies of the CBAM model with five particular maps, for the substance identification task, and the breast cancer identification task, respectively.

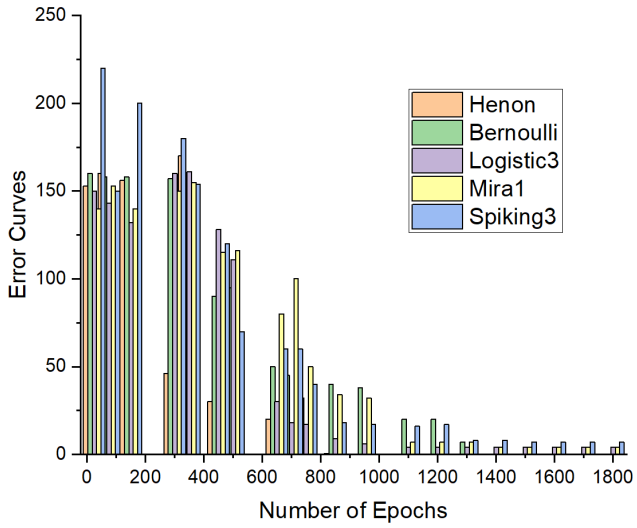


Figure 13: The CBAM error curves for the breast cancers' recognition task with the best 5 performing chaotic maps.

As one can notice in the above graphs, the accuracy is total in both problem domains, particularly with the Henon function.

5. Conclusion

Three different ANN models were developed in the present work to deal with two different real life problems. Both of those problems focus on pattern recognition, the first task concerns substance identification while the second is about breast cancer detection. On the one hand, the MLNN model reached a good performance at recalling the substance identification data despite the problem of the huge space dimension (56 is the vectors' size) and the multiclass criterion (19 different classes). The rate was less good for breast cancer identification with the same model in spite of this; the problem of space dimension was diminished to 10. This fact is indicative of the poor generalization capacity of the Multi-Layered Neural Network when data contains noise. In fact, as mentioned in the *Data Characteristics* section, noise and missing values are two properties that are kept in the datasets. On the other hand, the BAM recurrent model provides a good overall recall for breast cancer identification seeking more that 96% of exactitude. However, the error increased with the multi-class problem relative to fluorescent-based measurements. The BAM results highlight its good resilience to noise; nevertheless, it was less good at facing the large problem space of the fluorescence-based measurements.

In addition, the presence of chaos in the brain-inspired memory-based model provided remarkable results particularly with certain chaotic maps. The recognition accuracy of the CBAM facing the breast cancer detection task was acceptable with six functions employed in a chaotic mode [70% to 89%]; while eight maps varied from good to perfect, reaching a 100% of correct recognition with the Henon and Bernoulli map. The performance was less good with regard to the substance identification problem compared to the first one, and that is with almost all the chaotic maps except for the Henon map that kept the overall accuracy total. The pattern recognition system employing that particular map was remarkable dealing with both substance identification and breast cancer detection problems. Accordingly, we can state that this particular model encompasses assets which allow it an excellent generalization capacity and a great resilience to noise leading to perfect pattern recognition performance.

It is crucial to line up future artificial neural networks investigations with the dynamical characteristics of the Henon chaotic function, among which, the fractal dimensions of the Henon attractors. In essence, fractals and modern chaos theory radically question the dynamical concepts in all contexts and more particularly in nature and its mimetic artificial systems, among which, artificial neural networks. Besides, and to conclude, since it is the era of big data; chaos must imperatively be in the perspectives of deep learning techniques in artificial neural network models' analytics.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This study was funded by the Algerian ministry of higher education and scientific research and the General Directorate for Scientific Research and Technological Development (DG-RSDT). This work was also possible thanks to the financial support of excellence scholarships foundation of the University of Quebec at Montreal (UQAM) for graduate studies, Montreal, Quebec, Canada. Many thanks to the Texas-center of Superconductivity and Advanced Materials (TcSAM) in the Physics Department of the University of Houston, Houston, Texas, USA, for allowing us to use their data in our experimentations.

References

[1] H. Naoum, S. Benslimane, M. Boukadoum, "Classical and Brain-inspired Neural Networks for Substance Identification and Breast Cancer Detection: The Chaos Challenge," The first international conference on Cyber Management and Engineering (CyMaEn`21), IEEE, 1–6, 2021, doi: 10.1109/CyMaEn50288.2021.9497280.

[2] K. T. Alligood, T. D. Sauer, J. A. Yorke, "Chaos: An Introduction to Dynamical Systems," Textbook in Mathematical Sciences. Springer, New York, NY, 105–147, 1996, doi:10.1007/b97589.

[3] W. J Freeman, "Simulation of chaotic EEG patterns with dynamic model of the olfactory system," Biological Cybernetics, **56**(2–3), 139–150, 1987.

[4] H. Korn, P. Faure, "I there chaos in the brain? Experimental evidence and related models," Neurosciences, **326**(9), 787–840, 2003.

[5] A. Combs, S. Krippner. W. Freeman, "III and the chaotic nature of deams," Nonlinear Dynamics, Psychology and life sciences, **21**(4), 475-484, 2017.

[6] V. V. Kozlova, V. A. Galkin, M. A. Filatov, "Diagnostics of brain neural network states from the perspective of chaos," Journal of physics: conference series. 1889 052016, 2021, doi:10.1088/1742-6596/1889/5/052016.

[7] M. AR. Thabet, "Quantum chaos and the brain," IBCHN - Imperial Collage - Michael Crawford's Lab, 2020, doi:10.13140/RG.2.2.20160.48645.

[8] B. Yan, S. Mukherjee, A. Saha, "Exploring noise-induced chaos and complexity in a red blood cell system," Springer, The European Physical Journal Special Topics, **230**, 1517–1523, April 2021, doi:10.1140/epjs/s11734-021-00030-2.

[9] N. B. Harikrishnan, N. Nagaraj, "When noise meets chaos: stochastic resonanace in neurochaos learning," Elseivier, **143**, 425–435. Special Issue, 2021, doi:10.1016/j.neunet.2021.06.025.

[10] G. Eason, B. Noble, I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, **A247**, 529–551, April 1955, doi:10.1098/rsta.1955.0005.

[11] A. Babloyantz, C. Lourenc-o, "Computation with chaos: A paradigm for cortical activity," Proceedings of the National Academy of Sciences, **91**, 9027–9031, 1994, doi:10.1073/pnas.91.19.9027.

[12] M. P. Dafilis, D. T. J. Liley, P. J. Cadusch, "Robust chaos in a model of the electroencephalogram: Implications for brain dynamics," Chaos, **11**, 474–478, 2001, doi:10.1063/1.1394193.

[13] H. Korn, P. Faure, "Is there chaos in the brain? II. Experimental evidence and related models," Comptes Rendus Biologies, **326**(9), 787–840, 2003, doi:10.1016/j.crvi.2003.09.011.

[14] M. A Rozhnova, E. V Pankratova, S. V Stasenko, V. B Kazantsev, "Bifurcation analysis of multistability and oscillation emergence in a model of brain extracellular matrix," Elseivier, Chaos, Solitons & Fractals, **151**, October 2021, doi:10.1016/j.chaos.2021.111253.

[15] A. Wu, Y. Chen, Z. Zeng, "Quantization synchronization of chaotic neural networks with time delay under event-triggered strategy," Springer Verlag, Cognitive Neurodynamics, **15**, 897–914, 2021, doi: 10.1007/s11571-021-09667-0.

[16] M. Negnevitsky, "Artificial Intelligence: A guide to Intelligent Systems," Addison Wesley, 3rd edition, 2011.

[17] K. Saravanan, S. Sasithra, "A review on Classification Based on Artificial Neural Networks," International Journal of Ambient Systems and Applications (IJASA), **2**(4), 11–18, 2014, doi:10.5121/ijasa.2014.2402.

[18] S. Haykin, "Neural networks: A comprehensive foundation," Englewood Cliffs, NJ: Prentice-Hall, 1999.

[19] T. Kohonen, "Correlation matrix memories," IEEE Trans. Comput. , **C-21**, 353–359, Dec. 1972, doi: 10.1109/TC.1972.5008975.

[20] S. Chartier, M. Boukadoum, "A bidirectional Heteroassociative Memory for binary and Grey-Level Patterns, " IEEE Transactions on Neural Networks, **17**(2), March 2006, doi: 10.1109/TNN.2005.863420.

[21] S. Chartier, M. Boukadoum, "Encoding static and temporal patterns with a bidirectional heteroassociative memory,"Journal of applied mathematics, **20011**, 1–34, 2011, doi: 10.1155/2011/301204.

[22] S. Chartier, S. Helie, M. Boukadoum, R. Proulx, "SCRAM: statistically converging recurrent associative memory," IEEE International Joint Conference on Neural Networks, IJCNN, 2005, doi: 10.1109/IJCNN.2005.1555941.

[23] S. Chartier, M. Renaud, M. Boukadoum, "A nonlinear dynamic artificial neural network model of memory," New Ideas in Psychology, **26**(2), 252–277, 2008, doi:10.1016/j.newideapsych.2007.07.005.

[24] M. Adachi, K. Aihara, "Associative dynamics in a chaotic neural network. Neural Networks," **10**(1), 83–98, 1997, doi:10.1016/S0893-6080(96)00061-5.

[25] K. Aihara, T. Takabe, M. Toyoda, "Chaotic neural networks," Physics Letters A, **144**(6–7), 333–340, 1990, doi:10.1016/0375-9601(90)90136-C.

[26] H. Imai, Y. Osana, M. Hagiwara, "Chaotic analog associative memory," Systems and Computers in Japan, **36**(4), 82–90, 2005, doi: 10.1109/IJCNN.2001.939522.

[27] R. S. T. Lee, "e-associator: A chaotic auto-associative network for progressive memory recalling," Neural Networks, **19**(5), 644–666, 2006, doi: 10.1016/j.neunet.2005.08.017.

[28] Y. Osana, M. Hagiwara, "Knowledge processing system using improved chaotic associative memory," Proceeding of the International Joint Conference on Neural Networks (IJCNN'00), **5**, 579–584, 2000.

[29] U. Ozdilek, "Value order in disorder," Springer, International Journal of Dynamics and Control, 2022, doi:10.1007/s40435-021-00903-3.

[30] H. Lin, C. Wang, Q. Deng, C. Xu, Z. Deng, C. Zhou, "Review on chaotic dynamics of memristive neuron and neural network," Nonlinear Dynamics, **106**, 959–973, 2021, doi:10.1007/s11071-021-06853-x

[31] Y. Zhang, Y. He, F. Long, "Augmented two-side-looped Lyapunov functional for sampled-data-based synchronization of chaotic neural networks with actuator saturation," Elsevier, Journal of Neurocomputing, **422**, 287–294, 2021, doi:10.1016/j.neucom.2020.09.018.

[32] C. Chen, A. Abbott, D. Stilwell, "Multi-Level generative chaotic recurrent network for image inpainting," Proceedings of the IEEE CVF winter conference on applications of computer vision (WACV), 3626-3635, 2021, doi: 10.1109/WACV48630.2021.00367.

[33] H. Kaur, A. Bhosale, S. Shrivastav, "Biosensors: Classification, fundamental characterization and new trends: A Review," International Journal Of Health Sciences and Research, **8**(6), 315–333, 2018.

[34] M. H. Mozaffari, L. Tay, "A review of 1D Convolutional Neural Networks toward Unknown Substance Identification in Portable Raman Spectrometer," arXiv:2006.10575 [eess.SP] (2020).

[35] R. Fleh, M. Othman, S. Gomri, "WO3 sensors array coupled with pattern recognition method for gases identification," 13th International Multi-Conference on systems, Signals and Devices, IEEE, 147–152, 2016.

[36] D. Karakaya, O. Ulucan, M. Turkan, "Electronic nose and its applications: A survey," International Journal of Automation and Computing, **17**, 179–209, 2020, doi:10.1007/s11633-019-1212-9.

[37] B. Podola, M. Melkonian, "Genetic programming as a tool for identification of analyte-specificity from complex response patterns using a non-specific whole-cell biosensor," Biosensors and Bioelectronics **33**, 254-259, 2012, doi: 10.1016/j.bios.2012.01.015.

[38] M. Kukade, T. Karve, D. Gharpure, "Identification and classification of spices by Machine Learning," IEEE International Conference on Intelligent Systems and Green Tchnology (ICISGT), 2019, doi:

10.1109/ICISGT44072.2019.00015.
[39] A. L. Vazquez, M. M. Domenech Rodriguez, T. S. Barrett, S. Schwartz, N. G. Amador Buenabad, M. N. Bustos Gamino, M. L. Gutierrez Lopez, J. A. Villatoro Velazquez, "Innovative Identification of Substance Use Predictors: Machine Learning in a National Sample of Mexican Children," Journal of Society for Prevention Research, Springer-Verlag, 2020, doi: 10.1007/s11121-020-01089-4.
[40] F. L. Melquiades, A. Mattos Alves da Silva, "Identification of sulphur in nail polish by pattern recognition methods combined with portable energy dispersive X-ray fluorescence spectral data," Journal of Analytical Methods, **8**, 3920–3926, 2016, doi:10.1039/C6AY00195E.
[41] Z. Almheiri, M. Meguid, T. Zayed, "Intelligent Approaches for predicting failure of water mains," Journal of Pipeline Systems Engineering and Practice, **11**(4), 1949–1190, 2020, doi:10.1061/(ASCE)PS.1949-1204.0000485.
[42] F. Hu, M. Zhou, P. Yan, K. Bian, R. Dai, "PCANet: A common solution for laser-induced fluorescence spectral classification," IEEE Access **7**, 2169–3536, 2019, doi: 10.1109/ACCESS.2019.2933453.
[43] L. G. Zhang, X. Zhang, L. J. Ni, Z. B. Xue, X. Gu, S X. Huang, "Rapid identification of adulterated cow milk by non-linear pattern recognition methods based on near infrared spectroscopy," Food Chemistry **145**, 342–348, 2014, doi:10.1016/j.foodchem.2013.08.064.
[44] P. T. Hernandez, S. Hailes, I. P. Parkin, "Cocaine by-product detection with metal oxide semiconductor sensor arrays," Royal Society of Chemistry, **10**, 28464–28477, 2020, doi:10.1039/D0RA03687K.
[45] Y. Hui, X. Xue, Z. Xuesong, W. Yan, Z. Junjun, "Bacteria strain identification with fluorescence spectra," Applied Mechanics and Materials, **865**, 630–635, 2017, doi:10.4028/www.scientific.net/AMM.865.630.
[46] L. Poryvkina, V. Aleksejev, S. M. Babichenko, T. Ivkina, "Spectral pattern recognition of controlled substances in street samples using artificial neural network system," Optical Pattern Recognition, Proceedings of SPIE, **8055**, 2011, doi:10.1117/12.883408.
[47] H. Naoum, M. Boukadoum, C. Joseph, D. Starikov, A. Bensaoula, "Intelligent Classifier Module for Fluorescence Based Measurements," Proc. International Workshop on Signal Processing and its Applications (WoSPA 2008), Sharjah (UAE), 18–20 2008.
[48] N. Nouaouria, M. Boukadoum, "A Particle Swarm Optimization Approach for Substance Identification," The Genetic and Evolutionary Computation Conference (GECCO), 1753–1754, 2009, doi:10.1145/1569901.1570142.
[49] A. R. Vaka, B. Soni, R. K. Sudheer, "Breast Cancer Detection by leveraging Machine Learning," The Korean Institute of Communication and Information Sciences (KICS), 10–1016, 2020, doi:10.1016/j.icte.2020.04.009.
[50] H. Jouni, M. Issa, A. Harb, G. Jacquemod, Y. Leduc, "Neural Network Architecture for Breast Cancer Detection and Classification," IEEE International Multidisciplinary Conference on Engineering Tchnology (IMCET), 987-1-5090-5281-3, 2016, doi: 10.1109/IMCET.2016.7777423.
[51] S. Sharma, A. Aggarwal, T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," International conferfence on computational techniques, electronics and mechanical systems (CTEMS), IEEE, 2019, doi: 10.1109/CTEMS.2018.8769187.
[52] J. Sivapriya, V. Aravind Kumar, S. Siddarth Sai, S. Sriram, "Breast cancer prediction using machine learning," International Journal of Recent Technology and Engineering (IJRTE), **8**(4), 2019, doi:10.35940/ijrte.D8292.118419.
[53] K. Santhosh, T. Daniya, J. Ajayan, "Breast Cancer Prediction Using Machine Learning Algorithms," International Journal of Advanced Science and Technology, **29**(3), 7819–7828, 2020.
[54] B. Karthikeyan, G. Sujith, H. V. Singamsetty, P. V. Gade, S. Mekala, "Breast cancer cetection using machine learning," International Journal of Advanced Trends in Computer Science and Engineering, **9**(2), 981–984, 2020, doi:10.30534/ijatcse/2020/12922020.
[55] A. E. Bayrak, P. Kirci, T. Ensari, "Comparison of machine learning methods for breast cancer diagnosis," Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science (EBBT), Istambul, Turkey, 1-3, 2019, doi: 10.1109/EBBT.2019.8741990.
[56] M. Karabatak, " New classifier for breast cancer detection based on Naïve Bayesian," Measurement. **72**, 32–36, 2015.
[57] B. Li, C. Delpha, D. Diallo, A. Migan-Dubois, "Application of artificial neural networks to photovoltaic fault detection and diagnosis: A review," Elsevier, Renewable and Sustainable Energy Reviews, **138**, 2021, doi:10.1016/j.rser.2020.110512.
[58] M. Riedmiller, H. Braun, "A direct adaptive method for faster backpropagation learning: the RPROP algorithm,"IEEE International Conference on Neural Networks, 1993, doi: 10.1109/ICNN.1993.298623.
[59] J. Han, C. Moraga, "The influence of the sigmoid function parameters on the speed of backpropagation learning," Springer Verlag, International Workshop on Artificial Neural Networks, 195–201, 1995, doi:10.1007/3-540-59497-3_175.
[60] B. Kosko, "Bidirectional associative memories," IEEE Trans. Syst., Man, Cybern., **18**(1), 49–60, Jan.-Feb, 1988, doi: 10.1109/21.87054.
[61] I. Gumowski, C. Mira, "Recurrences and discrete dynamic systems," Lecture notes in mathematics, book series, **809**, 1980, doi:10.1007/BFb0089135.
[62] J. P. England, B. Krauskopf, H. M. Osinga, "Computing One-Dimensional Stable Manifolds and Stable Sets of Planar Maps without the Inverse," SIAM J. Applied Dynamical Systems, Society for Industrial and Applied Mathematics, **3**, 161–190, 2004, doi:10.1137/030600131.
[63] A. L. Shilnikov, N. F. Rulkov, "Origin of Chaos in a Two-Dimensional Map Modeling Spiking-Bursting Neural Activity," International Journal of Bifurcation and Chaos (IJBC), 13(11), 3325–3340, 2003, doi:10.1142/S0218127403008521.
[64] S. J. Baek, E. Ott, "Onset of synchronization in systems of global1y coupled chaotic maps," Physical Review Letters, 69(6), 066210 2004, doi: 10.1103/PhysRevE.69.066210.
[65] V. Patidar, G. Purohit, K.K Sud, "Dynamical behavior of q-deformed Hénon map". International journal of bifurcation and chaos, **21**(05), 1349–1356, 2011, doi:10.1142/S0218127411029215.
[66] T. Kohonen, "Self-Organizing Maps," In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, 2011, doi:10.1007/978-0-387-30164-8_746.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 3, 44-50 (2022)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Effectiveness of Gamified Instructional Media to Improve Critical and Creative Thinking Skills in Science Class

Neni Hermita [*,1], Rian Vebrianto[2], Zetra Hainul Putra[1], Jesi Alexander Alim[1], Tommy Tanu Wijaya[3], Urip Sulistiyo[4]

[1]*Faculty of Teacher Training and Education, Universitas Riau, Pekanbaru, 28154, Indonesia,*

[2]*Fakultas Tarbiyah, UIN SUSKA Riau, Pekanbaru, 28154, Indonesia*

[3]*School of Mathematical Sciences, Beijing Normal University, Beijing, 065001, China*

[4]*Faculty of Teacher Training and Education, Universitas Jambi, Jambi, 36131, Indonesia*

ARTICLE INFO

Article history:
Received: 05 December, 2021
Accepted: 04 May, 2022
Online: 25 May, 2022

Keywords:
Critical thinking skills
Creative thinking skills
Gamification
Genially
Heat transfer concept

ABSTRACT

Gamified Instructional Media has recently been widely used in the education sector to improve students' abilities. Using Gamified Instructional Media at the elementary school level becomes more interesting because it is in accordance with the way children learn K1-K6. The research aims to identify the gamified instructional using Genially to improve students' critical and creative thinking skills. A quasi-experimental method was applied using a nonequivalent control group research design. The research subject is 40 students of Public Primary School in Pekanbaru. The results show a significant effect of the gamified instructional learning using Genially toward students' critical and creative thinking skills. Besides, there is a significant difference in students' critical and creative thinking skills between the control and experimental group. This study implies that gamified instructional media with Genially can support teachers and teaching practices.

1. Introduction

Artificial intelligence contributes to education, particularly to the implementation of educational process, particularly in the teaching and learning scheme in the industrial revolution of the 21st century. It is necessary to conduct an online or blended learning program to respond the health situation we live in [1]. Despite the students appear to miss the face-to-face meetings interaction in a physical classroom setting, they can also adapt well to the sudden changes from offline to online settings [2]. This phenomenon further indicates that most of the current university students are ready to participate in an innovative educational procedure primarily based on blended learning activities. The strength of such educational setting is that the students are able to learn technical skills in their personal environment without any pressure as offered by the online platform, while simultaneously obtaining social resources in a classroom environment [3]. There were various 21st-century skills that students can improve through online platform, including critical and creative thinking skills [4, 5].

According to [6,7], critical thinking requires three components: (1) a disposition to thoughtfully analyze the issues and subjects experienced by someone, (2) knowledge concerning logical exploration and argumentation procedures, and (3) certain competence in using those approaches. Paul and Binker (1990) state that critical thinking skill is the ability and disposition to critically evaluate a belief, what assumptions are based on it and on what basis these assumptions can survive [8]. In [9, 10], the author stated that critical thinking skill is defined as the ability to make decisions by considering the facts available, the situation's context, and the concept raised. In [11], the authors then added that critical thinking is a form of a rational reflective thinking process that focuses on determining what to believe or what to do. Students are expected to have the ability to think critically if they are able to ask something and find information appropriately. Based on the analysis of the information and knowledge they have, students try to answer problems logically and creatively with conclusions that are acceptable to common sense [12].

Creative thinking skill is the ability of students to identify, solve or find solutions, and solve various problems by looking for alternative problem solving based on their own abilities and thoughts [13–15]. The ability to think creatively can be measured

*Corresponding Author: Neni Hermita, Email: neni.hermita@lecturer.unri.ac.id

by several indicators including (1) fluency which means the capacity of sharing ideas, (2) flexibility which means the skill to suggest several problem solutions, (3) originality which is the capability to develop new ideas as a result of their own thought process, and (4) elaboration which is the capability to describe something in detail [16].

In practice, there are some factors causing students' low critical and creative thinking skills one of which is teacher-centered learning approach that the teacher function in the classroom is more like a lecturer presenting instructional materials and the students are expected to passively receive the knowledge being presented. Science teachers should applied student-centered learning approach in which students are allowed to hone and use their critical and creative thinking skills [17]-[20]. In the industrial revolution of the 21st century, direct learning model increasingly makes students not accustomed to think critically and creatively. Therefore, by using digital technology teachers can present science instructional materials which can allow the students to have a better critical thinking skill [21].

In the Industrial revolution of the 21st century, teachers usually use technology in presenting the instructional materials. The role of technology in education is very significant [16]. Almost all learning processes in elementary schools today involve technology such as using applications as instructional media. This is commonly known as blended learning, and the use of technology aims to support in achieving learning objectives and in creating a different learning experience for students.

Gamification is the use of game in a non-game environment to increase students' learning motivation [22]. Gamification increases students' participation in the learning process [23]. Children generally enjoy playing games and such phenomenon may be applied in learning context, prompting the development of a new teaching technique: gamification [24–27]. There are several applications that can be implemented to innovate interactive learning media one of which is Genially.

Genially is a learning media creation platform that has been widely used in education. vidergor [28] uses genially to design a digital escape room to increase elementary school students' collaboration and motivation. Genially has the advantage of being easy to use and accessible [29, 30]. The features in Genially are suitable for beginner developers, so teachers at schools can design learning media using Genially according to their needs [31]. Genially can also increase engagement in learning and allow students to may share their knowledge and improve their communication abilities [32].

The study conducted by [24] found that a game increased the knowledge of Serbian fifth graders in recognizing plants. In education, Gamification may help students improve their computational thinking skills and motivate them to develop their learning capacities on their own by assisting them in increasing their insight, processing the information, communication, and community awareness skills [33]. During the learning process, students get innovation and new insight. In other words, students get a new atmosphere in the learning process. Therefore, students easily understand the learning materials [34]. Science learning involves students directly in acquiring knowledge as a result of student curiosity. The effectiveness of gamification is different from each student. Gamification must be studied and implemented with care, by paying attention to several factors such as including individual learning styles and personality characteristics [35]. Summary of the findings reported Gamified learning experiences developed using the software package Genially were shown to enhance the students' critical thinking competence.

Through information technology, blended learning is able to enhance students' critical thinking skills [36]. In this case, the teachers conduct the learning process by employing learning media in the forms of technology [37]. In particular, gamification can improve the performance of academic skills [38]. Academic skills can be developed through oral discussion, critical thinking, vocabulary development, oral interpretation, creative acting, observation and recording, information research, graphs and graph interpretation, and summaries. In a classroom setting, a physical and intellectual setting provided has the ability to supports the development critical thinking through a spirit of discovery [39]. Critical and creative thinking skill are two skills that must be included in school curriculum in the 21st century.

Gamification such as digital escape room was implemented as a teaching approach, particularly in science classes for the fourth-graders students of primary school. This approach has been proven to affect students effectively, so that they become more motivated and able to solve problems they face [40]. In addition, both critical and creative thinking skills significantly affected the cognitive learning outcomes [41]. Compared from urban environments, higher levels of realism were believed to boost the restorative effects of viewing natural environments and promote creative thinking [42].

The learning objective of learning science in primary schools is to develop students' process skills in investigating nature, solving problems and making decisions, and applying the learning experiences gained in the previous learning process that has been done [43]. Science learning at the fifth-grade primary school level includes several materials that are human and animal organs, green plants, adaptation of living things and their environment, the building blocks of objects and their properties, changes in the properties of objects, forces, simple machines, light, earth, and the universe. The instructional materials of heat transfer are learnt in the sub unit of changes in the properties of objects. The instructional materials are basic materials at the primary school level so that the instructional materials must be presented as attractively as possible [44]. The presentation of the heat transfer materials should be done in an interesting way. For example, it can be done by using Genially.

Gamification is expected to make students able to remember science concepts and be able to apply learning in everyday life [38]. According to [45], Improvements in critical thinking abilities can be identified both from novel biological and non-biological daily issues, indicating that thinking skills can be implemented in various aspects. Furthermore, knowledge test, the experimental students outperformed the control group, implying that "knowledge of facts" and "learning to think" both as educational purposes should be able to interact with each other.

In science learning, teachers can design learning that can improve the critical and creative thinking abilities of the students by using gamified instructional media using Genially since it has many features that support interactive learning. Therefore, by using gamified instructional media with Genially to teach science, it is expected to improve students' critical and creative thinking skills.

In accordance with the background above, this study aimed to used Genially to create interactive science learning media in elementary schools. The researchers make an educational video game that helps students in learning science and in improving their critical thinking. The research question underpinning this study is: Can gamified instructional using Genially improv students' critical and creative thinking skills?

2. Methodology

This study investigated the effectiveness of gamified instructional media with Genially to improve fifth-grade primary school students' critical and creative thinking skills in science class. The researchers applied a quasi-experimental nonequivalent control group research design to conduct this study. The researchers taught the experimental class by using gamified instructional media with Genially while the control class had conventional learning. The sample comprised 40 students of Public Primary School in Pekanbaru, 20 students in the experimental class and 20 students in the control class. This study was conducted from March to April 2021. To collect the data, the researchers used an instrument consisted of 7 short-answered questions to assess the students' critical thinking and 4 short-answered questions to assess their creative thinking skills.

The researchers conducted a pretest and posttest before and after the treatment by using the instrument. The validity and reliability of the instrument had been tested before it was used. After assessing the students' critical thinking and creative thinking skills, the data collected was analyzed. The link to the gamified instructional media is; https://view.genially/605214085ec620fd0b41406/interactive-content-hasil-final.

The data obtained were analyzed by using SPSS, a computer program used for statistical analysis, to know the effectiveness of gamified instructional media with Genially to improve critical and creative thinking skills of the students in fifth grade of primary school in science class on the topic of heat. This research was conducted from March to April 2021 in the fifth grade of Public Primary School in Pekanbaru. Among these populations, samples were selected based on those who have obtained legality based on the certificate of doing research no: 422/SDN192PKU/2021/277.

3. Results and Discussion

The results and discussion sections are divided into several sections, analyze of critical thinking skills, analyze of creative thinking skills and the correlation between creative thinking skill and critical thinking skill to get more in-depth data. In the end the study results are discussed to analyze the effects of Gamified Instructional Media more deeply.

3.1. Analysis of Critical Thinking Skills

Pretest

After the researchers had conducted normality and homogeneity test, the researchers conducted an independent sample t-test. Table 1 below presents the results of the independent sample t-test.

Table 1: The output table of the independent sample t-test

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	Df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Critical Thinking Skills	Equal variances assumed	8.758	.005	-.871	38	.389	-.300	.345	-.997	.397
	Equal variances not assumed			-.871	27.428	.391	-.300	.345	-1.006	.406

Table 2: The output table of the independent sample t-test

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	Df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Critical Thinking Skills	Equal variances assumed	1.055	.311	4.887	38	.000	3.450	.706	2.021	4.879
	Equal variances not assumed			4.887	36.929	.000	3.450	.706	2.020	4.880

Levene's Test for Impartiality of Discrepancies had an implication rate of 0.005 0.05. Because the value indicated that the two variances were not the same, the interpretation of the above output table was based on the values in the equal variances not assumed column. In the equal variances assumed column, the Sig. (2-tailed) value was 0.391> 0.05. There was no substantial difference in critical thinking skills among students in the experimental and control groups, according to the results (pretest). In other words, before treatment, the critical thinking skills of students in the experimental and control classes were the same.

Posttest

After the researchers had conducted normality and homogeneity test, the researchers conducted an independent sample t-test. Table 2 below offers the outcomes of the independent sample t-test.

The consequence rate of Levene's Test for Equivalence of Discrepancies was 1.055 > 0.05. The value indicated that the two variances were identical, so the interpretation of the output table above is based on the values in the identical variances assumed column. The Sig. (2-tailed) value in the equal variances assumed column was 0.000 < 0.05. It indicated a significant difference in critical thinking skills between students in experimental and control classes (posttest). In other words, the critical thinking skills between students in experimental and control class before treatment was dissimilar.

This study proved that the gamified instructional media with Genially media can progress students' critical thinking skills. This benefit arises from agreeing on the separate mastery of procedural skills in the secluded and stress-free situation provided by the operational stage and admission to social properties in the classroom situation. Instruction and education consuming online platform have carried confident influences specifically in emerging the 21st-century assistances. Critical thinking skills are one of the 21st-century skills that must be included in the world of education. However, the main challenge is how to teach thinking or critical thinking and how to stimulate students to reflect on their own thinking ways. [46].

3.2. Analysis of Creative Thinking Skills

Posttest

After the researchers had conducted normality and homogeneity test, the researchers conducted independent sample t-test. Table 3 below presents the results of the independent sample t-test.

The significance value of Levene's Test for Equality of Variances was 0.005 0.05. The value indicated that the two variances were not the same, so the interpretation of the output table above was based on the values in the equal variances not assumed column. The Sig. (2-tailed) value in the equal variances assumed column was 0.391> 0.05. It revealed that there was no significant difference in critical thinking skills between students in the experimental and control groups (pretest). In other words, before treatment, the critical thinking abilities of students in the experimental and control classes were the same.

According to [47], games can be used as a supplement to traditional teaching methods to improve learners' learning experiences while also teaching other skills such as following rules, adaptation, problem solving, interaction, critical thinking skills, creativity, teamwork, and good sportsmanship. Furthermore in [48], contend that digital games can assist students in developing higher-order thinking skills and 21st-century skills, as well as making learning more enjoyable and engaging.

3.3. Analysis of the correlation between students' critical and creative thinking skills Experiment class

To know if there was a correlation between students' critical and creative thinking skills, the researchers conducted a (Pearson) bivariate correlation. Table 4 below is the output table of the (Pearson) bivariate correlation.

Table 3: The output table of the independent sample t-test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	Df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Creative Thinking Skills	Equal variances assumed	4.672	.037	2.148	38	.038	1.550	.722	.089	3.011
	Equal variances not assumed			2.148	27.782	.041	1.550	.722	.071	3.029

Table 4: the output table of the (Pearson) bivariate correlation

		Critical thinking skills	Creative thinking skills
Critical thinking skills	Pearson Correlation	1	.588**
	Sig. (2-tailed)		.006
	N	20	20
Creative thinking skills	Pearson Correlation	.588**	1
	Sig. (2-tailed)	.006	
	N	20	20

**. Correlation is significant at the 0.05 level (2-tailed).

Table 5: The output table of the (Pearson) bivariate correlation

Correlations			
		Critical thinking skills	Creative thinking skills
Critical thinking skills	Pearson Correlation	1	.425
	Sig. (2-tailed)		.062
	N	20	20
Creative thinking skills	Pearson Correlation	.425	1
	Sig. (2-tailed)	.062	
	N	20	20

**. Correlation is significant at the 0.05 level (2-tailed).

Based on the output table, the correlation coefficient was 0.588, and the Sig. (2-tailed) value was 0.006 < 0.05. Therefore, there was a positive correlation between students' critical thinking skills and creative thinking skills, and the correlation was at a moderate level.

Control class

To know if there was a correlation between students' critical and creative thinking skills, the researchers conducted a (Pearson) bivariate correlation. Table 5 below is the output table of the (Pearson) bivariate correlation.

The correlation coefficient was 0.425, and the Sig. (2-tailed) value was 0.062 > 0.05, according to the output table. As a result, there was no positive correlation between students' critical and creative thinking skills, and the correlation was moderate.

This study demonstrated that students in the experimental class demonstrated a positive correlation between critical thinking skills and creative thinking skills, whereas students in the control group demonstrated no positive correlation between critical thinking skills and creative thinking skills.

This finding is supported by [41] claim that there is a significant correlation between critical thinking skills and creative thinking skills on cognitive learning outcomes. The restorative properties of nature are most visible for creativity when viewing stimuli indoors; however, being outdoors in general may be enough to stimulate creativity, regardless of whether it is surrounded by nature or a busy urban environment [42, 49]. The application of Digital Escape Room with Science Teaching in Primary School has Problem Solving Ability. Especially in science subjects [50].

Figure 1: Interesting animation

There is a significant effect because gamification of instructional media makes the learning fun and interesting (see figure 1) and helps students develop higher-order thinking skills [48]. Moreover, [44] and [51] state that presenting instructional materials by using interactive instructional media provides convenience to students. In addition, the animations displayed and interactive simulations that must be done by students through discussion sheets can train students' logical thinking in physics problems solving related to the concepts of temperature and heat. That evidence proves that there is a significant role of using gamified instructional media with Genially to teach science on the topic of heat transfer.

Based on the test results above, the results of this study have the same results with the study proposed by [44] showing on the topic of temperature and heat, interactive instructional media affects the students' conceptual understanding and critical thinking skills. Gamified instructional media can help students improve their computational thinking competence and motivate them to develop their learning capacities on their own by assisting them in expanding their insight, processing information they have, communication, and community awareness skills [33].

Figure 2: interesting games in the learning media

Learning media on heat transfer material with genius has an interesting game. In addition, another previous research project also claimed that games can also be implemented teach other skills including critical thinking, problem solving, sportsmanship, interaction and peers-collaboration [47]. Moreover, great potential is provided by games for training because they affect the learning process of users significantly [52]. Hikmah and Ngazizah (2020) state that creative thinking skills are part of higher order thinking

skills. Therefore, gamified instructional media with Genially can improve students' creative thinking skills which belongs to higher order thinking skills.

The existence of technology and various innovations helps overcome learning problems, especially at the primary school level [40]. That indicates that gamified instructional media helps students to identify and solve problems. Instructional media is one of the most important things that support learning. Gamified instructional media is physical means in delivering instructional materials from teachers to students in a more sophisticated and efficient way [53], [54]. Therefore, gamified instructional media is necessary to be considered carefully as it has long-term effects on education [55].

In conclusion, gamified instructional media using Genially in science class, particularly on heat transfer topic improve students' critical and creative thinking skills in terms of the ability to generate ideas, the ability to propose various solutions to problems, and the ability to create different ideas.

4. Conclusions and Suggestions

This study discusses in depth the effects of Gamified Instructional Media on critical thinking and creative thinking skills, then analyse of correlation between critical thinking and creative thinking skills. The results of the study prove that Gamified Instructional Media has an influence on the critical thinking and elementary school students' creative thinking skills. While the education is focusing on improving high order thinking skills, this study adds new literature on Gamified Instructional Media at the elementary school level which can convince teachers that using Gamified Instructional Media is to increase HOTS at the elementary school level effectively.

Based on the results, the researchers suggest that teachers have to improve their technology literacy so that they can involve technology in their teaching. With technology, teachers can develop many forms of instructional media. The researchers also suggest facilitating teachers to develop instructional media with technology to others stakeholders in education.

Acknowledgements

The authors express gratitude to the Ministry of Research and Technology/National Research and Innovation Agency Indonesia for supporting this study under the Grant of DRTPM 2022

References

[1] García-Peñalvo, F.J, A.J. Mendes, "Exploring the computational thinking effects in pre-university education," Computers in Human Behavior Journal, 2018.

[2] C. Giovannella, "Effect Induced by the Covid-19 Pandemic on Students' Perception About Technologies and Distance Learning," In Ludic, Co-Design and Tools Supporting Smart Learning Ecosystems and Smart Education, 105–116, 2021, doi:10.1007/978-981-15-7383-5.

[3] L. Warren, D. Reilly, A. Herdan, Y. Lin, "Self-efficacy, performance and the role of blended learning," Journal of Applied Research in Higher Education, 2019, doi:10.1108/JARHE-08-2019-0210.

[4] R. Vebrianto, R.U. Rery, K. Osman, "BIOMIND Portal for Developing 21st Century Skills and Overcoming Students' Misconception in Biology Subject," International Journal of Distance Education Technologies, **14**(4), 55–67, 2016, doi:10.4018/IJDET.2016100105.

[5] T.T. Wijaya, W. Hidayat, Y. Zhou, "Development of Interactive Learning Video on Linear Program," Universal Journal of Educational Research, **8**(12A), 7530–7538, 2020, doi:10.13189/ujer.2020.082537.

[6] E. Glaser, "An experiment in the development of critical thinking," Teachers College Record, **43**(5), 409–410, 1942.

[7] T.T. Wijaya, J. Tang, L. Li, A. Purnama, "Implementing Dynamic Mathematics Software in Calculus II for Engineering Students: Quadratic Surfaces," Software Engineering and Algorithms, **230**, 480–491, 2021, doi:10.1007/978-3-030-77442-4_41.

[8] R.W. Paul, A.J.A. Binker, "Critical thinking: What every person needs to survive in a rapidly changing world," Center for Critical Thinking and Moral Critique, Sonoma State University, Rohnert Park, CA 94928., 1990.

[9] P.A. Facione, "Critical Thinking: What It Is and Why It Counts," Insight Assessment, **2007**(1), 1–23, 2011.

[10] M. Hutajulu, T.T. Wijaya, W. Hidayat, "the Effect of Mathematical Disposition and Learning Motivation on Problem Solving: an Analysis," Infinity Journal, **8**(2), 229, 2019, doi:10.22460/infinity.v8i2.p229-238.

[11] R. Ennis, "Critical thinking: Reflection and perspective Part II," Inquiry: Critical Thinking across the Disciplines, **26**(2), 5–19, 2011.

[12] R. Anjani, "Analisis Kemampuan Berpikir Kritis Siswa Gaya Belajar Accomodator dalam Menyelesaikan Soal Pemecahan Masalah Matematika di Kelas VIII Smp Negeri 6 Muaro Jambi," Jurnal Pendidikan Matematika, **3 Oktober**, 2017.

[13] R. Nuraini, Suparman, "Deskripsi Kemampuan Berpikir Kritis dan Kreatif Siswa Melalui Penerapan Pendekatan Saintifik," in Prosiding Seminar Nasional Etnomatnesia ISBN: 978-602-6258-07-6, 2019.

[14] N. Hermita, A. Suhandi, E. Syaodih, A. Samsudin, Isjoni, H. Johan, F. Rosa, R. Setyaningsih, Sapriadil, D. Safitri, "Constructing and Implementing a Four Tier Test about Static Electricity to Diagnose Pre-service Elementary School Teacher' Misconceptions," Journal of Physics: Conference Series, **895**(1), 2017, doi:10.1088/1742-6596/895/1/012167.

[15] J. Pereira, T. Jianlan, T.T. Wijaya, A. Purnama, Neni, Hermita, M. Tamur, "Using Hawgent Mathematics Software to Help Primary School Students to Read Clocks," Journal of Physics: Conference Series, **2049**(1), 2021, doi:10.1088/1742-6596/2049/1/012049.

[16] Yustina, W. Syafii, R. Vebrianto, "The Effects of Blended Learning and Project-Based Learning on Pre-Service Biology Teachers' Creative Thinking Skills through Online Learning in the Covid-19 Pandemic," Jurnal Pendidikan IPA Indonesia, **9**(3), 2020, doi:10.15294/jpii.v9i3.24706.

[17] N. Azriani, N. Islami, N. Hermita, M. Nor, E. Syaodih, H. Handayani, Z. Zulirfan, A. Suhandi, A. Malik, K. Mahbubah, A. Samsudin, "Implementing inquiry learning model to improve primary school students' critical thinking on earth and universe concept," Journal of Physics: Conference Series, **1227**(1), 2019, doi:10.1088/1742-6596/1227/1/012033.

[18] A.R. Ningsih, A. Suhandi, E. Syaodih, B. Maftuh, N. Hermita, A. Samsudin, "Fourth-grade elementary students critical thinking skills: A preliminary study on magnetic force," in Journal of Physics: Conference Series, 2019, doi:10.1088/1742-6596/1157/3/032045.

[19] E. Syaodih, L. Kurniawati, H. Handayani, D. Setiawan, I. Suhendra, N. Hermita, "Critical Thinking Skills of Fifth Grade Elementary School Students in Bandung City on the Topic of Water Cycle in Natural Science Subjects," in Journal of Physics: Conference Series, 2019, doi:10.1088/1742-6596/1351/1/012073.

[20] D.A.I. Wijayanti, K. Pudjawan, I.G. Margunayasa, "Analisis Kemampuan Berpikir Kritis Siswa Kelas V dalam Pembelajaran IPA di 3 SD Gugus X Kecamatan Buleleng," E-Journal PGSD Universitas Pendidikan Ganesha, **3**(1), 2015.

[21] F.Z. Firdaus, Suryanti, U. Azizah, "Pengembangan Multimedia Interaktif Berbasis Pendekatan SETS untuk Meningkatkan Kemampuan Berpikir Kritis Siswa Sekolah Dasar," Jurnal Basicedu, **4**(3), 681–689, 2020, doi:10.31004/basicedu.v4i3.417.

[22] M. Kuo, T. Chen, T.-Y. Chuang, B.-Y. Cheng, Y.-N. Su, "What Are the Better Gamification Tools for Elementary School Teachers?," International Congress on Advanced Applied Informatics (IIAI-AAI), (1), 346–349, 2018, doi:10.1109/IIAI-AAI.2018.00074.

[23] Q. Zhang, Z. Yu, "A literature review on the influence of Kahoot! On learning outcomes, interaction, and collaboration," Education and Information Technologies, 2021.

[24] E. Borsos, "The gamification of elementary school biology : a case study on increasing understanding of plants increasing understanding of plants," Journal of Biological Education, 1–14, 2018, doi:10.1080/00219266.2018.1501407.

[25] M.A. Çiftçi, M. Aykaç, "The effect of creative drama activities in early childhood on the executive functions of children," Education 3-13, **0**(0), 1–14, 2020, doi:10.1080/03004279.2020.1849343.

[26] S.B. Kert, M.K. Büyükimdat, A. Uzun, "Comparing active game-playing scores and academic performances of elementary school students," **4279**(March), 2016, doi:10.1080/03004279.2016.1140800.

[27] K. Jayantilal, N.O. Leary, "The factors influencing two primary teachers ' interpretation of games," Education 3-13, **0**(0), 1–17, 2020,

doi:10.1080/03004279.2020.1810094.
[28] H.E. Vidergor, "Effects of digital escape room on gameful experience, collaboration, and motivation of elementary school students," Computers & Education, **166**(February), 104156, 2021, doi:10.1016/j.compedu.2021.104156.
[29] M. Khoiron, Harmanto, A. Kasdi, A.R. Wardani, "Development of digital social studies teaching materials in the era of pandemic emergency learning," The Indonesia Journal of Social Studies, **4**(1), 36–44, 2021.
[30] Â. Musskopf, D.N.F.D.N.F. Barbosa, P.B.S.P.B.S. Bassani, A. Jefferies, "Using Digital Resources to Boost English Writing Development," Communications in Computer and Information Science, **1011**, 337–348, 2019, doi:10.1007/978-3-030-20798-4_29.
[31] P.M. Manuel, A.M. Pilar, R.M. María Dolores, D. MP, P. Sara, M.J. M. Pilar, R.M. Dolores, D. MP, P. Sara, M.-J.M. Pilar, "Characterization of biodiesel using virtual laboratories integrating social networks and web app following a ubiquitous- and blended-learning," Journal of Cleaner Production, **215**, 399–409, 2019, doi:10.1016/j.jclepro.2019.01.098.
[32] G. Cheung, K. Wan, K. Chan, "Efficient use of clickers: A mixed-method inquiry with university teachers," Education Sciences, **8**(1), 1–15, 2018, doi:10.3390/educsci8010031.
[33] E. Choi, Y. Jung, N. Park, "Strategies to Teach Elementary School Students the Principles of Blockchain Technology by Implementing Gamification," Ilkogretim Online, **20**(3), 1205–1211, 2021, doi:10.17051/ilkonline.2021.03.134.
[34] M.T. Alshammari, "Evaluation of Gamification in E-Learning Systems for Elementary School Students," TEM Journal, **9**(2), 806–813, 2020, doi:10.18421/TEM92.
[35] P. Buckley, E. Doyle, "Individualising gamification: An investigation of the impact of learning styles and personality traits on the efficacy of gamification using a prediction market," Computers & Education, **106**, 43–55, 2017, doi:10.1016/j.compedu.2016.11.009.
[36] B.N. Nirbita, S. Joyoatmojo, S. Sudiyanto, "ICT Media Assisted Problem Based Learning for Critical Thinking Ability," International Journal of Multicultural and Multireligious Understanding, **5**(4), 341–348, 2018.
[37] W. Rajagukguk, E. Simanjuntak, "Problem-Based Mathematics Teaching Kits Integrated With ICT to Improve Student's Critical Thinking Ability In Junior High Schools In Medan," Cakrawala Pendidikan, (3), 2015.
[38] Ü. Cakiroğlu, B. Başibüyük, M. Güler, M. Atabay, B.Y. Memiş, "Gamifying an ICT Course: Influences on Engagement and Academic Performance," Computers in Human Behavior, 2017, doi:10.1016/j.chb.2016.12.018.
[39] B. Potts, "Strategies for Teaching Critical Thinking," Practical Assessment, Research, and Evaluation, **4**(4), 1995.
[40] S. Huang, Y. Kuo, H. Chen, "Applying digital escape rooms infused with science teaching in elementary school : Learning performance , learning motivation , and problem-solving ability," Thinking Skills and Creativity, **37**(129), 100681, 2020, doi:10.1016/j.tsc.2020.100681.
[41] J. Siburian, A.D. Corebima, Ibrohim, M. Saptasari, "The Correlation Between Critical and Creative Thinking Skills on Cognitive Learning Results," Eurasian Journal of Educational Research, **81**, 99–114, 2019, doi:10.14689/ejer.2019.81.6.
[42] A. Palanica, A. Lyons, M. Cooper, A. Lee, Y. Fossat, "A comparison of nature and urban environments on creative thinking across different levels of reality," Journal of Environmental Psychology, **63**, 44–51, 2019.
[43] N. Hendracipta, "Menumbuhkan Sikap Ilmiah Siswa Sekolah Dasar Melalui Pembelajaran IPA Berbasis Inkuiri," JPSD, **2**(1), 109–116, 2016, doi:10.31226/osf.io/etg5n.
[44] S. Husein, L. Herayanti, Gunawan, "Pengaruh Penggunaan Multimedia Interaktif terhadap Penguasaan Konsep dan Keterampilan Berpikir Kritis Siswa pada Materi Suhu dan Kalor," Jurnal Pendidikan Fisika Dan Teknologi, **I**(3), 2015.
[45] A. Zohar, Y. Weinberger, P. Tamir, "The Effect of the Biology Critical Thinking Project on the Development of Critical Thinking," Journal Of Research in Science Teaching, **31**(2), 183–196, 1994.
[46] E. AlJaafil, M. Sahin, "Critical Thinking Skills for Primary Education: The Case in Lebanon," Online Submission, **1**(1), 1–7, 2019.
[47] V. Zirawaga, A. Olusanya, T. Maduki, "Gaming in education: Using games a support tool to teach History," Journal of Education and Practice, **8**(15), 55–64, 2017.
[48] Y. An, L. Cao, "The Effects of Game Design Experience on Teachers' Attitudes and Perceptions regarding the Use of Digital Games in the Classroom," TechTrends, 2016, doi:10.1007/s11528-016-0122-8.
[49] H. Huang, G.-J. Hwang, "Facilitating inpatients' family members to learn: A learning engagement-promoting model to develop interactive e-book systems for patient education," Educational Technology and Society, **22**(3), 74–87, 2019.
[50] M. Kalogiannakis, S.J. Papadakis, "Evaluating pre-service kindergarten teachers ' intention to adopt and use tablets into teaching practice for natural sciences Evaluating pre-service kindergarten teachers ' intention to adopt and use tablets into teaching practice for natural sciences Mic," International Journal of Mobile Learning and Prganisation, **13**(1), 24–28, 2019, doi:10.1504/IJMLO.2019.096479.
[51] A. Syawaludin, G. Gunarhadi, P. Rintayati, "Development of Augmented Reality-Based Interactive Multimedia to Improve Critical Thinking Skills in Science Learning," International Journal of Instruction, **12**(4), 2019, doi:10.29333/iji.2019.12421a.
[52] P.-M. Noemí, S.H. Máximo, "Educational games for learning," Universal Journal of Educational Research, **2**(3), 230–238, 2014, doi:10.13189/ujer.2014.020305.
[53] S. Hikmah, N. Ngazizah, "Profil Kemampuan Higher Order Thinking Skills dan Karakter Siswa pada Materi Panas dan Perpindahannya pada Kelas 5 Sekolah Dasar," in Seminar Nasional Pendidikan Dasar, 2020.
[54] N. Hermita, H.S. Ningsih, J.A. Alim, M. Alpusari, Z.H. Putra, T.T. Wijaya, "Developing Science Comics for Elementary School Students on Animal Diversity. Solid State Technology," **63**(1s), 2020.
[55] M. Kalogiannakis, S. Papadakis, A.-I. Zourmpakis, "Gamification in Science Education. A Systematic Review of the Literature," Education Sciences, **11**(22), 2021.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 1, 60-78 (2022)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Thermoelectric Generators (TEGs) and Thermoelectric Coolers (TECs) Modeling and Optimal Operation Points Investigation

Nganyang Paul Bayendang*, Mohamed Tariq Khan, Vipin Balyan

Department of Electrical, Electronic and Computer Engineering (DEECE), Cape Peninsula University of Technology (CPUT), Bellville Campus, Cape Town, Western Cape, 7535, South Africa

A R T I C L E I N F O

Article history:
Received: 11 September, 2021
Accepted: 05 December, 2021
Online: 10 February, 2022

Keywords:
Alternative Energy
Energy Efficiency
Energy Harvesting
Thermoelectric Coolers (TECs)
Thermoelectric Generators (TEGs)
TEGs and TECs Modeling
TEGs/TECs Optimal Operation
Thermoelectricity

A B S T R A C T

Sustainable energy is gradually becoming the norm today due to greenhouse warming effects; as a result, the quests for different renewable energy sources such as photovoltaic cells as well as energy efficient electrical appliances are becoming popular. Therefore, this article explores the alternative energy case for thermoelectricity with focus on the steady-state mathematics, mixed modelings and simulations of multiple TEGs and TECs modules to study their performance dynamics and to establish their optimal operation points using Matlab and Simulink. The research substantiates that the output current from TEGs or input current to TECs, initially respectively increases the output power of TEGs and the cooling power of TECs, until the current reaches a certain maximum optimal point, after which any further increase in the current, decreases the TEGs' and or TECs' respective output and cooling powers as well as efficiencies, due to Ohmic heating and or entropy change caused by the increasing current. The research main contributions are elaborate easy to understand TEGs/TECs theoretical formulations as well as static and dynamic simulated models in Matlab/Simulink, that can be used initially to dynamically investigate an infinite quantity of TEG and TEC modules connections, be it in series and or in parallel. This is to assist system designers grasp TEGs and TECs theoretical operations better and their limits, when designing energy efficient waste heat recovery (using TEGs)/cooling (using TECs) systems for industrial, residential, commercial and vehicular applications.

1. Introduction

According to [1], energy security and green economy are becoming paramount today; as a result, the demands for renewable and alternative energy sources such as solar, wind, hydro energy, bio-fuels and fuel cells, as well as energy efficient loads, are on the rise in an effort to ensure energy sustainability and carbon free environment. In this regards, we investigated thermoelectricity as a potential alternative energy for sustainable energy source and loads − that is, as a clean DC power source for low energy lighting/applications and as well to provide clean cooling/heating in various human habitats. Thermoelectricity as reviewed in [2], practically focuses on the Seebeck and Peltier effects. Seebeck effect is basically converting heat to DC electricity and the device that does this is a thermoelectric generator (TEG). The reversed phenomenon is a Peltier effect − which is basically the production of cold from DC electricity and if the direction of current flow changes (swap voltage polarity), heat is also produced and the device that does this is called a thermoelectric cooler (TEC). Therefore, by efficiently applying thermoelectricity prudently, a clean alternative energy source for DC low power applications using TEGs and or energy efficient loads in the forms of heat pumps, air conditioners, refrigerators etc using TECs; can be passably implemented to help sustain some human habitats basic energy consumption such as lighting, cooling and heating; as well as reduce environmental pollution.

As already examined in [2], thermoelectricity lends itself to various applications with focus on how TEGs and TECs can be used respectively as a power source and as a load. Furthermore, studied in [3], is a re-configurable TEG DC-DC converter for maximum TEG energy harvesting in a battery-powered wireless sensors network (WSN). Described in [4], is the analysis and design of a thermoelectric energy harvester (TEH) prototype for

*Corresponding Author: NP Bayendang, CPUT DEECE, +27765404896, bayendangn@cput.ac.za

powering up outdoor sensors and devices. Solar energy was harvested using different TEG arrays in [5] and a theoretical analysis of implementing a re-configurable TEG was researched in [6]. Electronic cooling was investigated in [7] and the findings revealed that the TEC cooling capacity could be increased by increasing its cold side junction temperature and decreasing its temperature difference. A multi-stage TEC module in cascade was examined in [8]; whereas in [9], an extensive mathematical analyses were articulated for TEG and TEC design and materials. A TEG model was developed in [10] for maximum power point tracking but lacks the detailed underlining maths and the parallel TEG combinations was limited to just 2. A comprehensive TEG and TEC models with the detailed maths supporting the TEG and TEC models, were presented respectively in [11] and [12]. In [13], a modeling of TEG using Modelica is asserted but deficient in the comprehensive maths, especially considering modeling infinite multiple TEGs and as well TECs modules —which were not articulated. A parametric ANSYS study of TEG and TEC was presented in [14]; however, the detailed maths and especially for the case for infinite TEG and TEC modules use/ connection, was inadequate. In addition, for large scale TEGs and TECs applications, the following studies were examined. In [15], 600 TEGs with a temperature difference of ~120 ℃, were applied to harvest and generate up to 1 kW of DC power from geothermal heat. It was further indicated a 2 kW power could be achieved with a higher temperature difference and also the TEG cost is much lower to generate equivalent amount of power than using photovoltaic. However, the study lacks the theoretical details to substantiate it. TEG harvesting of waste thermal energy from household heat sources such as a generator exhaust pipe and a kerosene stove, were performed in [16] and various parameters measurements were made but without detailing the maths to calculate these parameters. Light and heat from the Sun are the most common forms of energy abundant on Earth; as a result, [17] reviewed the possibility of integrating photovoltaic and TEG in a hybrid photovoltaic-TEG system and further examined the efficiency improvement. A 128 TEGs system was assembled in [18] to generate ~684 W of power from radiation heat transfer at a temperature difference of ~125 K and with a corresponding power density of 845 W/m^2. Their results further justified that with a greater practical temperature difference of 200 K, the respective generated power and power density of their TEGs system could attain 1.23 kW and 1.51 kW/m^2. Their TEGs system open circuit voltage, its output power, its power density and its conversion efficiency were investigated in details at different temperature differences; however, the underlining maths was not elaborated. A grid-tied 20 W TEG experimental model using 24 modules in series with the heat harvested from a waste incinerator, was experimented in a lab and the preliminary and analytical models of the electric output power as a function of specific temperatures, were investigated in [19]. A micro combined cold, heat and power system for a small household with a TEC as the cooler and achieving a cooling power of 26.8 W, was presented in [20]. In [21], a 3D printable TEG device architecture with a high thermocouple density of 190 per cm² by using a thin substrate as an electrical insulation between the thermoelectric elements, resulted in a high-power output of 47.8 μW/cm^2 from a 30 K temperature difference. A stove-powered TEG (SPTEG) was used in [22] to generate power from waste heat released during cooking. They researched series and parallel TEGs connections and the effect of pressure to address low power output due to irregular temperature. Finally, an experimental and a numerical investigations on TECs for comparing air-to-air and air-to-water refrigeration were investigated in [23], with the findings revealing that air-to-water achieves 30-50% efficiency, compared to air-to-air cooling.

These are just a few noted studies; however, lacking in the TEGs/TECs literature are comprehensive details on their maths, modeling and operations when connected in series and also in parallel to increase the output power (in the case of TEG) and the cooling power (in the case of TEC). This article therefore, expands on i) developing and expressing further the theoretical maths covering TEGs and TECs various parameters/modules with focus on the total internal resistance, ii) the modeling of multiple TEGs and TECs modules focusing on their electrical parameters and finally iii) their static and dynamic simulations with focus on the optimal operation points investigation as well as the interpretations thereof. The results are then validated with established published studies and concluding remarks are drawn.

2. TEGs and TECs Mathematical Analyses and Modeling

In [9], [11] and [12], the standard static mathematics defining various TEG and TEC parameters as well as their modeling are demonstrated. We developed further and present in the following sections: i) TEGs and TECs maths and ii) the implemented models (based on their maths) using Matlab/Simulink and the simulations of TEG/TEC modules, be it in series and or in parallel connections.

2.1. TEGs and TECs Steady-state Mathematical Analyses

The derivations thus far of the TEG and TEC parameters have been based-on the p-n junction thermoelement resistance at the thermocouple level and by extension at the module level as indicated in [9], [11] and [12]. However, in practice, more than one TEG and TEC modules will be needed for more power production and this will take the form of series and or parallel connections; as a result, the electrical resistance will often change. This section redefines the change in R to R_t and is articulated next.

(I) *TEGs Steady-state Mathematical Analysis*

The following TEG parameters mathematics are developed and presented step-wise for multiple TEGs case as follows:

- Thermoelectric (TE) device p-n junction thermocouple resistance (r)

The TE device p-n thermocouple resistance r in ohm is:

$$r = \rho L/A \quad (\Omega) \tag{1}$$

with ρ being the TEG/TEC electrical resistivity in Ω.m, L is the length in (m) of the TEG/TEC p-n thermocouple and the TEG/TEC p-n thermocouple area is A in metre squared (m^2).

- TE device (TEG and TEC) module resistance (R)

The resistance in (Ω) of a TEG/TEC module is computed as:

$$R = nr \quad (\Omega) \tag{2}$$

where n (which differs, could be 100, 127, 199, 255 etc) is a TEG/TEC manufacturer p-n thermocouples amount used in a TEG/TEC. The more the n, the more powerful is the TEG/TEC.

- TEG/TEC module(s) total resistance (R_t)

The total resistance R_t in (Ω) of a TEG/TEC module(s) is simply calculated as:

$$R_t = n\frac{T_s}{T_p}r = R\frac{T_s}{T_p} \quad (\Omega) \tag{3}$$

with T_p being the TEGs/TECs (TEG/TEC modules) amount connected in parallel and T_s the TEGs/TECs (TEG/TEC modules) amount connected in series. NB: all the TEGs/TECs used in (3), have to be identical model to make sure the R of each TEG/TEC is not vastly different; if not, (3) would be inaccurate.

- TEG(s) output voltage (V_o)

The TEG(s) voltage generated in volt, can be derived as:

$$V_o = nS\Delta T - IR_t \quad (V) \tag{4}$$

with S being the TE device Seebeck coefficient in V/K, $\Delta T = T_h - T_c$ the TEG(s) temperature difference in kelvin or °C and the output current of the TEG(s) is I in ampere.

- TEG(s) output current (I)

The TEG(s) generated current I in ampere is deduced as:

$$I = \frac{nS\Delta T}{R_L + R_t} \quad (A) \tag{5}$$

with R_L being the resistance of the electrical load connected to the TEG(s) output. NB: more I causes the TEG(s) more Joule heating, which negatively affects the TEGs efficiency.

- TEG(s) hot-side heat absorbed (Q_h)

TEG(s) produce DC power when their hot-side is at a high temperature T_h, during which the TEG(s) becomes hotter and the absorbed heat in watt is Q_h, given as:

$$Q_h = n[(SIT_h) + (K\Delta T)] - 0.5I^2R_t \quad (W) \tag{6}$$

with K being the TEG(s) thermal conductance in W/K.

- TEG(s) cold-side heat emitted (Q_c)

TEG(s) produce DC power when the cold-side of the TEG(s) is at a low temperature T_c releasing the heat Q_c in watt.

$$Q_c = n[(SIT_c) + (K\Delta T)] + 0.5I^2R_t \quad (W) \tag{7}$$

- TEG(s) output power (P_o)

The TEG(s) modules generated power P_o in watt, is found variously as follows:

$$P_o = Q_h - Q_c \quad (W) \tag{8}$$
$$P_o = IV_o = n\,[(SI\Delta T)] - I^2R_t \quad (W) \tag{9}$$

- TEG(s) electrical/conversion/thermal efficiency (η)

η is the TEG(s) power output P_o divided by the TEG(s) hot-side heat absorbed Q_h. η being a performance parameter is:

$$\eta = P_o/Q_h \tag{10}$$

The conversion efficiency details is presented later.

- TEG/TEC Carnot's efficiency (η_c)

Carnot efficiency is the efficiency determined based-on the temperatures T_h and T_c.

$$\eta_c = \frac{\Delta T}{T_h} = \frac{T_h - T_c}{T_h} = 1 - \frac{T_c}{T_h} \tag{11}$$

- TEG(s) conversion efficiency expression (η_e)

Simply, η_e is the raw expression of η. That is, when equations of Q_h and P_o (respectively (6) and (8) or (9)) are both substituted in (10).

$$\eta_e = \eta_c \frac{(nR_L/R_t)}{[(1+nR_L/R_t) - 0.5\eta_c + ((1/(2Z\bar{T}))(1+nR_L/R_t)^2(1+T_c/T_h))]} \tag{12}$$

with $Z\bar{T}$ being the TE device average dimensionless merit figure. NB: Z is the TE device merit figure in per K (K^{-1}) and $\bar{T} = (T_h + T_c) / 2$, is the TE device average temperature in K.

- TEG(s) maximum conversion efficiency (η_m)

η_m is the efficiency of the TEG(s) at $R_t / R_L = \sqrt{1 + Z\bar{T}}$. The η_m expression as a function of TEG temperatures and Z is:

$$\eta_m = \eta_c\left(\frac{(\sqrt{1+Z\bar{T}}) - 1)}{(\sqrt{1+Z\bar{T}} + (T_c/T_h))}\right) \tag{13}$$

- TEGs maximum power conversion efficiency (η_{mp})

As a function of temperatures and Z, η_{mp} is the efficiency of the TEG at its maximum output power P_o − that is, at $R_t = R_L$.

$$\eta_{mp} = \eta_c/[2 - 0.5\eta_c + (2/Z\bar{T})(1+T_c/T_h)] \tag{14}$$

- TEG(s) maximum power output (Po_{max})

The TEG(s) maximum transfer of power theoretically happens at $R_t = R_L$. NB: in practice, $R_t = R_L$ is hardly ever the case.

$$Po_{max} = (nS\Delta T)^2(R_L/R_t)/R(1+(R_L/R_t))^2 \quad (W) \tag{15}$$

- TEG(s) maximum voltage output (Vo_{max})

TEG(s) Vo_{max} happens at open circuit, that is when R_L is not connected or R_L is infinity (extremely large), $I = 0$A.

$$Vo_{max} = nS(T_h - T_c) = nS\Delta T \quad (V) \tag{16}$$

- TEG(s) maximum current output (I_{max})

TEG(s) I_{Max} happens at short circuit – meaning, when the load R_L is 0Ω. NB: R_t will therefore ideally be the sole resistance.

$$I_{Max} = nS\Delta T/R_t = nS(T_h - T_c)/R_t \quad (A) \tag{17}$$

- TEG(s) generated current normalized (I_n)

I_n is the normalized current of the TEG(s) in the range $0 \leq I_n \leq 1$. At the TEG(s) maximum transfer of power ($R_t = R_L$), I_n =

0.5. Simply, I_n is the TEG(s) generated current divided by the TEG(s) maximum current output. It is calculated as:

$$I_n = \frac{I}{I_{Max}} = \frac{R_t}{R_t + R_L} \quad (18)$$

- TEG(s) generated voltage normalized (V_n)

V_n is the normalized voltage of the TEG(s) ranging from 0 ≤ V_n ≤ 1. At the TEG(s) maximum transfer of power (i.e. R_L=R_t), V_n = 1/2. V_n is the TEG(s) voltage generated divided by the TEG(s) maximum (ideal) voltage generated. It is given as:

$$V_n = \frac{V_o}{Vo_{max}} = \frac{R_L}{R_L + R_t} \quad (19)$$

- TEG(s) output power normalized (P_n)

P_n is the normalised TEG(s) power bounded between 0 ≤ P_n ≤ 1. P_n = 1 at the TEG(s) maximum transfer of power (R_L=R_t). P_n is the TEG(s) power generated divided by the TEG(s) maximum output power. It is expressed as:

$$P_n = \frac{P_o}{Po_{max}} = \frac{4(R_L/R_t)}{[(R_L/R_t)+1]^2} \quad (20)$$

- TEG(s) conversion efficiency normalized (η_n)

η_n is the conversion efficiency of the TEG(s) in the region 0≤ η_n ≤1. η_n depends on R_t/R_L, T_c/T_h and $\overline{ZT}$. η_n is the conversion efficiency of the TEG(s) divided by the maximum conversion efficiency of the TEG(s), deduced as:

$$\eta_n = \eta/\eta_m \quad (21)$$

- TEG(s) effective Seebeck coefficient (S_e)

S_e measured in volt/kelvin, is expressed as:

$$S_e = 4Po_{max}/(nI_{max}\Delta T) \quad \text{(V/K)} \quad (22)$$

- TEG(s) effective electrical resistivity (ρ_e)

ρ_e measured in ohm metre, is found using:

$$\rho_e = 4[(A/L)Po_{max}]/nI_{max}^2 \quad (\Omega.\text{m}) \quad (23)$$

- TEG(s) effective figure of merit (Z_e)

Z_e measured in per kelvin, is computed as:

$$Z_e = [(2/\overline{T})(1+(T_c/T_h))]/[\eta_c((1/\eta_{mp})+0.5)-2] \quad (K^{-1}) \quad (24)$$

- TEG(s)/TEC(s) effective thermal conductivity (k_e)

k_e measured in watt per metre kelvin, is expressed as:

$$k_e = S_e^2/(\rho_e Z_e) \quad \text{(W/mK)} \quad (25)$$

TEGs/TECs effective parameters enable researchers to factor in TEGs/TECs system losses using maximum parameters to bridge the theoretical and measured specifications differences [9].

- TEG(s) Heat Flux Density (HFD)

HFD is the amount of heat absorbed per TEGs hot-side surface area (*TEGsa*) in watt per centimetre square.

$$HFD = Q_h/TEG_{sa} \quad (\text{W/cm}^2) \quad (26)$$

This concludes the TEG(s) modules static mathematical analysis.

(II) *TECs Steady-State Mathematical Analysis*

The following TEC parameters mathematics are examined and developed step-wise for multiple TECs case as follows:

- TEC(s) voltage input (V_{in})

The TEC(s) applied voltage in volt, is expressed as:

$$V_{in} = n[S(T_h - T_c)] + I_{in}R_t \quad \text{(V)} \quad (27)$$

where I_{in} is the TECs input current from the power supply.

- TEC(s) input current (I_{in})

The TECs input current in ampere is derived as:

$$I_{in} = \frac{nS\Delta T}{R_s - R_t} \quad \text{(A)} \quad (28)$$

where R_s is the internal source electrical resistance of the power supply connected to the TECs.

- TEC(s) cold-side heat absorbed (Q_c)

TECs create cold when their cold-side is at a low temperature T_c to absorb heat and supply a steady cooling power Q_c in W.

$$Q_c = n[(SI_{in}T_c) - (K\Delta T)] - 0.5I_{in}^2R_t \quad \text{(W)} \quad (29)$$

- TEC(s) hot-side heat emitted (Q_h)

TECs produce cold when their hot-side is at a high temperature T_h emitting the heat Q_h in watt.

$$Q_h = n[(SI_{in}T_h) - (K\Delta T)] + 0.5I_{in}^2R_t \quad \text{(W)} \quad (30)$$

- TEC(s) power input (P_{in})

The applied power P_{in} in watt required to power the TECs, is calculated variously as follows:

$$P_{in} = Q_h - Q_c = n[(SI_{in}\Delta T)] + I_{in}^2R_t \quad \text{(W)} \quad (31)$$

$$P_{in} = I_{in}V_{in} \quad \text{(W)} \quad (32)$$

- TEC(s) coefficient of performance (*CoP*)

This is TECs cooling power Q_c divided by its input power P_{in}.

$$CoP = Q_c/P_{in} \quad (33)$$

- TEC(s) CoP Expression (CoP_e)

CoP_e is the raw expression of CoP when the equations of Q_c and P_{in} (respectively (29) and (31) or (32)) are put in (33).

$$CoP_e = \frac{[(SI_{in}T_c) - (K\Delta T) - (0.5I_{in}^2R_t/n)]}{[(SI_{in}\Delta T) + (I_{in}^2R_t/n)]} \quad (34)$$

- TEC(s) current to yield CoP (I_{cop})

I_{cop} is the TECs input current in (A) needed to attain *CoP*.

$$I_{cop} = \frac{nS\Delta T}{R_t[(\sqrt{1+Z\bar{T}})-1]} \quad (A) \qquad (35)$$

- TECs maximum CoP (CoP_{max})

CoP_{max} is the TECs maximum CoP that can be achieved.

$$CoP_{max} = \frac{[T_c/\Delta T]\left((\sqrt{1+Z\bar{T}}) - \frac{T_h}{T_c}\right)}{((\sqrt{1+Z\bar{T}})+1)} \qquad (36)$$

- TEC(s) maximum cooling power current (Icp_{max})

Icp_{max} is TECs current in ampere needed to realise max Q_c.

$$Icp_{max} = nST_c/R_t \quad (A) \qquad (37)$$

- TEC(s) I_{cop} maximum cooling power (Qcp_{max})

Qcp_{max} in (W), is TECs maximum Q_c attained based-on I_{cop}.

$$Qcp_{max} = n[(SI_{cop}T_c) - (K\Delta T)] - 0.5I_{cop}^2R_t \quad (W) \qquad (38)$$

- TEC(s) maximum temperature difference (ΔT_{max})

TEC(s) ΔT_{max} in (K), occurs at maximum I_{in} and at $Q_c = 0$W.

$$\Delta T_{max} = \left(T_h + \frac{1}{Z}\right) - \sqrt{(T_h + \frac{1}{Z})^2 - {T_h}^2} \quad (K) \qquad (39)$$

- TEC(s) maximum input current (I_{max})

I_{max} is TEC(s) maximum input current in (A) at $Q_c = 0$W.

$$I_{max} = nS(T_h - \Delta T_{max})/R_t \quad (A) \qquad (40)$$

- TEC(s) maximum input voltage (Vin_{max})

Vin_{max} is the maximum V_{in} in volt, that produces maximum ΔT_{max} when $I_{in} = I_{max}$, R_t=0, $T_c = 0$, $Q_c = 0$ and T_h is maximum.

$$Vin_{max} = nST_h \quad (V) \qquad (41)$$

- TEC(s) maximum cooling power (Qc_{max})

Qc_{max} is the maximum absorbable heat or cooling power in watt, at $I_{in} = I_{max}$ and $\Delta T = 0°C$.

$$Qc_{max} = (nS)^2(T_h^2 - \Delta T_{max}^2)/2R_t \quad (W) \qquad (42)$$

- TEC(s) input current normalized (Iin_n)

TEC(s) Iin_n is I_{cop} divided by I_{max}.

$$Iin_n = I_{cop}/I_{max} \qquad (43)$$

- TEC(s) input voltage normalized (Vin_n)

TEC(s) Vin_n is V_{in} divided by Vin_{max}.

$$Vin_n = V_{in}/Vin_{max} \qquad (44)$$

- TEC(s) cooling power normalized (Qc_n)

TEC(s) Qc_n is Q_c divided by Qc_{max}.

$$Qc_n = Q_c/Qc_{max} \qquad (45)$$

- TEC(s) CoP normalized (CoP_n)

TEC(s) CoP_n is CoP divided by CoP_{max}.

$$CoP_n = CoP/CoP_{max} \qquad (46)$$

- TEC(s) normalized temperature difference (ΔT_n)

TECs ΔT_n, is ΔT divided by ΔT_{max} and it is expressed as:

$$\Delta T_n = \Delta T/\Delta T_{max} \qquad (47)$$

Normalized parameters give dimensionless parameters.

- TEC(s) effective Seebeck coefficient (S_e)

TECs S_e measured in VK^{-1}, is defined as:

$$S_e = 2Qc_{max}/[nI_{max}(T_h + \Delta T_{max})] \quad (V/K) \qquad (48)$$

- TEC(s) effective electrical resistivity (ρ_e)

TECs ρ_e measured in ohm metre, is written as:

$$\rho_e = AS_e(T_h - \Delta T_{max})/LI_{max} \quad (\Omega.m) \qquad (49)$$

- TEC(s) effective figure of merit (Z_e)

TECs Z_e measured in per kelvin, is given as:

$$Z_e = 2\Delta T_{max}/(T_h - \Delta T_{max})^2 \quad (K^{-1}) \qquad (50)$$

- TEC(s) midpoint current (I_{mid})

I_{mid} measured in ampere, is the mean of Icp_{max} and I_{cop}.

$$I_{mid} = 0.5(Icp_{max} + I_{cop}) \quad (A) \qquad (51)$$

- TEC(s) midpoint cooling power (Qc_{mid})

Qc_{mid} measured in watt, is expressed as:

$$Qc_{mid} = n[(SI_{mid}T_c) - (K\Delta T)] - 0.5I_{mid}^2R_t \quad (W) \qquad (52)$$

- TEC(s) midpoint input power (Pin_{mid})

Pin_{mid} measured in watt, is deduced as:

$$Pin_{mid} = n[(SI_{mid}\Delta T)] + I_{mid}^2R_t \quad (W) \qquad (53)$$

- TEC(s) midpoint CoP (CoP_{mid})

CoP_{mid} is computed as:

$$CoP_{mid} = Qc_{mid}/Pin_{mid} \qquad (54)$$

Midpoint parameters ascertain safer optimal TECs design.

- TEC(s) cold flux density (CFD)

CFD is the cold amount produced (heat absorbed) per TECs cold-side surface area ($TECsa$) in W/cm^2. It is computed as:

$$CFD = Q_c/TECsa \quad (W/cm^2) \qquad (55)$$

2.2. TEGs and TECs Modelling and Simulations

Covered in Section 2.1., are the TEGs and TECs parameters of interests —which were extensively expressed mathematically with emphasis/basis on the total internal resistance R_t—which was derived and the regular TEG/TEC equations re-expressed based-on R_t to now cover TEG(s)/TEC(s). The above equations are herein further modeled in Matlab and Simulink, to institute the TEGs and TECs models that can now be utilized to simulate and investigate many connected TEGs and TECs optimal performance.

Exemplified in Figures 1a and 1b, are the TEGs static and dynamic simulated model GUIs, from which the TEGs parameters expressed in Section 2.1.I, can all be statically and dynamically configured for an infinite amount of TEGs connections and then simulated to obtain the TEG(s) optimum operation points. Figures 1c and 1d zoom-in on the TEGs internal modeling. Figure 1e expands on the TEGs R_t modeling − this must be matched to the load resistance R_L − which can be changed before or while the simulation is running to match the TEGs R_t for maximum power transfer simulation. Figure 2 exemplifies the TECs simulator user interface. Also, multiple TECs combinations in T_s and T_p and the various parameters presented in Section 2.1.II, can be optimally simulated. Likewise, maximum power will be transferred also from the DC power supply to the TECs by matching its R_t to R_s.

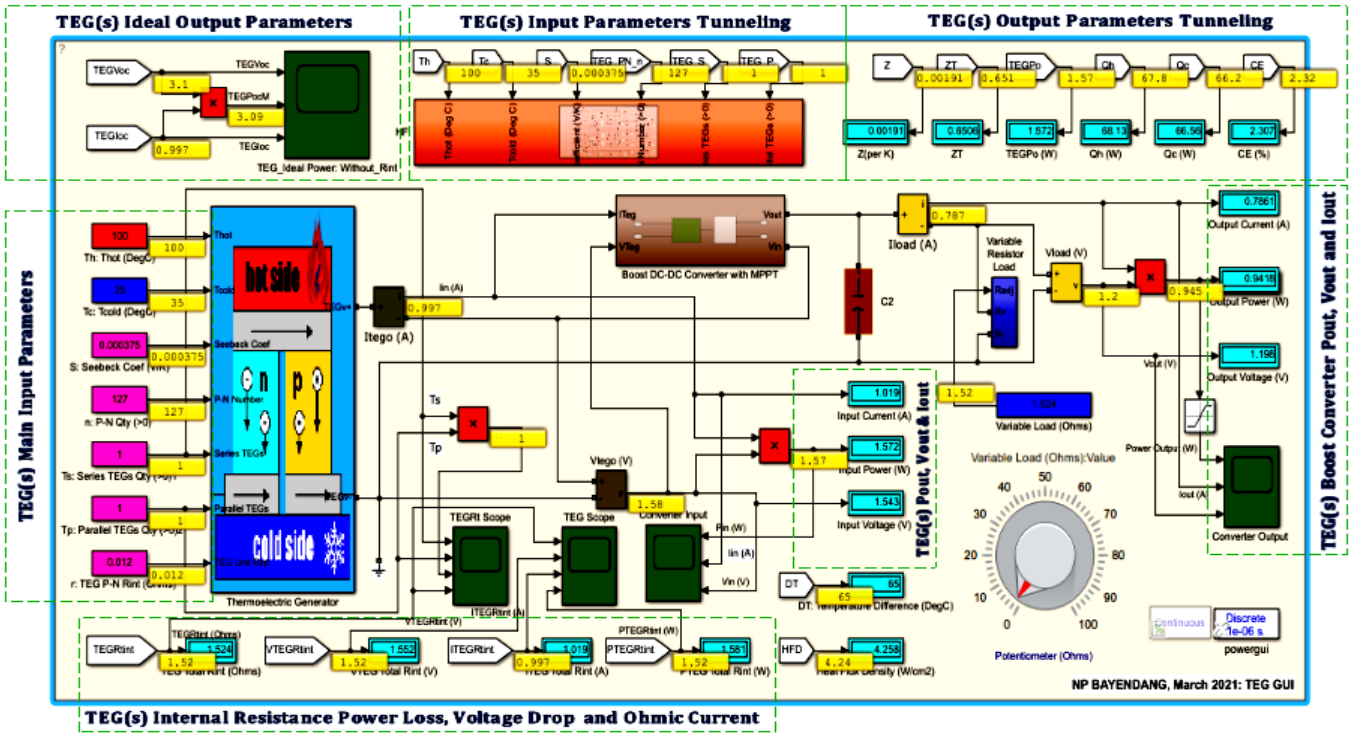


Figure 1a: TEG(s) static simulator user's interface − shows the steady-state simulation with all the input parameters fixed (though can change) over-time

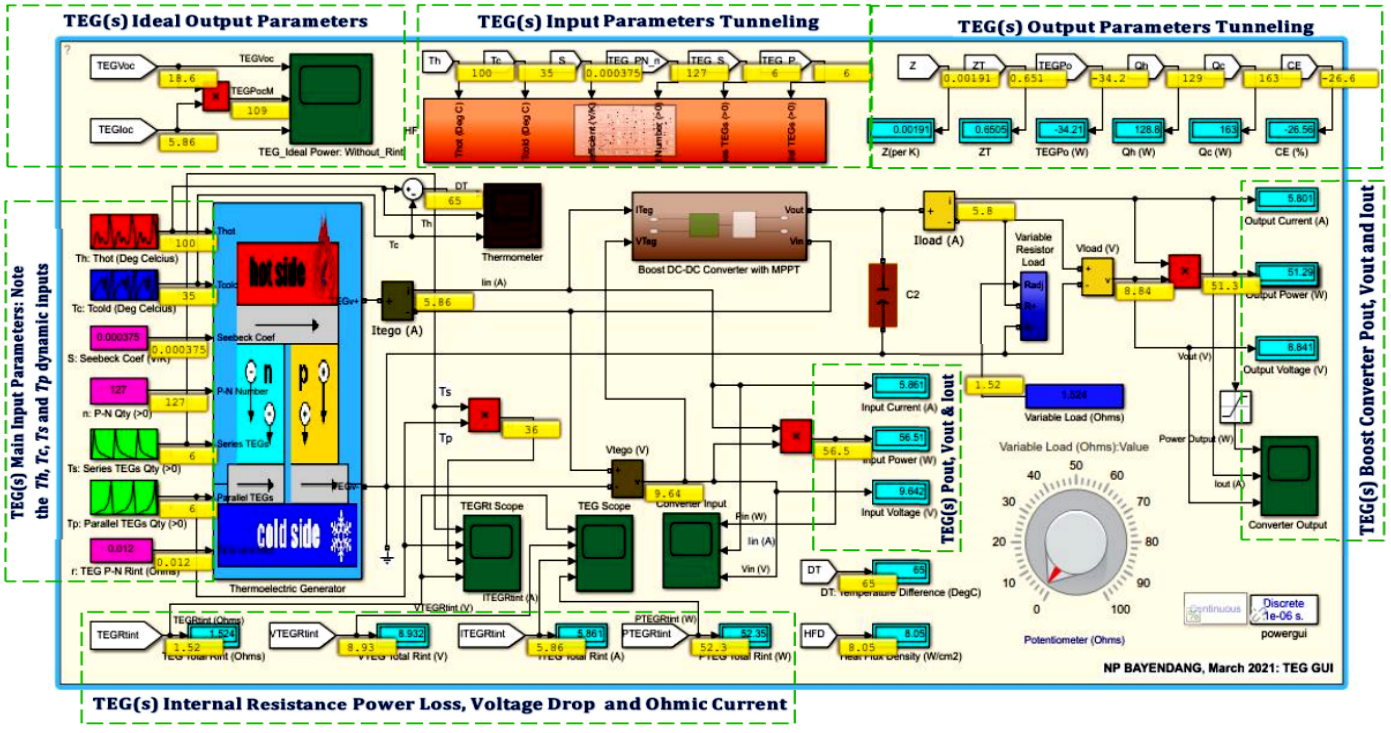


Figure 1b: TEG(s) dynamic simulator user's interface − shows the transient simulation with the T_h, T_c, T_s *and* T_p input parameters auto changing with time

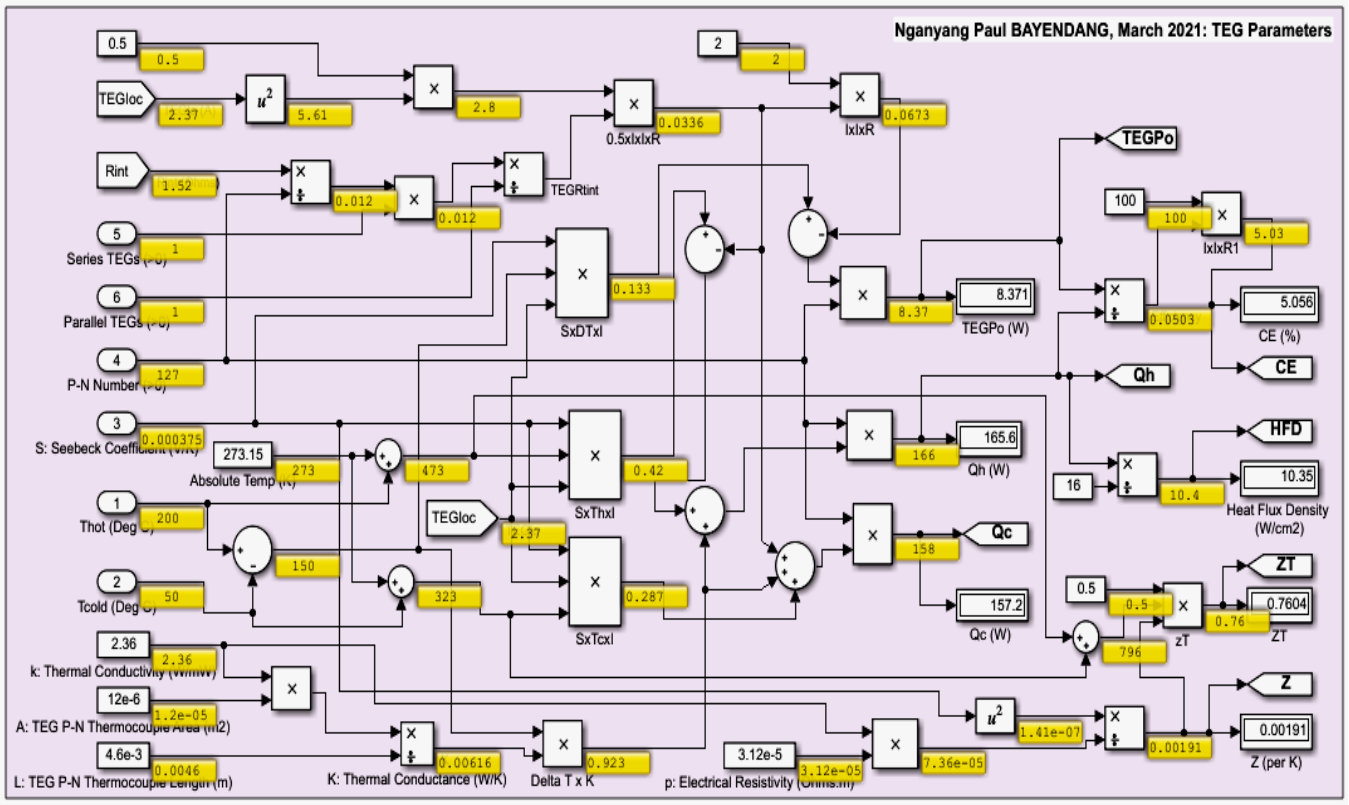


Figure 1c: TEG(s) modeling and simulation − TEG(s) parameters

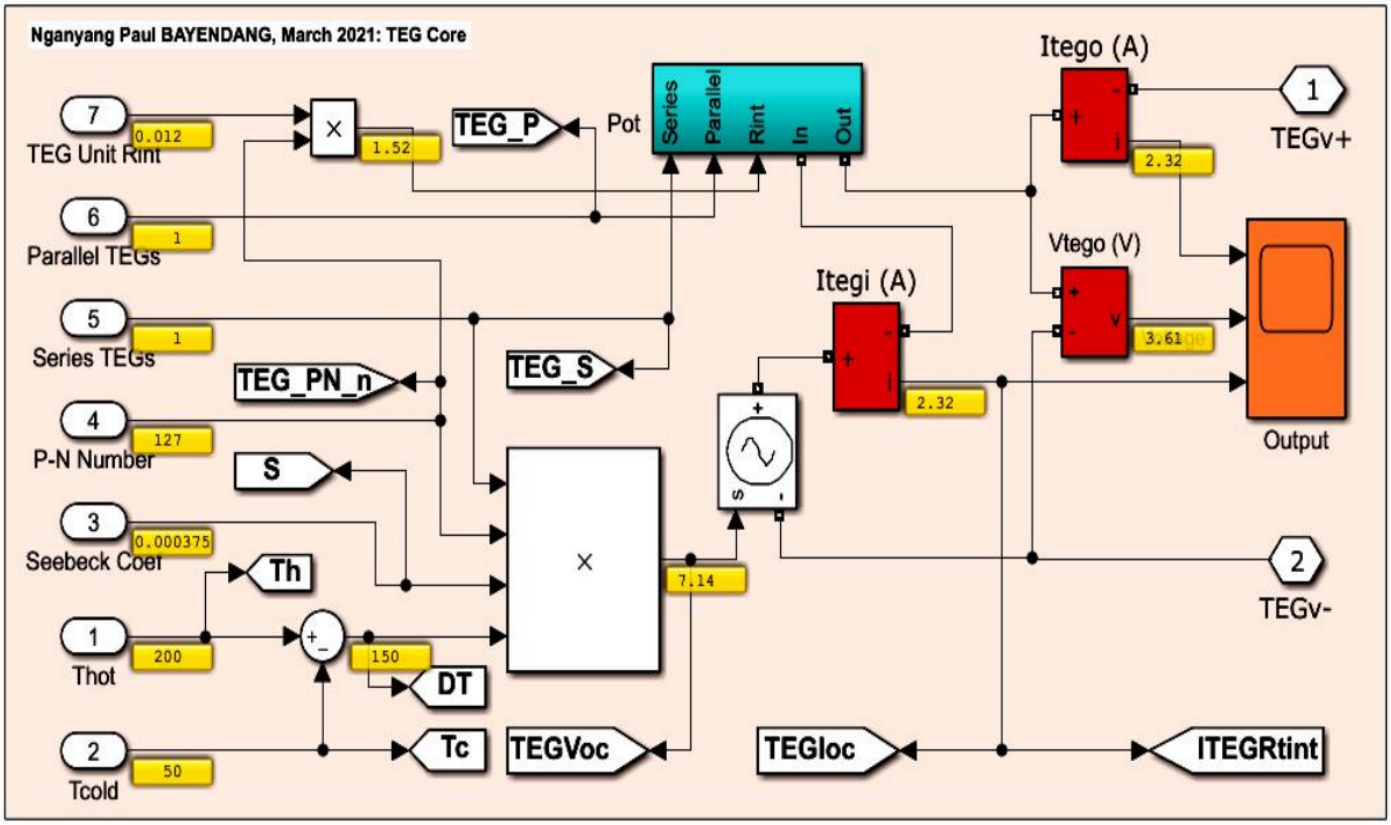


Figure 1d: TEG(s) modeling and simulation − TEG(s) engine

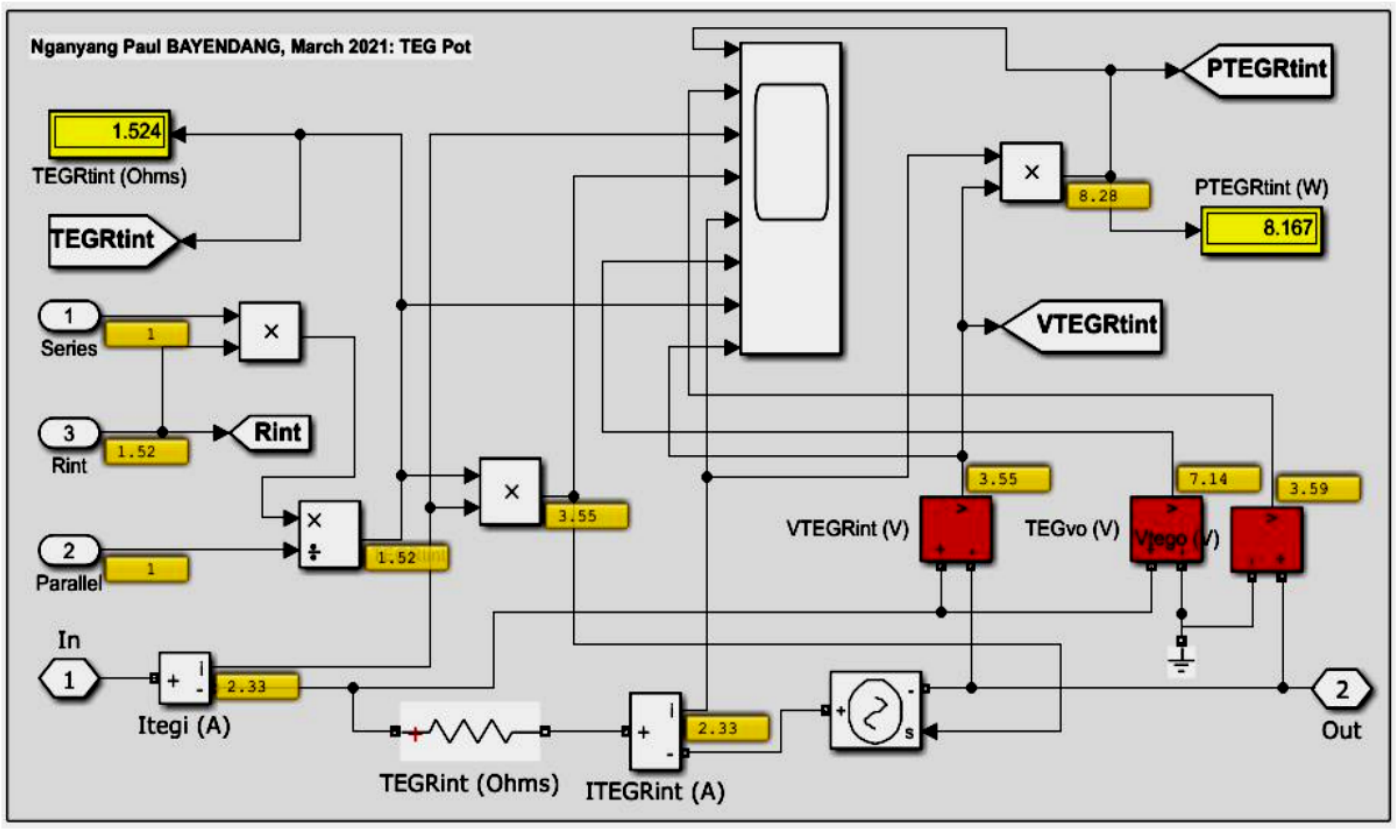


Figure 1e: TEG(s) modeling and simulation − TEG(s) automatic internal source total electrical resistance R_t

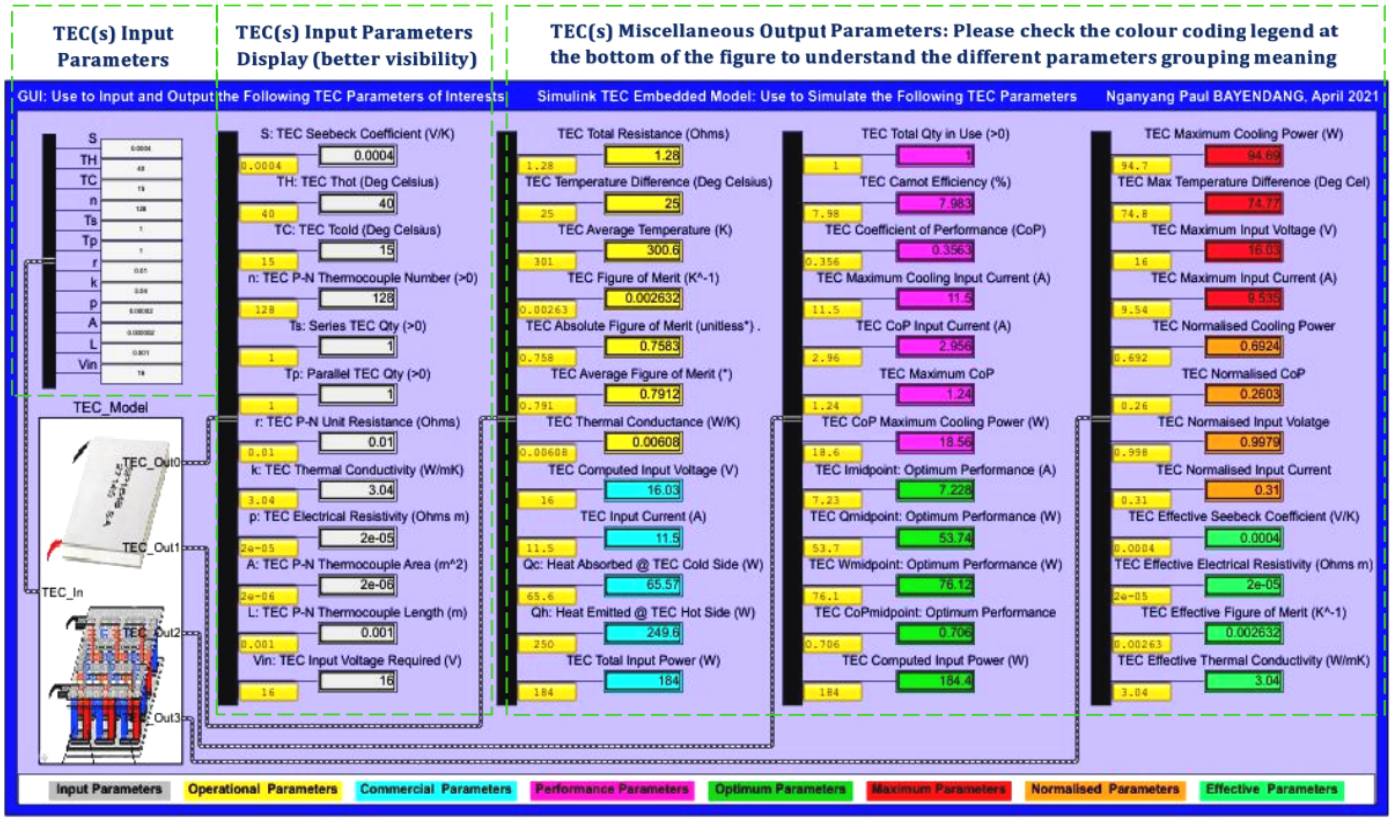


Figure 2: TEC(s) simulator − simulates TECs various parameters by inputting a TEC specific data sheet parameters and calculates its theoretical outputs

3. TEGs and TECs Simulations Results

The TEGs and TECs simulations results are presented in three parts as follows, the i) TEGs parameters static simulation results ii) TECs parameters static simulation results and iii) TEGs parameters dynamic simulation results. Understanding these parameters operation is very paramount; otherwise, doing the physical design would just be a matter of taking chances and hoping for the best − which is sometimes the case, as most designers have reported very bad design results, likely from not understanding TE devices dynamic operations and limitations. The results from investigating the TEG(s) and TEC(s) parameters optimal operation points are discussed in details in Section 4.

3.1. TEGs Parameters Static Simulation Results

Figures 3 − 6 expound the TEGs parameters simulated to determine their optimal operation points − marked in green.

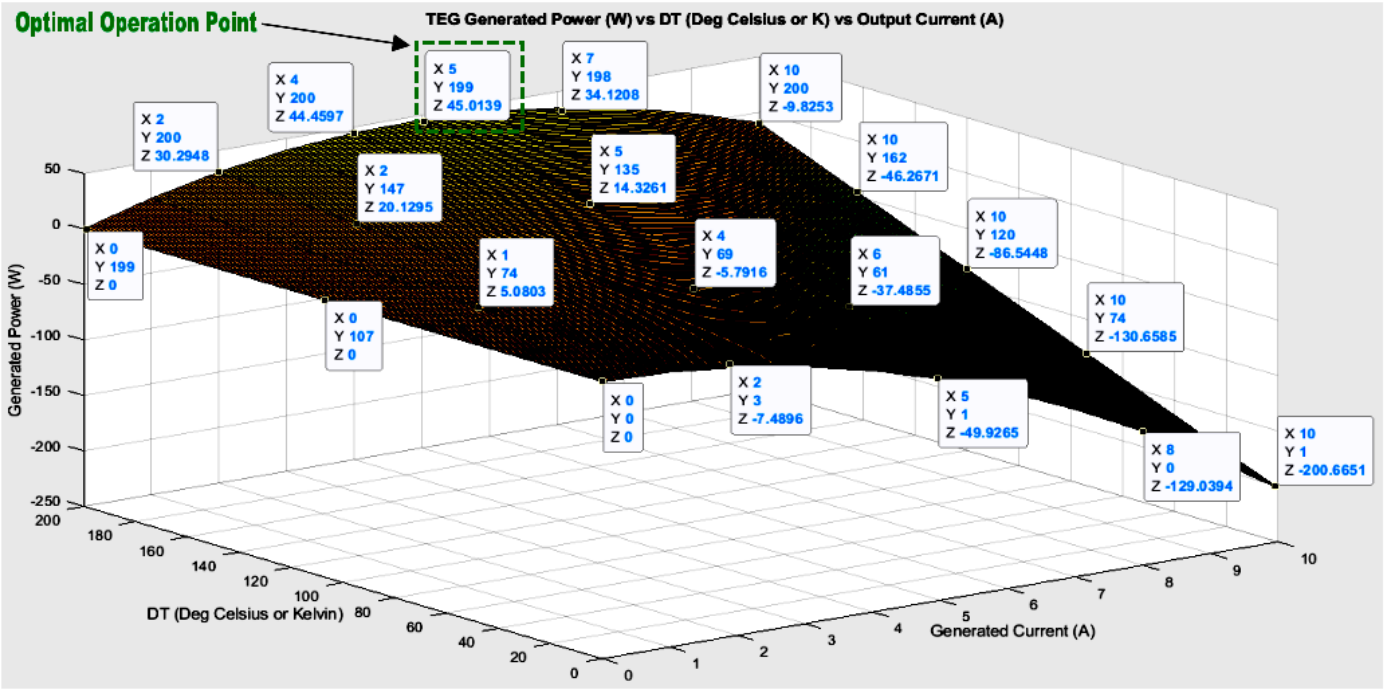


Figure 3: TEG power output P_o (W) vs temperature difference ΔT (°C) vs current output I (A)

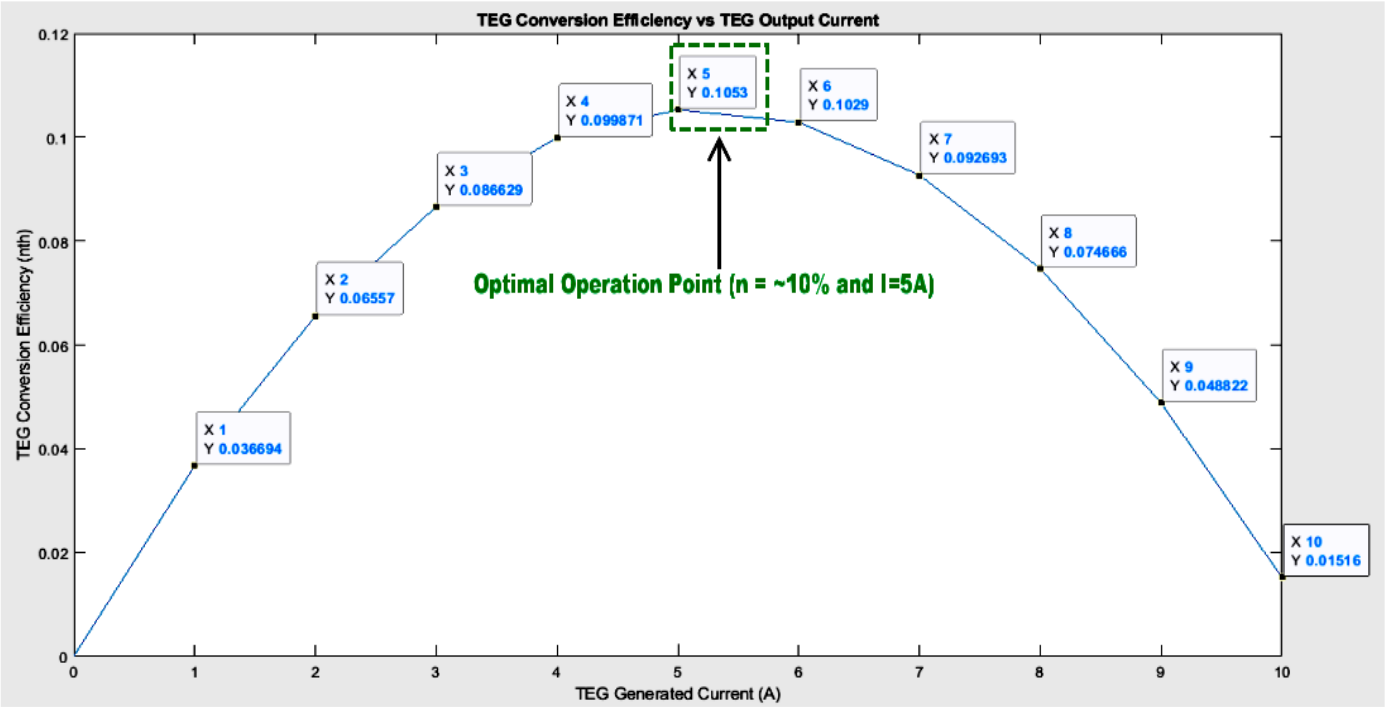


Figure 4: TEG conversion efficiency η vs current output I (A)

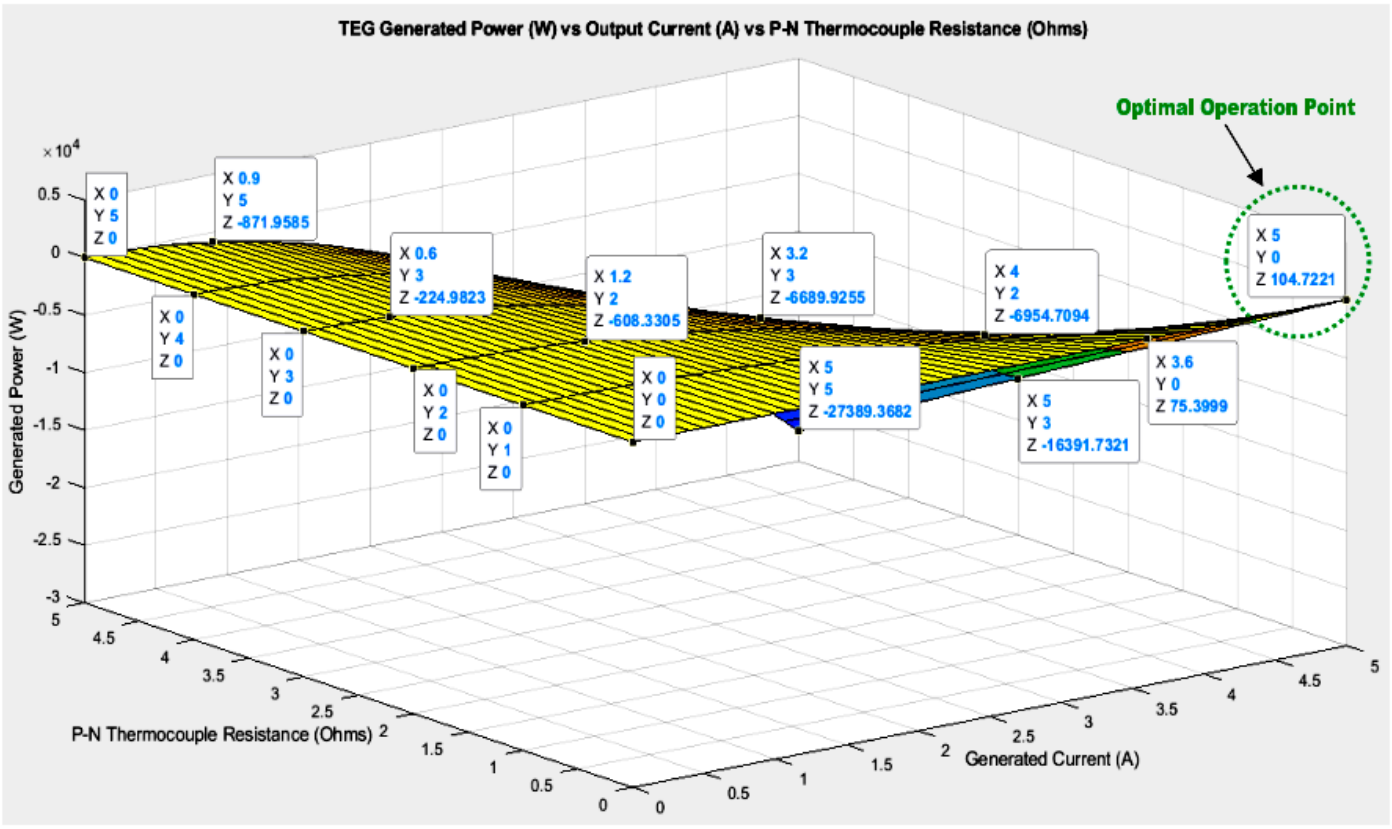


Figure 5: TEG power output P_o (W) vs r or R or R_t (Ω) vs current output I (A)

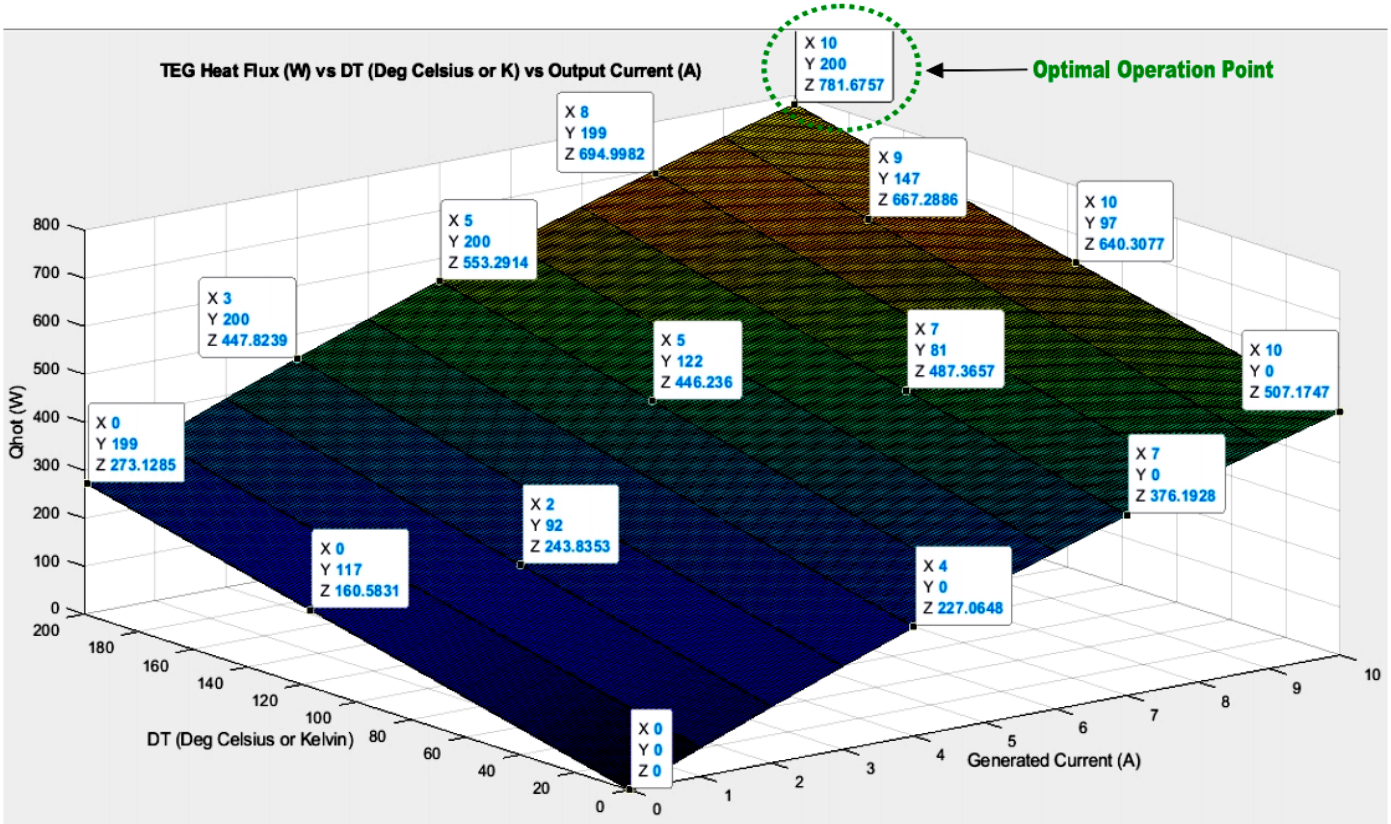


Figure 6: TEG absorbed heat Q_h (W) vs temperature difference ΔT (°C) vs output current I (A)

3.2. TECs Parameters Static Simulation Results

TECs parameters are simulated in Figures 7 - 10 to determine their possible optimal operation points − shown highlighted in red.

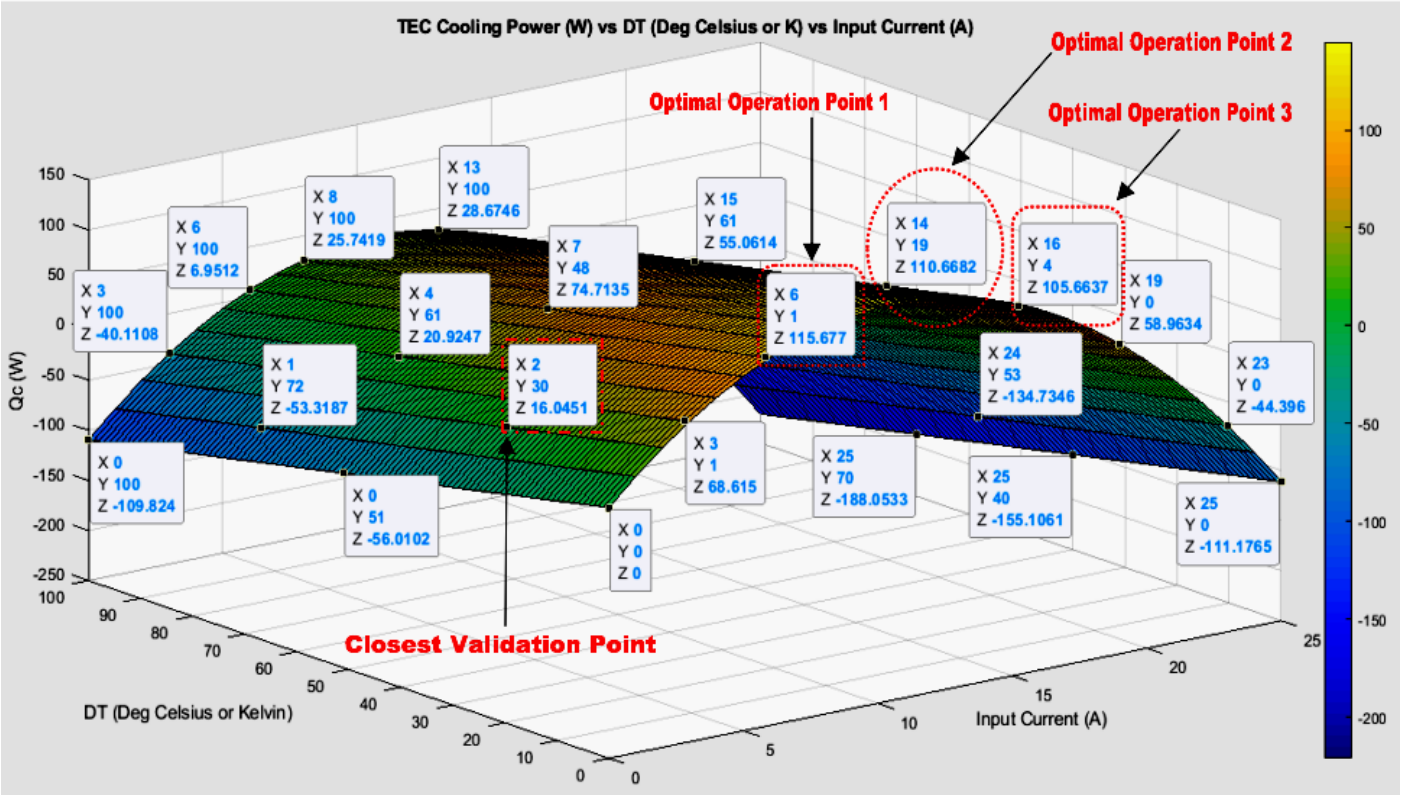


Figure 7: TEC cooling power or heat absorbed Q_c (W) vs temperature difference ΔT (°C) vs input current I_{in} (A)

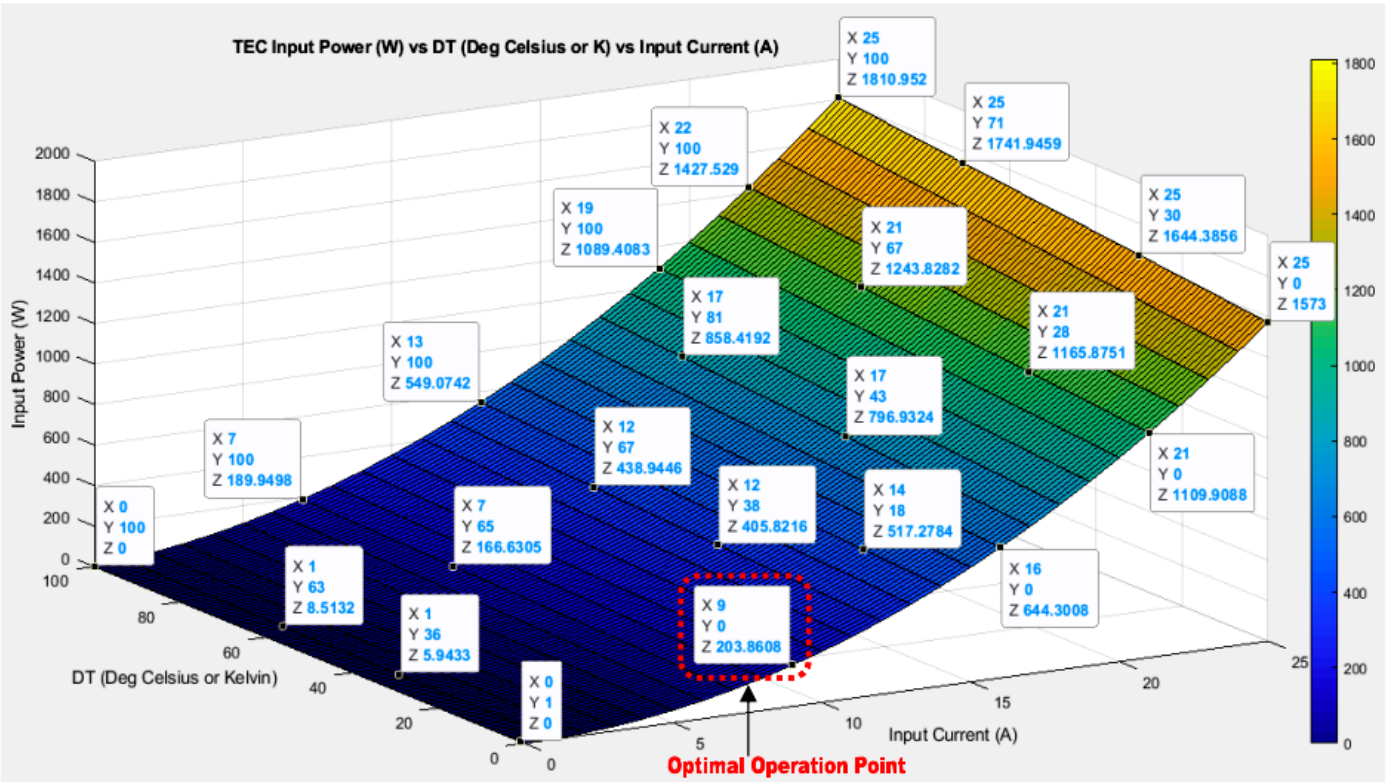


Figure 8: TEC input power P_{in} (W) vs temperature difference ΔT (°C) vs input current I_{in} (A)

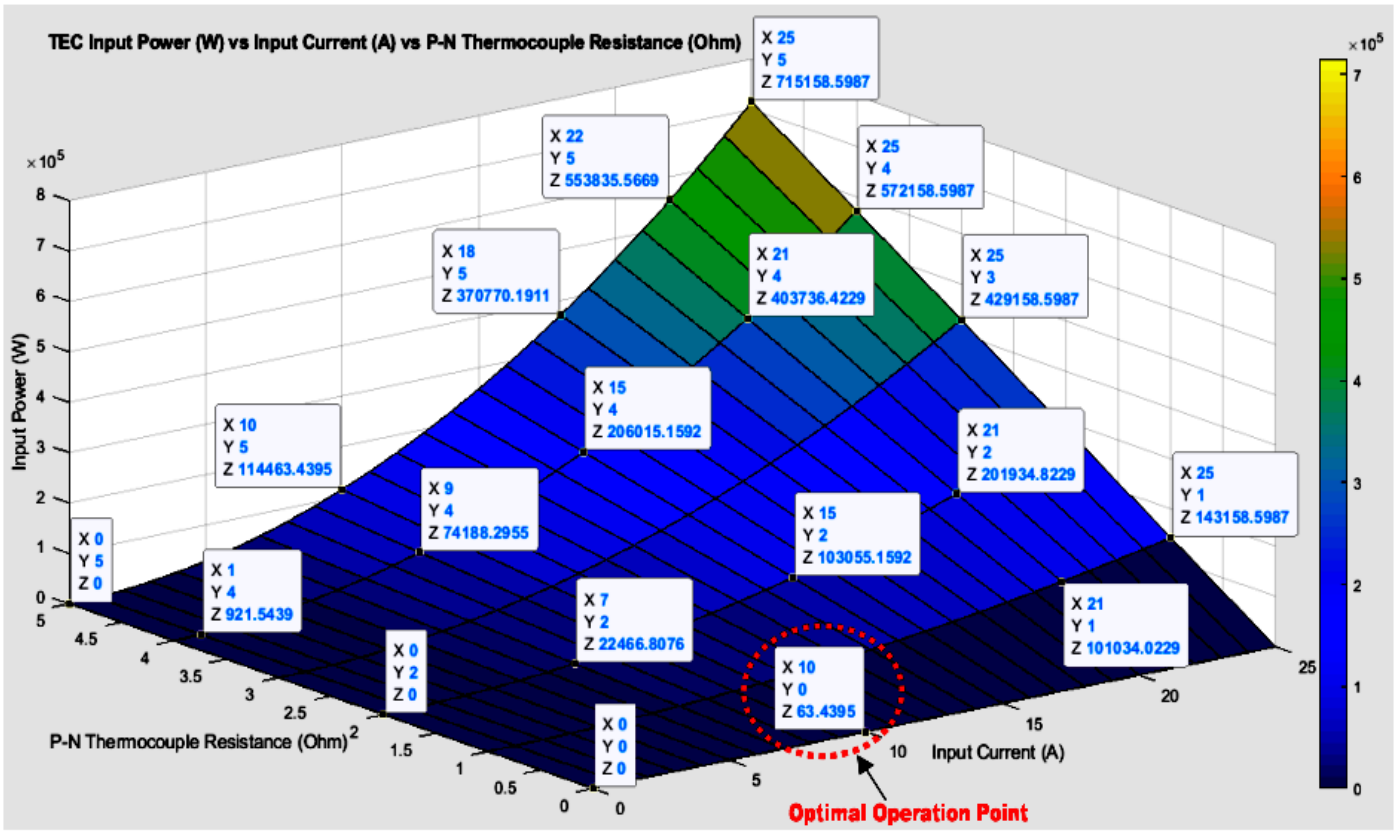


Figure 9: TEC input power P_{in} (W) vs internal resistance r or R or R_t (Ω) vs input current I_{in} (A)

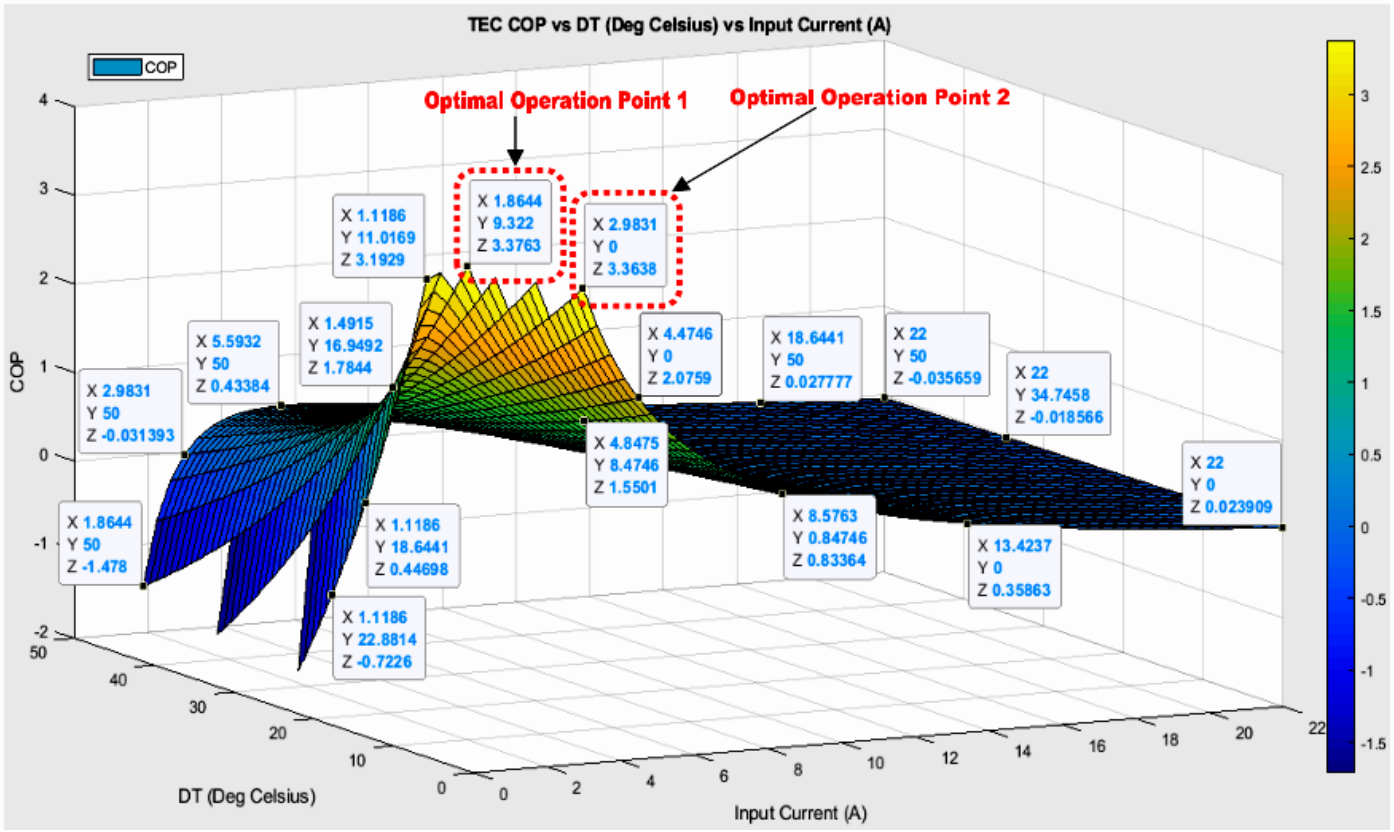


Figure 10: TEC coefficient of performance CoP (%) vs temperature difference ΔT (°C) vs input current I_{in} (A)

3.3. TEGs Parameters Dynamic Simulation Results

TEG modules temperatures, its series and parallel connections dynamic simulation results are depicted in Figures 11a - 11f.

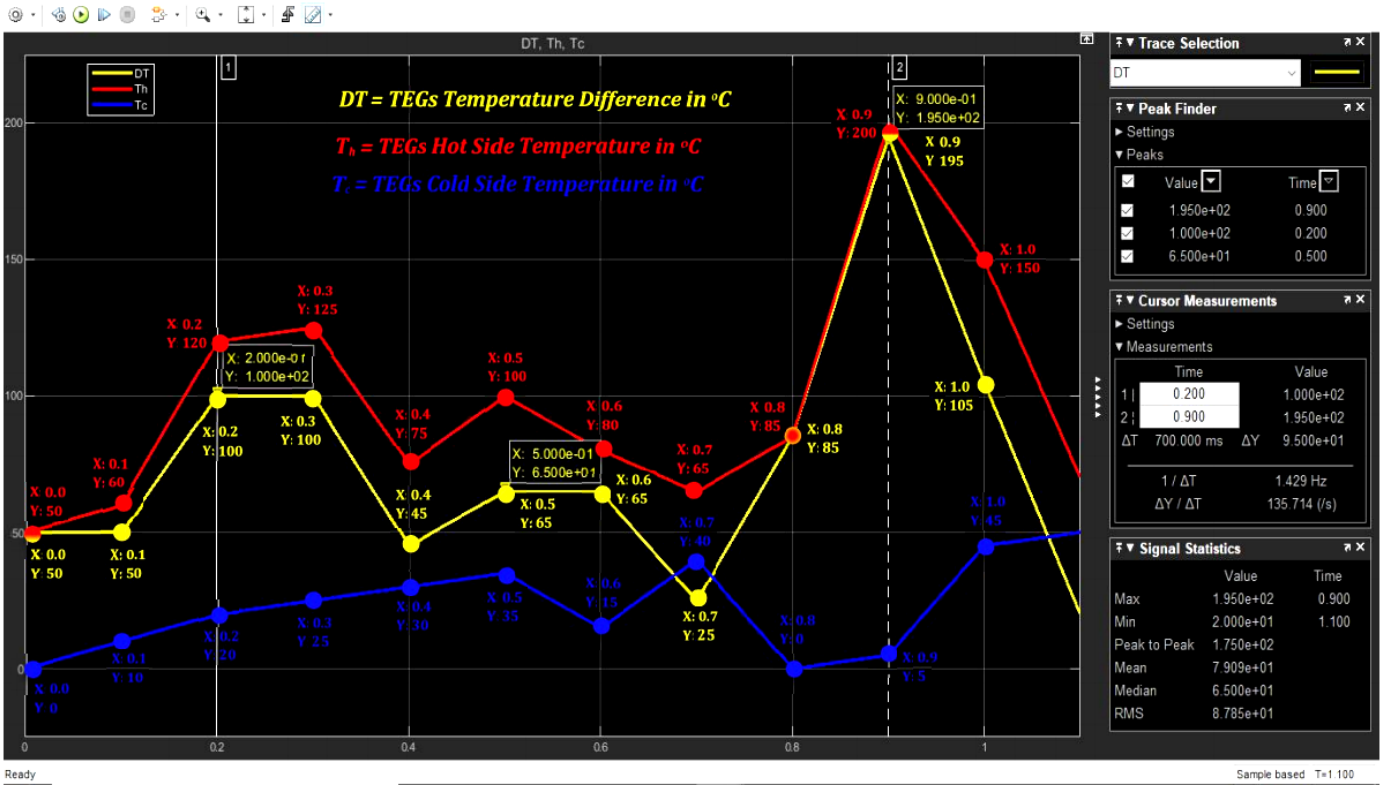


Figure 11a: 36 TEGs hot (T_h) and cold (T_c) temperatures as well as temperature difference (DT) dynamics − temperature changes as simulation progresses

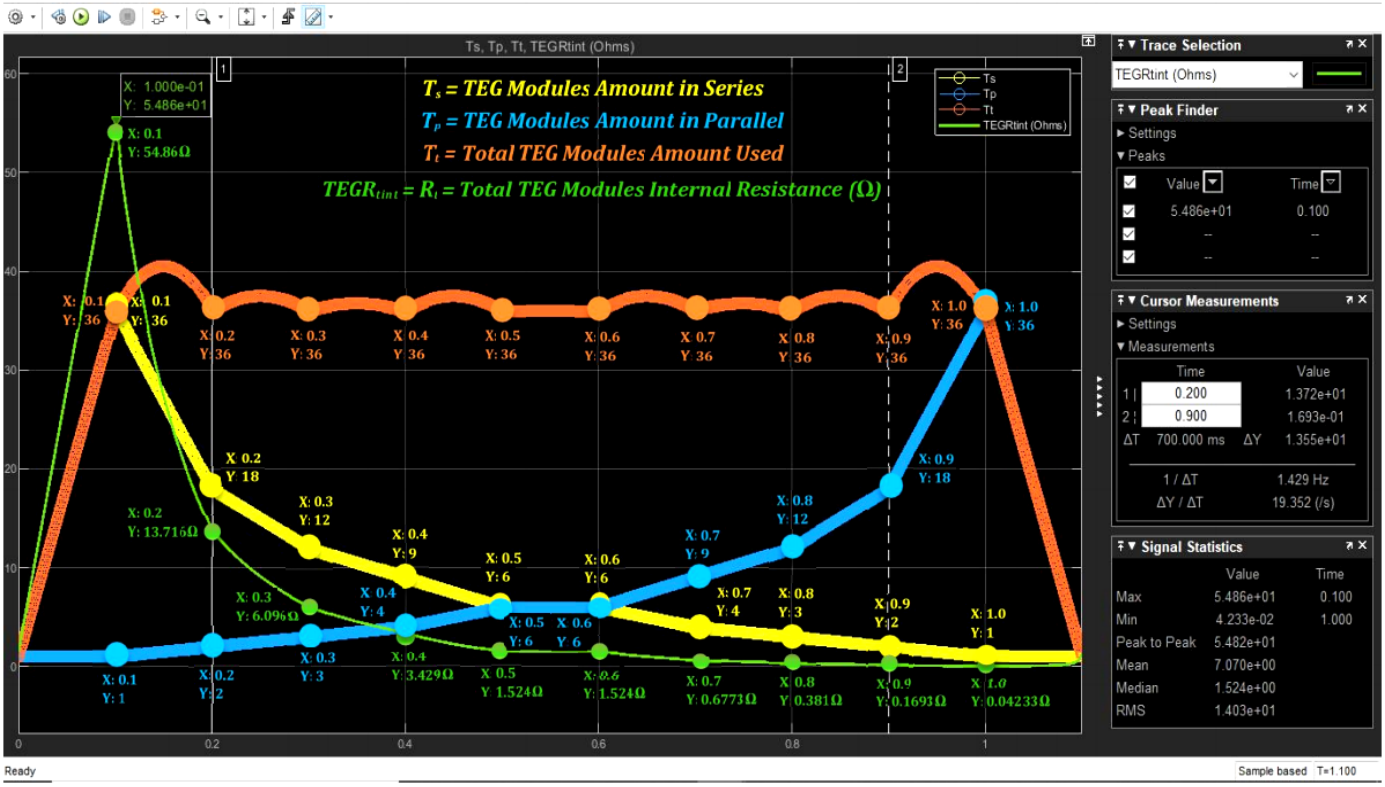


Figure 11b: TEGs in series (T_s), parallel (T_p) and total internal resistance (R_t) dynamics − 36 TEG modules simulated in 10 different auto reconfiguration

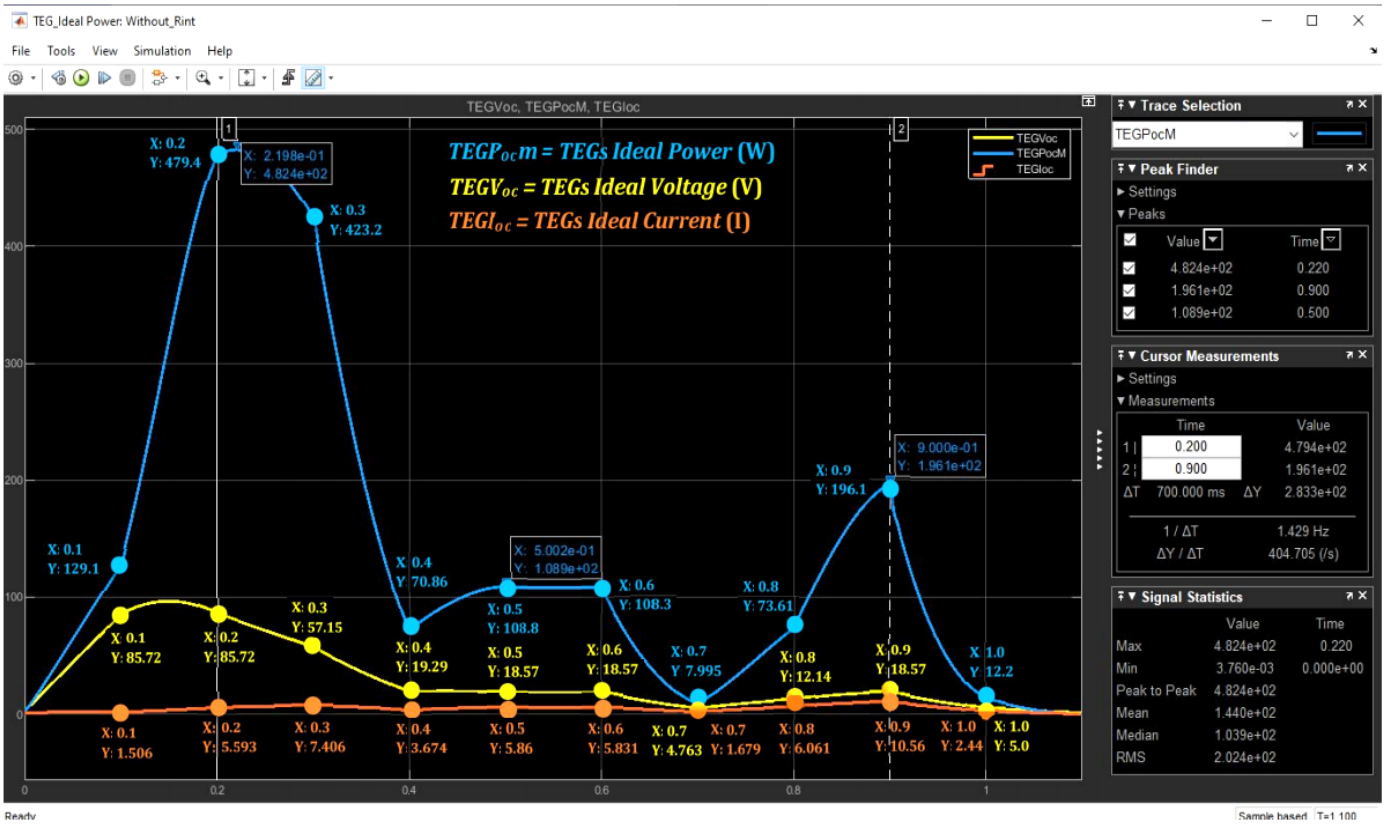


Figure 11c: 36 TEGs ideal output power, voltage and current dynamics; as TEGs temperatures and its 10 configurations change as simulation progresses

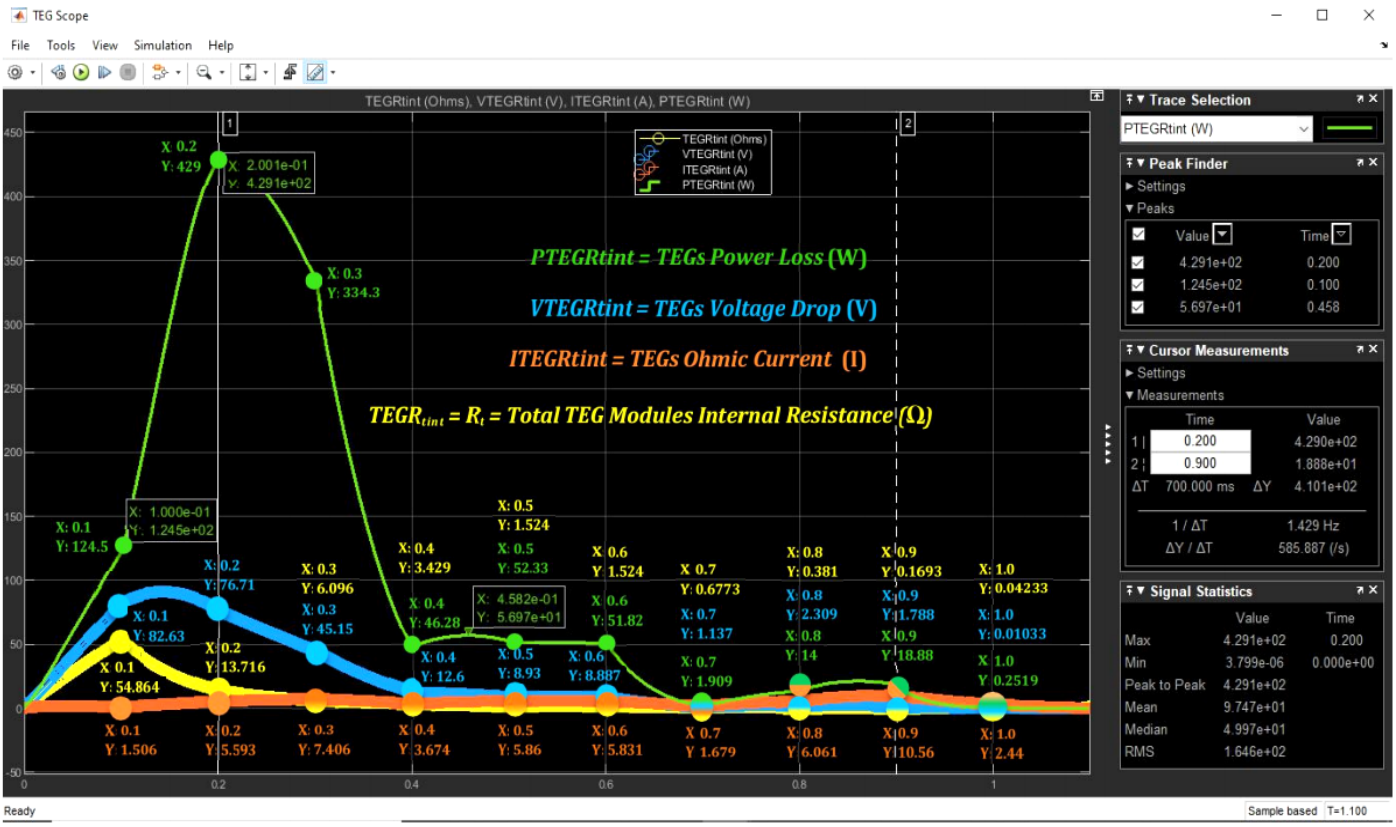


Figure 11d: 36 TEGs total internal resistance current, voltage and power losses dynamics; as the TEGs 10 configurations and temperatures auto change

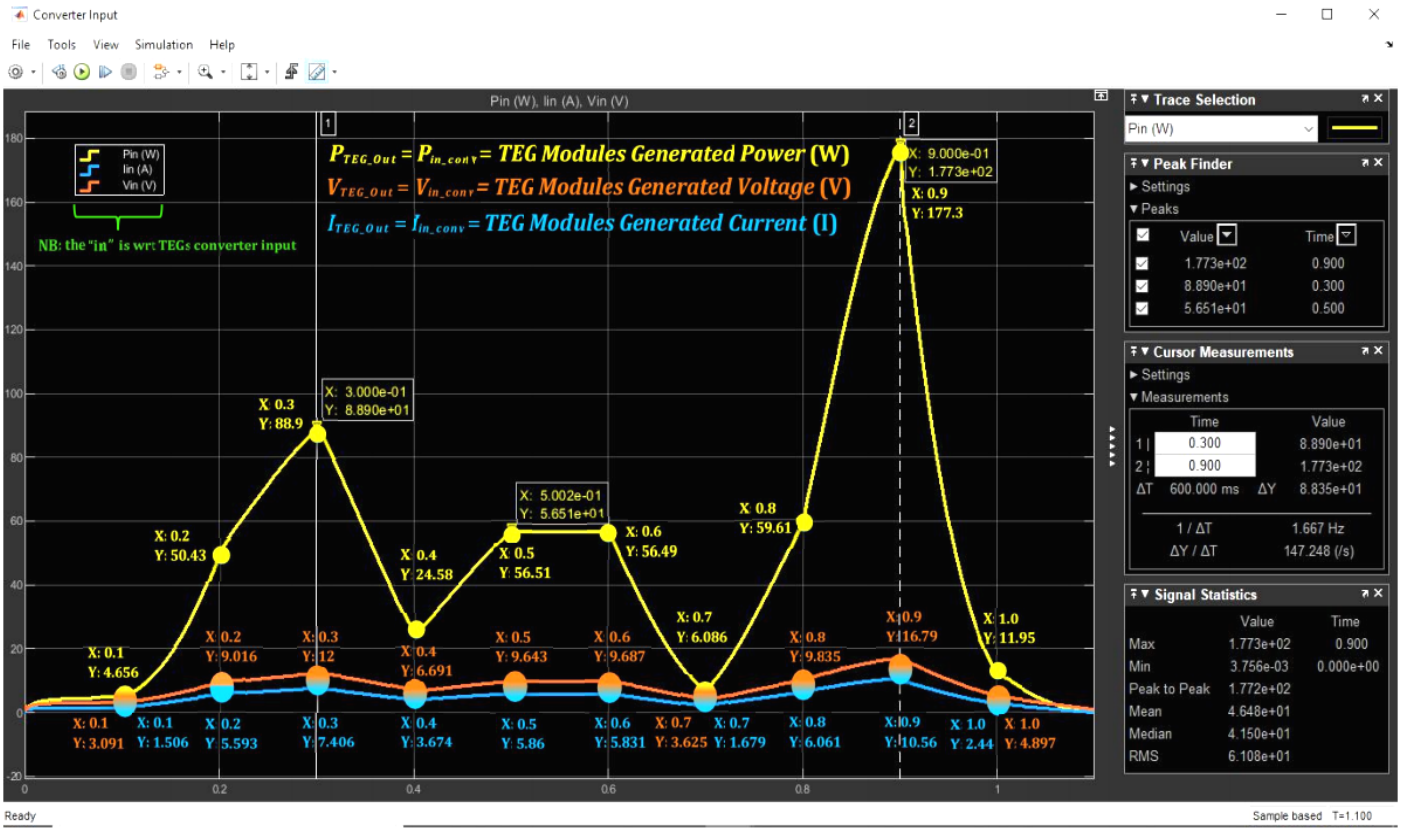


Figure 11e: 36 TEGs output power, voltage and current dynamics as the TEGs temperatures and 10 configurations auto change as simulation progresses

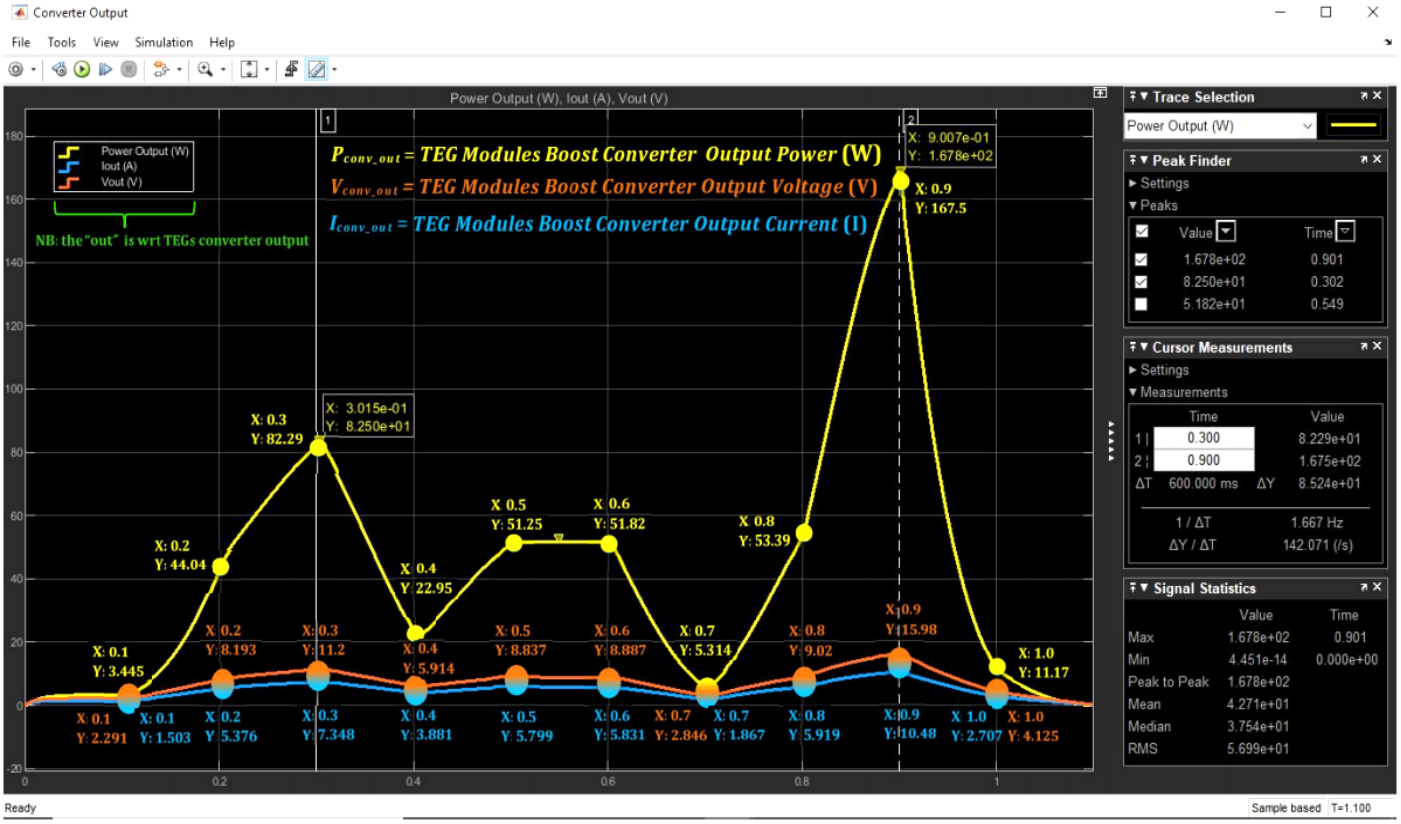


Figure 11f: 36 TEGs boost converter output power, voltage and current dynamics as the TEGs temperatures and the TEGs 10 configurations auto change

4. TEGs and TECs Simulations Results Discussions

The TEGs/TECs simulations results demonstrated in Section 3, are engaged below in their following respective sub-sections.

4.1. TEGs Parameters Static Simulation Results Discussion

Some of the crucial TEGs parameters simulated in Section 3.1. and the significance of the results are herein asserted. As exemplified in Figure 3, a TEGs generated power P_o is proportional to its temperature difference ΔT and output current I; however, I above 5A (in this case) will decrease P_o − which is because of the TEGs internal Ohmic heating as a result of the increasing output current I. The ΔT, P_o and I optimum operation points are emphasized in green in Figure 3. In Figure 4, a TEGs conversion efficiency η is directly proportional to current output I up to ~5A max (in this case) and decreases later as highlighted in green. It should be noted that η is as well directly proportional to P_o. However, a TEG P_o is reciprocally proportional to its p-n thermocouple junction resistance r and as well to its total internal resistance R_t (more than one connected TEG modules), though pro rata to I up to ~5A (in this case) as portrayed in Figure 5. At this optimal point; R_t or R is 0Ω, I is ~5A maximum and P_o is ~105W as highlighted in green. In Figure 6, the TEGs current output I is proportional directly to the TEGs absorbed heat Q_h (at temperature T_h on the TEG hot-side) which in turn is directly dependent on the TEG ΔT. Figure 6 pictured the optimum point stressed-out in green. It should be noted that these results are not specific to a particular TEGs' connections – the results just fundamentally give a holistic theoretical understanding on what TEGs physical parameters must be taken into considerations, how they are interrelated, their associated dynamics and technical limitations and how they can be practically traded-off or optimized for optimal performance when designing TEGs power supply systems.

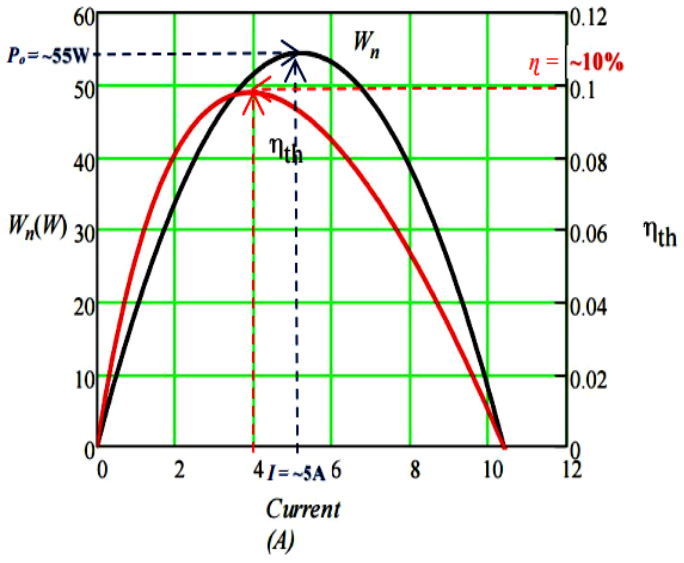


Figure 12: Validating our model with [9] − TEG (i) output power P_O = ~55W vs output current I = ~5A validating our Figure 3 result and (ii) conversion efficiency η = ~10% vs output current I = ~ 4A validating our Figure 4 result.

Depicted in Figure 12, is a result of a typical TEG model simulated with Mathcad using TEG standard specifications from typical manufacturers data-sheet as presented in [9]. This was used as a benchmark to validate our TEG model simulation accuracy − which is very close, besides a few discrepancies due to minor simulation settings differences. In light of this, our implemented TEG model can be used and developed further to simulate TEGs, including infinite series and parallel connections, which are central to our research and in large scale TEGs uses.

4.2. TECs Parameters Static Simulation Results Discussion

Some of the critical TECs parameters simulated in Section 3.2. and the importance of the results are herein articulated. Figure 7 illustrates that TECs Q_c on TECs cold-side T_c, is reciprocally proportional to ΔT but proportional directly to I_{in} up to a maximum point, after which Q_c starts dropping. The reasons are due to i) Joule heating (the more I_{in}, the more the internal heating effect) and also ii) the second law of thermodynamics − simply put, heat flows from a hotter to a colder body; in this regards, the heating caused by the increasing I_{in}, increases the TECs internal temperature up to a temperature greater than that surrounding the TECs hot-side T_h; consequently, heat now starts to flow from the TECs hot-side to its cold-side, thus making the cooling process (heat pumping) on the TECs cold-side inefficient. In Figure 7 and highlighted in red, the Q_c, ΔT and I_{in}; display three optimal operation points depending on the TECs design constraints/ priorities. In option 1, Q_c is 115.677W with a ΔT of 1°C and I_{in} of 6A. In option 2, Q_c is 110.668W with a ΔT of 19°C and I_{in} of 14A. In option 3, Q_c is 105.664W with a ΔT of 4°C and I_{in} of 16A. As evident, either ΔT and or I_{in} depending on the design constraints, can be optimized by either minimizing the TECs ΔT and or maximizing TECs I_{in} to increase Q_c within max operational limits. In Figure 8, P_{in} and I_{in} are directly proportionally, which will initially increase Q_c until a certain maximum limit, after which increasing P_{in} and I_{in} drop Q_c − contrary to ΔT which is inversely proportional to Q_c. The optimal operation point is highlighted in red. Figure 9 shows a TECs P_{in} vs I_{in} vs R. Normally R is set fixed when designed by the manufacturer but now, with R_t introduced, R can be fairly altered and if it is matched to R_s, maximum power will be transferred to the TEC(s); thereby, optimizing P_{in} and maximizing Q_c as highlighted in red. Figure 10 demonstrates how *CoP* akin to Q_c; increases with decreasing ΔT and initially with increasing I up to a maximum value and then starts decreasing, as current I increases as shown variously in Figure 10. Depending on the design constraints, two optimal *CoP* operation points are evident as highlighted in red − in optimal operation point 1, a *CoP* of 3.3763 is achievable by minimizing I_{in} to 1.8644A and maximizing ΔT to 9.322°C; whereas in optimal operation point 2, a *CoP* of 3.3638 is attainable by maximizing I_{in} to 2.9831A and minimizing ΔT to 0°C. Finally, our TECs model is reasonably validated by comparing a specific Q_c of Figure 7 with that of Figure 13, as shown. The discrepancy is due to different TECs parameters setting. In sum, understanding the theory of TECs parameters and taking the various operational dynamics involved into considerations are very crucial in TEC(s) design/performance.

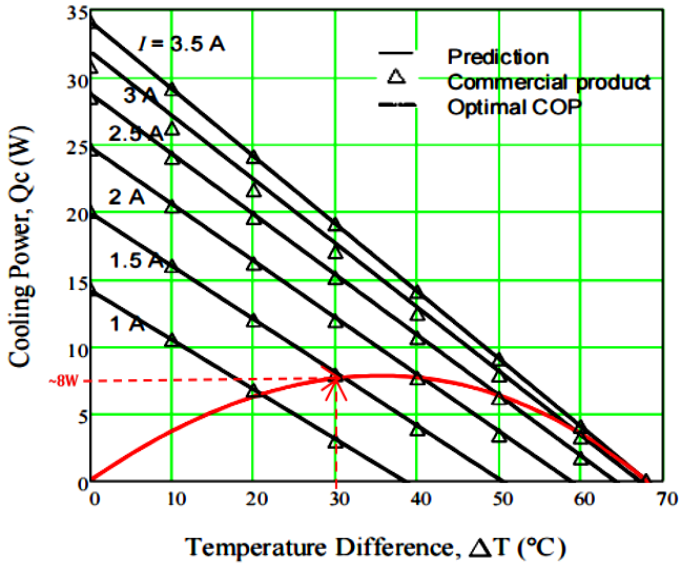


Figure 13: Validating our model with Lee, 2016 [9] − using TEC cooling power Q_c = ~8W vs input current I = ~1.5A vs ΔT= ~30°C to validate our TECs Q_c in Figure 7 result with cooling power Q_c = ~16W vs input current I = ~2A vs ΔT= ~30°C.

4.3. TEGs Dynamic Simulation Results Discussion

Some of the critical TEGs dynamic simulated in Section 3.3. and the importance of the results are herein discussed. The TEGs temperatures and modules electrical connections (series, parallel, series/parallel) dynamics were simulated. In which beginning with the TEGs temperature dynamics, various arbitrary temperatures on the TEGs hot and cold sides as demonstrated in Figure 1b and Figure 11a, as well as summarized in Table 1, were simply dynamically simulated using time-series inputs. As expected, the TEGs dynamically generated power, voltage and current; increased with increasing T_h and DT but with decreasing T_c.

Table 1: TEGs time-series inputs dynamic simulations results summary

Parameters	Matlab / Simulink Simulation Time									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Figure 11a	**TEG modules T_h, T_c and DT dynamic temperature inputs in ℃**									
TEGs T_h	60	120	125	75	100	80	65	85	200	150
TEGs T_c	10	20	25	30	35	15	40	0	5	45
TEGs DT	50	100	100	45	65	65	25	85	195	105
Figure 11b	**36 TEG modules in 10 dynamic T_s, T_p, T_t and R_t auto configuration**									
T_s	36	18	12	9	6	6	4	3	2	1
T_p	1	2	3	4	6	6	9	12	18	36
T_t	36	36	36	36	36	36	36	36	36	36
R_t (Ω)	54.86	13.72	6.096	3.429	1.524	1.524	0.677	0.381	0.169	0.0423
Figure 11c	**36 TEG modules ideal (if TEG_{Rtint} = 0) power, voltage and current**									
TEG_{PocM} (W)	129.1	479.4	423.2	70.86	108.8	108.3	7.995	73.61	196.1	12.2
TEG_{Voc} (V)	85.72	85.72	57.15	19.29	18.57	18.57	4.763	12.14	18.57	5
TEG_{Ioc} (A)	1.506	5.593	7.406	3.674	5.86	5.831	1.679	6.061	10.56	2.44
Figure 11d	**36 TEG modules internal resistance power, voltage and current**									
TEG_{Rtint} (Ω)	54.86	13.72	6.096	3.429	1.524	1.524	0.677	0.381	0.169	0.0423
$PTEG_{Rtint}$ (W)	124.5	429	334.3	46.28	52.33	51.82	1.909	14	18.88	0.2519
$VTEG_{Rtint}$ (V)	82.63	76.71	45.15	12.6	8.93	8.887	1.137	2.309	1.788	0.0103
$ITEG_{Rtint}$ (A)	1.506	5.593	7.406	3.674	5.86	5.831	1.679	6.061	10.56	2.44
Figure 11e	**36 TEG modules generated (terminal) power, voltage and current**									
P_{TEG_Out} (W)	4.656	50.43	88.9	24.58	56.51	56.49	6.086	59.61	177.3	11.95
V_{TEG_Out} (V)	3.091	9.016	12	6.691	9.643	9.687	3.625	9.835	16.79	4.897
I_{TEG_Out} (A)	1.506	5.593	7.406	3.674	5.86	5.831	1.679	6.061	10.56	2.44
Figure 11f	**36 TEG modules boost converter output power, voltage and current**									
P_{conv_out} (W)	3.445	44.04	82.29	22.95	51.25	51.82	5.314	53.39	167.5	11.17
V_{comv_out} (V)	2.291	8.193	11.2	5.914	8.837	8.887	2.846	9.02	15.98	4.125
I_{conv_out} (A)	1.503	5.376	7.348	3.881	5.799	5.831	1.867	5.919	10.48	2.707

The TEG modules quantity used and most vitally in series, parallel and mixed connection were simulated, whereby as shown in Figure 1b and Figure 11b, as well as summarized in Table 1; 36 TEGs were arbitrary chosen and then arranged in 10 different combinations to study the effects of the various arrangements and when matched to a 1.524Ω electrical load. Each arrangement gives a different R_t, consequently giving different generated powers, voltages and currents. Figure 11c depicts the TEGs ideal power, voltage and current generated − assuming the TEGs R_t or TEG_{Rtint} is trivial. Figure 11d shows the power loss, voltage drop and Ohmic current due to the presence of TEG_{Rtint}. Finally, Figures 11e and 11f, show the resultant output power, voltage and current supplied to the DC-DC boost converter and from it. As apparent, more TEG modules increased the output values; however, what is more insightful is how TEGs opt to be connected and matched to a R_L − to ensure maximum power is transferred between R_t and a R_L.

5. Conclusions

Sustainable energy is becoming popular to supplement the traditional grid and for private use, as well as for green economy. In view of this, we proffer thermoelectricity as an alternative energy source (TEGs) as well as an energy efficient load (TECs) for assorted applications that require low DC power, cooling and heating. However, TEG and TEC require multiple units connected in series and or in parallel to provide decent output and cooling powers respectively. Usually, the uninformed perception would be trying to utilize more TEGs and TECs with the hope to get more output and cooling powers respectively. However, our findings asserted this is not really the case, since i) TEG and TEC are not entirely linear devices, especially with increasing current, ii) TEG and TEC temperature difference ΔT and current parameters have performance dynamics which must be operated within very strict optimal operation limits to guarantee efficiency and iii) TEGs and TECs total electrical resistance R_t changes − increases when connected in series and decreases when connected in parallel. Thus, the overall power and efficiency will be affected, especially if the source and load resistances are not matched to transfer maximum power. In essence, our research major contributions include formulas developed for various TEGs/TECs parameters with focus on the TEG and TEC modules total resistance R_t variations − when more than one TEG and or TEC modules are connected in infinite series and or in parallel combinations. Further contributions include detailed TEGs and TECs theoretical simulated models using Matlab/Simulink, whereby the TEGs and TECs models were used to easily simulate and investigate some thermoelectricity profound parameters performance dynamics, R_t losses and to validate some of their operation points with industry standard models. Assorted large scale practical applications of TEGs and TECs were examined and in light of their results, our future work will include embarking on an actual lab design, testing our implemented models using them and refining ours accordingly while taking the physical dynamics into account. Thereafter, a practical pilot 1kW implementation shall be devised for a low energy combined cooling, heating and power (CCHP) system − as an alternative energy green option for private use.

Nomenclature/Symbols

A	TEG/TEC p-n junction thermocouple area in m^2
$CCHP$	Combined cooling, heating and power
CFD	TEC(s) cold flux density in W/m
CoP	TEC(s) coefficient of performance
CoP_e	TEC(s) CoP expression
CoP_{max}	TECs maximum CoP
CoP_{mid}	TEC(s) midpoint CoP
CoP_n	TEC(s) normalized CoP TEC(s)
ΔT	TEG(s) temperature difference ($T_h - T_c$) in °C or K
ΔT_{max}	TEC(s) maximum temperature difference in °C
ΔT_n	TEC(s) normalized temperature difference
HFD	TEG(s) heat flux density in W/m^2
I	TEGs output current in ampere through the TEG(s)
I_{conv_out}	TEGs booster converter output current
I_{cop}	TEC(s) current in ampere to yield CoP
Icp_{max}	TEC(s) maximum cooling power current in ampere
I_{in}	TEC module(s) input current in ampere
Iin_n	TEC(s) normalized input current is the ratio of I_{cop} and I_{max}
I_{Max}	TEG(s) maximum output current in ampere
I_{max}	TEC(s) maximum input current in ampere when $Q_c = 0$
I_{mid}	TEC(s) midpoint current in ampere
I_n	TEG(s) normalized output current
$ITEG_{Rtint}$	TEG ohmic current − results to TEG Ohmic or Joule heating
I_{TEG_Out}	TEGs generated current (input current to the boost converter)
K	TEC/TEG(s) thermal conductance in (W/K)
k_e	TEG(s)/TEC(s) effective thermal conductivity in W/mK
L	TEG/TEC p-n junction thermocouple length in meter
n	P-N thermocouples amount used in a TEG/TEC
η	TEG(s) thermal/electrical/conversion efficiency
η_c	Carnot efficiency
η_e	TEG(s) conversion efficiency expression
η_n	TEG(s) conversion efficiency normalized
η_m	TEG(s) maximum conversion efficiency
η_{mp}	TEGs max power conversion efficiency at the TEGs maximum P_o
ρ	TEG/TEC electrical resistivity in Ω.m
ρ_e	TEG(s)/TEC(s) effective electrical resistivity in Ω.m
P_{conv_out}	TEGs booster converter output power
P_{in}	TEC module(s) input power in watt
Pin_{mid}	TEC(s) midpoint input power in watt
Po	TEG(s) output power in watt − which is $Q_h - Q_c$
Po_{max}	TEG(s) maximum output power in watt
P_n	TEG(s) normalized output power
$PTEG_{Rtint}$	TEG generated power loss − due to TEG internal resistance
P_{TEG_Out}	TEGs generated power (input power to the boost converter)
Q_c	TEC module(s) cooling power on its cold-side in (W)
Q_c	TEG module(s) heat emitted on its cold-side in watt
Q_h	TEC module(s) heat emitted on its hot-side in watt
Q_h	TEG module(s) heat absorbed on its hot-side in watt
Qcp_{max}	TEC(s) Icop maximum cooling power in watt
Qc_{max}	TECs maximum absorbable heat in watt, when $\Delta T = 0$°C
Qc_{mid}	TEC(s) midpoint cooling power in watt
Qc_n	TEC(s) normalized cooling power is the ratio of Q_c and Qc_{max}
r	TE device p-n thermocouples unit resistance in ohm
R	TE device (TEG and TEC) module unit resistance in ohm
R_L	TEGs electrical load resistance in Ω connected to the TEG(s)
R_s	Power source resistance in ohm connected to the TECs
R_t	TEG/TEC module(s) total resistance in ohms
S	TE device Seebeck coefficient in V/K
Se	TEG(s)/TEC(s) effective Seebeck coefficient in V/K
$\bar{T}$	TE device average temperature ($T_h + T_c$)/2 in K or °C
T_c	Temperature on TEG/TEC cold-side in °C
T_h	Temperature on TEG/TEC hot-side in °C
TE	Thermoelectric
TEC	Thermoelectric cooler
$TECsa$	TEC cold-side surface area
TEG	Thermoelectric generator
TEG_{Ioc}	TEG ideal generated current − assuming there is no TEG_{Rtint}
TEG_{PocM}	TEG ideal generated power − assuming there is no TEG_{Rtint}
TEG_{Rtint}	TEG internal resistance (R_t) − responsible for the power loss
$TEGsa$	TEG hot-side surface area
$TEGs\ T_c$	TEGs cold side temperature
$TEGs\ DT$	TEGs temperature difference
$TEGs\ T_h$	TEGs hot side temperature
TEG_{Voc}	TEG ideal generated voltage − assuming there is no TEG_{Rtint}
TEH	Thermoelectric Energy Harvester
T_p	TEGs/TECs module quantity connected in parallel
T_s	TEGs/TECs module quantity connected in series
T_t	TEG/TEC modules total quantity connected
V_{comv_out}	TEGs booster converter output voltage
Vin	TEC module(s) input voltage in volt
Vin_{max}	TEC's max V_{in} in (V) that produces max ΔT_{max} when $I_{in}=I_{max}$
Vin_n	TEC(s) normalized input voltage is the ratio of V_{in} and Vin_{max}
Vo	TEG module(s) output voltage in volt
Vo_{max}	TEG(s) maximum output voltage in volt
Vn	TEG(s) normalized output voltage
V_{TEG_Out}	TEGs generated voltage (input voltage to the boost converter)
$VTEG_{Rtint}$	TEG generated voltage drop − due to TEG internal resistance
WSN	Wireless Sensors Network
Z	TE device figure of merit in per K
Ze	TEG(s)/TEC(s) effective figure of merit in per K

$Z\bar{T}$ TE device average dimensionless figure of merit

Acknowledgment

This work was supported in parts by the Cape Peninsula University of Technology CPGS and the University of the Western Cape HySA Systems.

Data availability

Research in progress − data available upon completion.

Conflict of Interest

Authors declare no conflict of interest.

References

[1] M.L. van der Walt, J. van den Berg, M. Cameron, "State of Renewable Energy in South Africa", South Africa Department of Energy, Pretoria, 2017, http://www.energy.gov.za/files/media/Pub/2017-State-of-Reewable-Energy-in-South-Africa.pdf.

[2] N.P. Bayendang, M.T. Kahn, V. Balyan, "A structural review of thermoelectricity for fuel cells CCHP applications," Hindawi Journal of Energy 2020, 1-23, 2020, https://doi.org/10.1155/2020/2760140/

[3] Y.-S. Noh, J.-I. Seo, W.-J. Choi, J.-H. Kim, H.V. Phuoc, H.-S. Kim, S.-G. Lee, "17.6 A Re-configurable DC-DC Converter for Maximum TEG Energy Harvesting in a Battery-Powered Wireless Sensor Node," 2021 IEEE International Solid-State Circuits Conference (ISSCC),San Francisco, CA, USA, 266-268, 2021, doi: 10.1109/ISSCC42613.2021.9365811.

[4] D. Charris, D. Gómez, M. Pardo, "A Portable Thermoelectric Energy Harvesting Unit to Power Up Outdoor Sensors and Devices," 2019 IEEE Sensors Applications Symposium (SAS), Sophia Antipolis, France, 1-6, 2019, doi: 10.1109/SAS.2019.8705985.

[5] J. Singh, P. Kuchroo, H. Bhatia, E. Sidhu, "Floating TEG based solar energy harvesting system," 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Pune, India, 763-766, 2016, doi: 10.1109/ICACDOT.2016.7877689.

[6] Q. Wan, Y. Teh, P.K.T. Mok, "Analysis of a Re-configurable TEG Array for High Efficiency Thermoelectric Energy Harvesting," 2016 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), Jeju, Korea (South), 662-665, 2016, doi: 10.1109/APCCAS.2016.7804084.

[7] R. Chein, G. Huang, "Thermoelectric cooler application in electronic cooling", Applied Thermal Engineering, **24**(14–15), 2207-2217, 2004, https://doi.org/10.1016/j.applthermaleng.2004.03.001.

[8] I.R. Belovski, A.T. Aleksandrov, "Examination of the Characteristics of a Thermoelectric Cooler in Cascade," 2019 X National Conference with International Participation (ELECTRONICA), 1-4, 2019, doi: 10.1109/ELECTRONICA.2019.8825631.

[9] H. Lee, Thermoelectrics: design and materials, John Wiley & Sons, Inc., Wiley, New Jersey, USA, 2016.

[10] H. Mamur, Y. Çoban, "Detailed Modeling of a Thermoelectric Generator for Maximum Power Point Tracking", Turkish Journal of Electrical Engineering & Computer Sciences, **28**, 124–139, 2020, https://doi.org/10.3906/elk-1907-166.

[11] N.P. Bayendang, M.T. Kahn, V. Balyan, I. Draganov, S. Pasupathi, "A Comprehensive Thermoelectric Generator (TEG) Modelling", AIUE Congress 2020: Energy and Human Habitat Conference, Cape Town, South Africa, 1-7, 2020; Publisher Zenodo: Geneva, Switzerland, Available online 2020, http://doi.org/10.5281/zenodo.4289574.

[12] N.P. Bayendang, M.T. Kahn, V. Balyan, I. Draganov, S. Pasupathi, "A Comprehensive Thermoelectric Cooler (TEC) Modelling", AIUE Congress 2020: International Conference on Use of Energy, Cape Town, South Africa, 1-7, 2020; Publisher SSRN: Rochester, NY, USA, Available online 2021, https://ssrn.com/abstract=3735378 or http://dx.doi.org/10.2139/ssrn.3735378.

[13] F. Felgner, L. Exel, M. Nesarajah, G. Frey, "Component-oriented modeling of thermoelectric devices for energy system design," in IEEE Transactions on Industrial Electronics, **61**(3), 1301-1310, 2014, doi: 10.1109/TIE.2013.2261037.

[14] C. Luo, R. Wang, W. Yu, W. Zhou, "Parametric study of a thermoelectric module used for both power generation and cooling," Renewable Energy, **154**, 542-552, 2020, https://doi.org/10.1016/j.renene.2020.03.045.

[15] C. Liu, P. Chen, K. Li. "A 1kW Thermoelectric Generator for Low-temperature Geothermal Resources," PROCEEDINGS, 39th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California, USA, 24-26, 2014, SGP-TR-202. https://pangea.stanford.edu/ERE/pdf/IGAstandard/SGW/2014/Li.pdf [Date accessed: August 3, 2021].

[16] S.O. Giwa, C.N. Nwaokocha, A.T. Layeni, O.O. Olaluwoye, "Energy harvesting from household heat sources using a thermoelectric generator module," Nigerian Journal of Technological Development, **16**(3), 2019, http://dx.doi.org/10.4314/njtd.v16i3.6.

[17] M.W. Aljibory, H.T. Hashim, W.N. Abbas, "A Review of Solar Energy Harvesting Utilizing a Photovoltaic–Thermoelectric Integrated Hybrid System," IOP Conference Series: Materials Science and Engineering, 4th International Conference on Engineering Sciences (ICES 2020), 1067, 2021, Kerbala, Iraq, doi:10.1088/1757-899X/1067/1/012115.

[18] X. Hu, C. Jiang, X. Fan, B. Feng, P. Liu, Y. Zhang, R. Li, Z. He, G. Li, Y. Li, "Investigation on waste heat recovery of a nearly kilowatt class thermoelectric generation system mainly based on radiation heat transfer," Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 2020, https://doi.org/10.1080/15567036.2020.1829190.

[19] M.Y. Fauzan, S.M. Muyeen, S. Islam, "Experimental modelling of grid-tied thermoelectric generator from incinerator waste heat," International Journal of Smart Grid and Clean Energy, 2020, doi: 10.12720/sgce.9.2.304-313.

[20] M. Ebrahimi, E. Derakhshan, "Design and evaluation of a micro combined cooling heating and power system based on polymer exchange membrane fuel cell and thermoelectric cooler", Energy Conversion and Management, 171, 507-517, 2018, https://doi.org/10.1016/j.enconman.2018.06.007.

[21] A.G. Rösch, A. Gall, S. Aslan, M. Hecht, L. Franke, M.M. Mallick, L. Penth, D. Bahro, D. Friderich, U. Lemmer, "Fully printed origami thermoelectric generators for energy-harvesting," NPJ Flex Electrons, **5**(1), 2021, https://doi.org/10.1038/s41528-020-00098-1.

[22] V.B. Abhijith, K. Narayanaswamy, C.M. Pooja, S.V. Prasad, R. Sambhu, "Household Application of Thermoelectric Generator in the Field of Propagating Power from Waste Heat," AIP Conference Proceedings 2297, 020010, 2020, https://doi.org/10.1063/5.0031699.

[23] F. Afshari, "Experimental and numerical investigation on thermoelectric coolers for comparing air-to-water to air-to-air refrigerators," Journal of Thermal Analysis and Calorimetry, **144**, 855–868, 2021, https://doi.org/10.1007/s10973-020-09500-6.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 2, 63-72 (2022)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Stability Analysis of a DC Microgrid with Constant Power Load

Sarah Ansari*, Kamran Iqbal

University of Arkansas, Little Rock, Department of Systems Engineering, Little Rock, 72204, USA

ARTICLE INFO

Article history:
Received: 14 November, 2021
Accepted: 20 February, 2022
Online: 18 March, 2022

Keywords:
DCMG
Constant Power Load
PI Controller
Buck Converter
Cascaded Network

ABSTRACT

DC Microgrids (DCMGs) aggregate and integrate various distribution generation (DG) units through the use of power electronic converters (PECs) that are present on both the source side and the load side of the DCMGs. Tightly regulated PECs at the load side behave as constant power loads (CPLs) and may promote instability in the entire DCMG. Previous research has mostly focused on devising stabilization techniques with ideals CPLs that may not be feasible to realize; few publications that emulate DCMG stability with practical CPLs are restricted in application because they add components that considerably increase the cost of the DCMGs. This study aims at stabilizing the DCMG in the presence of practical CPL in a way that is economically feasible, i.e., without the addition of complex compensators. This paper presents a Simulink model of the smallest DCMG, i.e., a cascaded DC-DC power converter network with a practical CPL assumed at the load side of the network. Using theoretical calculations and computer simulations, we have determined the suitable CPL power level and the bandwidth of the current controller at which the smallest DCMG is stable. We have performed the stability analysis of the source side buck converter and the CPL with the derived power level and bandwidth, and found that individual converter systems are stable, thereby proving that the entire DCMG is stable despite the presence of a CPL.

1. Introduction

In recent times, the demand for DCMGs is surging. With this, there are significant issues related to the distribution networks in the power systems. The preliminary analysis and results of one such issue i.e., caused due to CPL is done in [1]. There are other associated problems like voltage fluctuations, there is a need to aggregate DG units and provide proper coordination. Thus, the development of microgrids becomes indispensable to integrate and coordinate different power systems. US department of energy, DoE defines microgrids as "Locally confined and independently controlled electric power grids in which distribution architecture integrates loads and distributed energy resources which allows the microgrid to operate connected or isolated to a main grid" [2].

While microgrids can be developed for both AC and DC supplies, DCMGs are considered superior to the AC microgrids due to several factors. The DC networks sidestep reactive power issues, which simplifies the control loop design [3]. It also results in reduced hardware (power cables), thereby reducing the overall equipment cost. Further, DCMG implementation eliminates long transmission and distribution lines that aids in providing reliable and efficient DG system [4]. Also, in recent times, integration of renewable energy sources, fuel cells, and energy storage devices with conventional power systems has become indispensable. The urgency of these issues has brought DC power systems back into picture through DCMGs. DCMGs consist of power electronic elements that are used for various purposes. For example, they can be used to isolate the microgrid from the main power system, or to make a network of distributed generation systems that need to be synchronized. These are termed as multi-converter power electronic systems [5, 6] that employ power converters to control various grid parameters like voltage, current, power, etc.

A DC distribution system has two broad stages as shown in Figure 1 [7]. The first stage consists of two or more converters that are connected in parallel and feeding the DC bus [7]. These are switched mode power supplies (SMPS1) called line regulating converters (LRC) or source side converters. This converter system feeds the DC bus with a regulated voltage; the bus is further connected to another set of switched mode power supplies (SMPS2) called point of load (POL) converters, or load side converters. It has been shown that tightly regulated POLs behave as constant power loads (CPLs). Theoretically, the power supplied to the CPL equals the product of output voltage of the CPL and the

*Corresponding Author: Sarah Ansari, sxansari@ualr.edu

current flowing through it. When the power supplied in a CPL is constant, then the voltage varies inversely with respect to the current change. Thus, the voltage increases when the current decreases and vice versa thereby resulting in negative incremental impedance. This concludes that the constant power loads exhibit a non-linear phenomenon that causes instability in DCMGs. Moreover, solving the stability issue becomes challenging when at least two power converters are cascaded to each other. Previous literature has considered load side power converters to be behaving as ideal CPLs. As a result, study of power levels and dynamic performance of CPL and how they affect system stability has been mostly neglected. Hence, it becomes important to investigate the system stability and evaluate the technical restrictions of CPL.

This paper seeks to study what technical restrictions can be levied on CPL to ensure stability of the DCMG. To do the analysis, a simulation scheme of source side buck converter and CPL is designed. The choice of this buck topology has been reinforced by two main reasons: i) Buck topology has simple construction and dynamic performance, and ii) It has higher system stability than boost or buck-boost topology. The rest of this paper is organized as follows: sections II and III discusses the design and design of source side buck converter. Sections IV and V discusses the stability analysis and design of the CPL. Section VI discusses the cascading of the source side buck converter and CPL to form the smallest DCMG. Appendix and section VII show the simulation models developed and their corresponding results respectively. The conclusion is mentioned in section VIII.

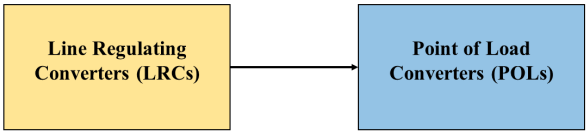


Figure 1: Major Components of DC distribution system

2. Controller Design for Source Side Buck Converter

In the research literature, linear droop control is realized through a virtual or an actual resistor in series with the DC-DC converters that are modeled as voltage sources. While droop control is a practical and viable voltage control scheme to regulate a constant DC voltage supply, it may work for buck converter topology [7]. Whereas the equivalent circuit of a converter in other topologies consists of transformers with nonlinear turn ratios, this will hinder the use of linear droop control for such converters [8]. Thus, to implement linear droop scheme, the voltage source is modeled using the DC-DC buck converter topology. Moreover, a PI controller is designed as a fast controller for the current flow through the power converters. Both controllers are proposed in this section and integrated with DC-DC buck converters to analyze their dynamic performance.

A buck converter is a power converter that steps down the DC voltage from higher input to lower output value. A buck converter with the predefined parameters is shown Figure 2. The stepping down is governed by an adjustable duty ratio which is realized by designing a suitable controller for a given buck converter.

The source side buck converter has two controllers: the voltage controller and the current controller. The load side buck converter or the CPL has a current controller and a constant power supply. This section discusses the controller design of the source side buck converter.

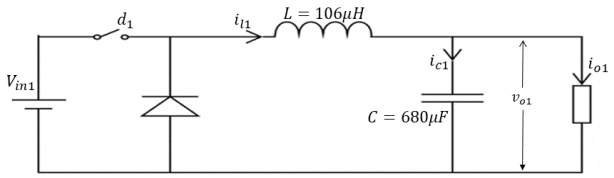


Figure 2: Schematic diagram of the source side buck converter

From the control point of view, the buck converter is considered as a power stage (shown in Figure 3), which is controlled by a PI controller. Figure 3 shows the complete control model for the buck converter with specified parameters. The objective is to design a controller to achieve the desired output voltage of 120V from an input supply of 140V. This can be done by controlling the voltage and current flowing through the power stage. Thus, the goal is to design voltage and current controllers (shown in Figure 4). In the diagram, the voltage controller compares the output voltage with the setting voltage V_set1 (120V). Using the droop control governed by droop characteristic shown in Figure 5, the setting value of load current 〖i_l1〗^*can be obtained which is then input to the current controller. The current controller controls an inner loop consisting of a PI controller and the power stage as shown in Figure 6.

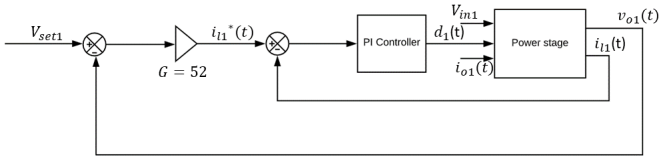


Figure 3: Control model of the source side buck converter

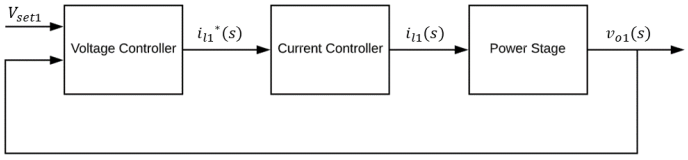


Figure 4: Voltage and current controllers for the source side buck converter

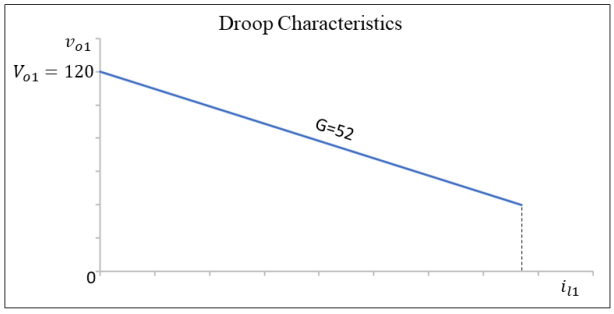


Figure 5: Droop characteristic with a droop gain of 52

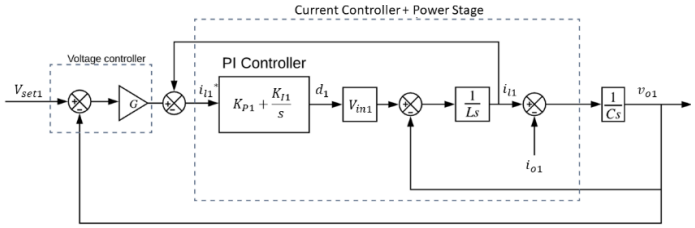


Figure 6: Closed loop current and voltage control scheme for the source side buck converter

The PI controller parameters include the proportional gain K_{P1} and the integral gain K_{I1}, which need to be determined. For design purposes, the open loop transfer function for the current control is (from Figure 6):

$$H_{ol1} = \frac{(K_{P1}s+K_{I1})V_{in1}-v_{o1}}{Ls^2} \quad (1)$$

Since the dynamic change in inductor current is faster than that of the voltage across the capacitor, thus v_{o1} can be considered as a disturbance and can be neglected. Then, the modified open loop transfer function becomes

$$H_{ol1} = \frac{(K_{P1}s+K_{I1})V_{in1}}{Ls^2} \quad (2)$$

The closed loop transfer function comprising of PI controller, power stage and unity feedback is

$$H_{cl1} = \frac{H_{ol1}}{1+H_{ol1}} = \frac{(K_{P1}s+K_{I1})}{\left({}^{L}/_{V_{in1}}\right)s^2+K_{P1}s+K_{I1}} \quad (3)$$

3. Stability Analysis of Source Side Converter

The model of the source side buck converter was implemented in Simulink. The stability analysis of the source side buck converter was undertaken in two distinct ways, as described below:

Approach 1: the current loop design

In this approach the current loop is analyzed and the current controller consisting of the power stage and the PI controller is considered. The complete model is shown in Figure 7.

The transfer function of the PI controller $G1_{c1}(s)$ is:

$$G1_{c1}(s) = K_{P1} + \frac{K_{I1}}{s} = K_{P1}\left(\frac{s+\left({}^{K_{I1}}/_{K_{P1}}\right)}{s}\right) \quad (4)$$

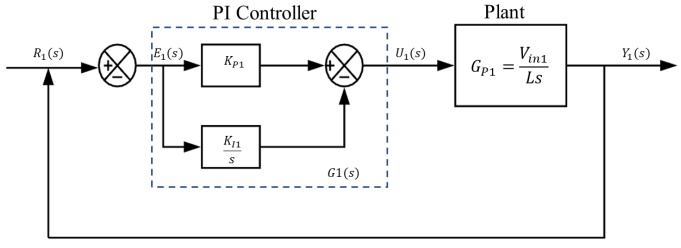


Figure 7: Control system implementation of the source side buck converter

The open-loop transfer function is

$$G1_{ol1}(s) = G1_{P1}(s)G1_{c1}(s) = \frac{V_{in1}K_{P1}\left(s+\left({}^{K_{I1}}/_{K_{P1}}\right)\right)}{Ls^2} \quad (5)$$

Let ${}^{K_{I1}}/_{K_{P1}} = K1$, then Equation (5) becomes

$$G1_{ol1}(s) = \frac{V_{in1}K_{P1}(s+K1)}{Ls^2} \quad (6)$$

The characteristic equation of the closed-loop system is given as $1 + G1_{ol1}(s)$, where

$$s^2 + \frac{V_{in1}K_{P1}}{L}s + \frac{V_{in1}K_{I1}}{L} = 0 \quad (7)$$

The closed-loop transfer function is

$$G1_{cl1} = \frac{Y_1(s)}{R_1(s)} = \frac{\left({}^{V_{in1}}/_{L}\right)(K_{P1}s+K_{I1})}{s^2+\left({}^{V_{in1}K_{P1}}/_{L}\right)s+\left({}^{V_{in1}K_{I1}}/_{L}\right)} \quad (8)$$

Approach 2: voltage and current loop design

In this approach, the entire system including voltage and current controllers is considered. The controller structure as shown in Figure 6 will be considered for the analysis.

The open-loop transfer function system for the design of PI controller is derived from Matlab:

$$G1_{c2}(s) = \frac{1.32\times10^6 s+1.011\times10^{11}}{s^2+1.387\times10^7} \quad (9)$$

The loop transfer function with PI controller in the loop is (where, $K1 = {}^{K_{I1}}/_{K_{P1}}$)

$$G1_{ol2}(s) = \frac{1.32\times10^6 s+1.011\times10^{11}}{s^2+1.387\times10^7} \times \left(\frac{K_{P1}(s+K1)}{s}\right) \quad (10)$$

The stability analysis and design of both approaches is performed using Root locus method, Routh Hurwitz criterion, and Nyquist criterion.

3.1. Root Locus Method

The root locus-based design aims to find suitable values for the proportional and integral gains.

Approach 1: The open-loop transfer function (Equation 6) of the current controller is studied by varying gain K_{P1}. The characteristic equation of (6) is

$$s^2 + \frac{V_{in1}K_{P1}}{L}s + \frac{V_{in1}K_{I1}}{L} = 0 \quad (11)$$

The resulting root loci of Equation (5) with ${}^{K_{I1}}/_{K_{P1}} = 2\times10^5$ are shown in Figure 8. From the root loci plot, when gain $K_{P1} = 0.568$, the damping ratio is 0.968, which is considered reasonable for the converter. The corresponding value of $K_{I1} = 113600$. The characteristic equation has two complex roots (also shown in Figure 8) at:

$s = -3.75 \times 10^5 + j9.69 \times 10^4$ and $s = -3.75 \times 10^5 - j9.69 \times 10^4$ (12)

These are the poles of the closed loop system. Since, these poles are located in the LHP, the closed loop system is stable with reasonable damping.

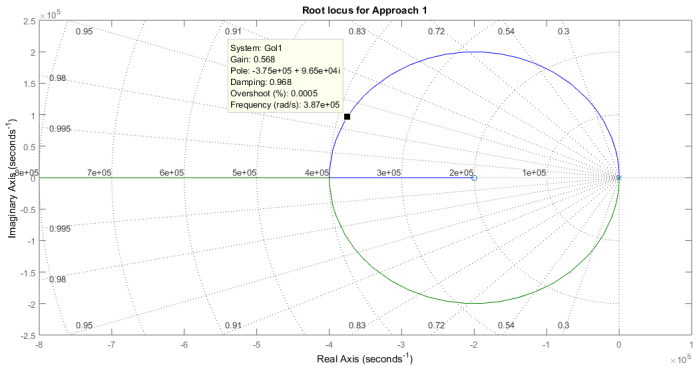


Figure 8: Root Loci of equation (11) with ${K_{I1}}/{K_{P1}} = 2 \times 10^5$; K_{P1} varies

Approach 2: In this case, similarly, the gain ratio ${K_{I1}}/{K_{P1}} = 2 \times 10^5$ is considered. Equation (10) gives the open-loop transfer function of the system with voltage and current controllers. The root loci of (10) are shown in Figure 9. Clearly, the poles of the closed-loop system are located in the LHP. Thus, the closed loop system is stable.

The closed loop transfer function with $K_{P1} = 0.568$ *and* $K_{I1} = 113600$ is

$$G1_{cl2}(s) = \frac{7.498\times10^5 s^2 + 2.073\times10^{11} s + 1.147\times10^{16}}{s^3 + 7.498\times10^5 s^2 + 2.073\times10^{11} s + 1.147\times10^{16}} \quad (13)$$

Thus, the closed loop system is stable.

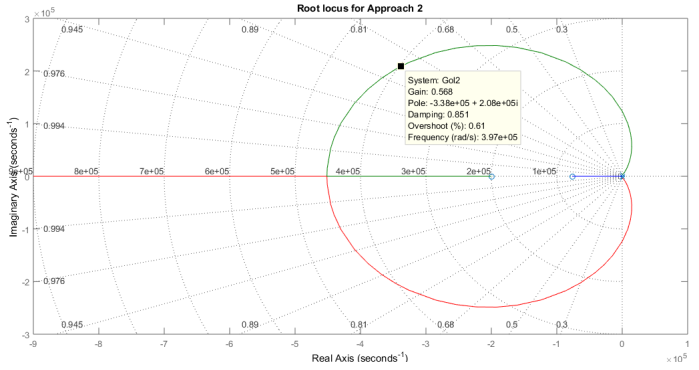


Figure 9: Root Loci of equation (10) with ${K_{I2}}/{K_{P2}} = 2 \times 10^5$; K_{P2} varies

3.2. Routh-Hurwitz Criterion

The Routh-Hurwitz criterion algebraically ascertains the stability requirements for a linear time-invariant (LTI) system modeled with constant coefficients. The criterion tests whether any roots of the characteristic equation lie in the right half *s*-plane.

Approach 1: The characteristic equation of the closed-loop system is given as $1 + G1_{ol1}(s)$ and is

$$s^2 + \frac{V_{in1}K_{P1}}{L}s + \frac{V_{in1}K_{I1}}{L} = 0 \quad (14)$$

Applying the Routh Hurwitz's stability criterion to equation (7) yields that the system is stable for $K_{P1} > 0$ and $K_{I1} > 0$. Thus, the chosen parameter values of $K_{P1} = 0.568$ *and* $K_{I1} = 113600$ stabilize the system.

Approach 2: The characteristic equation of the closed-loop system is

$$s^3 + 7.5 \times 10^5 s^2 + 2.1 \times 10^{11} s + 1.1 \times 10^{16} = 0 \quad (15)$$

Then, for the closed-loop system to be stable, it should meet the following constraints:

$$7.5 \times 10^5 \times 2.1 \times 10^{11} > 1.1 \times 10^{16}$$

$$1.5 \times 10^{17} > 1.1 \times 10^{16} \quad (16)$$

The above design satisfies these constraints; hence, the system is stable.

3.3. Nyquist Criterion

The Nyquist criterion graphically reveals information about the number of poles and zeroes of the closed-loop transfer function that are in the right half *s*-plane. The Nyquist criterion is applied to the two design approaches as follows.

Approach 1: The Nyquist plot of the open loop transfer function $G1_{ol1}(s)$ (from equation (6)) with $K_{P1} = 0.568$ *and* $K_{I1} = 113600$ is shown in Figure 10, where

$$G1_{ol1}(s) = \frac{79.52s + 1.59\times10^7}{106\times10^{-6}s^2} \quad (17)$$

For a minimum phase transfer function, the closed-loop system is stable if there are no encirclements of the critical point $(-1 + j0)$. From Figure 10, since there are no encirclements of the critical point, thus the closed-loop system is stable. This result is also verified by Matlab.

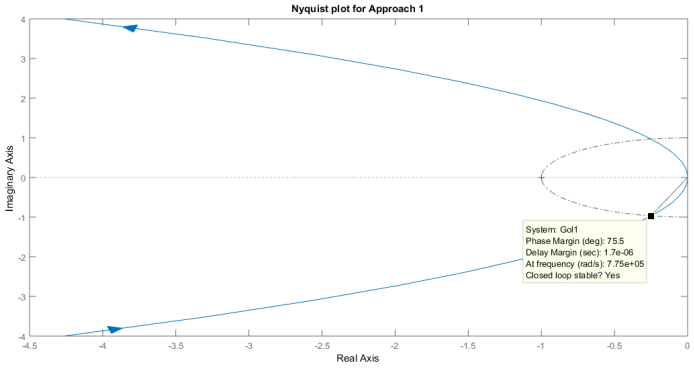


Figure 10: Nyquist plot for Approach 1

Approach 2: The Nyquist plot of the open loop transfer function $G1_{ol2}(s)$ (from (10)) with $K_{P1} = 0.568$ *and* $K_{I1} = 113600$ is shown in Figure 11, where the loop transfer function is given as

$$G1_{ol2}(s) = \frac{7.498\times10^{5}s^{2}+2.073\times10^{11}s+1.147\times10^{16}}{s^{3}+1.387\times10^{7}s} \quad (18)$$

From Figure 11, there are no encirclements of the critical point, hence the closed-loop system is stable. This result is also verified by Matlab.

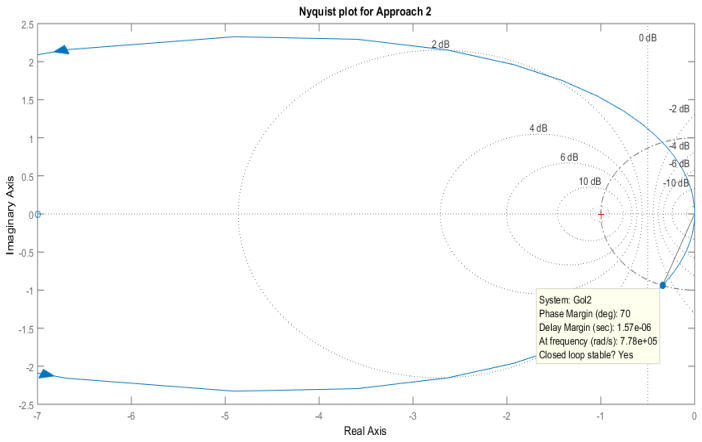


Figure 11: Nyquist plot for Approach 2

Based on the above stability criteria, the values of K_P1 and K_I1 (reported in Table 1 below) stabilize the source side buck converter. The simulation model and corresponding results are shown in appendix and section VII respectively.

Table 1: Values of K_{P1} and K_{I1} for the given source side buck converter

K_{P1}	0.568
K_{I1}	113600

4. Design of CPL

Buck converters can be emulated as instantaneous constant power loads when cascaded with at least one source side DC-DC power converters. For the study, one source side buck converter is considered, and its controller design is proposed in the previous section. Figure 12 shows the control model for a power stage (here buck converter) emulated as CPL.

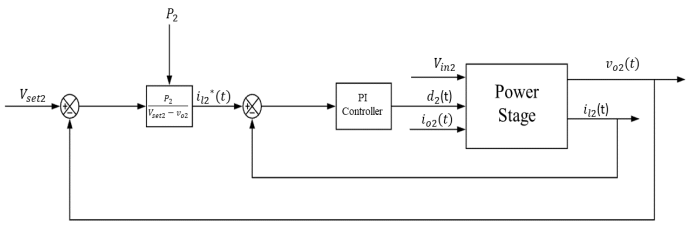


Figure 12: Control Model of CPL

Also, the power stage is supplied with a constant supply of power P_2 which characterizes the non-linear nature of the CPL. It thus becomes important to linearize the system about an equilibrium point.

4.1. Linearization of Load Side Converter (CPL)

From figure 12 the relationship between setting value of inductor current $i_{l2}{}^*$ and incoming voltage of the CPL is non-linear and is given as,

$$i_{l2}{}^* = \frac{P_2}{v_{o2}} \quad (19)$$

Here, V_{set2} is not considered because the component is added to the simulation model to the closed loop system. Theoretical analysis of the CPL that involves linearization of CPL and its related calculation is based on the open loop circuitry of the CPL which does not have V_{set2}. Each of the parameters in Figure 12 can also be represented as the sum of steady state value at equilibrium point and the small signal perturbation around the equilibrium as shown in equation 20.

$$\left.\begin{aligned} i_{l2}{}^* &= I_{l2}{}^* + \tilde{\imath}_{l2}{}^* \\ v_{o2} &= V_{o2} + \tilde{v}_{o2} \\ P_2 &= P_2 + \tilde{p}_2 \end{aligned}\right\} \quad (20)$$

Using Taylor series as explained in [9], the linearized equation is

$$i_{l2}{}^* = \frac{1}{V_{o2}}\tilde{p}_2 - \frac{P_2}{{V_{o2}}^2}\tilde{v}_{o2} \quad (21)$$

Hence, the linearized model of CPL is shown in Figure 13.

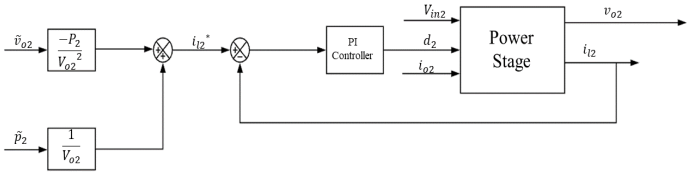


Figure 13: Linearized version of CPL

4.2. PI Control for Linearized CPL (Load Side Buck Converter)

The PI controller parameters namely proportional gain K_{P2} and integral gain K_{I2} need to be determined. The open loop transfer function for the current control is (from Figure 14):

$$H_{ol2} = \frac{(K_{P2}s+K_{I2})V_{in2}-v_{o2}}{Ls^2} \quad (22)$$

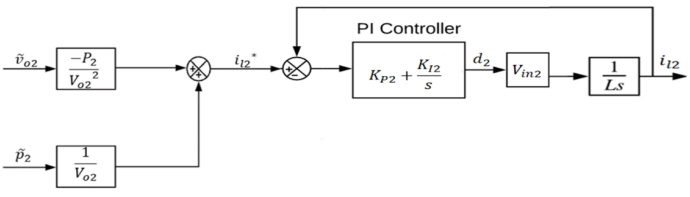


Figure 14: PI control for linearized CPL

Since the dynamic change in inductor current is faster than that of the voltage across the capacitor, thus v_{o2} can be considered

as a disturbance and can be neglected. Thus, the modified open loop transfer function becomes

$$H_{ol2} = \frac{(K_{P2}s+K_{I2})V_{in2}}{Ls^2} \quad (23)$$

The closed loop transfer function comprising of PI controller, power stage and unity feedback is

$$H_{cl2} = \frac{H_{ol2}}{1+H_{ol2}} = \frac{(K_{P2}s+k_{I2})}{\left({}^{L}/_{V_{in2}}\right)s^2+K_{P2}s+K_{I2}} \quad (24)$$

Equation (24) is the closed loop transfer function of the current controller which is the PI controller and the power stage.

4.3. Power and Bandwidth of CPL

The power stage of the CPL used in this study is the same as that of the source side buck converter. The function of the CPL is to step down the voltage from 120V to 100V. In order to design a PI controller for such a CPL, we have assumed the values of $K_{P2} = K_{P1}$ and $K_{I2} = K_{I1}$. This is done, due to two reasons:

1. In practical DCMGs, it becomes favorable to have similar current controllers for the source side converter and CPL, as it reduces the complexity of the cascaded network.
2. By doing so, the dynamic behavior of both the converters can be compared in order to better understand the working of the CPL.

Thus, $K_{P2} = 0.568\ and\ K_{I2} = 113600$.

Since in a CPL, the power supplied is constant, thus $\tilde{p}_2 = 0$. Thus, equation (21) is modified and is given as,

$$i_{l2}{}^{*} = -\frac{P_2}{{V_{o2}}^2}\tilde{v}_{o2} \quad (25)$$

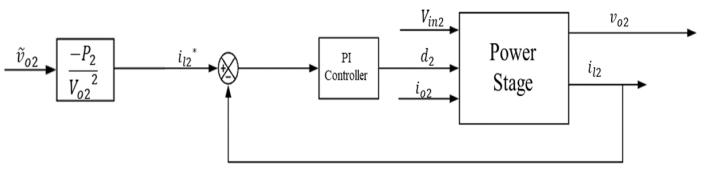


Figure 15: Linearized CPL, ignoring $\frac{1}{V_{o2}}\tilde{p}_2$, since $\tilde{p}_2 = 0$

Notice, in Figure 15, the value of the incremental impedance is positive as now the CPL emulates a resistive load, thereby ensuring stability to the DCMG, by keeping its property intact.

Now,

$$I_{l2}{}^{*} = \frac{P_2}{{V_{o2}}^2} \quad (26)$$

Since, the parameters of the source side buck converter and that of the CPL is considered the same, thus the droop control of the source side buck converter is analogous to the $\frac{P_2}{{V_{o2}}^2}$ factor. Thus, assuming $\frac{P_2}{{V_{o2}}^2} = G = 52$ and V_{o2} is the desired output voltage of CPL, which is 100V, thus, we get

$$P_2 = 52\ X\ {V_{o2}}^2 = 52\ X\ (100)^2 = 520kW \quad (27)$$

Using this value of power (derived in Equation (26)), we have done the stability analysis of the CPL to verify that at $P_2 = 520kW$ the CPL is stable.

5. Stability Analysis of CPL (Load Side Converter)

Stability analysis of the Simulink model of the buck converter is similarly done in two distinct ways, as described below:

Approach 1: the current loop design

In this approach the current controller consisting of the power stage and PI controller is considered. The complete model is shown in Figure 16. The transfer function of the PI controller $G2_{c1}(s)$ is:

$$G2_{c1}(s) = K_{P2} + \frac{K_{I2}}{s} = K_{P2}\left(\frac{s+\left({}^{K_{I2}}/_{K_{P2}}\right)}{s}\right) \quad (28)$$

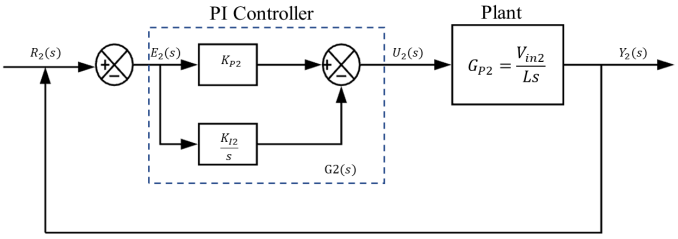


Figure 16: Control system of the CPL

The forward-path transfer function of the feedback control system is

$$G2_{ol1}(s) = G2_{P2}(s)G2_{c1}(s) = \frac{V_{in2}K_{P2}\left(s+\left({}^{K_{I2}}/_{K_{P2}}\right)\right)}{Ls^2} \quad (29)$$

Let ${}^{K_{I2}}/_{K_{P2}} = K2$, thus Eq (29) becomes

$$G2_{ol1}(s) = \frac{V_{in2}K_{P2}(s+K2)}{Ls^2} \quad (30)$$

The characteristic equation of the closed-loop system as given by $1 + G2_{ol1}(s)$ is

$$s^2 + \frac{V_{in2}K_{P2}}{L}s + \frac{V_{in2}K_{I2}}{L} = 0 \quad (31)$$

The closed-loop function is

$$G2_{cl1} = \frac{Y_2(s)}{R_2(s)} = \frac{\left({}^{V_{in2}}/_{L}\right)(K_{P2}s+K_{I2})}{s^2+\left({}^{V_{in2}K_{P2}}/_{L}\right)s+\left({}^{V_{in2}K_{I2}}/_{L}\right)} \quad (32)$$

Approach 2: Current and voltage loop designs

In this approach, the entire system with linearized CPL (having $\frac{P_2}{{V_{o2}}^2}$) and current controller (shown in Figure 15) is considered. The open loop transfer function system with analysis point as the PI controller, is derived from Matlab:

$$G2_{c2}(s) = \frac{7.208\times10^{-4}s^3+979.2s^2+5.396\times10^{9}s+1.04\times10^{15}}{7.208\times10^{-4}s^3+s} \quad (33)$$

The open loop transfer function of the system with PI controller is given below, where $K2 = {K_{I2}}/{K_{P2}}$

$$G2_{ol2}(s) = \frac{7.208\times10^{-4}s^3+979.2s^2+5.396\times10^{9}s+1.04\times10^{15}}{7.208\times10^{-4}s^3+s} \times \left(\frac{K_{P2}(s+K2)}{s}\right) \quad (34)$$

Three stability analysis criteria are employed toward controller design. These include: The Root Locus method, the Routh Hurwitz criterion and the Nyquist criterion.

5.1. Root Locus Method

The root locus method is aimed to find suitable values for the proportional and integral gains.

Approach 1: Equation (30) gives the open loop transfer function of the current controller with varying gain K_{P2}. The closed-loop characteristic equation for (30) is

$$s^2 + \frac{V_{in2}K_{P2}}{L}s + \frac{V_{in2}K_{I2}}{L} = 0 \quad (35)$$

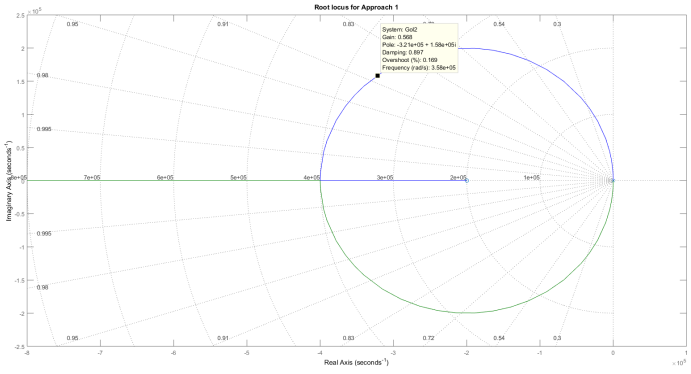


Figure 17: Root Loci of equation (35) with ${K_{I2}}/{K_{P2}} = 2\times10^5$; K_{P2} varies

The root loci of (30) with ${K_{I2}}/{K_{P2}} = 2\times10^5$ are shown in Figure 17. From the root loci, when gain $K_{P2} = 0.568$, the damping ratio is 0.897, which is considered reasonable for the converter. Consequently, the value of $K_{I2} = 113600$ is selected. The two characteristic equation roots are (shown in Figure 17) are at

$$s_1 = -3.21\times10^5 + j1.59\times10^5 \quad \text{and} \quad s_2 = -3.21\times10^5 - j1.59\times10^5 \quad (36)$$

Since these closed-loop poles lie in the LHP, the closed loop system is stable.

Approach 2: In this case, similarly, the design ratio ${K_{I2}}/{K_{P2}} = 2\times10^5$ is considered. Equation (34) gives the open loop transfer function of the linearized CPL with the current controller. The root loci of (34) are shown in Figure 18. From the graph, the poles of the closed-loop system lie on the LHP. The closed loop transfer function with $K_{P2} = 0.568$ *and* $K_{I2} = 113600$ is

$$G2_{cl2}(s) = \frac{4.094\times10^{-4}s^4+638.1s^3+3.176\times10^{9}s^2+1.204\times10^{15}s+1.181\times10^{20}}{1.13\times10^{-3}s^4+638.1s^3+3.176\times10^{9}s^2+1.204\times10^{15}s+1.181\times10^{20}} \quad (37)$$

Thus, the closed loop system is stable.

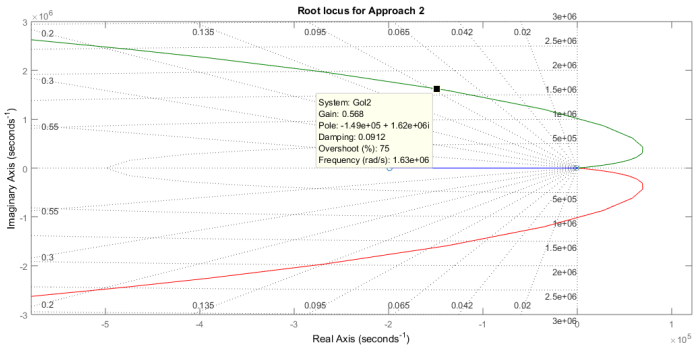


Figure 18: Root Loci of equation (34) with ${K_{I2}}/{K_{P2}} = 2\times10^5$; K_{P2} varies

5.2. Routh-Hurwitz Criterion

Routh-Hurwitz criterion is similarly applied to determine the conditions for closed-loop stability.

Approach 1: The characteristic equation of the closed-loop system, given as $1 + G2_{ol1}(s)$, is

$$s^2 + \frac{V_{in2}K_{P2}}{L}s + \frac{V_{in2}K_{I2}}{L} = 0 \quad (38)$$

Applying the Routh Hurwitz's stability criterion to equation (33) yields the result that the system is stable for $K_{P2} > 0$ and $K_{I2} > 0$. Thus, for the chosen values of $K_{P2} = 0.568$ *and* $K_{I2} = 113600$, the closed-loop system is stable.

Approach 2: The characteristic equation of the closed-loop system is

$$1.1\times10^{-3}s^4 + 638.1s^3 + 3.2\times10^{9}s^2 + 1.2\times10^{15}s + 1.2\times10^{20} = 0 \quad (39)$$

For the system to be stable, each of the diagonal minors $(\Delta_1, \Delta_2, \Delta_3)$ should be zero, i.e.,

$$\Delta_1 = 638.1 > 0$$
$$\Delta_2 = 7.2192\times10^{11} > 0$$
$$\Delta_3 = 8.16\times10^{26} > 0 \quad (40)$$

From the conditions in equation (40), it is evident that for the selected parameter values, each of the diagonal minors are greater than 0, thus the closed-loop system is stable.

5.3. Nyquist Criterion

The Nyquist criterion is similarly applied to ascertain the stability of the closed-loop system.

Approach 1: The Nyquist plot of the open loop transfer function $G2_{ol1}(s)$ (from equation (30)) with $K_{P2} = 0.568$ *and* $K_{I2} = 113600$ is given as

$$G2_{ol1}(s) = \frac{79.52s+1.59\times10^{7}}{106\times10^{-6}s^{2}} \quad (41)$$

For a minimum phase transfer function, the closed-loop system is stable if there no encirclements of the critical point $(-1 + j0)$. Since, from Figure 19, there are no encirclements of the critical point, thus the closed loop system is stable. It is also verified by Matlab.

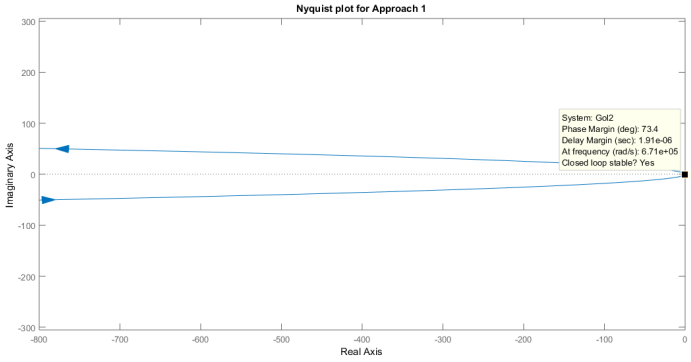


Figure 19: Nyquist plot for Approach 1

Approach 2: The Nyquist plot of the open loop transfer function $G2_{ol2}(s)$ (from (34)) with $K_{P2} = 0.568$ *and* $K_{21} = 113600$ is shown in Figure 20, where the loop transfer function is given as

$$G2_{ol2}(s) = \frac{4.094\times10^{-4}s^{4}+638.1s^{3}+3.176\times10^{9}s^{2}+1.204\times10^{15}s+1.181\times10^{20}}{7.208\times10^{-4}s^{4}+s^{2}} \quad (42)$$

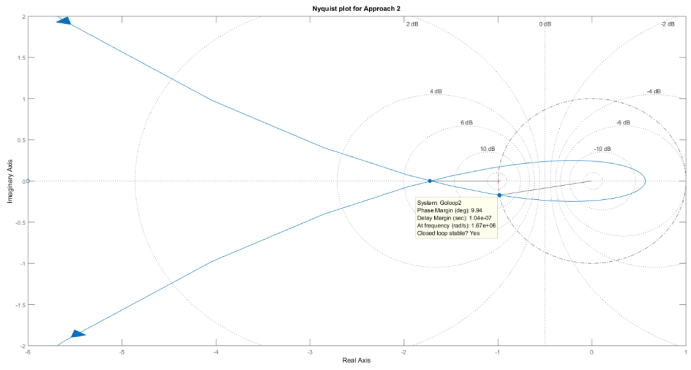


Figure 20: Nyquist plot for Approach 2

The Approach 1 takes into consideration only the current controller and the power stage. The resulting bandwidth BW_2 of the closed loop transfer function is the bandwidth when the CPL is stable. The values of P_2 and BW_2 are shown below in Table 2. The simulation model of CPL is shown in Appendix and the results are shown in section VII.

Table 2: Derived values of P_2 and BW_2 for the CPL

P_2	520kW
BW_2	8.321×10^5rad/sec

6. Cascaded Network of Source Side Converter and CPL

Cascading two power converters means that the load side converter behaves as the load of the source side converter. In other words, the output voltage of source side converter acts as the input voltage of load side converter, and the output inductor current of source side converter is fed as an input to the load side converter [10].

In our study, we have cascaded the source side buck converter with the linearized CPL. Cascading can be done only when both the power converters are independently stable. In our study, we have shown that the source side buck converter and the CPL are independently stable. Thus, they can be cascaded, thereby forming the smallest DCMG.

In our study, the output voltage of the source side buck converter, V_{o1}, is fed as the input voltage to the CPL as V_{in2}. Also, the output inductor current of the source side buck converter, I_{L1} is fed as input current to the CPL, I_{o2}.

Table 3 gives the values of all the parameters of the cascaded power converters, also considered as the smallest DCMG. The simulation model of the cascaded network is shown in Appendix and the results are shown in Sections VII.

Table 3: Values of all the parameters of cascaded source side buck converter and CPL

V_{in1}	140V
I_{o1}	20A
V_{o1}	120V
I_{l1}	20A
$V_{in2} = V_{o1}$	120V
$I_{o2} = I_{l1}$	20A
V_{o2}	100V
I_{l2}	20A
P_2	520kW
BW_2	8.321×10^5rad/sec
L	106µH
C	680µF

7. Simulation Results of DCMG

The Simulink model for the DCMG is shown in Figure 27 (see Appendix). When the model is simulated, the inductor currents of the source side buck converter and CPL will charge their respective capacitors in order to increase the voltage across the capacitor from 0 to 120V (in case of source buck converter) and from 0 to 100V (in case of CPL). When the load current of 20A is supplied, the voltage across the capacitor decreases slightly, resulting in the inductor current exceeding 20A. The output voltage of 120 V (shown in Figure 21) from the source side buck

converter is fed as an input voltage to the CPL. The resulting output voltage of the CPL is 100 V (shown in Figure 23). The output inductor current of 20A (shown in Figure 22) from the source side buck converter is fed as an input current to the CPL. The resulting output inductor current of the CPL also comes out to be 20A (shown in Figure 24). This confirms the cascading of the source side converter with the CPL, to form the smallest DCMG. Clearly, there is no overshoot and no oscillations observed in Figures 21-24.

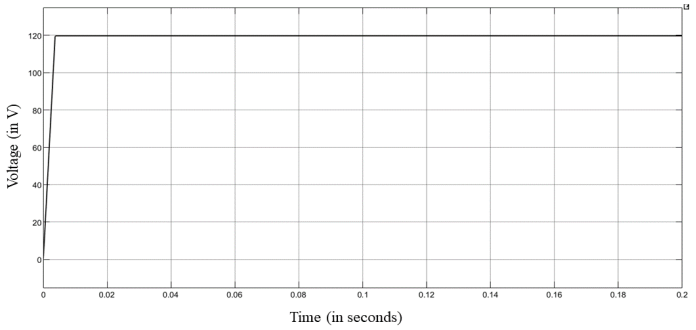


Figure 21: Oscilloscope result of output voltage of source side buck converter

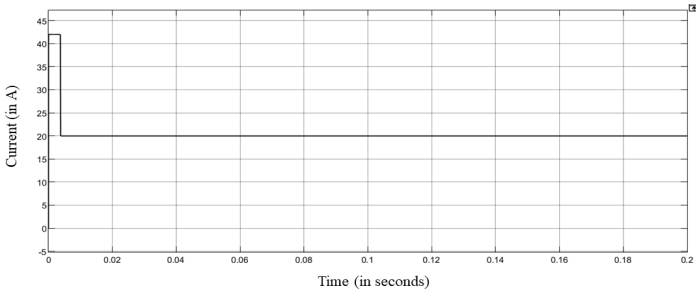


Figure 22: Oscilloscope result of output inductor current of source side buck converter

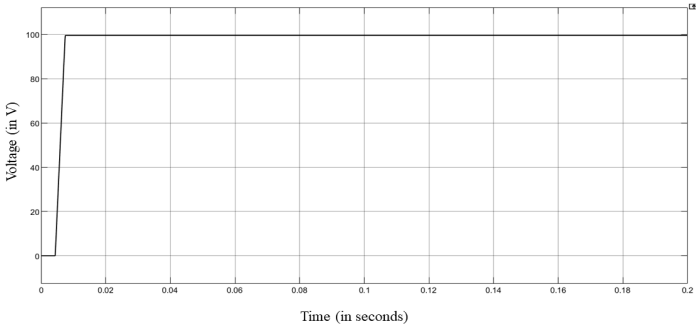


Figure 23: Oscilloscope result of output voltage of CPL

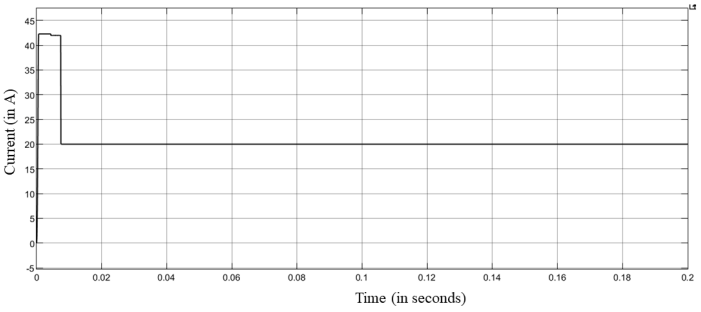


Figure 24: Oscilloscope result of inductor current of CPL

8. Conclusion

This paper discussed the stability analysis of the smallest DCMG that consists of a source side buck converter and a CPL. The cascaded power converters are abundantly found in the DCMGs, and power converters located at the load side that act as CPLs have the potential to cause instability to the entire DCMGs. Thus, it is important to eliminate the instability caused due to CPLs, so that the entire DCMG is stable. Keeping this in mind, the stability analysis of cascaded DC-DC power converters was done. This research proved that despite the presence of CPL, the DCMG can still be made stable. The stability analysis done on the individual components of the cascaded network draws interesting conclusions that support the fact that the DCMG can be stable at certain power level and bandwidth of the CPL controller.

The following are the main results that can be drawn from this research:

1. The CPL, that causes instability to the entire DCMG is stabilized at a power level of 520kW and bandwidth of $8.321 \times 10^5 rad/sec$.
2. The individual components of the cascaded network, consisting of source side converter and CPL (load side converter) are stable in steady state, thereby making the DCMG stable.
3. The DCMG consisting of CPL in cascaded DC-DC power converter network is stable at a certain power level of the load. The power of the load is found out to be 520kW.
4. The DCMG consisting of CPL in cascaded DC-DC power converter network is stable with controller bandwidth of $8.321 \times 10^5 rad/sec$, which is the bandwidth of the current controller of the CPL.

It is important to note that the stability analysis of the DCMG with CPL is done with specific parameter values used in this study. The stability analysis can be repeated with a different set of controller design parameters.

Appendix

This paper describes the design and stability analysis of:

1. Source side buck converter.
2. CPL (emulated as buck converter) and
3. Cascaded network of source side buck converter and CPL.

In order to verify that the theoretical results and calculations align well, we have simulated the models using Simulink software. The result of the cascaded network is shown in section VII.

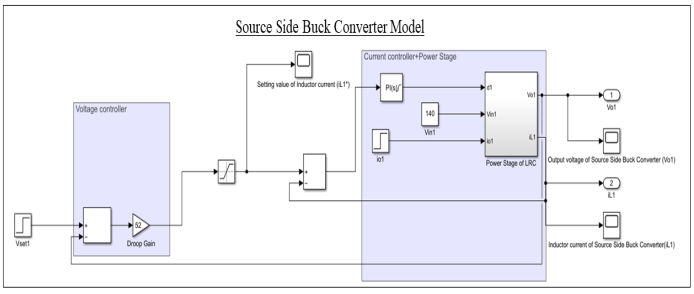


Figure 25: Simulink Model of Source Side Buck Converter

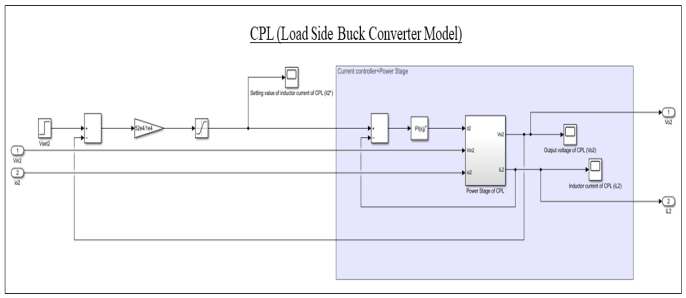


Figure 26: Simulink Model of CPL

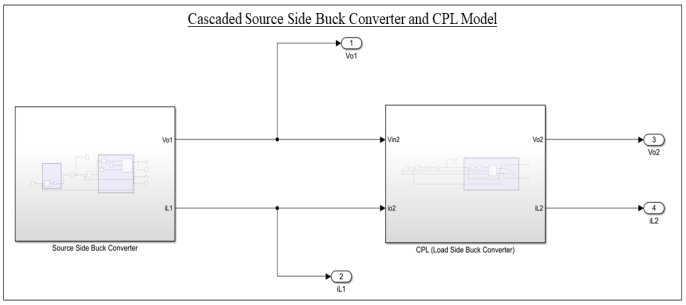


Figure 27: Simulink Model of the cascaded network of source side buck converter and CPL

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors are grateful to the Department of Systems Engineering, University of Arkansas, Little Rock, USA for providing necessary facilities to perform the experiments.

References

[1] S. Ansari, J. Zhang, K. Iqbal, "Modeling, Stability Analysis and Simulation of Buck Converter in a DC Microgrid," in 2021 IEEE Kansas Power and Energy Conference (KPEC), 1-4, 2021, doi: 10.1109/KPEC51835.2021.9446255.

[2] T. Dragičević, X. Lu, J. C. Vasquez, J. M. Guerrero, "DC microgrids— Part I: A review of control strategies and stabilization techniques", IEEE Trans. Power Electron., **31**(7) 4876-4891, 2016, doi: 10.1109/TPEL.2015.2478859.

[3] D. Olson, "Current market electricity supply issues & trends. It is all about the Peak," in 2008 IEEE Intl. Telecom. Energy Confer. (INTELEC), 1–5, 2008, doi: 10.1109/INTLEC.2008.4664017.

[4] S. Luo, "A review of distributed power systems part I: DC distributed power system" IEEE Aerosp. Electron. Syst. Mag. **20**(8), 5–16, 2005, doi: 10.1109/MAES.2005.1499272.

[5] D.K. Fulwani, S. Singh, "Mitigation of Negative Impedance Instabilities in DC Distribution Systems: A Sliding Mode Control Approach", Springer: Berlin, 2016.

[6] V. Grigore, J. Hatonen, J. Kyyra, T. Suntio, "Dynamics of a buck converter with a constant power load," in IEEE Power Electron. Spec. Conf. (PESC 98), vol. 1, 72-78, 1998, doi: 10.1109/PESC.1998.701881.

[7] M. Srinivasan, "Hierarchical Control of DC Microgrids with Constant Power Loads," Ph.D. Thesis, The University of Texas, 2017.

[8] R.W. Erickson and D. Maksimovic, Fundamentals of Power Electronics Norwell, MA: Kluwer, 1997.

[9] R. Gavagsaz-Ghoachani, L.-M. Saublet, M. Phattanasak, J.-P. Martin, B. Nahid-Mobarakeh, S. Pierfederici, "Active stabilisation design of DC–DC converters with constant power load using a sampled discrete-time model: stability analysis and experimental verification." IET Power Electronics **11**(9), 1519-1528, 2018, doi: 10.1049/iet-pel.2017.0670.

[10] M. Cupelli, L. Zhu, A. Monti, "Why ideal constant power loads are not the worst case condition from a control standpoint." IEEE Transactions on Smart Grid **6**(6), 2596-2606, 2014, doi: 10.1109/TSG.2014.2361630.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 2, 73-93 (2022)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

On the Construction of Symmetries and Retaining Lifted Representations in Dynamic Probabilistic Relational Models

Nils Finke*, Ralf Möller

Institute of Information Systems, Universität zu Lübeck, 23562 Lübeck, Germany

ARTICLE INFO

Article history:
Received: 11 January, 2022
Accepted: 05 March, 2022
Online: 18 March, 2022

Keywords:
Relational Models
Lifting
Ordinal Pattern
Symmetry

ABSTRACT

Our world is characterised by uncertainty and complex, relational structures that carry temporal information, yielding large dynamic probabilistic relational models at the centre of many applications. We consider an example from logistics in which the transportation of cargoes using vessels (objects) driven by the amount of supply and the potential to generate revenue (relational) changes over time (temporal or dynamic). If a model includes only a few objects, the model is still considerably small, but once including more objects, i.e., with increasing domain size, the complexity of the model increases. However, with an increase in the domain size, the likelihood of keeping redundant information in the model also increases. In the research field of lifted probabilistic inference, redundant information is referred to as symmetries, which, informally speaking, are exploited in query answering by using one object from a group of symmetrical objects as a representative in order to reduce computational complexity. In existing research, lifted graphical models are assumed to already contain symmetries, which do not need to be constructed in the first place. To the best of our knowledge, we are the first to propose symmetry construction a priori through a symbolisation scheme to approximate temporal symmetries, i.e., objects that tend to behave the same over time. Even if groups of objects show symmetrical behaviour in the long term, temporal deviations in the behaviour of objects that are actually considered symmetrical can lead to splitting a symmetry group, which is called grounding. A split requires to treat objects individually from that point on, which affects the efficiency in answering queries. According to the open-world assumption, we use symmetry groups to prevent groundings whenever objects deviate in behaviour, either due to missing or contrary observations.

1 Introduction

This paper is an extension of two works originally presented in *KI 2021: Advances in Artificial Intelligence* [1] and in *AI 2021: Advances in Artificial Intelligence – 34rd Australasian Joint Conference* [2]. Both papers study the approximation of symmetries using an ordinal pattern symbolisation approach to prevent groundings in dynamic probabilistic relational models (DPRMs)[1].

In order to cope with uncertainty and relational information of numerous objects over time, in many real-world applications, probabilistic temporal (also called dynamic) relational models (DPRMs) need often be employed [3]. Reasoning on large probabilistic models, like in data-driven decision making, often requires evaluating multiple scenarios by answering sets of queries, e.g., regarding the probability of events, probability distributions, or actions leading to a maximum expected utility (MEU). Further, reasoning on large probabilistic models is often performed under time-critical conditions, i.e., where computational tractability is essential [4]. In this respect, DPRMs, together with lifted inference approaches, provide an efficient formalism addressing this problem. DPRMs describe dependencies between objects, their attributes and their relations in a sparse manner. To encode uncertainty, DPRMs encode probability distributions by exploiting in-dependencies between random variables (randvars) using factor graphs. Factor graphs are combined with relational logic, using logical variables (logvars) as parameters for randvars to compactly represent sets of randvars that are considered indistinguishable (without further evidence). This technique is also known as *lifting* [5, 6]. A lifted representation of a probabilistic graphical model allows for a sparse representation to restrain state complexity and enables to decrease runtime complexity in inference.

To illustrate the potential of lifting, let us think of creating a probabilistic model for navigational route planning and congestion

*Corresponding Author: Nils Finke, Institute of Information Systems, Universität zu Lübeck, 23562 Lübeck, Germany, finke@ifis.uni-luebeck.de
[1] pronounced *deeper* models

avoidance in dry-bulk shipping. Dry-bulk shipping is one of the most important forms of transportation as part of the global supply chain [7, 3]. Especially the last year 2020, which was marked by the coronavirus pandemic, shows the importance of good supply chain management. An important sub-challenge in supply chain management is congestion avoidance, which has been studied in research ever-since [8]-[10]. Setting up a probabilistic model to improve planning and to avoid congestion requires identifying features, such as demand for commodities and traffic volume, affecting any routing plans. Commodities are unevenly spread across the globe due to the different mineral resources of countries. In case of excessive demand, regions where the demanded commodities are mined and supplied are excessively visited for shipping, resulting in congestion in those regions. If such a model includes only a few objects, here regions from which commodities are transported, the model might still be considerably small, but once including more regions to capture the whole market, i.e., with an increasing domain size, the complexity of the model increases. However, with an increase in complexity due to an increase in the domain size also the likelihood of keeping redundant information within the model increases. For example, in the application of route planning, multiple regions may exist that are similar in terms of features of the model. Intuitive examples are regions offering the same commodities, i.e., regions with similar mineral resources.

Lifting exactly exploits that existence: Regions which are symmetrical with respect to the features used in the model can be treated by one representative for a group of symmetrical objects to obtain a sparser representation of the model. Further, by exploiting those occurrences, reasoning in lifted representations has no longer a complexity exponential in the number of objects represented by the model, here regions, but is limited to the number of objects with asymmetries only [11, 12]. More specifically, symmetries across objects of a models domain, i.e., objects over randvars of the same type, are exploited by means of performing calculations in inference only once for groups of similarly behaving objects, instead of performing the same calculations over and over again for all objects individually. The principle of lifting applies not only to logistics but also to many other areas like politics, healthcare, or finance – just to name a few.

DPRMs encode a temporal dimension and can be used in any *online scenario*, i.e., new knowledge is on the fly encoded to enable for continuous query answering without relearning the model. In existing research, it is assumed that lifted graphical models already contain symmetries, i.e., simply speaking, a model is setup so that all objects behave according to the same probability distribution. New knowledge is then incorporated in the model with new observations for each object. Observations are encoded within the model as realisations of randvars, resulting in a split off from a symmetrical consideration, called *grounding*. Of course, if the same observation is made for multiple objects, those objects are split off together and continue to be treated as a group. Over time, models dissolve into groups of symmetrically behaving objects, i.e., symmetries are implicitly exploited. Note that in the worst case, the models are split in such way that all objects are treated individually, i.e., no symmetries are available in the model so that lifted inference can no longer be applied and all its advantages disappear.

To the best of our knowledge, existing research has not yet focused on constructing symmetries in advance instead of deriving symmetries implicitly. Constructing symmetries in advance has benefits in application and results from the characteristics of real-world applications:

(i) Certain information about objects of the model may not be available at runtime but only become known downstream. In such cases it is beneficial to infer information according to the open-world assumption from the behaviour of other object which tend to behave similar, i.e., applying an intrinsic default. On the one hand wrong information can be introduced in the model, but on the other it is likely that objects continue to behave the same as per other objects.

(ii) Symmetrical objects can behave the same in the long term, but may deviate for shorter periods of time. Even already small deviations lead to groundings in the model, which, if prevented, introduce a small error in the model, but which also is negligible in the long term.

In both cases a model grounds, which must be prevented in order to keep reasoning in polynomial time. We construct groups of objects with similar behaviour, which we denote as *symmetry clusters*, through a symbolisation scheme to approximate temporal symmetries, i.e., objects that tend to behave the same over time. Using the symmetry clusters, it is possible to selectively prevent groundings, which helps to retain a lifted representation.

This work contributes with a summary on approximating model symmetries in DPRMs based on multivariate ordinal pattern symbolisation and clustering to obtain groups of objects with approximately similar behaviour. Behaviour is derived from the realisations of randvars, which generate either a univariate or multivariate time series depending on the number of interdependent randvars in the model. In the first original conference paper [1], we present multivariate ordinal pattern symbolisation for symmetry approximation (MOP_4SA) for the univariate case and introduce symmetry approximation for preventing groundings (SA_4PG) as an algorithm to prevent groundings in inference a priori using the learned entity symmetry clusters. In the second original conference paper [2], we extend MOP_4SA to the multivariate case and motivate the determination of related structural changes and periodicities in symmetry structures. Further, this work contributes with an extension of original works in [1] and [2] by

- a **comprehensive review** of MOP_4SA and SA_4PG with additional applications from dry-bulk shipping,
- a **complement** of the existing theoretical and experimental investigations of MOP_4SA and SA_4PG in the variation of its parameters, i.e., we fill in the gaps by investigating different orders and delays for ordinal patterns while also examining different thresholds in the symmetry approximation with respect to the introduced error in inference, and
- a **new approach** named MOP_4SCD to detect changes in symmetry structures based on a models similarity graph, an intermediate step of MOP_4SA, to re-learn symmetries on demand.

Together MOP_4SA, SA_4PG and multivariate ordinal pattern symbolisation for symmetry change detection (MOP_4SCD) combine into a rich toolset to identify model symmetries as part of the model

construction process, use those symmetries to maintain a lifted representation by preventing groundings a priori and detect changes in model symmetries after the model construction process.

This paper is structured as follows. After presenting preliminaries on DPRMs in Section 2, we continue in Section 3 with an review on how to retain lifted solutions through approximation and its relation to time series analysis and common approaches to determine and approximate similarities in that domain. In the following, we recapitulate MOP$_4$SA, an approach that encodes entity behaviour through ordinal pattern symbolisation in Section 4.1 and summarise approximating entity symmetries based on the symbolisation in Section 4.2. In Section 5 we outline SA$_4$PG and elaborate how to prevent groundings in DPRMs a priori with the help of entity symmetry clusters. In Section 6 we evaluate both MOP$_4$SA and SA$_4$PG in a shipping application and provide a detailed discussion on its various parameters and its effect on the accuracy in inference. In Section 7 we introduce a new approach to detect changes in symmetry structures to uncover points in time at which relearning of symmetry clusters becomes beneficial. In Section 9 we conclude with future work.

2 Background

In the following, we recapitulate DPRMs [5] in context of an example in logistics, specifically dry-bulk shipping. Dry-bulk shipping is one of the most important forms of transportation as part of the global supply chain [7, 3]. Especially the last year 2020, which was marked by the coronavirus pandemic, showed the importance of good supply chain management. The global supply chain was heavily affected as a result of required lock-downs all over the world, which has led to disruptions and significant delays in delivery. An important sub-challenge in supply chain management is to avoid congestion in regions/zones in which cargo is loaded, i.e., making sure that vessels arrive when those regions are not blocked by too many other vessels being anchored up in same. Congestion avoidance has always been an important topic in research [8]-[10]. As follows, we setup a DPRM to infer idle times related to global supply for commodities. As such, we formally define DPRMs and elaborate on sparse representations and more efficient inference by exploiting symmetries to enable for faster decision making.

2.1 A Formal Model on Congestion in Shipping

We setup a simplified DPRM to model congestion resulting in idle times in different regions/zones across the globe. To infer idle times in certain zones, we use freight rates, a fee per ton, which is paid for the transportation of cargoes and differs across zones, as a driver for operators to plan their vessel movements. E.g., an increase in idle time in a zone can be caused by a high freight rate in that same zone, resulting in an higher interest for sending vessels due to the potential to gain higher profits due to high freight rates. Of course, even though freight rates might be higher, not every vessel operator will be able to conclude business in zones which are over-crowed or have higher waiting times increasing costs for lay time. Hence, to describe the interaction between waiting times and freight rates, the idle condition and freight rates in a zone can be represented by one randvar. Freight rates itself are driven by the supply of commodities in zones represented by another randvar. Since idle conditions, freight rates and supply can be similar in multiple zones, we can develop a much smaller model and combine all randvars into one and parameterise these with a logvar to represent the set of all zones respectively. In this example one zone from the set of all zones is referred to as an object or entity, which we use interchangeably moving forward. Such a parameterised random variable is referred to as PRV for short.

Definition 2.1 (PRV) *Let* $\mathbf{R}$ *be a set of randvar names,* $\mathbf{L}$ *a set of logvar names,* Φ *a set of factor names, and* $\mathbf{D}$ *a set of entities. All sets are finite. Each logvar* L *has a domain* $\mathcal{D}(L) \subseteq \mathbf{D}$*. A* constraint *is a tuple* $(\mathcal{X}, C_{\mathcal{X}})$ *of a sequence of logvars* $\mathcal{X} = (X_1, \ldots, X_n)$ *and a set* $C_{\mathcal{X}} \subseteq \times_{i=1}^{n} \mathcal{D}(X_i)$*. A* PRV $A(L_1, \ldots, L_n), n \geq 0$ *is a construct of a randvar name* $A \in \mathbf{R}$ *combined with logvars* $L_1, \ldots, L_n \in \mathbf{L}$*. Then, the term* $\mathcal{R}(A)$ *denotes the (range) values of a PRV* A*. Further, the term* $lv(P)$ *refers to the logvars and* $rv(P)$ *to the randvars in some element* P*. The term* $gr(P_{|C})$ *denotes the set of instances of* P *with all logvars in* P *grounded w.r.t. constraint* C*.*

The idea behind PRVs is to enable for combining objects with similar behaviour in a single randvar to come up with a sparse representation, introducing a technique called *lifting*. To model the interaction between idle times, freight rates and supply in zones across the globe, we use randvars *Idle*, *Supply* and *Rate* parameterised with a logvar Z representing zones, building PRVs $Idle(Z)$, $Supply(Z)$ and $Rate(Z)$. The domain of Z is $\{z_0, z_1, \ldots, z_n\}$ and range values for all PRVs are $\{high, medium, low\}$[2]. A constraint $C = (Z, \{z_1, z_2\})$ for Z allows to restrict Z to a subset of its domain, such as here to z_1 and z_2. Using this constraint, the expression $gr(Idle(Z)_{|C})$ evaluates to $\{Idle(z_1), Idle(z_2)\}$. To represent independent relations, PRVs are linked by a parametric factor (parfactor) to compactly encode the full joint distribution of the DPRM.

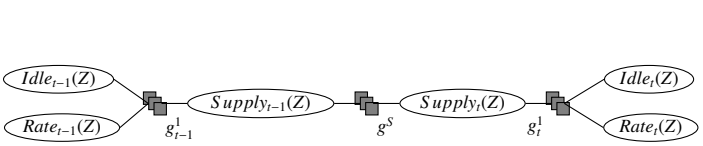
Definition 2.2 (Parfactor) *We denote a parfactor* g *by* $\phi(\mathcal{A})_{|C}$ *with* $\mathcal{A} = (A^1, \ldots, A^n)$ *a sequence of PRVs,* $\phi : \times_{i=1}^{n} \mathcal{R}(A^i) \mapsto \mathbb{R}^+$ *a function with name* $\phi \in \Phi$*, and* C *a constraint on the logvars of* $\mathcal{A}$*. A PRV* A *or logvar* L *under constraint* C *is given by* $A_{|C}$ *or* $L_{|C}$*, respectively. An instance is a grounding of* P*, substituting the logvars in* P *with a set of entities from the constraints in* P*. A parameterized model* PRM G *is a set of parfactors* $\{g^i\}_{i=1}^{n}$*, representing the full joint distribution* $P_G = \frac{1}{Z} \prod_{f \in gr(G)} f$*, where* Z *is a normalizing constant.*

All PRVs are dependent on each other and therefore are combined through one parfactor

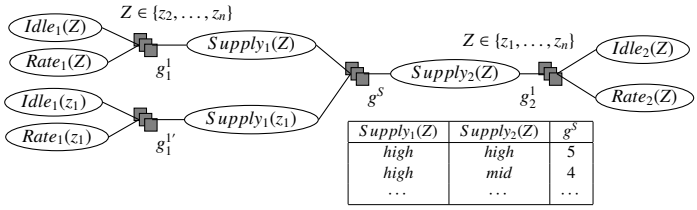
$$g^1 = \phi^1(Idle(Z), Rate(Z), Supply(Z)), \tag{1}$$

which denotes their joint probability distribution. We omit the concrete mappings of potentials to range values of ϕ^1. To encode temporal behaviour, DPRMs follow the same idea as dynamic Bayesian networks (DBNs) with an initial model and a temporal copy pattern to describe model changes over time. DPRMs model a stationary process, i.e., changes from one time step to the next follow the same distribution.

[2] for sake of simplicity we only consider three range values here



(a) Two-slice parameterized probabilistic model (without encoded evidence).



$Supply_1(Z)$	$Supply_2(Z)$	g^S
high	*high*	5
high	*mid*	4
...	...	...

(b) Setting evidence as part of the query $P(S_2(z_1) \mid S_1(z_1) = high$ for $t = 1$.

Figure 1: Graphical representation of a slice of a dynamic probabilistic graphical model illustration how to encode evidence.

Definition 2.3 (DPRM) *A DPRM is a pair of PRMs* $(G_0, G_\rightarrow)$ *where* G_0 *is a PRM representing the first time step and* $G_\rightarrow$ *is a two-slice temporal parameterized model representing* $\boldsymbol{A}_{t-1}$ *and* $\boldsymbol{A}_t$ *where* $\boldsymbol{A}_\pi$ *is a set of PRVs from time slice* π*. An inter-slice parfactor* $\phi(\mathcal{A})_{|C}$ *has arguments* $\mathcal{A}$ *under constraint* C *containing PRVs from both* $\boldsymbol{A}_{t-1}$ *and* $\boldsymbol{A}_t$*, encoding transitioning from time step* $t-1$ *to* t*. A DPRM* $(G_0, G_\rightarrow)$ *represents the full joint distribution* $P_{(G_0,G_\rightarrow),T}$ *by unrolling the DPRM for* T *time steps, forming a PRM as defined above.*

Figure 1a shows the final DPRM. Variable nodes (ellipses) correspond to PRVs, factor nodes (boxes) to parfactors. Edges between factor and variable nodes denote relations between PRVs, encoded in parfactors. The parfactor g^S denotes a so-called inter-slice parfactor that separates the past from the present. The submodel on the left and on the right of this inter-slice parfactor are duplicates of each other, with the one on the left referring to time step $t-1$ and the one on the right to time step t. Parfactors reference time-indexed PRVs, namely $Idle_t(Z)$, $Rate_t(Z)$ and $Supply_t(Z)$.

2.2 Query Answering under Evidence

Given a DPRM, one can ask queries for probability distributions or the probability of an event given evidence.

Definition 2.4 (Queries) *Given a DPRM* $(G_0, G_\rightarrow)$*, a ground PRV* Q_t*, and evidence* $\boldsymbol{E}_{0:t} = \{\{E_{s,i} = e_{s,i}\}_{i=1}^n\}_{s=0}^t$ *(set of events for time steps* 0 *to* t*), the term* $P(Q_\pi \mid \boldsymbol{E}_{0:t})$*,* $\pi \in \{0, \ldots, T\}$*,* $t \leq T$*, denotes a query w.r.t.* $P_{(G_0,G_\rightarrow),T}$*.*

In context of the shipping application, an example query for time step $t = 2$, such as $P(Supply_2(z_1) \mid Supply_1(z_1) = high)$, which asks for the probability distribution of supply at time step $t = 2$ in a certain zone z_1, given that in the previous time step $t = 1$ the supply was high, contains an observation $Supply_1(z_1) = high$ as evidence. Sets of parfactors encode evidence, one parfactor for each subset of evidence concerning one PRV with the same observation.

Definition 2.5 (Encoding Evidence) *A parfactor* $g_e = \phi_e(E(X))_{|C_e}$ *encodes evidence for a set of events* $\{E(x_i) = o\}_{i=1}^n$ *of a PRV* $E(X)$*. The function* ϕ_e *maps the value* o *to 1 and the remaining range values of* $E(X)$ *to 0. Constraint* C_e *encodes the observed groundings* x_i *of* $E(X)$*, i.e.,* $C_e = (X, \{x_i\}_{i=1}^n)$*.*

Figure 1b depicts how evidence for $t = 1$, i.e., to the left of the interslice parfactor g^S, is set within the lifted model.

Evidence is encoded in parfactors $g_1^{1'}$ by duplicating the original parfactor g_1^1 and using $g_1^{1'}$ to encode evidence and g_1^1 to represent all sets of entities that are still considered indistinguishable. Each parfactor represents a different set of entities bounded by the use of constraints, i.e., limiting the domain for the evidence parfactor $g_1^{1'}$ to $\{z_1\}$ and the domain for the original parfactor g_1^1 to $\mathcal{D}(Z) \setminus \{z_1\}$. The parfactor that encodes evidence is adjusted such that all range value combinations in the parfactors distribution ϕ for $Supply_1(z_1) \neq high$ are dropped. Groundings in one time step are transferred to next time steps, i.e., also apply to further time steps, which we discuss as follows.

2.3 The Problem of Model Splits in Lifted Variable Elimination for Inference

As shown in Fig. 1b, evidence leads to groundings in the lifted model. Those model splits are carried over in message passing over time when performing query answering, i.e., in inference. Answering queries, e.g., asking for the probability of an event, results in joining dependent PRVs, or more specifically, joining those parfactors with *overlapping* PRVs. Figure 1b shows a sample of the probability distribution for the interslice parfactor g^S which separates time steps $t-1$ and t. Answering the query $P(Supply_2(z_1) \mid Supply_1(z_1) = high$ as per the example in Section 2.1 to obtain the probability distribution over supply in time step $t = 2$, requires to multiply parfactors $g_1^{1'}$ with the interslice parfactor g^S and the parfactor g_2^1. However, since in time step $t = 1$ a grounding for z_1 exists, the grounding is carried over to the following time step $t = 2$ as the PRV $Supply_1(z_1)$ is connected to the following time step via the interslice parfactor g^S. Therefore, evidence, here $Supply_1(z_1) = high$, is also set within g^S dropping all range values for $Supply_1(z_1) \neq high$. This step is necessary to obtain an exact result in inference. For any other queries, this evidence is carried over to all future time steps, accordingly. Thus, under evidence a model $G_t = \{g_t^i\}_{i=1}^n$ at time step t is split w.r.t. its parfactors such that its structure remains

$$G_t = \{g_t^{i,1}, \ldots, g_t^{i,k}\}_{i=1}^n \tag{2}$$

with $k \in \mathbb{N}^+$. Every parfactor g_t^i can have up to $k \in \mathbb{N}^+$ splits $g_t^{i,j} = \phi_t^{i,j}(\mathcal{A}^i)_{|C^{i,j}}$, where $1 \leq j \leq k$ and $\mathcal{A}^i$ is a sequence of the same PRVs but with different constraint $C^{i,j}$ and varying functions $\phi_t^{i,j}$ due to evidence. Note that moving forward we use the terms *parfactor split* or *parfactor group* interchangeably.

In our example, the model is only splitted with regards to evidence for the entity z_1. All other entities are still considered to be indistinguishable, i.e., lifted variable elimination (LVE) can still exploit symmetries for those instances. To do so, lifted query answering is done by eliminating PRVs, which are not part of the

query, by so called *lifted summing out*. Basically, variable elimination is computed for one instance and exponentiated to the number of isomorphic instances represented. In [5], the author introduce the lifted dynamic junction tree (LDJT) algorithm for query answering on DPRMs, which uses LVE [6, 13] as a subroutine during its calculations. For a full specification of LDJT, we recommend to read on in [5]. In the worst case a model is fully grounded, i.e., a model as defined in Eq. (12) contains

$$k = \prod_{L \in lv(\mathcal{A})} |L| \tag{3}$$

splits for every parfactor $g_t^i = \phi_t^i(\mathcal{A})_{|C^i}$ such that each object $l \in L$ is in its own parfactor split. The problem of model splits, i.e., groundings, can generally be traced back to two aspects. Groundings arise from

(a) partial evidence or unknown evidence, i.e., certain information about objects of the model may not be available at runtime and either never or only become known downstream, which we denote as *unknown inequality*, or

(b) from different observations for two ore more objects, i.e., objects show different behaviour requiring to consider those individually moving forward, which we denote as *known inequality*.

Once the model is split, those splits are carried forward over time, potentially leading to a fully ground model. By doing so, the model remains exact as new knowledge (in form of observations) is incorporated into the model in all details. Over time, however, distinguishable entities might align and can be considered as one again (in case of known inequality) or entities might have ever since behaved similarly without knowing due to less frequent evidence (in case of unknown inequality). To retain a lifted representation the field of approximate inference, i.e., approximating symmetries, has emerged in research.

3 Related Work on Retaining Lifted Solutions Through Approximation and the Connection to Time Series Analysis

Lifted inference approaches suffer under the dynamics of the real world, mostly due to asymmetric or partial evidence. Handling asymmetries is one of the major challenges in lifted inference and crucial for its effectiveness [14, 15]. To address that problem, approximating symmetries has emerged in related research that we discuss in the following.

3.1 Approximate Lifted Models

For static (non-temporal) models, in [16] the author propose to approximate model symmetries as part of the lifted network construction process. They perform Lifted Belief Propagation (LBP) [17], which constructs a lifted network, and apply Belief Propagation (BP) to it. The lifted network is constructed by simulating message passing and identifying nodes sending the same message. To approximate the lifted network, message passing is stopped at an earlier iteration to obtain an approximate instance. In [18], the author also approximate symmetries using LBP, but propose piecewise learning [19] of the lifted network. That means that the entire model is divided into smaller parts which are trained independently and then combined afterwards. In this way, evidence only influences the factors in each part, yielding a more liftable model. Besides approaches using LBP, in [20] the author propose evidence-based clustering to determine similar groundings in an Markov Logic Network (MLN). They measure the similarity between groundings and replace all similar groundings with their cluster centre to obtain a domain-reduced approximation. Since the model becomes smaller, also inference in the approximated lifted MLN is also much faster. In [15], the author propose so-called over-symmetric evidence approximation by performing low-rank boolean matrix factorisation (BMF) [21] on MLNs. They show, that for evidence with high boolean rank, a low-rank approximate BMF can be found. Simply put, finding a low-rank BMF corresponds to removing noise and redundant information from the data, yielding a more compact representation, which is more efficient as more symmetries are preserved. As with any existing approach to symmetry approximation, inference is performed on the symmetrised model, ignoring the introduction of potentially spurious marginals in the model. In [12], the author propose a general framework that provides improved probability estimates for an approximate model. Here, a new proposal distribution is computed using the Metropolis-Hasting algorithm [22, 23] on the symmetrised model to improve the distribution of the approximate model. Their approach can be combined with any existing approaches to approximate model symmetries.

Still, most of the existing research is based on static models and requires to get evidence in advance. However, the problem of asymmetric evidence is particularly noticeable in temporal models, and even more so in an *online setting*, since performing the symmetry approximation as part of the lifted network construction process is not feasible [24]. That means that it is necessary to construct a lifted temporal model once and to encode evidence as it comes in over time. Continuous relearning, i.e., reconstructing the lifted temporal model before performing query answering, is too costly. For temporal models in [25], the author propose to create a new lifted representation by merging groundings introduced over time. They perform clustering to group sub-models and perform statistical significance checks to test if groups can be merged.

In comparison to that and to the best of our knowledge, no-one has focused on preventing groundings before they even occur. To this end, we propose to learn entity behaviour in time and cluster entities that behave approximately similar in the long run and use them to accept or reject incoming evidence to prevent the model from grounding. Clustering entity behaviour requires approaches which find symmetries in entity behaviour, i.e., clustering entities which tend to behave the same according to observations made for them. As observations collected over time result in a time series our problem comes down to identify symmetries across time series.

3.2 From DPRMs to Time Series

In a DPRM, (real-valued) random variables observed over time are considered as time series. Let Ω be a set containing all possible states of the dynamical system, also called state space. Events are

taken from a σ-*algebra* $\mathcal{A}$ on Ω. Then $(\Omega, \mathcal{A})$ is a measurable space. A sequence of random variables, all defined on the same probability space $(\Omega, \mathcal{A}, \mu)$, is called a *stochastic process*. For real-valued random variables, a stochastic process is a function

$$X : \Omega \times \mathbb{N} \to \mathbb{R}, \tag{4}$$

where $X(\omega, t) := X_t(w)$ depending on both, coincidence and time. Note that in the most simple case Ω matches with $\mathbb{R}$ and X with the identity map. Then the observations are directly related to iterates of some ω, i.e., there is no latency, and the X itself is redundant. Over time, the individual variables $X_t(\omega)$ of this stochastic process are observed, so-called realisations. The sequence of realisations is called *time series*. With the formalism from above and fixing of some $\omega \in \Omega$, a time series is given by

$$(X_1(\omega), X_2(\omega), X_3(\omega), \dots) = (x_t)_{t\in\mathbb{N}}. \tag{5}$$

In the case $x_t \in \mathbb{R}$ the time series is called univariate, while in the case $x_t \in \mathbb{R}^m$ it is called multivariate. Note that for stochastic processes we use the capitalisation $(X(t))_{t\in\mathbb{N}}$, while for observations, i.e., paths or time series, we use the small notation $(x(t))_{t\in\mathbb{N}}$. In summary, evidence in a DPRM encoding stochastic processes $(X(t))_{t\in\mathbb{N}}$ forms a time series $(x_t)_{t\in\mathbb{N}}$ that is the subject of further consideration.

3.3 Symmetry Approximation in Time Series

In time series analysis, the notion of similarity, known as symmetry in DPRMs, has often been discussed in the literature [26]-[28]. In general, approaches for finding similarities in a set of time series are either (a) value-based, or (b) symbol-based. *Value-based* approaches compare the observed values of time series. By comparing the value of each point x_t, $t = 1, \dots, T$ in a time series X with the values of each other point $y_{t'}$, $t' = 1, \dots, T'$ in another time series Y (warping), they are able to include shifts and frequencies. Popular algorithms such as dynamic time warping (DTW) [29] or matrix profile [30] are discussed, e.g., in [28]. As DPRMs can encode interdependencies between multiple variables, respective multivariate procedures should be used to assess similarities. The first dependent multivariate dynamic time warping (DMDTW) approach is reported by [31], in which the authors treat a multivariate time series with all its m interdependencies as a whole. The flexibility of warping in value-based approaches leads to a high computational effort and is therefore unusable for large amounts of data. Although there are several extensions to improve runtime [32] by limiting the warping path or reducing the number of data points, e.g., FastDTW [32] or PrunedDTW [33], the use of dimensionality reduction is inevitable in context of DPRMs. For dimensionality reduction, *symbol-based* approaches encode the time series observations as sequences of symbolic abstractions that match with the shape or structure of the time series. Since DPRMs encode discrete values, depending on the degree for discretisation, symbol-based approaches are preferred as they allow for discretisation directly. As far as research is concerned, there are two general ways of symbolisation. On the one hand, *classical symbolisation* partitions the data range according to specified mapping rules in order to encode a numerical time series into a sequence of discrete symbols. A corresponding and well-know algorithm is Symbolic Aggregate ApproXimation (SAX) introduced by [34]. On the other hand, as introduced by Bandt and Pompe [35] *ordinal pattern symbolisation* encodes the total order between two or more neighbours ($x < y$ or $x > y$) into so-called ordinal symbols ((0, 1) or (1, 0)). In [36], the author extend univariate ordinal patterns to the multivariate case, taking into account not only the dependencies of neighbouring values over time, but also the dependencies between spatial variables in a time series.

Specifically here, an ordinal approach has notable advantages in application: (i) The method is conceptionally simple, (ii) the ordinal approach supports robust and fast implementations [37, 38], and (iii) compared to classical symbolisation approaches such as SAX, it allows an easier estimation of a good symbolisation scheme [39, 40]. In the following, we introduce ordinal pattern symbolisation to classify similar entity behaviour.

4 Multivariate Ordinal Pattern for Symmetry Approximation (MOP$_4$SA)

In this section we recapitulate MOP$_4$SA, an approach for the approximation of symmetries over entities in the lifted model. MOP$_4$SA consists of two main steps, which is (a) encoding entity model behaviour through an *ordinal pattern symbolisation* approach, followed by (b) clustering entities with a similar symbolisation scheme to determine groups of entities with *approximately similar behaviour*. We have introduced MOP$_4$SA in [1] for the univariate case and extended same in [2] to the multivariate case.

4.1 Encoding Entity Behaviour through Ordinal Pattern Symbolisation

As mentioned in Section 3.3, approximating entity behaviour corresponds to finding symmetries in time series.

4.1.1 Gathering Evidence

To find symmetries in (multivariate) time series, we use evidence which encode model entity behaviour w.r.t. a certain context, i.e., w.r.t. a parfactor. In particular, this means: Every time-index PRV $P_t(X)$ represents multiple entities $x_0, \dots, x_n$ of the same type at a specific point in time t. That is, for a PRV $Supply_t(Z)$, zones $z_0, \dots, z_n$ are represented by a logvar Z with domain $\mathcal{D}(Z)$ and size $|\mathcal{D}(Z)|$. Note that a PRV can be parameterised with more than one logvar, but for the sake of simplicity we introduce our approach using PRVs with only one logvar throughout this paper. Symmetry detection for m-logvar PRVs works similarly to one-logvar PRVs, with the difference, that in symmetry detection, entity pairs, i.e., m-tuples, are used. As an example, for any 2-logvar PRV $P_t(X, Y)$, an entity pair is a 2-tuple (x_1, y_1) with $x_1 \in \mathcal{D}(X)$ and $y_1 \in \mathcal{D}(Y)$.

A DPRM, as introduced in Section 2.1, encodes temporal data by unrolling a DPRM while observing evidence for the models PRVs, e.g., the PRV $Supply_t(Z)$ encodes supply at time t in various zones Z on the globe. In addition, a DPRM exploits (conditional) independencies between randvars by encoding interdependencies in parfactors. As such, parfactors describe interdependent data through its linked PRVs, e.g., the correlation between supply $Supply_t(Z)$, idle times $Idle_t(Z)$ and freight rates $Rate_t(Z)$ within a common zone

Z encoded by the parfactor g_t^1. For each entity $z_i \in \mathcal{D}(Z)$ from the PRVs $P = \{Supply_t(Z), Idle_t(Z)$ and $Rate_t(Z)\}$ observations are made over time, i.e., a time series $((x_t^i)_{i=1}^m)_{t=1}^T$ with $x_t^i \in \mathcal{R}(P^i)$ is generated. In this work, the time series is to be assumed multivariate, containing interdependent variables, i.e., $m > 1$. Note that in [1] we consider the case $m = 1$. Having $|\mathcal{D}(Z)|$ entities in Z, we consider $|\mathcal{D}(Z)|$ samples of multivariate time series

$$\mathcal{X} = (((x_t^i)_{i=1}^m)_{t=1}^T)_{j=1}^{|\mathcal{D}(Z)|} \in \mathbb{R}^{m \times T \times |\mathcal{D}(Z)|}, \tag{6}$$

e.g., for $m = 3$ with observations $(x_t^1, x_t^2, x_t^3) = (Supply_t(z_j), Idle_t(z_j), Rate_t(z_j))$ for every $z_j \in \mathcal{D}(Z)$ in time $t \in \{1, \ldots, T\}$. As such, a multivariate time series is defined for several PRVs linked in a parfactor, while a univariate time series is defined for a single PRV. Identification of symmetrical entity behaviour is done on a sets of (multivariate) time series, i.e., across different (multivariate) time series that are observed for every entity individually.

4.1.2 Multivariate Ordinal Pattern (MOP) Symbolisation

To encode the behaviour of a time series, we use ordinal pattern symbolisation based on works from Bandt and Pompe [35]. For this, let $X_t \in \mathbb{R}^{m \times T}$ be a (multivariate) time series and $X_t \in \mathbb{R}^{m \times T \times n}$ be the reference database of $n \in \mathbb{N}$ (multivariate) time series. In case of $m = 1$, the time series is univariate. For a better understanding, we start with univariate ordinal patterns that encode the up and downs in a time series by the total order between two or more neighbours. The encoding gives a good abstraction, an approximation, of the overall behaviour or generating process. Univariate ordinal patterns are formally defined as follows.

Definition 4.1 (Univariate Ordinal Pattern) *A vector* $(x_1, ..., x_d) \in \mathbb{R}^d$ *has ordinal pattern* $(r_1, ..., r_d) \in \mathbb{N}^d$ *of order* $d \in \mathbb{N}$ *if* $x_{r_1} \geq ... \geq x_{r_d}$ *and* $r_{l-1} > r_l$ *in the case* $x_{r_{l-1}} = x_{r_l}$.

Figure 2 shows all possible ordinal patterns of order $d = 3$ of a vector $(x_1, x_2, x_3) \in \mathbb{R}^3$.

(2 1 9) (0 1 2) (1 2 0) (0 2 1) (2 0 1) (1 0 2)

Figure 2: All $d!$ possible univariate ordinal patterns of order $d = 3$.

For a *multivariate* time series $((x_t^i)_{i=1}^m)_{t=1}^T$, each variable x^i for $i \in 1, ..., m$ depends not only on its past values but also has some dependency on other variables. To establish a total order between two time points $(x_t^i)_{i=1}^m$ and $(x_{t+1}^i)_{i=1}^m$ with m variables is only possible if $x_t^i > x_{t+1}^i$ or $x_t^i < x_{t+1}^i$ for all $i \in 1, ..., m$. Therefore, there is no trivial generalisation to the multivariate case. An intuitive idea, based on some theoretical discussion in [41, 42] and introduced in [36], is to store univariate ordinal patterns of all variables at a time point t together into a symbol.

Definition 4.2 (Multivariate Ordinal Pattern) A matrix $(x_1, ..., x_d) \in \mathbb{R}^{m \times d}$ has *multivariate ordinal pattern* (MOP) of order $d \in \mathbb{N}$

$$\begin{pmatrix} r_{11} & \cdots & r_{1d} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{md} \end{pmatrix} \in \mathbb{N}^{m \times d} \tag{7}$$

if $x_{r_{i1}} \geq ... \geq x_{r_{id}}$ for all $i = 1, ..., m$ and $r_{il-1} > r_{il}$ in the case $x_{r_{il-1}} = x_{r_{il}}$.

For $m = 1$ the multivariate case matches with the univariate case in Definition 4.1. Figure 3 shows all $(d!)^m$ possible multivariate ordinal patterns (MOPs) of order $d = 3$ and number of variables $m = 2$.

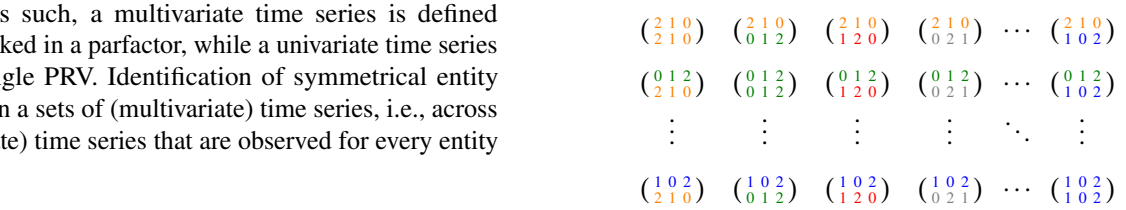


Figure 3: All $(d!)^m$ possible multivariate ordinal patterns of order $d = 3$ with $m = 2$ variables.

The number of possible MOPs $d!$ increases exponentially with the number of variables m, i.e., $(d!)^m$. Therefore, if d and m are too large, depending on the application, each pattern occurs only rarely or some not at all, resulting in a uniform distribution of ordinal patterns [36]. This has the consequence that subsequent learning procedures can fail. Nevertheless, for a small order d and sufficiently large T the use of MOPs can lead to higher accuracy in learning tasks, e.g., classification [36] because they incorporate interdependence of the spatial variables in the multivariate time series.

To symbolise a multivariate time series $X_t \in \mathbb{R}^{m \times T}$ each pattern is identified with exactly one of the *ordinal pattern symbols* $o = 1, 2, ..., d!$, before each point $t \in \{d, ..., T\}$ is assigned its ordinal pattern symbol of order $d \ll T$. The order d is chosen much smaller than the length T of the time series to look at small windows in a time series and their distributions of up and down movements. To assess long-term trends, delayed behaviour is of interest, showing various details of the structure of the time series. The time delay $\tau \in \mathbb{N}_{>0}$ is the delay between successive points in the symbol sequences.

4.1.3 MOP Symbolisation with Data Range Dependence

We assume that for each time step $t = \tau(d-1)+1, \ldots, T$ of a multivariate time series $((x_t^i)_{i=1}^m)_{t=1}^T) \in \mathcal{X}$, MOP is determined as described in Section 4.1.2. Ordinal patterns are well suited to characterise an overall behaviour of time series, in particular their application independent of the data range. In some applications, however, the dependence on the data range can be also relevant, i.e., time series can be similar in terms of their ordinals patterns, but differ considering their y-intercept. In other words, transforming a sequence

$$x = (x_t^i)_{a \leq t \leq b} \tag{8}$$

as $y = x + c$, where $c \in \mathbb{R}$ is a constant, should change y's similarity to other sequences, although the shape is the same. To address the

dependence on the data range, we use the arithmetic mean

$$\overline{x}_t^{d,\tau} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{d}\sum_{k=1}^{d} x_{i,t-(k-1)\tau} \tag{9}$$

of the multivariate time series' values corresponding to the ordinal pattern, where $x_{i,t-(k-1)\tau}$ is min-max normalised, as an additional characteristic or feature of behaviour. If one of the variables changes its behaviour significantly along the intercept, the arithmetic mean uncovers this. There are still other features that can be relevant. For simplicity, we only determine ordinal patterns and their means for each parfactor g^1 with, e.g., PRVs ($Supply_t(Z)$, $Idle_t(Z)$, $Rate_t(Z)$), yielding a new data representation

$$\mathcal{X}' = \langle o, \overline{x}\rangle^{(T-(\tau(d-1))\times|\mathcal{D}(Z)|} \tag{10}$$

where $\langle o, \cdot\rangle_{tj} \in \mathcal{X}'$ represents the MOP and $\langle \cdot, \overline{x}\rangle_{tj} \in \mathcal{X}'$ represents the corresponding mean $\overline{x}_t^{d,\tau}$ for entity z_j at time step t. The order d and delay τ are passed in from the outside and might depend on, e.g., the frequency of the data, to capture the long-term behaviour.

4.2 Clustering Entities with Similar Symbolisation Scheme

After encoding the behaviour of the entities through ordinal pattern symbolisation, we identify similar entities using clustering. For this purpose, based on the derived symbolisation representation in Eq. (10), we create a similarity graph indicating the similarities based on a distance measure between entity pairs.

4.2.1 Creating a Similarity Graph

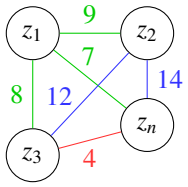
Entity similarity is measured per parfactor, i.e., per multivariate time series, separately. Therefore, multiple similarity graphs, more specifically one per parfactor, are constructed. A similarity graph for a parfactor g_t^1 connecting the PRVs $Supply_t(Z)$, $Idle_t(Z)$ and $Rate_t(Z)$ contains one node for each entity $z \in \mathcal{D}(Z)$ observed in form of multivariate time series. The edges of the similarity graph represent the similarity between two nodes, or more precisely, how closely related two entities of the model are. To measure similarity, we use the symbolic representation $\mathcal{X}'$, which contains tuples of multivariate ordinal numbers and mean values that describe the behaviour of an entity. The similarity of two entities z_i and z_j is given by counts w_{ij} of equal behaviours, i.e.,

$$w_{ij} = \sum_{t\leq T}\Big[\langle o, \cdot\rangle_{it} = \langle o, \cdot\rangle_{jt} \ \wedge\ |\langle \cdot, \overline{x}\rangle_{it} - \langle \cdot, \overline{x}\rangle_{jt}| < \delta\Big], \tag{11}$$

where $[x] = 1$ if x and, 0 otherwise. As an auxiliary structure, we use a square matrix $\mathcal{W} \in \mathbb{N}^{|\mathcal{D}(Z)|\times|\mathcal{D}(Z)|}$, where each $w_{ij} \in \mathcal{W}$ describes the similarity between entities z_i and z_j by simple counts of equal behaviour over time $t \in T$. Simply put, one counts the time steps t at which both multivariate time series of z_i and z_j have the same MOP and the absolute difference of the mean values of the corresponding MOPs is smaller than $\delta > 0$. Finally, as shown in Figure 4b the counts w_{ij} correspond to the weights of edges in the similarity graph $\mathcal{W}$, where zero indicates no similarity between two entities, while the larger the count, the more similar two entities are.

	z_1	z_2	z_3	...	z_n
z_1	0	9	8	...	7
z_2	9	0	12	...	14
z_3	8	12	0	...	4
...	...	...	...	...	...
z_n	7	14	3	...	0

(a) Auxiliary matrix with entity similarity counts.



(b) Auxiliary matrix in form of the Similarity Graph.

Figure 4: (a) Auxiliary matrix and (b) similarity graph.

Approximating symmetries based on the similarity graph leaves us with a classical clustering problem. That means, clustering entities into groups of entities showing *enough* similarities or leaving others independent if those are too different.

4.2.2 Derive Entity Similarity Clusters

Theoretically, any clustering algorithm can be applied on top of the similarity graph. Each weight in the similarity matrix, or each edge weight between an an entity pair, denotes the similarity between two entities, i.e., the higher the count, the more similar the two entities are to each other. Since this contribution focuses on identifying symmetries in temporal environments, we leave the introduction of a specific clustering algorithm out here, and compare two different ones, specifically Spectral Clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN), for the use in MOP$_4$SA as part of the evaluation in Section 6. After any clustering algorithm is run, we are left with n clusters of entities for each parfactor. Formally, a symmetry cluster is defined as follows.

Definition 4.3 (Symmetry Cluster) *For a parfactor $g_i \in G_t$ with G_t being the PRM at timestep t and $g_i = \phi(\mathcal{A})_{|C}$ containing a sequence of PRVs $\mathcal{A} = (A^1, \ldots, A^n)$, a symmetry cluster S^i contains entities $l \in \mathcal{D}(L)$ concerning the domain $\mathcal{D}(L)$ of one of the log-vars $L \in \boldsymbol{L}$ with $\boldsymbol{L} = \bigcup_{i=1}^{n} lv(A^i)$. Let the term $en(S)$ refer to the set of entities in any symmetry cluster S. Each parfactor $g_i \in G_t$ can contain up to m symmetry clusters $\mathcal{S}_{|g^i} = \{S^i\}_{i=0}^{m}$, such that $en(S^i) \cap en(S^j) = \emptyset$ for $i \neq j$ and $i, j \in \{1, \ldots, m\}$. $|\top$ may be omitted in $\mathcal{S}_{|\top}$.*

In the following section we propose how to utilise symmetry clusters to prevent any lifted model from unnecessary groundings.

5 Symmetry Approximation for Preventing Groundings (SA$_4$PG)

As described in Section 2.3, evidence leads to groundings in any lifted model. Further, those groundings are carried over in message passing when moving forward in time leading to a fully ground model in the worst case. As follows, we propose SA$_4$PG, which uses symmetry clusters as an outcome of MOP$_4$SA to counteract any unnecessary groundings, which occur due to evidence. Since symmetry clusters denote a sets of entities, for which entities in each group tend to behave the same, also observations for each entity individually within a cluster are expected to be similar. Regardless of our approach to prevent groundings, in DPRMs entities are

considered indistinguishable after the initial model setup. Under evidence entities split off from that indistinguishable consideration and are afterwards treated individually to allow for exact inference. Nevertheless, in case one observation was made for multiple entities, those all together split off and are considered individually, but still within the group of entities for which the that observation was made. Such groundings are encoded within the DPRM in parfactor groups as shown in Eq. (12). Symmetry clusters also denote parfactor groups, with the difference that those are determined as part of the model construction process in advance. Therefore, in the model construction process, i.e., before running inference, a model will be splitted according to the clusters into parfactor groups with each group containing only entities from the respective cluster. The only difference in creating parfactor groups without evidence is that no range values are set to zero, but get a different distribution representing the group the best. SA$_4$PG is based on the assumption, that symmetry clusters *stay valid* for a certain period of time after learning them, i.e., that entities within those clusters continue to behave similarity. More specifically and w.r.t. the two types of inequality (see Section 2.3), this means, that

(a) in case of *unknown inequality*, we assume that any entity without an observation most likely continues to behave similar to the other entities within the same cluster for which observations are present, and

(b) in case of *known inequality*, we assume that certain observations dominate one cluster and therefore will be applied for all entities within the cluster.

To make one example, lets assume a symmetry cluster contains entities z_1, z_2 and z_3. Groundings occur whenever observations differ across entities in a symmetry cluster, e.g., grounding occurs, if (a) the observation $(Supply_1(z_1) = high, Idle_1(z_1) = high)$ of entity z_1 differs from observations $(Supply_1(z_i) = low, Idle_1(z_i) = mid)$ of entities z_i for $i = 2, 3$, or (b) observations are only made for a subset of the entities, i.e., for entities z_2 and z_3, but not for entity z_1. In both cases, the entities z_2 and z_3 would be split off from their initial symmetry group, and are henceforth treated individually in a non lifted fashion. In SA$_4$PG we prevent such model splits until a certain extend. Algorithm 1 shows an outline of the overall preventing groundings approach. Preventing groundings works by consuming evidence and queries from a stream and dismissing or inferring evidence within symmetry clusters until an entity has reached an violation threshold H. The threshold H refers to the number of times evidence was inferred or dismissed. To do not force entities to stick to their initial symmetry clusters, we relieve entities from their clusters once the threshold H is received. To keep track on the number of violations, i.e., how often evidence was inferred or dismissed, we introduce a violation map as a helper data structure to store that number.

Definition 5.1 (Violation Map) *For a parfactor $g_i \in G_t$ with G_t being the PRM at timestep t and $g_i = \phi(\mathcal{A})_{|C}$ containing a sequence of PRVs $\mathcal{A} = (A^1, \ldots, A^n)$, a violation map $v_{|g_i} : \bigcup_{i=1}^{n} gr(A^i) \to 0$ is initialised with zero values for all entities in all PRVs $\mathcal{A}$ in g_i. In case a PRV A^i is is parameterised with more than one logvar, i.e., $m = |lv(A^i)|$ with $m > 1$, v contains m-tuples as entity pairs. A model contains up to n parfactors in G_t. The set of violation maps is denoted by $V = \{v_{|g_i}\}_{i=0}^{n}$. Let viol$(P)$ refer to the violation count of some entity m-tuple in V.*

SA$_4$PG continues by taking all evidence $\mathbf{E}_t$ concerning a timestep $t = 0, 1, \ldots, T$ from the evidence stream $\mathcal{E}$. To set evidence and to prevent groundings, for each observation $E_{s,i} \in \mathbf{E}_t$ with $\mathbf{E}_t = \{E_{s,i} = e_{s,i}\}_{i=1}^{n}$ so called *parfactor partitions* are identified. A parfactor partitions is a set of parfactor groups $g_t^{i,k}$ that are all affected by evidence $E_{s,i}(x_j)$ with $x_j \in \mathcal{D}(lv(E_{s,i}))$. A parfactor group is *affected*, if

(a) the parfactor g_t^i itself links the PRV $E_{s,i}$ for which an observation was made,

(b) and if the parfactor group g_t^i currently represents the distribution for the specific entity x_j for which the observation was made.

To make one example, observing $Supply_1(z_1) = high$, the evidence partition contains parfactor groups of the parfactors g_t^1 and g^S since the PRV is linked to both parfactors. Further, the parfactor partition is limited to only those i parfactor groups $g_t^{i,1}$ and $g_t^{i,S}$, which currently represent the distribution for the entity z_1. A parfactor partition containing all those parfactor groups is defined as follows.

Definition 5.2 (Parfactor Partition) *Every parfactor $g_t^i \in G_t$ can have up to $k \in \mathbb{N}^+$ splits such that*

$$G_t = \{g_t^{i,1}, \ldots, g_t^{i,k}\}_{i=1}^{n}. \tag{12}$$

Each parfactor g_t^i contains a sequence of PRVs $\mathcal{A}_t = (A_t^1, \ldots, A_t^n)$, which are afflicted with evidence $A_t^n(x_i) = a_{t,i}$ for any entity $x_i \in \mathcal{D}(X)$ with $X \in lv(\mathcal{A}_t)$ at timestep t leading to those splits. A parfactor partition P_t denotes a set of parfactors, which are affected by new evidence $E_t(x_i) = e_t$ with

$$P_t = \{g_t^{i,1}, \ldots, g_t^{i,l}\}_{i=1}^{n} \tag{13}$$

and $l \leq k$ such that any parfactor group $g_t^{i,l} \in P_t$ contain the randvar E_t, i.e., $E_t \in rv(g_t^{i,l})$ and $g_t^{i,l}$ is limited by constraints to at least the entity x_i for which the observation was made, i.e., $g_{t\,|C_e}^{i,l}$ with $C_e = (X, \{x_i\}_{i=1}^{n})$ and $x_i \in \{x_i\}_{i=1}^{n}$.

Considering all evidence $\mathbf{E}_t$ for a time step t, different observations $E_{t,i} \in \mathbf{E}_t$ can result in the same parfactor partition (before those observations are encoded within the model). This holds true for all observation, which are made for the same PRV with entities being in the same parfactor group, e.g., two observation $Supply_1(z_i) = high$ and $Supply_1(z_j) = mid$ for which $\{z_i, z_j\} \in gr(g_1^{1,l})$ and $\{z_i, z_j\} \in gr(g^{S,l})$. All observations that entail the same parfactor partition are treated in SA$_4$PG as one and those observations are informally denoted as an *evidence cluster*.

Therefore, in SA$_4$PG evidence $\mathbf{E}_t$ is rearranged in a sense such that $\mathbf{E}_t$ contains multiple collections of observations, i.e.,

$$\mathbf{E}_t = \{\{E_{t,l} = e_{t,l}\}_{l=0}^{m}, \ldots, \{E_{t,l} = e_{t,l}\}_{l=0}^{m}\}, \tag{14}$$

with each element $E_{t,l}$ originally being directly in $\mathbf{E}_t$ and each subset $\{E_{t,l} = e_{t,l}\}_{l=0}^{m}$ concerning the same parfactor partition P_t. SA$_4$PG proceeds by processing each evidence cluster separately. Evidence of each evidence cluster is processed in a sense such that known inequalities and any uncertainty about inequality is counteracted. This is being done by identifying the *dominating observation* within

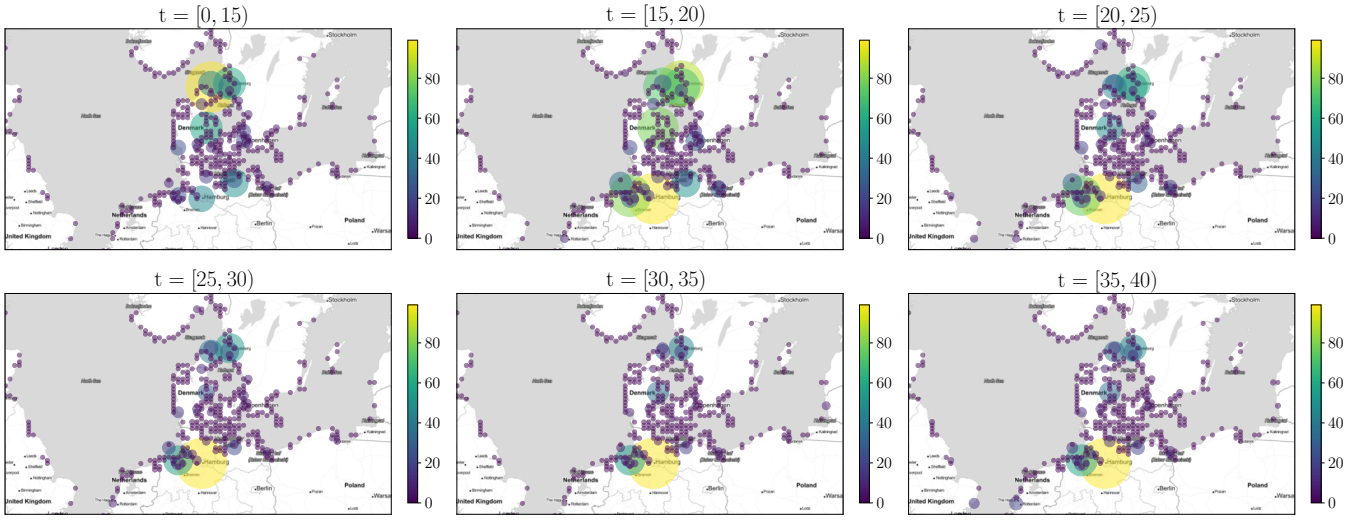


Figure 5: Pointmap showing the normalised average supply over time intervals $[t-5, t)$ in the Baltic Sea region. Best viewed in colour.

each evidence cluster $\{E_{t,l} = e_{t,l}\}_{l=0}^{m}$ and apply that observation to all entities within the respective parfactor partition P_t. The dominating observation $max(e_{t,l})$ is the observation that can be observed the most within the evidence cluster. Further, in case any other entities in the corresponding parfactor partition are unobserved, we also apply the dominating observation for those. For each entity for which evidence was inferred or dismissed, the violation counter is increased. In case the violation threshold H for an entity is already reached, evidence is no longer inferred or dismissed, but the entity is relieved from its symmetry cluster, i.e., split off from the parfactor group, and from then on considered individually.

In the following, we evaluate MOP$_4$SA and SA$_4$PG as part of a case study from the example shipping domain.

6 MOP$_4$SA and SA$_4$PG in Application

Since MOP$_4$SA consists of multiple steps, namely (a) encoding entity behaviour, (b) similarity counting, and (c) clustering, we evaluate each step separately before analysing the overall fitness in conjunction with SA$_4$PG, as introduced in Section 5.

6.1 The AIS Dataset

To setup a DPRM as shown in Figure 1a, we use historical vessel movements from 2020 based on automatic identification system (AIS) data[2] provided by the Danish Maritime Authority for the Baltic Sea. AIS data improves the safety and guidance of vessel traffic by exchanging navigational and other vessel data. It was introduced as a mandatory standard by the International Maritime Organisation (IMO) in 2000. Meanwhile, AIS data is used in many different applications in research, such as trade flow estimation, emission accounting, and vessel performance monitoring [43]. Preprocessing for retrieving variables *Supply* and *Idle* for 367 defined *Zones* can be found on GitHub[3]. Figure 5 gives an idea on how supply evolves over time t in the Baltic Sea region. Each point illustrates the normalised cargo supply amount in tons. For sake of simplicity, we only plot supply independent of idle times here. We can see, that in the beginning of the year (for $0 < t < 20$) the supply in the northern regions, i.e. the need for resources, is higher, while for the rest of the year (for $20 < t < 40$) the supply slowly decreases and increases in the southern regions. The important part here is, that the supply for $20 < t < 40$ in the respective regions is more or less constant over a longer period of time.

Algorithm 1: Preventing Groundings

Input: DPRM $(G_0, G_{\rightarrow})$, Evidence- $\mathcal{E}$ and Querystream Q, Order d, Delay τ, Symmetry Clusters C

for *each parfactor* $g_i \in G_0$ **do**
  $v_{|g_i} \leftarrow$ init violoation map `// see Definition 5.1`

for $t = 0, 1, ..., T$ **do**
  $\mathbf{E}_t \leftarrow$ get evidence from evidence stream $\mathcal{E}$
  Rearrange $\mathbf{E}_t$ to create evidence clusters according to parfactor partition P_t `// see Definition 5.2`
  for *each evidence cluster* $\{E_{t,l} = e_{t,l}\}_{l=0}^{m} \in \boldsymbol{E}_t$ **do**
    $max(e_{t,l}) \leftarrow$ get dominating observation
    `// Align Evidence`
    for *each observation in* $E_{t,l}(x_i) \in \{E_{t,l} = e_{t,l}\}_{l=0}^{m}$ **do**
      if $e_{t,l} \neq max(e_{t,l})$ *and* $viol(x_i) < H$ **then**
        Dismiss observation $e_{t,l}$
        $viol(x_i) \leftarrow viol(x_i) + 1$
    `// Infer Evidence`
    for *each unobserved entity* x_j *in* P_t **do**
      Set $E_{t,l}(x_j) = max(e_{t,l})$
  Answer queries Q_t from query stream Q

The idea behind MOP$_4$SA is to identify periods of time with similar behaviour for multiple entities. That means in our application, identifying zones with similar supply (or more specifically in the multivariate context supply/idle times) over a period of time.

[2] https://www.dma.dk/SikkerhedTilSoes/Sejladsinformation/AIS/
[3] https://github.com/FinkeNils/Processed-AIS-Data-Baltic-Sea-2020-v2

Next, we look into clustering based on the similarity graph as an outcome of similarity counting after applying the symbolisation.

6.2 Multivariate Symbolisation Scheme for Temporal Similarity Clustering

According to the procedure as introduced in Section 4.1.3, we apply the symbolisation scheme on the multivariate supply/idle-time time series as encoded in the parfactor g_t^i and create one similarity graph as the basis for clustering. We compare different clustering algorithms as follows. Unfortunately, classical clustering methods do not achieve good results in high-dimensional spaces, like for DPRMs, which are specifically made to represent large domains. Problems that classical clustering approaches have is, that the smallest and largest distances in clustering differ only relatively slightly in a high-dimensional space [44]. For DPRMs, a similarity graph, representing the similarity of entities $z \in \mathcal{D}(Z)$, contains

$$\binom{|\mathcal{D}(Z)|}{2} \tag{15}$$

fully-connected nodes in the worst case, where Z is a logvar representing a set of entities whose entity pairs share similar behaviour for least one time step. Here, Eq. (15) also corresponds to the number of dimensions a clustering algorithm has to deal with.

6.2.1 An Informal Introduction to Clustering

Generally, clustering algorithms can be divided into the four groups (a) centroid-based clustering, (b) hierarchical clustering, (c) graph-based clustering, and (d) density-based clustering.

We already pointed out the problem that classical clustering algorithms suffer due to their distance measures, which do not work well in high dimensional spaces. Especially centroid-based clustering approaches, like the well-known k-means algorithm or Gaussian Mixture Models, suffer, as they expect to find spherical or ellipsoidal symmetry. More specifically, in centroid-based clustering the assumption is that the points assigned to a cluster are spherical around the cluster centre and therefore no good clusters can be found due to the relatively equal distances. In hierarchical clustering time and space complexity is especially bad since the graph is iteratively split into clusters. Graph-based clustering algorithms, like spectral clustering, is known as being especially robust for high-dimensional data due to performing dimensionality reduction before running clustering [45]. One disadvantage, which also applies to clustering algorithms above, is that the number of clusters need to be specified as a hyperparameter in advance. In contrast, in density-based clustering approaches, like DBSCAN, the number of clusters are determined automatically while also handling noise. DBSCAN is based on a high-density of points. That means, clusters are dense regions, which are identified by running with a sliding window over dense points, making DBSCAN cluster shape independent.

For these reasons, we will compare spectral clustering and DBSCAN as part of MOP$_4$SA as follows. We start by informally introducing Spectral Clustering and DBSCAN.

DBSCAN works by grouping together points with many nearby neighbours, denoting points lying outside those regions as noise. In DBSCAN the two parameters ϵ and *minPoint* need to be provided from the outside, which correspond to the terms *Density Reachability* and *Density Connectivity* respectively. The idea behind DBSCAN is to identify points, that are reachable from another if it lies within a specific distance from it (Reachability), identifying core, border and noise points as the result of transitively connected points (Connectivity) [46]. More specifically, a core point is a point that has m points within a distance of n from itself, whilst a border point has at least one core point within the distance of n. All other points are considered as noise. The algorithm itself proceeds by randomly picking up a point from the dataset, that means, picking one node from the similarity graph, until every point was visited. All *minPoint*-points within a radius of ϵ around the randomly chosen point are considered as one cluster. DBSCAN proceeds by recursively repeating the neighbourhood calculations for each neighbouring point, resulting in n clusters.

Spectral Clustering involves dimensionality reduction in advance before using standard clustering methods such as k-means. For dimensionality reduction, the similarity graph $\mathcal{W}$ is transformed into the so-called graph Laplacian matrix L, which describes the relations of the nodes and edges of a graph, where the entries are defined by

$$L_{ij} := \begin{cases} \deg(z_i) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and } w_{ij} > 0\,, \\ 0 & \text{else} \end{cases} \tag{16}$$

with $\deg(z_i) = \sum_{j=1}^{|\mathcal{D}(Z)|} w_{ij}$. For decorrelation, data in the graph Laplacian matrix L is decomposed into its sequence of eigenvalues and the corresponding eigenvectors. The eigenvectors form a new uncorrelated orthonormal basis and are thus suitable for standard clustering methods. The observations of the reduced data matrix whose columns contain the smallest k eigenvectors can now be clustered using k-means. An observation assigned to cluster C_i with $i = 1, ..., k$ can then be traced back to its entity $z \in \mathcal{D}(Z)$ by indices.

We evaluate both clustering approaches as part of SA$_4$PG in Section 6.3. To improve comparability, we compare both clustering approaches as described in the next Section.

6.2.2 Clustering Comparison Approach

We compare DBSCAN and Spectral Clustering in MOP$_4$SA by identifying clusters with each clustering approach and use resulting clusters within SA$_4$PG respectively, i.e., run SA$_4$PG once using clusters determined by DBSCAN and once using clusters determined by Spectral Clustering.

Since DBSCAN is able to automatically determine the numbers of clusters, we use DBSCAN to identify same and provide the resulting number of clusters as a parameter when performing Spectral Clustering. As DBSCAN is capable to also classifies noise, i.e., entities, which cannot be assigned to a cluster, we use the number of points classified as noise plus the number of clusters as the number of total clusters in Spectral Clustering. Further, for DBSCAN we provide $minPoints = 2$ as the minimum number of entities in a cluster to allow for the maximum number of clusters in general. The eps parameter is automatically determined using the kneedle

algorithm [47]. For Spectral Clustering we just provide parameter k for the total number of clusters, which was previously determined by DBSCAN. Note that the total number of clusters n, which was determined by DBSCAN, does not necessarily corresponds to the best number of clusters for Spectral Clustering. Nevertheless, we obtain results that show which algorithm, given the same input n, is better at separating entities in the multidimensional space.

In the following, we perform a detailed comparison between the two clustering approaches as part of SA$_4$PG with different parameters for the symbolisation scheme in MOP$_4$SA using the approach to compare the two clustering mechanisms as described here.

6.3 Preventing Groundings

MOP$_4$SA is affected by (a) the efficiency of the clustering algorithm used, (b) the similarity measure itself, (c) and its hyper parameters such as order d, delay τ and δ for the arithmetic mean as defined in Eq. (9). We evaluate MOP$_4$SA as part of SA$_4$PG. Specifically, we approximate symmetry clusters using MOP$_4$SA with different settings and (i) perform inference using the the symmetry clusters to prevent the model from grounding, and (ii) compare it with exact lifted inference and calculate Kullback Leibler divergence (KLD) between query result to determine the error introduced through SA$_4$PG. A KLD with $D_{KL} = 0$ indicates that both distributions are equal. Inference in DPRMs is performed by the lifted dynamic junction tree algorithm. Details can be found in [5].

We ran 54 experiments in total with different parameter combinations $d \in \{2, 3, 4\}$, $\tau \in \{1, 2, 3\}$, $\delta \in \{0.05, 0.1, 0.15\}$ and clustering through Spectral Clustering and DBSCAN. For comparison, we perform query answering given sets of evidence, i.e., we perform inference by answering the prediction query $P(Supply_t(Z), Idle_t(Z))$ for each time step $t \in \{4, \ldots, 51\}$ and obtain a marginal distribution for each entity $z \in \mathcal{D}(Z)$. We repeat query answering three times, once without preventing any groundings, once with preventing groundings using the clusters determined by DBSCAN, and once again with preventing groundings but using clusters determined by Spectral Clustering for each parameter combinations. Note that we only discuss results for a sub-selection of the parameter combinations, which give good results in terms of accuracy in inference under preventing groundings, while Table 1 and Table 2 at the end of this paper show the full results for all parameter combinations. Table 1 and Table 2 show results for time intervals $t \in \{[5, 10), [10, 15), [15, 20), [20, 25), [25, 30), [30, 35), [35, 40), [40, 45), [45, 50)\}$.

We evaluate runtime in seconds s, the number of groundings $\#_{gr}$ and KLD D_{KL}. Note that, $\#_{gr}$ shows the number of clusters after time t, while n shows the initial number of clusters. Thus, the number of additional groundings at a specific timestep equals to $\#_{gr}$ minus n. Preventing groundings aims at keeping a lifted model as long as possible. A basic prerequisite for this is that similarities exists in the data. As to that, the variable $n_{\geq 1}$ shows the number of initial clusters, which contain more than one entity directly after clustering, i.e., clusters in which similarly behaving entities have been arranged. Note that similar to n, $n_{\geq 1}$ does not change over time. With increasing order d the number of neighbouring data points are increasing, i.e., the classification contains more long term patterns. With increasing delay τ, long-term behaviour is extended even further, while also allowing for temporary deviations. For data range dependence, in similarity counting we test different delta $\delta_{\leq}$.

The number of clusters with more than one entity $n_{>1}$ relative to the total number of clusters n are important in evaluating how well symmetries are exploited. When n is small, i.e. when n is *significantly* smaller than the total number of entities $|\mathcal{D}(Z)|$, a value of $n_{>1}$ close to n is desirable since it indicates that many entities show symmetries with each other. If $n_{>1}$ is significantly smaller than n, then only a few entities show symmetries, which on the one hand leads to a better accuracy in the inference since many entities are considered at a ground level, but on the other hand runtime will suffer greatly. As to that, Figure 6 shows a comparison for different parameter combinations and clustering approaches. The red line denotes the total number of clusters n independently of the number of entities included in a cluster, while the bars only show the number of clusters with more than one entity $n_{>1}$. Note that we also include entities, which are treated on a ground level already by the time after learning clusters, in the total number of cluster, i.e., clusters can also only include one entity. Since lifting highly depends on the degree of similarities, only those clusters with more than one entity are of interest. Each pair of bar plots correspond to a different experiment with different parameter combinations.

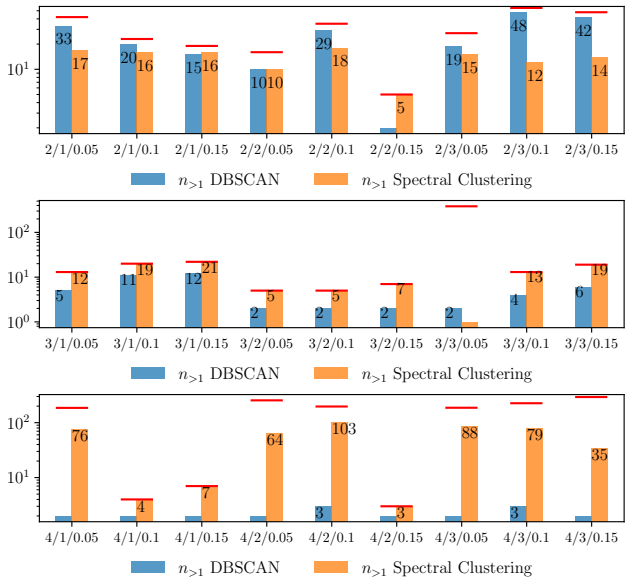
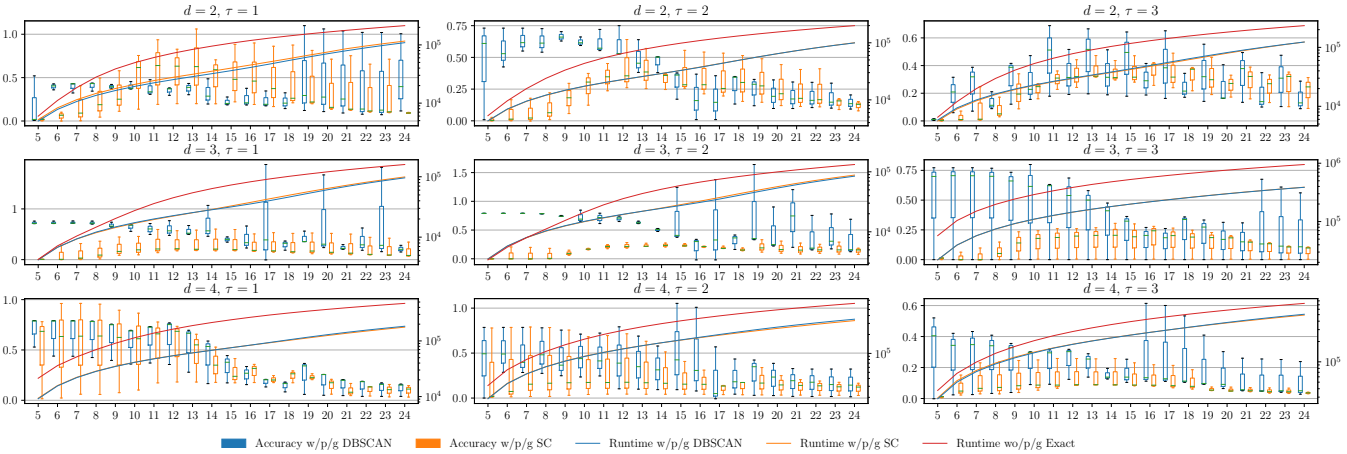
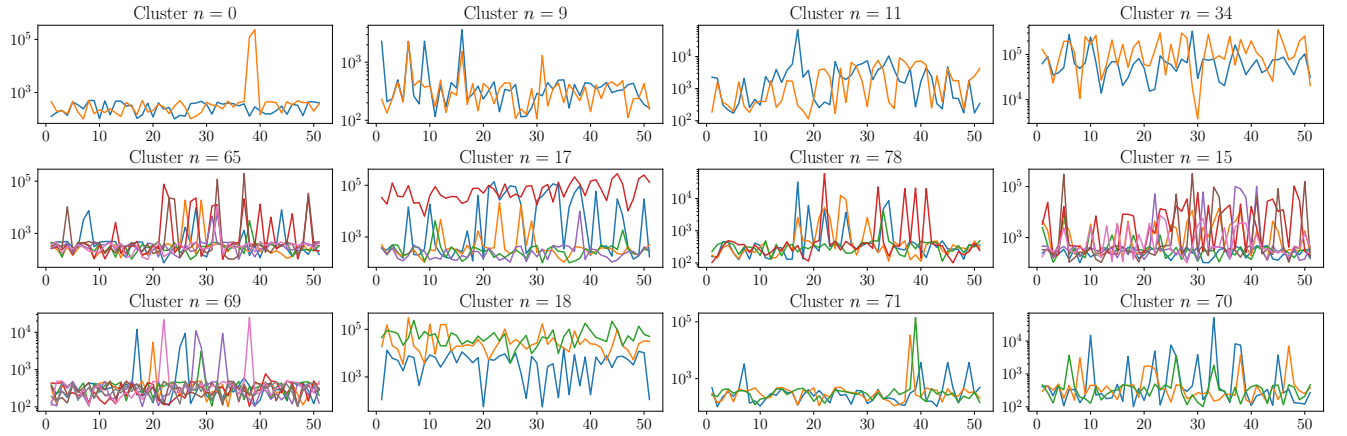


Figure 6: Comparison of number of clusters with more than one entity between DBSCAN and Spectral Clustering.

Due to space limitations we shorten order d, delay τ and delta δ_{leq} in the graph by the triple $d/\tau/\delta_{leq}$. From Fig. 6 it is most noticeable that the number of clusters with more than one entity $n_{\geq 1}$ for higher orders d is much less than for lower orders. This intuitively makes sense, since with higher orders d long term behaviour is captured much better than with lower orders and thus only a few clusters can be determined. For $d = 2$ much more clusters are identified since a smaller time span is considered resulting in a higher possibility of showing similarities. This observation also applies to increasing delays τ. The experiment with order $d = 3$, delay $\tau = 3$ and delta $\delta_{\leq} = 0.05$ is a good example for cases where not many similarities have been identified, but many entities are treated on a



(a) Accuracy and runtime in inference under SA$_4$PG for different parameter combinations.



(b) Supply over time t for a selection of clusters, which have been learned based on Supply/Idle time data for $0 < t \leq 4$ using Spectral Clustering.

Figure 7: Accuracy and runtime data based on query results under SA$_4$PG including raw supply data for entities within clusters.

ground level, i.e., accuracy will be good, but runtime will suffer. In the following, we look at accuracy results and will also come back to this example.

By comparing the KLD D_{KL} as a result of inference without preventing groundings and with preventing groundings based on clusters determined by DBSCAN and Spectral Clustering, it is noticeable that for clusters determined by Spectral Clustering in average a lower D_{KL} compared to DBSCAN results. This is explainable with better handling of higher dimensional data in Spectral Clustering. Figure 7a shows a comparison between the accuracy for both cases and different parameter combination. Each subplot corresponds to a different order d and delay τ, while the box-plot itself shows the variation of the accuracy over different deltas $\delta_{\leq} = \{0.05, 0.1, 0.15\}$ over time t. Note that we only plot data until $t = 24$ for better visibility and as the effect of any wrong evidence, which was brought in by preventing groundings, starts to level off. This happens since groundings are only prevented until the threshold H is reached, i.e., any other evidence afterwards at a later timestep balance out the effect of any wrong evidence at an early timestep after learning symmetry clusters. The blue box-plots in Fig. 7a correspond to the KLD D_{KL} with DBSCAN as the clustering approach in MOP$_4$SA, while the orange box-plots correspond to the KLD D_{KL} with on Spectral Clustering as the clustering approach in MOP$_4$SA. The solid blue, orange and red line correspond to the runtime for answering a query for the specific time step. From the plots, we can see that for higher orders and delays, i.e., with increasing time spans each ordinal represents, that D_{KL} is decreasing. Considering the total number of clusters for each experiment (see Fig. 6), this follows as not many similarities can be found in the data, but more entities are handled on a ground level, i.e., increasing accuracy. On the other hand, runtime drastically increases as symmetries are no longer exploited. Compared to exact reasoning, runtime is noticeably smaller in inference under SA$_4$PG. To lock again at the experiment with order $d = 3$, delay $\tau = 3$ (as highlighted above), the KLD D_{KL} is considerably small especially for Spectral clustering, but the runtime of the inference is very poor compared to all other experiments.

In SA$_4$PG, the violation threshold H is set to 5, i.e., groundings due to any inequalities are prevented for an entity H times. After $t = 10$ the number of groundings $\#_{gr}$ (see Table 1) are still the same as after learning the entity similarity cluster, i.e., all groundings are prevented. It is expected that the majority of groundings are prevented in the initial timesteps after learning the clusters. Still, if entities behave similarity in early timesteps, the threshold H is reached far later in time. Thus, if in clustering based on the similarity graph

the entities with similarities are identified better, then groundings will occur much later in time. The longer D_{KL} stays small, the better cluster fit, i.e, the error introduced in inference through preventing groundings is kept small. In Fig. 7a we see that the accuracy suffers approximately for all experiments around $t = 10$, i.e, after 4 more timesteps after learning the clusters. Figure 7b depicts raw supply data for a selection of clusters as a result of running MOP$_4$SA based on data for $t = [0, 4)$ with parameters $d = 2$, $\tau = 1$ and $\delta = 0.05$ for symbolisation and Spectral Clustering. Even though only providing a small amount of training data, we can see that symmetrical behaviour continues for most of the clusters until approx. $t = 10$, like especially for clusters $n \in \{0, 11, 34, 65, 78, 69, 71\}$ and therefore support the insight, which we have got based on Fig. 7a. For simplicity only raw supply data is plotted even though symmetry clusters are determined based on supply/idle data.

The best results are achieved with Spectral Clustering as part of SA$_4$PG for the parameter combinations $d = 2$, $\tau = 1$, $\delta_{\leq} = 0.1$, $d = 3$, $\tau = 3$, $\delta_{\leq} = 0.1$ and $d = 3$, $\tau = 2$, $\delta_{\leq} = 0.05$, which we will also further refer to in the following Section. Generally, when reasoning under time constraints, preventing grounds is a reasonable approach as it prevents groundings in the long term and therefore speeds up inference.

Entity similarity can change over time, i.e., to further prevent the model from grounding it is beneficial to relearn symmetry structures at some time. In the following we propose MOP$_4$SCD and use it to identify points in time when relearning clusters is beneficial.

7 Multivariate Ordinal Pattern for Symmetry Change Detection (MOP$_4$SCD)

Symmetries in temporal models can change over time as already seen in Fig. 5. Therefore, symmetry cluster, after they have been learned, may only stay valid for a certain period of time. Further, some are valid for a longer period of time, some not. To identify points in time when relearning symmetry clusters is reasonable, we use the similarity graph as an intermediate output of running MOP$_4$SA and check if the similarity graph has changed *significantly*. More specifically, we continue running MOP$_4$SA for every timestep, but instead of for continuously relearning symmetry clusters, we prevent relearning clusters in MOP$_4$SA after the initial sync run until the graph has changed *significantly enough*. To identify such points in time with a significant change, we introduce MOP$_4$SCD taking as inputs a similarity graph for two consecutive timesteps and calculating a distance measure between both. In case the distance measure is above a certain threshold we consider those points as change points to trigger the cluster relearning process. MOP$_4$SCD is based on the assumption, that clusters no longer stay valid, if entities within a cluster no longer show the same similarity to its cluster entities as in the previous timesteps, i.e., the similarity counts is no longer proportionally scaling as before. Those entities might transition to another cluster, since its showing more similarity with another cluster. Informally, if the similarity graph changes over time in a *constant and balanced way*, symmetry clusters stay valid, but if the similarity graph changes over time in an *unbalanced manner*, i.e., if similarity counts change significantly, there is a change in the structure of the symmetry clusters. To illustrate that, let us look at Figure 8. The Figure shows a similarity graph based on which two clusters have been identified.

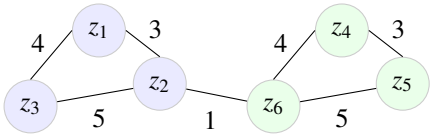


Figure 8: Overview of potential unbalanced changes in a similarity graph.

Nodes $S^1 = \{z_1, z_2, z_3\}$ (coloured in blue) denote a cluster S^1 and the nodes $S^2 = \{z_4, z_5, z_6\}$ (coloured in green) denote another cluster S^2. Both clusters are connected through nodes z_2 and z_6 since for both a similarity was measured at any timestep before learning clusters. Relearning clusters becomes necessary if the cluster structure itself changes. This happens either

(a) if similarities between entities of different clusters changes, e.g., if the similarity between z_2 and z_6 increases and might require to merge the clusters or even split them into more than two clusters, which we denote as a *unbalanced interclusteral change*,

(b) or if similarities within a cluster change disproportionately, e.g., if similarities for S^2 changes only for a subset of the entities such as for z_4 and z_5 but not proportionally for all entities such as z_4 and z_6 and z_6 and z_5 requiring to split the cluster even further, which we denote as a *unbalanced intraclusteral change*.

As follows we define both unbalanced interclusteral and intraclusteral change measures and combine both into a distance measure denoting the unbalanced change between consecutive timesteps. Both unbalanced inter- and intraclusteral changes are determined based on the similarity graph $\mathcal{W}^t$ from the current to the next time step $\mathcal{W}^{t+1}$ under current symmetry clusters $\mathcal{S}$, with interclusteral changes defined as

$$d_{inter}(\mathcal{W}^t, \mathcal{W}^{t+1}, \mathcal{S}) = \sum_{\substack{S^i \in \mathcal{S} \\ S^j = en(S^i) \cap en(\mathcal{S})}} \frac{\sum_{\substack{i \in en(S^i) \\ j \in en(S^j)}} \left[w_{ij}^{t+1} = w_{ij}^t + 1 \right]}{|en(S^i)| \cdot |en(S^j)|} \tag{17}$$

where $[x] = 1$ if x and, 0 otherwise for $en(S^i) \cap en(S^j) = \emptyset$ and intraclusteral changes defined as

$$d_{intra}(\mathcal{W}^t, \mathcal{W}^{t+1}, \mathcal{S}) = \sum_{S^i \in \mathcal{S}} \frac{\sum_{i,j \in en(S^i), i<j} \left[w_{ij}^{t+1} - w_{ij}^t = 0 \right]}{|en(S^i)| \cdot |en(S^i)|} \tag{18}$$

where $[x] = 1$ if x and, 0. Both d_{inter} and d_{intra} are merged into one combined measure with

$$d(\mathcal{W}^t, \mathcal{W}^{t+1}, \mathcal{S}) = \frac{d_{inter} + d_{intra}}{|\mathcal{S}|}. \tag{19}$$

Simply speaking, d_{inter} counts the number of increases in weights across different clusters S^i and S^j such as shown in Fig. 8 for entities z_2 and z_6. The resulting count is normalised by dividing through the number of comparison between entity pairs of the clusters $en(S^i)$ and $en(S^j)$, resulting in measure between 0 and 1 with a value close

to 1 denoting a maximum dissimilarity. Similarly, d_{intra} counts the occurrences of no weight increases within entity pairs of a similar cluster S^i. To ensure that entities within a cluster continue to behave the same, weights should proportionally increase equally distributed within the cluster. If there is no increase in weights most likely the entities discontinue to behave similarly. The resulting count is equally normalised with a value close to 1 denoting a maximum dissimilarity. Finally, both d_{inter} and d_{intra} are combined in a single measure also count normalised to determine a distance measure between 0 and 1. If $d(\mathcal{W}^t, \mathcal{W}^{t+1}, \mathcal{S}) = 0$, the change in the similarity graph is balanced, if $d(\mathcal{W}^t, \mathcal{W}^{t+1}, \mathcal{S}) > 0$, it is unbalanced. If $d(\mathcal{W}^t, \mathcal{W}^{t+1}, \mathcal{S}) > b$, $b \in \mathbb{N}_{>0}$ it may be worthwhile to (re)perform clustering and (re)build symmetry clusters.

As follows, we evaluate MOP$_4$SCD based on clusters determined by MOP$_4$SA for the same parameters as in the experiments performed in Section 6.

8 MOP$_4$SCD in Application

We evaluate MOP$_4$SCD based on clusters determined using MOP$_4$SA as described in Section 6. Since Spectral Clustering works better than DBSCAN in identifying clusters, we here only use clusters determined by Spectral Clustering as part of MOP$_4$SA. We run experiments 27 experiments in total for the same parameter combinations $d \in \{2, 3, 4\}$, $\tau \in \{1, 2, 3\}$ and $\delta \in \{0.05, 0.1, 0.15\}$ as in Section 6. For each experiment we calculate $d(\mathcal{W}^t, \mathcal{W}^{t+1}, \mathcal{S})$ for timesteps $t = 5, \ldots, 51$. Clusters are learned based on a similarity graph with data for $t = 1, \ldots 4$.

In this Section we discuss results for a sub-selection of the parameter combinations, which give good results in terms of accuracy in inference under preventing groundings as seen in Section 4, while Table 3 at the end of this paper shows detailed results for all parameter combinations. Each column in Table 3 shows the distance measure for consecutive timesteps, e.g., for $t = 5$, the distance is derived based on the similarity graph for timestep $t = 4$ to $t = 5$, i.e., $d(\mathcal{W}^4, \mathcal{W}^5, \mathcal{S})$. Note that since $d(\mathcal{W}^4, \mathcal{W}^5, \mathcal{S})$ is calculated for two consecutive timesteps, the distance measure has to be added up over time to derive the overall distance between more then two timesteps. Overall, the distance measure varies for different parameter combinations with in the optimal case showing an unbalanced change in weights of approximately 1.6% and in the worst case of approximately 22.4% between two consecutive timesteps. The distance measure is highly affected by the number of clusters n. In the case that the number of clusters with more than one entity $n_{>1}$ is considerably small compared to the total number of clusters n, the distance measure $d(\mathcal{W}^t, \mathcal{W}^{t+1}, \mathcal{S})$ is also considerably low since d_{inter} and d_{intra}, see Eq. (17) and Eq. (18), always return no unbalanced change for clusters with just one entity, i.e., for entities which are already being treated on a ground level. MOP$_4$SA aims at preventing groundings to speed up inference, i.e., lead to an increase in runtime. Therefore, choosing parameters d, τ and δ for MOP$_4$SA and consequently for MOP$_4$SCD is a trade-of between losses in accuracy and a speed up in inference.

The parameter combinations $d = 2$, $\tau = 1$, $\delta_{\leq} = 0.1$, $d = 3$, $\tau = 3$, $\delta_{\leq} = 0.1$ and $d = 3$, $\tau = 2$, $\delta_{\leq} = 0.05$ give good results in MOP$_4$SA as shown in Section 6. Results for MOP$_4$SCD also support this. Figure 9 shows the KLD D_{KL} in conjunction with results from MOP$_4$SCD. Each subplot corresponds to a different parameter combination with the blue line corresponding to the KLD D_{KL}, the solid red line to the distance measure $d(\mathcal{W}^t, \mathcal{W}^{t+1}, \mathcal{S})$ for two consecutive timesteps t and $t + 1$ and the dashed red line for the cumulative distance measure, i.e., from $t = 0$ until the current timestep t. Note that the cumulative distance is log scaled and can be read of from the right y-axis. The highlighted red area in each subplots mark the interval when the cumulative distance measure becomes greater then 50% until it has reached 100%, i.e., with a change of 100% that all relations between all entities have been affected.

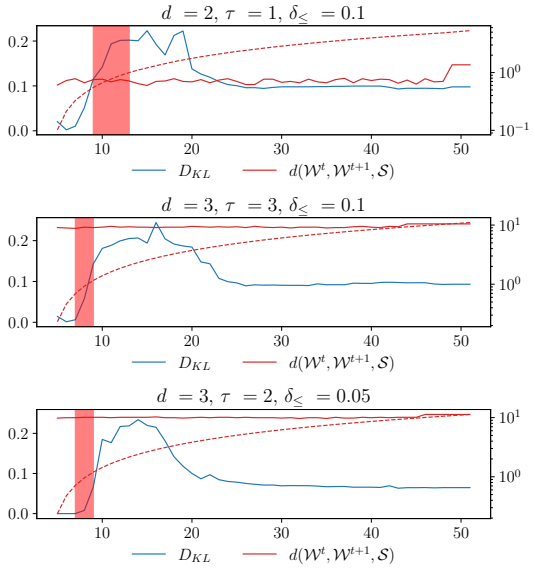


Figure 9: Results of MOP$_4$SCD for three parameter combinations which show the best results for MOP$_4$SA and SA$_4$PG. Further results can be found in the appendix.

Similar to the experiments in Section 6, clusters $\mathcal{S}$ have been determined by MOP$_4$SA based on data for $0 > t \leq 4$. For all upcoming timesteps the clusters $\mathcal{S}$ have been used to prevent groundings, i.e., execute SA$_4$PG as part of inference, see Algorithm 1. The KLD D_{KL} for all experiments as shown in Fig. 9 similarly raises up to a value of approx. 0.25 with its peak around $t = 15$. In contrast $d(\mathcal{W}^t, \mathcal{W}^{t+1}, \mathcal{S})$ varies across experiments and has for the experiment with parameter combination $d = 2$, $\tau = 1$ and $\delta_{\leq}0.1$ its best value of approx. 0.1, i.e, a unbalanced change of approx. 10% over time. For the two other experiments $d(\mathcal{W}^t, \mathcal{W}^{t+1}, \mathcal{S})$ is with 0.22 similar. Correspondingly, the cumulative distance reaches a value of 0.5 at timestep $t = 9$ until it reaches a value of 1 at $t = 14$ for the first experiment, while for the two other experiments the cumulative distance reaches a value of 0.5 at $t = 7$ and a value of 1 already at $t = 9$. I.e., clusters S are valid for a longer period of time using MOP$_4$SA with a parameter combination of $d = 2$, $\tau = 1$, $\delta_{\leq} = 0.1$. Further, for that parameter combination, D_{KL} settles off once a cumulative distance of 1 has reached. Settling off happens due to the amount of new evidence leading to more groundings removing the effect of any wrongly introduced evidence in previous timesteps. Relearning clusters at a threshold of 0.5 is here beneficial to prevent the D_{KL} from further increasing. That means relearning at $t = 9$, i.e., clusters are valid for approx. 4 timesteps after learning

them, which corresponds to a full month in our example application and therefore is a good result to considerably speed up inference while only introducing a small error in inference.

9 Conclusion and Future Work

Evidence lead to groundings in dynamic probabilistic relational models over time, negating runtime benefits in lifted inference. This paper provides MOP$_4$SA, SA$_4$PG and MOP$_4$SCD as a rich toolset to identify model symmetries as part of the model construction process, use those symmetries to maintain a lifted representation by preventing groundings a priori and detect changes in model symmetries after the model construction process. Preventing groundings a priori to maintain any lifted representation is important in lifted inference to preserve its runtime benefits. MOP$_4$SA detects symmetries across entities of the models domain using a multivariate ordinal pattern symbolisation approach and building a similarity graph for spectral clustering to identify sets of entities with symmetrical behaviour regarding a context of the model (symmetry clusters). Symmetry clusters are used in SA$_4$PG as part of query answering to prevent any unnecessary model splits by evidence, e.g., due to one time events. Symmetry structures can change over time, which MOP$_4$SCD detects based on the similarity graph, an intermediate output of MOP$_4$SA, and provide a distance measure denoting the degree of any unbalanced structural change to identify points in time when relearning symmetry clusters is beneficial.

The main contribution of this paper are the extension by theoretical and experimental results on the original papers [1, 2] and the introduction of MOP$_4$SCD as a mechanism to detect structural changes to complement MOP$_4$SA, SA$_4$PG as a rich toolset to prevent groundings a priori. We show, that MOP$_4$SA requires only a small amount of *training data* to come up with a good approximation of symmetry structures. Generally, MOP$_4$SA aims at determining symmetry structures which stay valid for shorter time periods. This follows, since MOP$_4$SA is not capable to capture any reoccurring patterns or periodicity, e.g., due to seasonality. MOP$_4$SA can be extended to capture such behaviour, but this would also increase the complexity of the overall approach. Due to this and since capturing symmetries for longer time spans, especially in real-world applications which normally change much faster, is not feasible, we focus with MOP$_4$SA as being a simple and easy to compute framework, requiring only few historical data points for learning, to identify symmetries for the short term future. In addition to MOP$_4$SA, MOP$_4$SCD supports in inference by identifying points in time when relearning clustering for SA$_4$PG is reasonable.

With preventing groundings a priori we complement existing approaches, which focus on retaining lifted representation after a model has already been splitted. In general, our approach works well with any other approach undoing splits after they occurred when moving forward in time, e.g., in message passing by merging sets of entities when those align again, denoted as *temporal approximate merging*, as proposed in [25]. Combining both kind of approaches brings together the best of both worlds: (a) While with *determining approximate model symmetries* a priori, we can use the full amount of historical training data to prevent groundings, (b) and with *temporal approximate merging*, we can merge non-preventable parfactor splits even after they occurred, i.e., a posterior.

Since MOP$_4$SA is designed to work with small amounts of data to provide symmetry clusters very quickly for the short term future, the overhead MOP$_4$SA and MOP$_4$SCD bring into query answering need to be kept to a minimum. Applying the symbolisation scheme to identify symmetries is already a suitable mechanism, but with the clustering approach we still depend on existing approaches, which are considerably costly. The investigation of more performant clustering approaches, e.g., taking advantage of some sort of incremental changes to clustering after the initial learning step, are left for future work.

List of Symbols

$\mathbf{R}$	set of random variables
$\mathbf{L}$	set of logical variables
Φ	set of factor names
$\mathbf{D}$	set of entities
$\mathcal{D}(L)$	domain of a logvar
$C, (\mathcal{X}, C_{\mathbf{X}})$	constraint restricting logical variables
$A(L_1, \ldots, L_n)$	parameterised logical variable (PRV)
$g, \phi(\mathcal{A})_{\vert C}$	parfactor
$gr(P)$	grounding
$lv(P)$	logical variables
$\mathcal{R}(A)$	range of a PRV
G	model
G_t	local model
$\mathbf{E}$	evidence, set of events
Q	query term
$\mathcal{X}$	multivariate time series
τ	delay between successive time points
d	order of ordinal pattern
w_{ij}	similarity count
$\mathcal{W}$	similarity graph
S	symmetry cluster
$en(S)$	objects in a symmetry cluster
$\mathbf{S}$	set of symmetry clusters
P	parfactor partition
L	Laplacian matrix
D_{KL}	Kullback-Leibler divergence
$\delta_{\leq}$	mean delta
$d(\mathcal{W}^t, \mathcal{W}^{t+1}, \mathcal{S})$	similarity change measure

References

[1] N. Finke, M. Mohr, "A Priori Approximation of Symmetries in Dynamic Probabilistic Relational Models," in S. Edelkamp, R. Möller, E. Rueckert, editors, KI 2021: Advances in Artificial Intelligence, 309–323, Springer International Publishing, Cham, 2021.

[2] N. Finke, R. Möller, M. Mohr, "Multivariate Ordinal Patterns for Symmetry Approximation in Dynamic Probabilistic Relational Models," in AI 2021: Advances in Artificial Intelligence - 34rd Australasian Joint Conference, Lecture Notes in Computer Science (LNCS), Springer International Publishing, In Press.

[3] N. Finke, M. Gehrke, T. Braun, T. Potten, R. Möller, "Investigating Matureness of Probabilistic Graphical Models for Dry-Bulk Shipping," in M. Jaeger, T. D.

Nielsen, editors, Proceedings of the 10th International Conference on Probabilistic Graphical Models, volume 138 of *Proceedings of Machine Learning Research*, 197–208, PMLR, 2020.

[4] Y. Xiang, K.-L. Poh, "Time-Critical Dynamic Decision Making," 2013.

[5] M. Gehrke, T. Braun, R. Möller, "Lifted Dynamic Junction Tree Algorithm," in Proceedings of the International Conference on Conceptual Structures, 55–69, Springer, 2018.

[6] D. Poole, "First-order Probabilistic Inference," in Proc. of the 18th International Joint Conference on Artificial Intelligence, 985–991, 2003.

[7] D. Akyar, "The Effects of Global Economic Growth on Dry Bulk Shipping Markets and Freight Rates," 2018.

[8] Z. Wang, X. Wu, K. L. Lo, J. J. Mi, "Assessing the management efficiency of shipping company from a congestion perspective: A case study of Hapag-Lloyd," Ocean & Coastal Management, **209**, 105617, 2021, doi: https://doi.org/10.1016/j.ocecoaman.2021.105617.

[9] C. Jiang, Y. Wan, A. Zhang, "Internalization of port congestion: strategic effect behind shipping line delays and implications for terminal charges and investment," Maritime Policy & Management, **44**(1), 112–130, 2017, doi: 10.1080/03088839.2016.1237783.

[10] T. Notteboom, "The Time Factor in Liner Shipping Services," Maritime Economics and Logistics, **8**, 19–39, 2006, doi:10.1057/palgrave.mel.9100148.

[11] M. Niepert, G. Van den Broeck, "Tractability through Exchangeability: A New Perspective on Efficient Probabilistic Inference," in AAAI-14 Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2467–2475, AAAI Press, 2014.

[12] G. V. den Broeck, M. Niepert, "Lifted Probabilistic Inference for Asymmetric Graphical Models," CoRR, **abs/1412.0315**, 2014.

[13] N. Taghipour, D. Fierens, J. Davis, H. Blockeel, "Lifted Variable Elimination: Decoupling the Operators from the Constraint Language," Journal of Artificial Intelligence Research, **47**(1), 393–439, 2013.

[14] K. Kersting, "Lifted Probabilistic Inference," in Proceedings of the 20th European Conference on Artificial Intelligence, ECAI'12, 33–38, IOS Press, NLD, 2012.

[15] G. Van den Broeck, A. Darwiche, "On the Complexity and Approximation of Binary Evidence in Lifted Inference," in C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 26, Curran Associates, Inc., 2013.

[16] P. Singla, A. Nath, P. Domingos, "Approximate Lifting Techniques for Belief Propagation," in Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI'14, 2497–2504, AAAI Press, 2014.

[17] P. Singla, P. Domingos, "Lifted First-Order Belief Propagation," in Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2, AAAI'08, 1094–1099, AAAI Press, 2008.

[18] B. Ahmadi, K. Kersting, M. Mladenov, S. Natarajan, "Exploiting symmetries for scaling loopy belief propagation and relational training," Machine Learning, **92**, 91–132, 2013.

[19] C. Sutton, A. Mccallum, "Piecewise Training for Structured Prediction," Machine Learning, **77**, 165–194, 2009, doi:10.1007/s10994-009-5112-z.

[20] D. Venugopal, V. Gogate, "Evidence-Based Clustering for Scalable Inference in Markov Logic," in T. Calders, F. Esposito, E. Hüllermeier, R. Meo, editors, Machine Learning and Knowledge Discovery in Databases, 258–273, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[21] P. Miettinen, T. Mielikäinen, A. Gionis, G. Das, H. Mannila, "The Discrete Basis Problem," in J. Fürnkranz, T. Scheffer, M. Spiliopoulou, editors, Knowledge Discovery in Databases: PKDD 2006, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[22] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, "Equation of State Calculations by Fast Computing Machines," The Journal of Chemical Physics, **21**(6), 1087–1092, 1953.

[23] W. K. Hastings, "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," Biometrika, **57**(1), 97–109, 1970.

[24] A. Nath, P. Domingos, "Efficient Lifting for Online Probabilistic Inference," volume 2, 2010.

[25] M. Gehrke, R. Möller, T. Braun, "Taming Reasoning in Temporal Probabilistic Relational Models," in Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020), 2020, doi:10.3233/FAIA200395.

[26] R. Agrawal, C. Faloutsos, A. Swami, "Efficient similarity search in sequence databases," in Lecture Notes in Computer Science, volume 730, Springer Verlag, 1993.

[27] E. Keogh, K. Chakrabarti, M. Pazzani, S. Mehrotra, "Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases," in Knowledge and Information Systems, 263–286, 2001, doi:10.1021/acsami.7b03579.

[28] S. Kramer, "A Brief History of Learning Symbolic Higher-Level Representations from Data (And a Curious Look Forward)," in Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, 4868–4876, 2020.

[29] J. B. Kruskal, M. Liberman, "The Symmetric Time Warping Problem: From Continuous to Discrete," in Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, Addison-Wesley Publishing Co., 1983.

[30] C.-C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, E. Keogh, "Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets," in 2016 IEEE 16th International Conference on Data Mining (ICDM), 1317–1322, 2016.

[31] F. Petitjean, J. Inglada, P. Gancarski, "Satellite Image Time Series Analysis Under Time Warping," IEEE Transactions on Geoscience and Remote Sensing, **50**(8), 2012.

[32] S. Salvador, P. Chan, "FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space," 70–80, 2004.

[33] D. F. Silva, G. E. A. P. A. Batista, "Speeding Up All-Pairwise Dynamic Time Warping Matrix Calculation," in Proceedings of the 2016 SIAM International Conference on Data Mining, 837–845, Society for Industrial and Applied Mathematics, 2016.

[34] B. Chiu, E. Keogh, S. Lonardi, "Probabilistic Discovery of Time Series Motifs," in Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 493–498, 2003.

[35] C. Bandt, B. Pompe, "Permutation Entropy: A Natural Complexity Measure for Time Series," Physical Review Letters, **88**(17), 4, 2002.

[36] M. Mohr, F. Wilhelm, M. Hartwig, R. Möller, K. Keller, "New Approaches in Ordinal Pattern Representations for Multivariate Time Series," in Proceedings of the 33rd International Florida Artificial Intelligence Research Society Conference, 2020.

[37] K. Keller, T. Mangold, I. Stolz, J. Werner, "Permutation Entropy: New Ideas and Challenges," Entropy, **19**(3), 2017.

[38] A. B. Piek, I. Stolz, K. Keller, "Algorithmics, Possibilities and Limits of Ordinal Pattern Based Entropies," Entropy, **21**(6), 2019.

[39] K. Keller, S. Maksymenko, I. Stolz, "Entropy Determination Based on the Ordinal Structure of a Dynamical System," Discrete and Continuous Dynamical Systems - Series B, **20**(10), 3507–3524, 2015.

[40] I. Stolz, K. Keller, "A General Symbolic Approach to Kolmogorov-Sinai Entropy," Entropy, **19**(12), 2017.

[41] A. Antoniouk, K. Keller, S. Maksymenko, "Kolmogorov-Sinai entropy via separation properties of order-generated σ-algebras," Discrete & Continuous Dynamical Systems, **34**(5), 1793–1809, 2014.

[42] K. Keller, "Permutations and the Kolmogorov-Sinai Entropy," Discrete & Continuous Dynamical Systems, **32**(3), 891–900, 2012.

[43] D. Yang, L. Wu, S. Wang, H. Jia, K. X. Li, "How big data enriches maritime research – a critical review of Automatic Identification System (AIS) data applications," Transport Reviews, **39**(6), 755–773, 2019, doi: 10.1080/01441647.2019.1649315.

[44] R. Bellman, Adaptive control processes: A guided tour, Princeton legacy library, Princeton University Press, 2015.

[45] A. L. Bertozzi, E. Merkurjev, "Chapter 12 - Graph-based optimization approaches for machine learning, uncertainty quantification and networks," in R. Kimmel, X.-C. Tai, editors, Processing, Analyzing and Learning of Images, Shapes, and Forms: Part 2, volume 20 of *Handbook of Numerical Analysis*, 503–531, Elsevier, 2019.

[46] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, X. Xu, "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN," ACM Trans. Database Syst., **42**(3), 2017, doi:10.1145/3068335.

[47] V. Satopaa, J. Albrecht, D. Irwin, B. Raghavan, "Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior," in 2011 31st International Conference on Distributed Computing Systems Workshops, 166–171, 2011, doi:10.1109/ICDCSW.2011.20.

Table 1: Accuracy scores of MOP_4SA in SA_4PG. Further results can be found in the appendix

				DBSCAN				Spectral Clustering				Exact		DBSCAN			Spectral Clustering			Exact	
d	τ	$\delta_{\leq}$	n	$n_{>1}$	$\#_{gr}$	s	D_{KL}	$n_{>1}$	$\#_{gr}$	s	D_{KL}	$\#_{gr}$	s	$\#_{gr}$	s	D_{KL}	$\#_{gr}$	s	D_{KL}	$\#_{gr}$	s
				t = [05, 10)										**t = [10, 15)**							
		0.05	42	33	42	19.6	0.342	17	42	19.1	0.317	306	38.7	66	37.8	0.294	81	37.5	0.621	357	95.2
	1	0.1	23	20	23	15.3	0.281	16	23	19.8	0.036	310	29.8	54	29.9	0.397	64	39.6	0.185	357	78.1
		0.15	19	15	19	14.5	0.441	16	19	14.5	0.126	268	25.1	46	28.4	0.401	64	28.8	0.857	351	69.6
		0.05	16	10	16	14.1	0.578	10	16	13.9	0.172	251	22.8	47	27.6	0.535	56	27.7	0.388	348	65.7
2	2	0.1	35	29	35	17.6	0.458	18	35	17.5	0.062	310	33.2	75	34.9	0.487	69	34.4	0.444	359	87.5
		0.15	5	2	5	12.5	0.724	5	5	12.4	0.014	180	15.8	29	24.4	0.655	48	24.8	0.224	342	50.9
		0.05	27	19	27	15.7	0.240	15	27	15.7	0.026	300	29.1	71	31.3	0.442	57	30.8	0.249	358	78.7
	3	0.1	54	48	54	22.0	0.236	12	54	21.8	0.093	317	41.1	81	42.5	0.506	85	42.4	0.307	361	103.9
		0.15	48	42	48	21.2	0.070	14	48	20.2	0.129	309	37.7	80	41.0	0.182	70	39.0	0.392	355	96.5
		0.05	13	5	13	13.6	0.756	12	13	13.6	0.021	177	17.1	31	26.5	0.682	63	27.3	0.152	339	55.8
	1	0.1	20	11	20	15.5	0.697	19	20	14.5	0.283	194	20.0	39	29.5	0.567	64	28.8	0.561	342	62.0
		0.15	22	12	22	15.2	0.706	21	22	15.1	0.058	195	20.3	40	29.3	0.620	67	30.2	0.207	344	63.7
		0.05	5	2	5	12.6	0.782	5	5	12.4	0.011	165	15.1	28	24.4	0.667	45	24.7	0.205	338	50.1
3	2	0.1	5	2	5	12.5	0.781	5	5	12.4	0.147	165	15.1	28	24.4	0.697	51	24.9	0.197	338	50.0
		0.15	7	2	7	12.8	0.776	7	7	12.7	0.019	168	15.4	25	24.9	0.599	48	25.2	0.246	340	51.3
		0.05	383	2	383	249.7	0.000	1	383	247.8	0.001	385	244.0	384	488.7	0.000	383	485.8	0.005	385	480.0
	3	0.1	13	4	13	13.6	0.763	13	13	13.7	0.040	174	16.8	36	26.4	0.574	60	27.3	0.194	338	55.3
		0.15	19	6	19	14.8	0.694	19	19	14.5	0.111	199	19.8	39	28.6	0.597	64	28.8	0.267	340	61.8
		0.05	186	2	186	79.0	0.445	76	186	79.9	0.047	296	90.9	205	154.3	0.306	203	155.1	0.163	358	214.2
	1	0.1	4	2	4	12.5	0.784	4	4	12.6	0.911	165	15.0	27	24.3	0.676	23	24.5	0.712	338	49.4
		0.15	7	2	7	12.8	0.779	7	7	12.8	0.644	167	15.4	25	24.9	0.694	38	25.0	0.571	338	51.1
		0.05	255	2	255	127.3	0.219	64	255	127.9	0.018	336	137.4	277	251.4	0.155	261	249.1	0.043	366	297.7
4	2	0.1	197	3	197	86.1	0.445	103	197	86.9	0.116	300	97.4	213	168.1	0.381	209	169.1	0.171	363	228.5
		0.15	3	2	3	12.5	0.779	3	3	12.4	0.779	165	14.9	21	24.2	0.699	27	24.2	0.647	338	48.9
		0.05	187	2	187	79.7	0.427	88	187	80.6	0.027	293	89.1	199	155.4	0.277	196	156.3	0.076	360	212.1
	3	0.1	226	3	226	105.5	0.353	79	226	107.6	0.056	315	114.9	240	205.7	0.263	235	208.1	0.088	371	261.3
		0.15	293	2	293	158.5	0.024	35	293	164.7	0.173	362	169.2	299	315.9	0.116	301	316.1	0.256	384	356.4
				t = [15, 20)										**t = [20, 25)**							
		0.05	42	33	231	71.7	0.170	17	232	73.7	0.355	359	154.5	302	123.9	0.150	316	133.1	0.117	362	214.0
	1	0.1	23	20	229	58.9	0.324	16	230	69.4	0.201	359	127.9	312	104.8	0.182	318	115.3	0.116	362	178.4
		0.15	19	15	236	55.2	0.374	16	227	56.3	0.867	353	116.6	305	98.9	1.029	314	100.4	0.584	356	164.2
		0.05	16	10	242	55.1	0.145	10	238	55.3	0.280	351	111.2	301	97.6	0.098	311	98.5	0.173	353	157.0
2	2	0.1	35	29	230	67.9	0.354	18	214	64.9	0.527	359	143.1	305	117.1	0.286	317	114.6	0.313	362	199.6
		0.15	5	2	227	46.7	0.203	5	229	49.4	0.193	348	90.1	301	83.8	0.162	308	87.3	0.102	350	129.9
		0.05	27	19	251	63.8	0.471	15	224	58.6	0.344	360	130.0	305	111.3	0.280	313	104.7	0.324	362	182.2
	3	0.1	54	48	211	78.1	0.392	12	239	80.5	0.237	362	168.6	298	131.4	0.346	330	139.4	0.132	364	233.9
		0.15	48	42	217	75.5	0.168	14	214	71.3	0.443	357	160.2	293	127.5	0.118	315	125.5	0.305	359	221.3
		0.05	13	5	222	49.6	0.264	12	219	53.8	0.147	345	98.8	303	90.3	0.187	307	94.7	0.091	348	142.5
	1	0.1	20	11	222	54.6	0.348	19	224	57.1	0.535	346	108.3	297	98.0	0.291	309	100.4	0.446	349	155.5
		0.15	22	12	225	55.0	0.753	21	222	58.8	0.176	347	111.1	300	99.4	0.908	303	103.0	0.110	351	159.5
		0.05	5	2	233	46.8	0.166	5	248	49.2	0.174	344	89.2	301	84.6	0.193	309	88.1	0.087	347	128.9
3	2	0.1	5	2	230	46.8	0.358	5	231	49.4	0.216	344	89.0	300	84.5	0.774	297	86.9	0.255	347	128.8
		0.15	7	2	217	46.3	0.995	7	228	49.8	0.217	345	91.5	305	84.8	0.402	319	89.4	0.161	348	132.3
		0.05	383	2	384	728.7	0.000	1	387	726.9	0.004	385	714.6	385	968.8	0.000	388	968.8	0.004	385	950.4
	3	0.1	13	4	234	51.0	0.309	13	236	55.0	0.202	344	98.2	301	91.9	0.493	310	96.3	0.135	347	141.8
		0.15	19	6	235	54.2	0.272	19	217	56.6	0.264	345	107.6	313	98.9	0.135	301	99.4	0.144	348	154.4
		0.05	186	2	319	257.8	0.121	76	269	245.3	0.181	361	343.5	337	382.2	0.043	329	361.1	0.128	364	472.5
	1	0.1	4	2	231	46.6	0.272	4	211	45.3	0.307	344	88.0	300	83.8	0.195	301	82.0	0.043	347	127.2
		0.15	7	2	217	46.3	0.307	7	229	49.6	0.257	344	93.4	304	84.7	0.192	295	87.4	0.156	347	134.2
		0.05	255	2	344	401.4	0.059	64	309	382.2	0.039	367	460.5	352	562.2	0.019	341	533.9	0.032	368	624.7
4	2	0.1	197	3	324	275.1	0.697	103	260	264.5	0.158	365	362.6	346	406.3	0.336	324	381.1	0.134	366	497.2
		0.15	3	2	215	45.0	0.251	3	225	45.8	0.273	344	87.1	303	81.6	0.179	297	82.3	0.197	347	126.1
		0.05	187	2	315	253.7	0.482	88	267	248.7	0.068	362	340.1	352	382.7	0.261	319	361.2	0.038	363	468.2
	3	0.1	226	3	324	325.9	0.104	79	290	321.7	0.082	373	413.8	357	472.8	0.052	339	460.3	0.063	373	564.5
		0.15	293	2	341	480.2	0.122	35	333	478.9	0.180	384	545.1	377	663.8	0.053	357	656.8	0.045	385	733.7

Table 2: Results of MOP$_4$SA for further timesteps with $t \geq 25$

				DBSCAN				Spectral Clustering				Exact		DBSCAN			Spectral Clustering			Exact	
d	τ	$\delta_{\leq}$	n	$n_{>1}$	$\#_{gr}$	s	D_{KL}	$n_{>1}$	$\#_{gr}$	s	D_{KL}	$\#_{gr}$	s	$\#_{gr}$	s	D_{KL}	$\#_{gr}$	s	D_{KL}	$\#_{gr}$	s
				t = [25, 30)										**t = [30, 35)**							
	1	0.05	42	33	328	182.7	0.371	17	338	193.2	0.078	363	274.0	334	248.7	0.356	345	255.4	0.073	366	334.9
		0.1	23	20	328	155.2	0.102	16	340	167.2	0.093	362	229.2	338	207.6	0.088	341	220.7	0.094	364	280.7
		0.15	19	15	318	146.5	0.988	16	332	150.4	0.064	356	212.3	322	195.5	0.442	338	201.1	0.052	358	261.9
2	2	0.05	16	10	318	143.7	0.068	10	333	146.2	0.128	355	203.5	322	191.5	0.052	340	195.7	0.110	358	250.6
		0.1	35	29	324	172.1	0.105	18	337	171.8	0.076	362	256.8	334	229.4	0.076	341	230.3	0.064	364	314.5
		0.15	5	2	312	124.5	0.127	5	320	128.8	0.070	352	170.1	314	167.1	0.136	325	171.7	0.068	355	210.8
	3	0.05	27	19	322	162.2	0.324	15	338	157.2	0.292	362	234.5	328	215.1	0.086	348	212.4	0.178	365	287.2
		0.1	54	48	331	194.4	0.233	12	348	205.7	0.070	364	299.1	342	261.1	0.119	352	275.6	0.065	366	365.4
		0.15	48	42	327	187.5	0.083	14	341	188.6	0.124	359	282.9	336	250.5	0.072	349	253.1	0.064	362	349.6
	1	0.05	13	5	313	134.9	0.143	12	323	140.1	0.064	349	187.2	317	180.7	0.134	329	187.2	0.063	352	232.2
		0.1	20	11	313	145.7	0.494	19	332	149.9	0.169	351	203.2	320	194.9	0.440	338	200.9	0.097	354	251.7
		0.15	22	12	314	148.0	0.656	21	326	152.9	0.079	352	208.4	319	198.6	0.189	330	204.3	0.077	355	258.0
3	2	0.05	5	2	311	125.5	0.149	5	320	131.4	0.071	349	169.9	313	167.4	0.145	325	174.4	0.066	352	211.0
		0.1	5	2	310	125.4	0.631	5	313	127.8	0.145	349	169.0	312	167.1	0.590	322	170.2	0.129	352	210.0
		0.15	7	2	311	126.6	0.142	7	335	133.6	0.112	350	173.6	316	169.5	0.122	340	179.1	0.092	353	215.5
	3	0.05	383	2	385	1214.0	0.000	1	388	1210.5	0.004	385	1186.4	385	1458.9	0.000	388	1454.6	0.004	385	1422.8
		0.1	13	4	309	136.1	0.523	13	322	142.5	0.090	349	186.0	312	181.3	0.492	327	189.5	0.089	352	231.0
		0.15	19	6	320	147.0	0.108	19	328	164.4	0.089	350	201.5	322	196.2	0.109	337	226.8	0.077	353	249.3
	1	0.05	186	2	341	508.9	0.037	76	360	490.5	0.108	364	602.1	342	635.8	0.036	372	627.3	0.042	367	732.7
		0.1	4	2	310	124.1	0.148	4	311	122.5	0.503	349	167.1	312	165.6	0.139	317	164.3	0.524	352	207.6
		0.15	7	2	312	126.6	0.143	7	313	129.2	0.105	349	176.1	317	169.8	0.135	319	172.4	0.075	352	218.2
4	2	0.05	255	2	355	725.2	0.015	64	361	698.5	0.025	369	789.8	356	889.8	0.016	368	873.8	0.021	370	956.1
		0.1	197	3	352	540.4	0.313	103	363	516.2	0.115	367	632.7	353	676.1	0.303	368	658.3	0.102	369	772.0
		0.15	3	2	312	122.2	0.134	3	309	122.6	0.156	349	165.4	317	163.8	0.131	311	163.8	0.152	352	205.5
	3	0.05	187	2	354	513.2	0.252	88	346	487.1	0.025	363	597.6	355	652.6	0.254	358	619.3	0.021	365	734.7
		0.1	226	3	360	623.2	0.050	79	361	611.6	0.030	373	715.9	362	775.4	0.047	368	769.2	0.028	375	868.0
		0.15	293	2	384	857.6	0.019	35	369	842.5	0.038	385	924.6	385	1054.2	0.013	374	1031.6	0.032	385	1115.7
				t = [35, 40)										**t = [40, 45)**							
	1	0.05	42	33	340	314.0	0.341	17	350	318.9	0.072	369	396.8	346	378.5	0.330	353	383.4	0.070	371	459.1
		0.1	23	20	346	261.6	0.078	16	342	274.7	0.095	367	333.0	349	316.7	0.074	349	329.3	0.093	370	385.8
		0.15	19	15	329	247.2	0.066	16	339	252.7	0.050	361	311.4	335	298.3	0.065	345	305.2	0.048	363	361.3
2	2	0.05	16	10	327	240.0	0.054	10	343	246.1	0.104	361	298.6	333	289.6	0.052	350	303.8	0.109	363	347.6
		0.1	35	29	338	288.0	0.070	18	345	290.0	0.063	367	373.8	343	347.7	0.069	348	350.8	0.064	369	432.8
		0.15	5	2	320	209.7	0.136	5	330	215.4	0.068	358	252.4	330	253.1	0.124	338	260.1	0.068	360	294.5
	3	0.05	27	19	334	269.6	0.081	15	355	269.0	0.087	367	341.2	340	324.8	0.078	359	326.6	0.082	370	395.3
		0.1	54	48	347	329.5	0.107	12	353	344.8	0.064	369	438.8	352	398.8	0.100	360	415.8	0.063	373	506.7
		0.15	48	42	346	316.1	0.055	14	349	319.3	0.061	365	414.6	353	383.1	0.052	352	385.6	0.063	368	479.4
	1	0.05	13	5	325	227.6	0.131	12	333	241.8	0.062	355	277.9	331	275.4	0.121	341	290.8	0.062	357	324.1
		0.1	20	11	325	245.1	0.416	19	343	253.5	0.066	357	300.9	333	296.5	0.386	347	306.7	0.065	359	350.9
		0.15	22	12	325	249.6	0.093	21	336	256.7	0.074	358	308.3	335	302.0	0.060	346	310.2	0.072	360	359.3
3	2	0.05	5	2	318	210.5	0.140	5	331	218.4	0.064	355	253.1	327	254.1	0.131	340	263.3	0.062	357	295.6
		0.1	5	2	318	209.7	0.568	5	326	213.6	0.132	355	251.9	328	253.4	0.527	332	257.9	0.134	357	298.2
		0.15	7	2	322	213.3	0.113	7	344	227.6	0.094	356	259.6	331	258.3	0.108	347	286.1	0.095	358	302.9
	3	0.05	383	2	385	1701.7	0.000	1	389	1699.4	0.003	385	1659.9	385	1944.8	0.000	390	1943.8	0.003	385	1898.4
		0.1	13	4	317	227.3	0.472	13	332	237.5	0.091	355	276.8	327	274.4	0.443	342	286.3	0.095	357	323.4
		0.15	19	6	328	246.2	0.111	19	341	285.6	0.077	356	299.6	335	298.8	0.101	349	338.4	0.075	358	349.2
	1	0.05	186	2	345	764.2	0.031	76	374	766.0	0.023	367	864.1	347	893.2	0.031	376	913.4	0.022	369	1003.2
		0.1	4	2	317	207.7	0.134	4	323	207.2	0.519	355	248.8	326	250.8	0.125	332	251.0	0.465	357	290.7
		0.15	7	2	323	213.8	0.134	7	323	216.7	0.076	355	261.9	331	259.1	0.123	330	261.7	0.149	357	305.5
4	2	0.05	255	2	356	1055.1	0.016	64	372	1045.3	0.021	370	1124.5	359	1221.2	0.015	376	1218.6	0.021	371	1291.4
		0.1	197	3	355	812.6	0.295	103	371	804.4	0.098	370	914.6	358	951.2	0.273	373	948.6	0.097	372	1067.9
		0.15	3	2	323	206.2	0.129	3	317	205.7	0.147	355	246.3	332	249.8	0.117	327	248.9	0.137	357	287.7
	3	0.05	187	2	356	785.1	0.245	88	360	753.9	0.020	368	865.9	358	917.9	0.238	366	889.9	0.019	368	998.0
		0.1	226	3	364	929.7	0.046	79	373	929.2	0.028	377	1021.8	366	1084.0	0.045	376	1090.0	0.027	378	1175.8
		0.15	293	2	385	1251.4	0.012	35	377	1223.1	0.030	385	1304.6	385	1447.7	0.012	379	1416.0	0.028	385	1494.3

Table 3: Distances as a result of running MOP$_4$SCD between consecutive timesteps

d	τ	$\delta_{\leq}$	$d(\mathcal{W}^t, \mathcal{W}^{t+1}, \mathcal{S})$ **5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
2	1	0.05	.067	.067	.068	.065	.066	.068	.067	.066	.063	.065	.069	.068	.062	.061	.064	.069	.064	.067	.068
		0.1	.102	.112	.116	.107	.114	.115	.109	.114	.111	.105	.101	.11	.111	.116	.11	.109	.114	.107	.113
		0.15	.135	.132	.142	.125	.136	.142	.142	.138	.14	.135	.137	.133	.139	.132	.129	.135	.131	.134	.143
	2	0.05	.105	.113	.114	.116	.113	.115	.113	.114	.115	.115	.116	.113	.117	.114	.115	.112	.113	.113	.116
		0.1	.068	.068	.069	.073	.068	.074	.061	.082	.071	.071	.077	.08	.073	.065	.075	.072	.082	.074	.081
		0.15	.177	.18	.18	.185	.18	.184	.178	.185	.18	.182	.184	.181	.181	.181	.178	.184	.185	.184	.188
	3	0.05	.079	.082	.081	.082	.086	.092	.087	.08	.09	.083	.079	.089	.08	.085	.084	.089	.086	.082	.088
		0.1	.037	.038	.037	.039	.039	.039	.037	.035	.038	.039	.034	.039	.038	.037	.036	.038	.038	.038	.038
		0.15	.044	.043	.043	.046	.046	.044	.043	.045	.044	.045	.044	.046	.048	.046	.044	.044	.046	.046	.05
3	1	0.05	.199	.201	.2	.2	.199	.201	.2	.201	.201	.201	.201	.19	.201	.2	.201	.2	.2	.2	.201
		0.1	.214	.219	.219	.218	.219	.218	.217	.216	.216	.218	.218	.217	.217	.217	.219	.217	.218	.217	.219
		0.15	.213	.218	.216	.217	.218	.219	.218	.219	.217	.218	.218	.218	.218	.217	.218	.219	.218	.218	.217
	2	0.05	.238	.239	.239	.24	.24	.24	.239	.24	.24	.24	.24	.241	.239	.239	.238	.239	.24	.239	.24
		0.1	.237	.239	.238	.239	.239	.239	.239	.24	.239	.24	.239	.239	.239	.238	.239	.239	.24	.239	.24
		0.15	.236	.237	.238	.238	.239	.238	.239	.238	.238	.239	.238	.238	.237	.237	.238	.238	.238	.238	.238
	3	0.05	.001	.001	.0	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001
		0.1	.232	.231	.23	.233	.232	.233	.235	.233	.234	.233	.233	.232	.233	.233	.233	.235	.234	.233	.234
		0.15	.223	.226	.223	.228	.227	.227	.226	.227	.23	.227	.227	.227	.229	.228	.228	.229	.226	.227	.228
4	1	0.05	.056	.055	.056	.055	.056	.056	.056	.055	.055	.056	.056	.056	.056	.056	.056	.056	.056	.056	.056
		0.1	.176	.178	.178	.178	.178	.178	.178	.176	.178	.178	.178	.178	.178	.178	.178	.178	.178	.178	.178
		0.15	.211	.21	.213	.213	.213	.213	.213	.213	.212	.212	.213	.213	.213	.213	.213	.213	.213	.213	.213
	2	0.05	.033	.034	.035	.034	.035	.034	.035	.035	.035	.035	.035	.035	.035	.035	.035	.035	.035	.035	.035
		0.1	.075	.076	.076	.076	.076	.076	.076	.075	.075	.076	.076	.076	.075	.076	.076	.075	.076	.075	.076
		0.15	.218	.216	.218	.218	.218	.218	.218	.218	.218	.218	.218	.218	.218	.218	.218	.218	.218	.218	.218
	3	0.05	.064	.064	.064	.065	.065	.064	.065	.065	.065	.065	.065	.065	.065	.065	.065	.064	.065	.065	.065
		0.1	.049	.05	.049	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.049	.05	.05	.05
		0.15	.016	.015	.015	.016	.016	.016	.016	.016	.016	.016	.016	.016	.016	.016	.016	.016	.016	.016	.016

d	τ	$\delta_{\leq}$	$d(\mathcal{W}^t, \mathcal{W}^{t+1}, \mathcal{S})$ **24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**
2	1	0.05	.066	.067	.064	.064	.064	.068	.067	.064	.063	.065	.066	.067	.062	.068	.066	.067	.066	.067	.067
		0.1	.116	.11	.103	.103	.115	.115	.108	.11	.107	.115	.109	.107	.114	.117	.107	.116	.112	.115	.114
		0.15	.135	.135	.139	.137	.139	.136	.132	.143	.133	.143	.132	.134	.129	.137	.137	.142	.13	.141	.14
	2	0.05	.111	.115	.113	.114	.111	.11	.11	.112	.108	.112	.11	.111	.112	.112	.115	.114	.11	.113	.109
		0.1	.078	.079	.073	.077	.077	.067	.074	.08	.068	.072	.065	.076	.072	.074	.076	.073	.068	.073	.067
		0.15	.184	.185	.182	.182	.182	.182	.178	.183	.172	.184	.175	.179	.183	.177	.184	.186	.176	.187	.18
	3	0.05	.075	.087	.08	.086	.09	.085	.091	.086	.089	.083	.091	.076	.086	.088	.082	.091	.087	.088	.092
		0.1	.039	.039	.036	.036	.039	.037	.039	.038	.037	.036	.036	.037	.036	.038	.035	.038	.037	.035	.039
		0.15	.046	.044	.045	.045	.048	.048	.048	.046	.045	.043	.047	.042	.045	.049	.048	.048	.049	.044	.048
3	1	0.05	.201	.2	.2	.2	.201	.198	.199	.2	.198	.201	.191	.2	.19	.2	.199	.2	.2	.2	.198
		0.1	.218	.218	.217	.217	.217	.218	.217	.217	.216	.217	.218	.217	.218	.217	.216	.217	.215	.219	.217
		0.15	.218	.218	.218	.218	.219	.217	.217	.218	.217	.216	.216	.217	.219	.217	.218	.216	.218	.219	.216
	2	0.05	.24	.239	.24	.239	.239	.239	.238	.239	.237	.239	.239	.237	.24	.238	.238	.24	.239	.24	.239
		0.1	.239	.239	.239	.239	.239	.239	.238	.239	.238	.237	.238	.238	.24	.237	.238	.239	.239	.239	.239
		0.15	.238	.237	.238	.237	.237	.238	.236	.238	.236	.237	.237	.237	.238	.236	.237	.238	.236	.238	.236
	3	0.05	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001
		0.1	.233	.234	.232	.235	.233	.232	.233	.231	.233	.233	.233	.231	.232	.232	.234	.235	.233	.232	.235
		0.15	.229	.229	.228	.228	.228	.227	.228	.227	.226	.228	.226	.227	.226	.225	.228	.228	.228	.228	.226
4	1	0.05	.056	.055	.056	.056	.055	.055	.055	.055	.055	.055	.054	.055	.056	.056	.056	.056	.056	.056	.056
		0.1	.178	.178	.178	.178	.178	.178	.178	.178	.178	.178	.178	.178	.178	.178	.178	.178	.178	.178	.178
		0.15	.213	.213	.213	.213	.213	.213	.213	.213	.213	.213	.213	.213	.213	.213	.213	.213	.213	.213	.213
	2	0.05	.035	.035	.035	.034	.035	.034	.035	.035	.035	.034	.035	.034	.035	.034	.035	.035	.035	.035	.035
		0.1	.075	.076	.076	.076	.076	.076	.076	.075	.076	.074	.076	.075	.076	.076	.076	.076	.076	.076	.075
		0.15	.218	.218	.218	.218	.218	.216	.218	.218	.218	.218	.218	.218	.216	.218	.218	.218	.218	.218	.218
	3	0.05	.065	.065	.063	.065	.065	.065	.065	.065	.065	.065	.065	.065	.064	.064	.065	.065	.064	.065	.065
		0.1	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05
		0.15	.016	.016	.016	.016	.016	.016	.016	.016	.016	.016	.016	.016	.016	.016	.016	.016	.016	.016	.016

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 1, 79-89 (2022)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

A Novel Algorithm Design for Locating Fault Distances on HV Transmission Lines

MK Ngwenyama[1,*], PF Le Roux[1], LJ Ngoma[2]

[1]Electrical Engineering Department, Tshwane University of Technology, Witbank, 1034, South Africa

[2]Electrical Engineering Department, Tshwane University of Technology, Pretoria, 0183, South Africa

ARTICLE INFO

Article history:
Received: 25 November, 2021
Accepted: 26 January, 2022
Online: 21 February, 2022

Keywords:
Conventional method
Electrical energy
Electrical fault
Impedance-based technique
MATLAB/SIMULINK
Transmission line
Transmission network

ABSTRACT

The transmission network has been considered among the globe's prevalent complex systems, comprised of hundreds of electrical transmission lines and other equipment used to transmit electrical energy from one location to another. Over a decade, power engineers have worked tirelessly to ensure that the transmission network operates reliably, transmitting electrical energy from the power station to the consumers without interruption. With growing generation capacity and the recent introduction of renewable energy systems (RES) such as wind turbines and solar energy, the transmission lines are increasingly being forced to run near their design limitations and greater unpredictability on the network operational configuration. As a result, the transmission network faces greater challenges than previously. As a worst-case scenario, large-scale electrical network power outages caused by electrical faults can disrupt electricity availability for several hours, impacting millions of customers and inflicting massive economic damage. These electrical faults must be repaired before electricity is restored to consumers. This necessitates a thorough grasp of the challenge and potential remedies to assure improved power efficiency. In the present work, an expansion of preceding work, a novel algorithm for estimating faults on transmission lines is presented. Impedance-based techniques are susceptible to producing errors or incorrect predictions. The presence of faults induced from high impedance sources produces an extra impedance to the ground, which negates the impedance calculation and produces errors in the distance to the fault. This results in inaccuracies that can affect a distance-to-fault estimation by 1-15 % of the overall line length. In this work, a design of a fault detection-location element (FDLE) algorithm is proposed. This algorithm relies on the dynamics of current and voltage signals on the transmission line while deserting impedance. Comparison research is undertaken against the impedance-based techniques to validate the proposed algorithm. Finally, the proposed algorithm findings are compared to fault location estimations using an impedance-based technique. Extensive trials on a simulated transmission line prove that the proposed algorithm is responsive to faults with an error as low as 1%, reaching a precision of 98.9%.

1. Introduction

This paper is an extension of work originally presented at the *2021 6th Asia Conference on Power and Electrical Engineering (ACPEE)* [1]. The transmission network has been considered among the globe's prevalent complex systems, comprised of hundreds of electrical transmissions such as the addition of wind turbines and solar energy. Therefore, the transmission network faces greater challenges than previously. Over a decade, power designers have worked diligently to guarantee that the transmission network runs adequately, transmitting electrical energy from the power station to customers without disruption. The demand for transmission lines is increasing as the penetration of renewable energy systems (RESs), electric vehicles (EVs), and energy storage systems (ESSs) grows in modern transmission networks [2]. To achieve dependable, efficient, and reliable RESs, significant improvement in the conventional transmission line protection method is necessary. The rapid increase of current during a fault doubles the potential risk of transmission network degradation and thus presents the requirement that faults

*Corresponding Author: MK Ngwenyama, South Africa, NgwenyamaMK@yahoo.com

in RESs be separated by a protection method with the shortest operational time. Due to the increased dependability and stability of RESs, the ring model is the ideal design for RESs because of the bidirectional flow of current in these configurations; the standard protection system would not perform adequately. As a worst-case scenario, large-scale electrical network power outages caused by electrical faults can disrupt electricity availability for several hours, impacting millions of customers and inflicting massive economic damage. These electrical faults must be repaired before electricity is restored to consumers. The introduction of the proposed algorithm may aid in the improvement of these systems.

Many electric utilities have opted to invest minimal, or no effort in electrical transmission fault detection devices in the past, alleging that faults are usually temporary thus do not necessitate location. Moreover, [3] suggested that the data produced by protection relays were not accurate enough to support deploying personnel to check temporary fault locations. As a result, fault location operations were commenced immediately after a fault had become permanent and a solution was required. With the restructuring and privatisation of electric power production enterprises across the globe, such sentiments are beginning to change [4, 5]. To attain the greater plant efficiency levels and increased service reliability regulations that a dynamic industry needs, electricity utilities are increasingly taking a more "pro-active" strategy to most business operations, notably in the field of transmission line fault location. The capability of impedance-based techniques to address the errors and constraints of protection relays was discovered in the mid-20th century when various trial designs were introduced; several of them were converted into industrial uses [6]. Although such designs are more accurate than other methods, they were discontinued due to dependability and maintenance issues, leading to a lack of involvement and credibility. Impedance-based fault location techniques are used repeatedly on transmission lines [7]. Many causes drive the resurgence of impedance-based techniques, the foremost of which is an industry need for quick and precise fault location on economically vital, extremely lengthy high-voltage transmission lines. Precise fault location lowers operational expenses by eliminating costly and time-consuming inspections [8]. Precise fault location speeds up line maintenance and recovery, decreasing revenue shortfall due to outages. Comparison research is undertaken against other techniques to validate the proposed algorithm.

In [9], the authors proposed an artificial intelligence-based approach for detecting faults in photovoltaic (PV) plants and transmission lines. The method is known as recurrent neural networks (RNN). The authors stress that faults in PV plants are usually foreseeable, and while conventional methods are precise, they are not cost-effective for mass application, allowing most PV plants to operate unmonitored. They developed an algorithm for precise fault diagnosis of PV plants that uses satellite weather data and low-frequency inverter sensors. The algorithm enables machine learning-based fault detection even for PV plants with the absence of detectors, and it utilises a recurrent neural network to detect all forms of defects based on the last 24 hours of observations, rather than just the latest detected fault. The authors also demonstrated that the proposed algorithm determines the generated power loss induced by the fault, for example, the fault level, whereas the traditionally utilised techniques are restricted to classifying the fault type. The results show that the algorithm is responsive to as low as 5 % intensity defects, with a precision of 96.9% utilising accurate climate data and 86.4% utilising satellite weather data. The results reveal that the algorithm could also detect unspecified defects, that is, defects that appeared to be not included in the training data.

The authors in [10] presented a closed-loop sinusoidal pulse width modulation (SPWM) monitoring device for the operations of converters utilised in wind turbine plants for utility grid purposes. Wind energy is a major form of clean energy; therefore, windmills are commonly deployed in electrical distribution systems and thus are coupled physically to electrical transmission systems. Wind energy significantly influences the performance of current configuration systems as network saturation improves, resulting in the probability of faults on the transmission lines transporting power to the main grid. The presented technique is utilised to investigate wind turbine efficiency. The authors state that voltage and current inverters produce discrete output waveform signals, harmonic pollution, extra power losses, and higher frequency noises. As a result, to achieve the required current waveforms, large inductors must be connected in line with the related load. The simulated results showed that the mentioned technique has a higher harmonic elimination strength. As a result, the transmission network will require fewer fault detecting and locating devices.

In [11], the authors presented a technique for performing serial fault repair in a wind turbine plant with synchronous actuator and sensor defects. The technique is known as a robust sliding mode observer (SMO). The authors used a diagonal transform matrix and a post-filtering system to design a novel improved model. The novel model then transforms the sensor fault into an actuator fault to locate the fault. The novel model allows the SMO technique to detect and pinpoint faults on transmission lines. The simulated tests indicate how the presented technique can properly restore actuator and sensor defects; thus, the active fault controller can obtain optimum wind power extraction. Although this technique works optimally for a microgrid that operates in an islanded mode, there is no concise evidence of how this technique functions when employed to a transmission line of a microgrid configuration integrated into the main grid.

The theories of the single and double-ended impedance-based fault location techniques are also discussed in this paper and explain the concepts related to fault location and present different impedance-based fault location algorithms. The work aims to investigate numerous network faults and assess the efficiency of fault locators in light of potential causes of inaccuracy [12].

In the present work, an expansion of preceding work [1], a novel algorithm for estimating faults on transmission networks is presented. Impedance-based techniques are susceptible to producing errors or incorrect predictions. The presence of faults induced from high impedance sources produces an extra impedance to the ground, which negates the impedance calculation and produces errors in the distance to the fault. This results in inaccuracies that can affect a distance-to-fault estimation by 1 -15 % of the overall line length. This will have a detrimental impact on the transmission network and renewable

energy systems. In this work, a design of a fault detection-location element (FDLE) algorithm is proposed. The model utilises MATLAB/SIMULINK software to study the electrical faults on a transmission line. The proposed algorithm relies on the dynamics of current and voltage signals on the transmission line while deserting impedance. The authors provide simulation results to demonstrate the efficiency of the proposed algorithm. The FDLE algorithm findings are compared to fault location estimations using an impedance-based approach. Extensive trials on a simulated transmission line prove that the proposed algorithm is responsive to faults with an error as low as 1%, reaching a precision of 98.9%.

2. Techniques and Specifications for Fault Diagnosis

Various techniques of predicting fault position are currently in practice in the industry [13]:

- Conventional methods:
  - Impedance-based methods [14]:
    - Single-ended method, and
    - Double-ended method (Synchronized, Unsynchronized, and Unsynchronized Current-only).
  - Travelling wave methods [15, 16]:
    - Single-ended method, and
    - Double-ended method.
- Artificial Intelligence methods:
  - Artificial Neural Network (ANN) [17],
  - Support Vector Machine (SVM) [18],
  - Fuzzy Logic [19], and
  - Matching Approach [20].

In this work, comparison research is undertaken against the impedance-based techniques to validate the proposed algorithm and presents the outcomes using real-world simulated faults.

3. Theoretical Foundations of Impedance-based Fault Location Techniques

Various impedance-based fault location techniques have been designed for transmission lines purposes. Single-ended techniques are fault-locating techniques that use information recorded by a fault location detection element at one end-point of the transmission line to locate faults [21]. Double-ended techniques use information from fault location detection elements at both end-points of the connection. The impedance between both the fault locator and the position of the fault is estimated using voltage and current signals collected by fault location detection elements during a fault [22]. The distance to the fault may be precisely determined if the transmission line impedance in ohms is known. When determining the proximity to a fault, each technique has different basic data needs and requires different estimates. These estimates may or may not be valid in a specific fault position case [23]. Selecting the appropriate technique for pinpointing faults within quite a diverse set of impedance-based fault location techniques is thus a challenging effort that demands a diligent grasp of the essential principles of an individual fault-locating technique. Constructed on the preceding knowledge, this paper discusses the fundamental concept of single-ended and double-ended impedance-based fault location techniques. The aims are to properly articulate each fault-finding technique's input data requirements, explain each technique's applicability in pinpointing actual transmission line faults, and offer advice on selecting the optimal fault-locating technique.

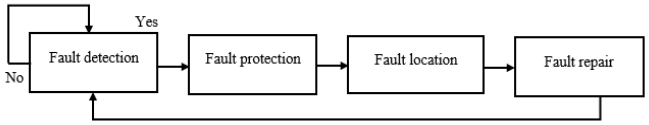


Figure 1: Flow Chart Procedure of a Conventional Technique.

4. Techniques and Prerequisites for Impedance-based Fault Location

The underlying analysis is required for impedance-based techniques:

- Voltage and current amplitudes should be recorded,
- Extract the essential elements,
- Identify the fault category and faulted phasor/s, and
- Employ the impedance-based technique.

Single-ended impedance techniques employ a simplified approach. Hence, links between communicating and off-site data are typically unnecessary [24]. Whereas double-ended techniques are extremely precise. However, they require data from both end-points of the transmission line. For this technique to be executed [25].

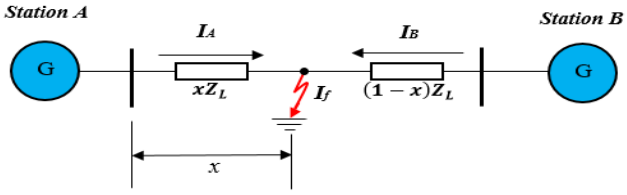


Figure 2: A Single-ended Impedance-based Algorithm.

Table 1: Different Formulae for Fault Types [1].

Fault Type	Positive-Sequence Impedance Equation *mZ1L is equal to:*
A-earth	$V_a/(I_a + (k \times 3 \times I_O))$
B-earth	$V_b/(I_b + (k \times 3 \times I_O))$
C-earth	$V_c/(I_c + (k \times 3 \times I_O))$
A-B / A-B-earth	V_{ab}/I_{ab}
B-C / B-C-earth	V_{bc}/I_{bc}
C-A / C-A-earth	V_{ca}/I_{ca}
A-B-C	*Any of the following:* V_{ab}/I_{ab}, V_{ab}/I_{ab}, V_{ca}/I_{ca}

The single-ended impedance-based fault location technique determines the fault position using the detected impedance through examining the transmission line from a single end-point. Every line's line-to-line and line-to-ground output signal must be recorded to pinpoint all fault forms [12]. The impedance equations shown in Table 1 may be utilised to approximate the fault position assuming zero fault resistance.

Where V_A, V_B, and V_C are the phase voltages,

I_A, I_B, and I_C are the phase currents,

Z_O the zero-sequence impedance,

Z_1 the positive sequence impedance, and

K_O the residual compensation factor and may be expressed as:

$$K_O = (Zo - Z_1)\, K * Z_1 \quad (1)$$

K may be 1 or 3, depending on the fault location detection element design. According to Kirchhoff's law, the zero-sequence Io in Equation (2) will be zero in a balanced system.

$$Io = (I_A + I_B + I_C)/3 \quad (2)$$

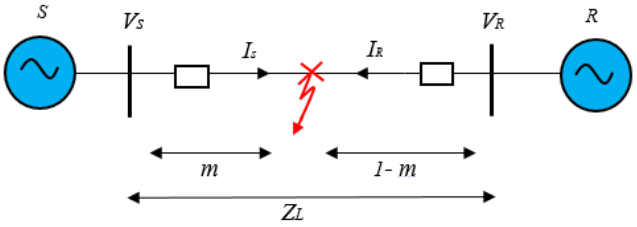


Figure 3: A Double-ended Impedance-based Algorithm.

Where m in Figure 3 denotes the distances to a fault per unit [26], V_S the sending voltage, V_R the receiving voltage, I_S the sending current, I_R the receiving current, and Z_L the impedance of the line.

The difficulties in accurately locating faults using single-ended techniques are recognised and documented in numerous publications [27].

To summarise, the preceding circumstances may result in single-ended impedance-based fault location techniques producing errors:

- Imprecise fault-type (faulted phase/s) detection.
- Incorrect line parameters that do not correspond to actual parameters.
- Lack of accuracy of the line model.
- Current and voltage transformer errors.

5. Single-ended Impedance-based Method

5.1. Simple Reactance Technique

The potential difference loss at the sending (S) end of the transmission line, illustrated in Figure 4 is:

$$V_S = (m \times Z_{1L} \times I_S) + (R_f \times I_f) \quad (3)$$

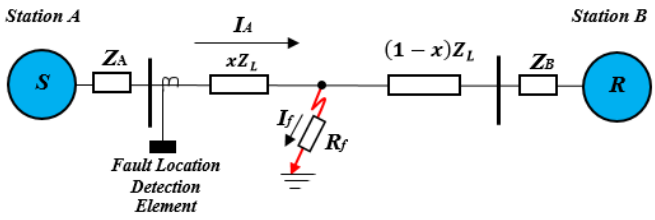


Figure 4: Faulted Transmission Network with Fault Location Detection Element.

For an A phase to earth fault, it must have the following characteristics,

$$V_S = V_{a-g} \text{ and } I_S = I_a + (k \times 3 \times I_O) \quad (4)$$

The objective is to reduce the impact of the $R_f \times I_f$ variable. The basic reactance algorithm divides all variables by I_S (I obtained at the fault location) and excludes the variable $R_f \times (I_f/I_S)$.

To achieve this, preserve the imaginary component and solve m.

$$I_m(V_s/I_s) = I_m(m \times Z_{1L}) = m \times X_{1L} \quad (5)$$

$$m = \frac{I_m\left(\frac{V_S}{I_S}\right)}{X_{1L}} \quad (6)$$

Error is 0 if $\angle I_s = \angle I_f$ or $R_f = 0$

5.2. Takagi Technique

The Takagi technique needs the collection of pre-fault and fault data. It mostly enhances the simple reactance technique by minimising the influence of fault resistance and lowering the impact of load flow [28].

$$V_s = (m \times Z_{1L} \times I_f) + (R_f \times I_f) \quad (7)$$

Superposition current (I_{sup}) may be used to identify a variable that is in conjunction with I_f:

$$I_{sup} = I_f - I_{pre} \quad (8)$$

$$I_f = Fault\ Current \quad (9)$$

$$I_{pre} = Pre - Fault\ Current \quad (10)$$

Potential difference loss produced by the sending bus:

$$V_s = (m \times Z_{1L} \times I_s) + (R_f \times I_f) \quad (11)$$

Multiply both sides of (7) with the complex conjugate of I_{sup} (I^*_{sup}) and keep the imaginary component. Secondly, solve m:

$$I_m(V_s \times I_{sup^*}) = [(m \times I_m)(Z_{1f} \times I_s \times I_{sup^*})] + [(R_f I_m)\ (I_f \times I_{sup^*})] \quad (12)$$

$$m = \frac{I_m(V_s \times I_{sup^*})}{I_m(Z_{1L} \times I_S \times I_{sup^*})} \quad (13)$$

The fact that I_s and I_f angles are the equivalents that are important to the Takagi technique's efficiency. These angles are the same in a perfect homogenous network. The inaccuracy in the fault position estimation improves as the angle between I_s and I_f improves [29].

5.3. Modified Takagi Technique

For earth faults, the modified Takagi utilises zero-sequence current from the sending end $(3 \times I_{os})$ rather than the superposition current. As a result, no pre-fault data is required for this technique. Angle rectification is similarly possible with the Modified Takagi technique. Provided that the user understands the network source impedances, he/she can change the zero-sequence current using angle T to enhance the fault position estimation for a specific transmission line [30].

$$m = \frac{I_m(V_S \times (3 \times I_{OS})^* \times e^{-jT})}{I_m(Z_{1L} \times I_S \times (3 \times I_{OS})^* \times e^{-jT})} \quad (14)$$

The defined angle T is only applicable for a single fault point within the transmission line. Figure 5 demonstrates methods to compute T.

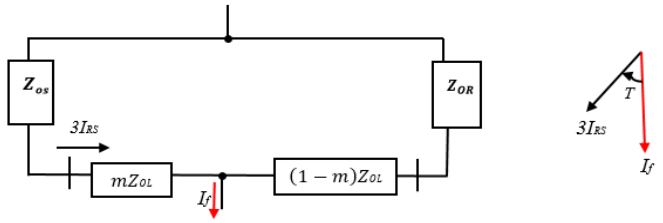


Figure 5: Correction of a Zero Sequence Angle.

$$\frac{I_f}{3 \times I_{RS}} = \frac{Z_{OS}+Z_{OL}+Z_{OR}}{(1-m) \times (Z_{OL}+Z_{OR})} = A \angle T \quad (15)$$

Even though the Modified Takagi technique outperforms the Takagi technique, the precision of position estimations is dependent on having the source impedance characteristics precisely.

5.4. Eriksson Technique

This technique employs the generated impedance characteristics to avoid any reactance inaccuracy induced from fault resistance, loads, or non-homogeneity network to determine the distances to the fault. Furthermore, such a technique evaluates the magnitude of fault impedance, which is important in detecting the main reason for a fault and verifying a transmission network short-circuit design [31].

Equation (16) summarises and updates the variables in Equation (14).

$$m^2 - (k_1 m) + k_2 - (k_3 R_F) = 0 \quad (16)$$

Wherein k_1, k_2, and k_3 represent complex algebraic expressions of a transmission line voltage, current, impedance, and supply impedances; thus, can be presented as follows:

$$k_1 = a + jb = 1 + \left(\frac{Z_{H1}}{Z_{L1}}\right) + \left(\frac{V_{G1}}{Z_{L1} \times I_G}\right)$$

$$k_2 = c + jd = \left(\frac{V_{G1}}{Z_{L1} \times I_G}\right)\left(1 + \frac{Z_{H1}}{Z_{L1}}\right)$$

$$k_3 = e + jf = \left(\frac{\Delta I_{G1}}{Z_{L1} \times I_G}\right)\left(1 + \frac{Z_{H1}+Z_{G1}}{Z_{L1}}\right)$$

By dividing (16) into actual and unreal components, total distance to fault m may be calculated using the quadratic formula (17).

$$m = \frac{\left(a - \frac{eb}{f}\right) \pm \sqrt{\left(a - \frac{eb}{f}\right)^2 - 4\left(c - \frac{ed}{f}\right)}}{2} \quad (17)$$

Should m be known, any other unknown variable may be obtained using Equation (17). Given that the fault position prediction has to be smaller than the entire transmission line distance, an m value of between 0 and 1 per unit must be utilised for the distance measurement.

Therefore, Equation (18) may be used to determine fault tolerance:

$$R_f = \frac{d - mb}{f} \quad (18)$$

Assuming that the supply impedance Z_{G1} is not known, and the load impedance Z_{H1} is precisely calculated, the impedance Z_{G1} can be estimated using fault occurrence records as:

$$Z_{G1} = -\frac{V_{G1} - V_{G1pre}}{I_{G1} - I_{G1pre}} \quad (19)$$

5.5. Novosel et al. Technique

The Novosel et al. technique is a revised edition of the Eriksson technique for pinpointing faults on a compact, radial transmission line. At downstream of the transmission line, all the loads supplied or connected to the line are combined. Given that there is a fixed load impedance design, the initial action is to determine the load impedance using the pre-fault current and voltage as inputs [14]; this is expressed as:

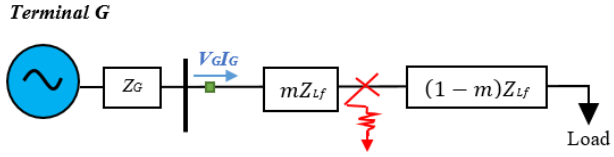


Figure 6: Novosel et al. Technique Applied to a Transmission Network with a Constant Impedance Load

$$Z_{Load} = R + jX = \frac{V_{G1pre}}{I_{G1pre}} - Z_{L1} \quad (20)$$

Resolving the quadratic Equation in (17) results in the distance to the fault per unit, whereby the variables are specified as:

$$k_1 = a + jb = 1 + \left(\frac{Z_{Load}}{Z_{L1}}\right) + \left(\frac{V_{G1}}{Z_{L1} \times I_G}\right)$$

$$k_2 = c + jd = \left(\frac{V_{G1}}{Z_{L1} \times I_G}\right) + \left(1 + \frac{Z_{Load}}{Z_{L1}}\right)$$

$$k_3 = e + jf = \left(\frac{\Delta I_{G1}}{Z_{L1} \times I_G}\right) + \left(1 + \frac{Z_{Load}+Z_{G1}}{Z_{L1}}\right)$$

Similar to the Eriksson technique, *m* can include any of these variables on the Novosel et al. technique. Given that the fault position prediction must be less than the total transmission line distance, several *m* between 0 and 1 per unit must be used for distance measurement.

6. Double-ended Impedance-based Method

6.1. Synchronised Two-ended Technique

This technique is built on the hypothesis that data from both end-points of a transmission line are coordinated to a similar period response using a global positioning system (GPS). During fault position analysis, any three symmetric variables can be employed. However, utilising negative-sequence elements is highly beneficial because it is unaffected by load current or zero-sequence connected loads feeding to the system. To demonstrate the fault-finding concept, a negative-sequence system amid an imbalanced fault must be considered [32].

Figure 7 illustrates how to compute V_{F2}, the negative-sequence potential difference at the fault location *F*, using Terminals *G* and *H*. The Equations are represented as:

$$\textit{Terminal G: } V_{F2} = V_{G2} - (mZ_{L2}I_{G2}) \tag{21}$$

$$\textit{Terminal H: } V_{F2} = V_{H2} - (1-m)(Z_{L2}I_{H2}) \tag{22}$$

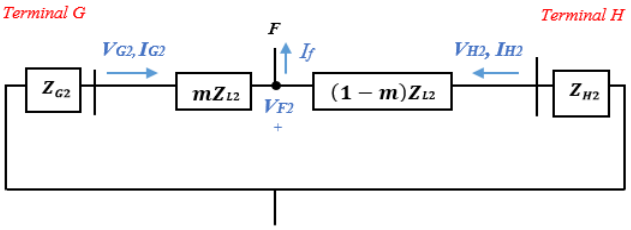


Figure 7: Negative Sequence System during Asymmetrical Fault.

The potential difference V_{F2} determined on either bus of the transmission line is equivalent. As a result of matching Equations (21) and (22), the distance measurement to a fault is calculated as:

$$m = \frac{V_{G2} - V_{H2} + (Z_{L2}I_{H2})}{(I_{G2} + I_{H2})Z_{L2}} \tag{23}$$

A symmetrical fault may be located using Equation (23). On the other hand, negative-sequence components do not occur when a symmetrical fault is present. In this scenario, the identical fault pinpointing approach is used for the positive sequence system, and the fault distance is calculated as:

$$m = \frac{V_{G1} - V_{H1} + (Z_{L1}I_{H1})}{(I_{G1} + I_{H1})Z_{L1}} \tag{24}$$

6.2. Unsynchronised Two-ended Technique

In this technique, current and voltage signals recorded by fault location detection elements at opposite ends of a transmission line could be out of synchronisation. The GPS sensor could be damaged or barely working properly. Fault location detection elements can also have various sample speeds or pinpoint the fault at marginally varying time intervals [33].

A phase shift can be caused by the communication route that transmits data from one fault location detection element to the other. As a result, $e^{j\delta}$ must be employed as a synchronising controller to synchronise the voltages and currents measured at Terminals G and H. This process is as follows:

$$\textit{Terminal G: } V_{Fi} = V_{Gie^{j\delta}} - \left(mZ_{Li}I_{Gie^{j\delta}}\right) \tag{25}$$

$$\textit{Terminal H: } V_{Fi} = V_{Hi} - (1-m)(Z_{Li}I_{Hi}) \tag{26}$$

Asymmetrical faults are calculated using the negative-sequence element, whereas symmetrical faults are calculated using the positive-sequence element. The sync-controllers, as indicated in Equation (27), are formed by matching Equations (25) and (26).

$$\left|e^{j\delta}\right| = 1 = \left|\frac{V_{Hi} - (1-m)(Z_{Li}I_{Hi})}{V_{Gi} - (mZ_{Li}I_{Gi})}\right| \tag{27}$$

The distance measurement to fault *m* is an algebraic expression obtained after simplification and arranging the variables.

$$m = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \tag{28}$$

Where the parameters are specified as follows:

$$A = |Z_{Li}I_{Gi}|^2 - |Z_{Li}I_{Hi}|^2$$

$$B = -2 \times Re[V_{Gi}(Z_{Li}I_{Gi})^* + (V_{Hi} - Z_{Li}I_{Hi})(Z_{Li}I_{Hi})^*]$$

$$C = |V_{Gi}|^2 - |V_{Hi} - Z_{Li}L_{Hi}|^2$$

When the algebraic expression in (28) is solved, two parameters of *m* are obtained. The fault position estimation must have a value of 0 to 1 per unit.

6.3. Unsynchronised Current-only Two-ended Technique

This technique is used when there are data accessibility constraints, assuming that merely current signals at buses *G* and *H* are provided for fault analysis and the voltage signals V_{G2} and V_{H2} are either absent or unavailable. V_{F2} is determined from the two buses using just the current as well as the supply impedance characteristics as:

$$\textit{Terminal G: } V_{F2} = -(Z_{G2} + mZ_{L2})I_{G2} \tag{29}$$

$$\textit{Terminal H: } V_{F2} = -(Z_{H2} + (1-m)Z_{L2})I_{H2} \tag{30}$$

V_{F2} is removed by dividing Equations (29) by (30). This is to prevent difficulties of synchronising data collections from all buses of the transmission line and simply use real numbers:

$$|I_{H2}| = \left|\frac{(Z_{G2} + mZ_{L2})}{(Z_{H2} + (1-m)Z_{L2})} \times I_{G2}\right| \tag{31}$$

The fault distance *m* is then calculated using the algebraic Equation (28), where the variables are specified:

$$a + jb = I_{G2}Z_G$$

$$c + jd = Z_{L2}I_{G2}$$

$$e + jf = Z_{H2} + Z_{L2}$$

$$g + jh = Z_{L2}$$

$$A = |I_{H2}|^2 \times (g^2 + h^2) - (c^2 + d^2)$$

$$B = -2 \times |I_{H2}|^2 \times (eg + fh) - 2(ac + bd)$$

$$C = |I_{H2}|^2 \times (e^2 + f^2) - (a^2 + b^2)$$

This technique is only convenient for pinpointing imbalanced faults. Given that current data is available, the position predictions' precision depends on having the supply impedance characteristics precisely [34].

7. Proposed Model

The proposed model is shown in Figure 8. The model is applied to a simulated transmission line design. The sequencing plan of the algorithm utilised to pinpoint the transmission line faults is indicated in Figure 9.

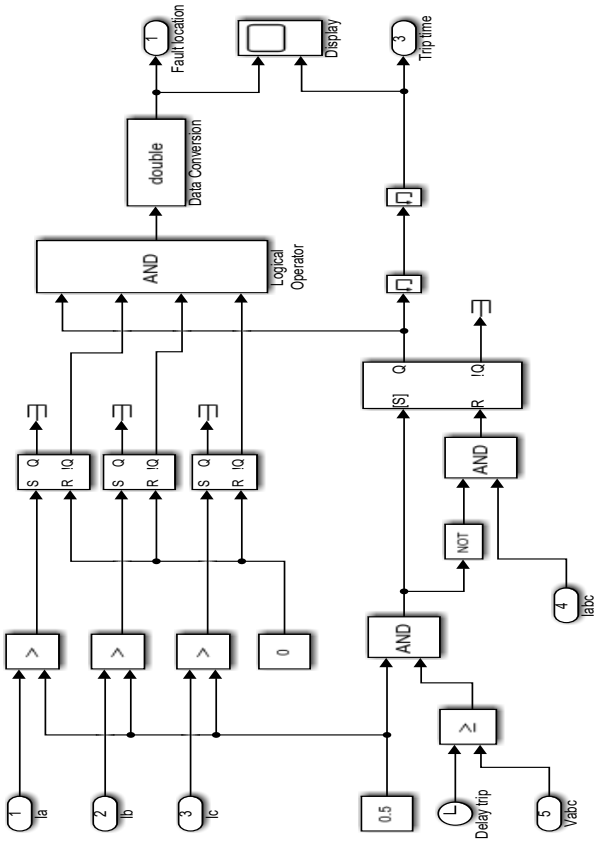


Figure 8: Proposed Algorithm Model.

Measurements will be obtained, parameters will be computed to construct an appropriate transmission line model utilising MATLAB/SIMULINK software, and several adjustments will be required to achieve a robust and efficient network.

The second process is implementing multiple faults on a selected line at various intervals from a designated end terminal. The pre-fault, fault, and post-fault voltage and current waveforms will be studied. These data can then be utilised to determine the location of the fault. Assuming that all the phases are completed successfully, the precise current and voltage signals are gathered and used to pinpoint the exact position of the fault, as illustrated in Figure 9.

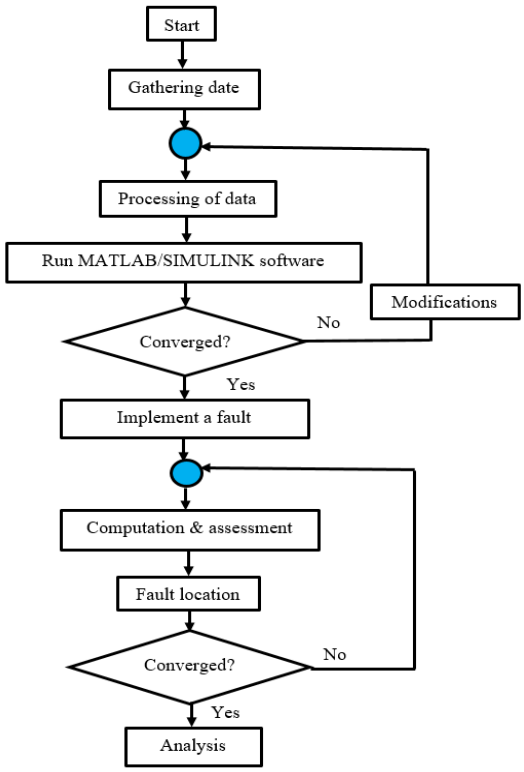


Figure 9: Proposed Algorithm Flow Chart.

The MATLAB/SIMULINK Model is presented in figure 10.

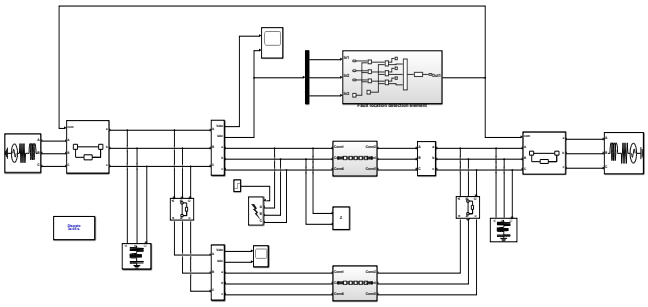


Figure 10: Transmission Line Design with a Single-ended Impedance-based Fault Location Technique.

8. Simulation Results

The Simulation results are as fellows

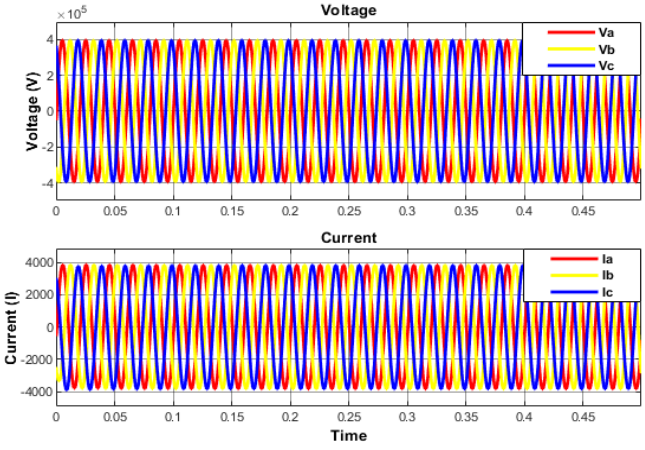


Figure 11: Balanced Voltage & Current Waveforms.

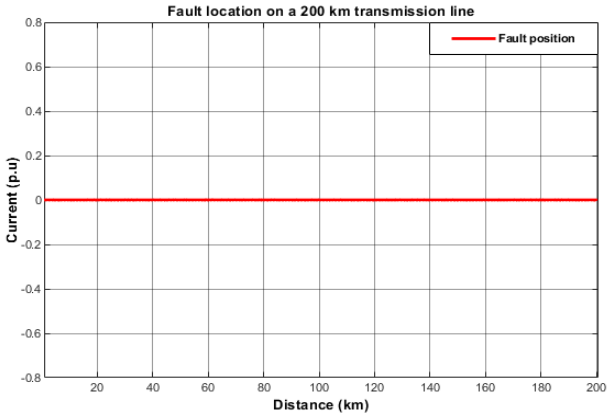


Figure 12: No Located Fault/s.

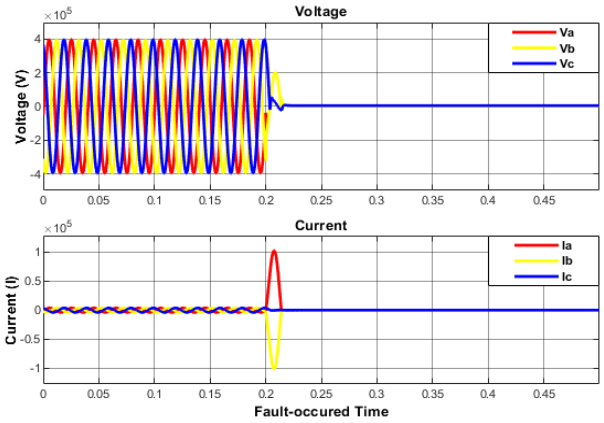


Figure 13: Distorted Voltage & Current Oscillations Induced by L-L Fault.

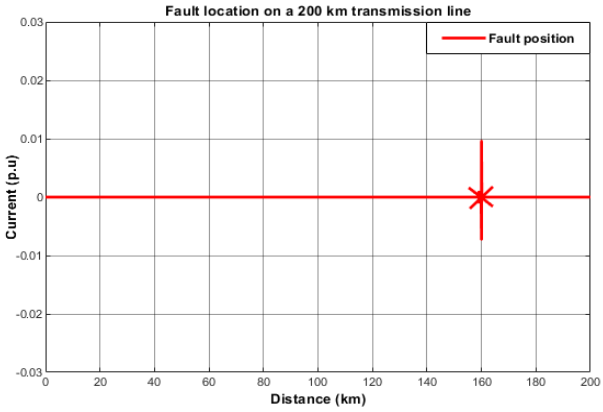


Figure 14: L-L Fault Pin-pointed at 160.5 km.

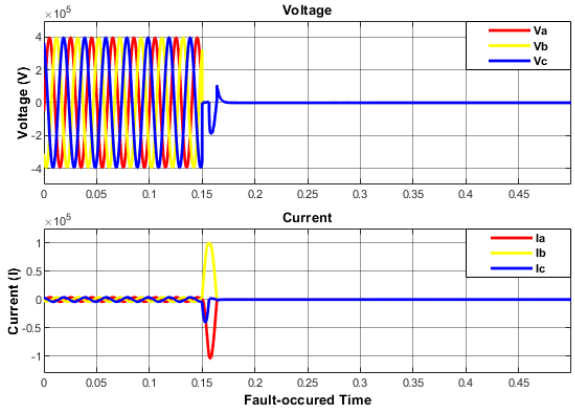


Figure 15: Distorted Voltage & Current Oscillations Induced by L-L-L Fault.

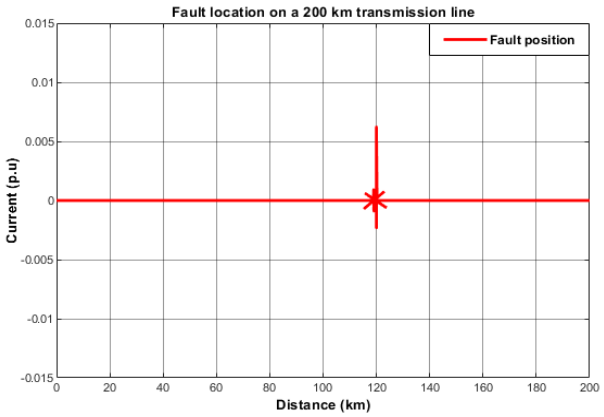


Figure 16: L-L-L Fault Pin-pointed at 119.7 km.

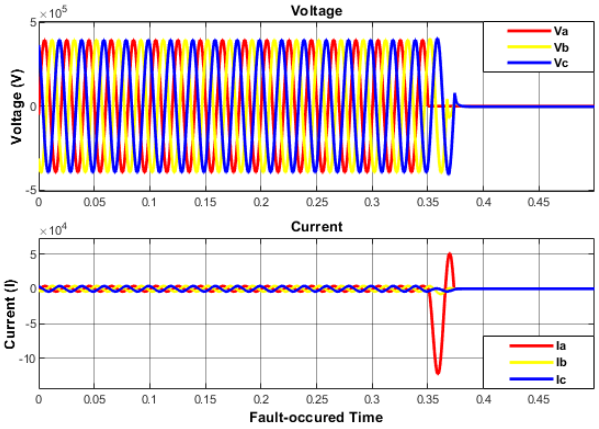


Figure 17: Distorted Voltage & Current Oscillations Induced by L-G Fault.

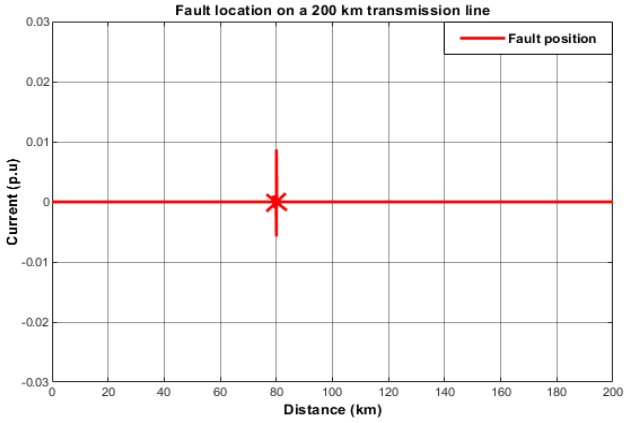


Figure 18: L-G Fault Pin-pointed at 79.9 km.

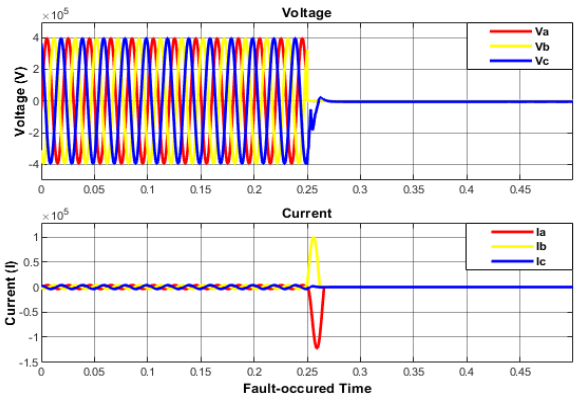


Figure 19: Distorted Voltage & Current Oscillations Induced by L-L-G Fault.

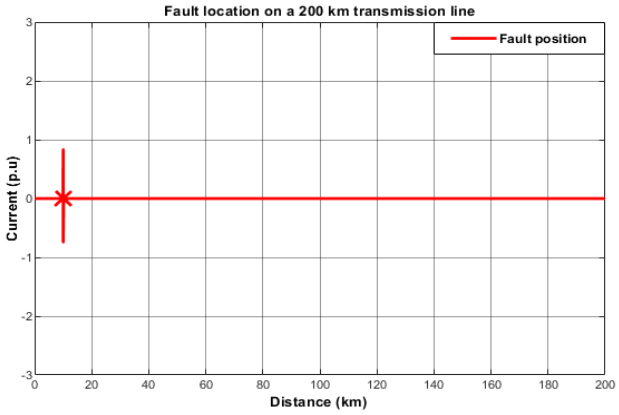


Figure 20: L-L-G Fault Pin-pointed at 10.6 km.

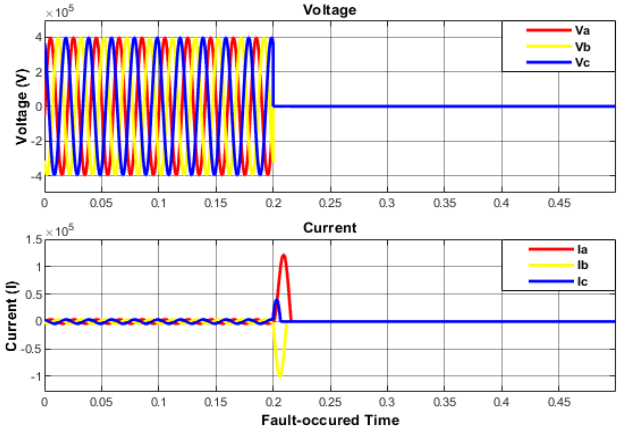


Figure 21: Distorted Voltage & Current Oscillations Induced by L-L-L-G Fault.

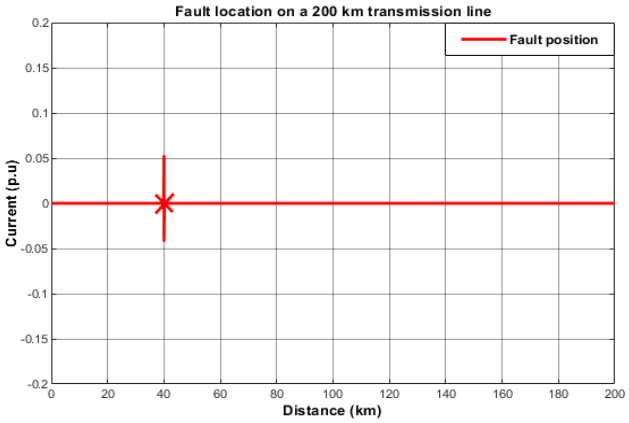


Figure 22: L-L-L-G Fault Pin-pointed at 41 km.

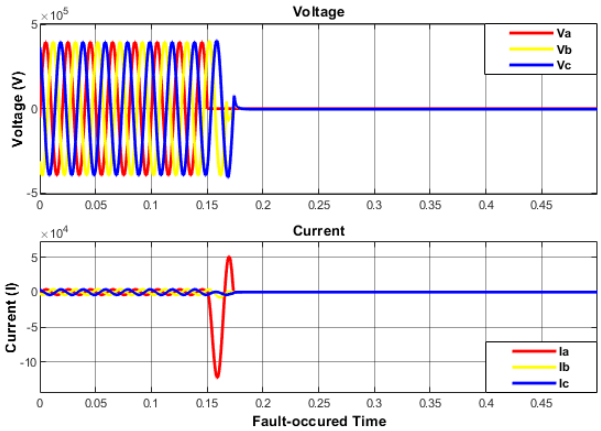


Figure 23: Distorted Voltage & Current Oscillations Induced by L-G Fault.

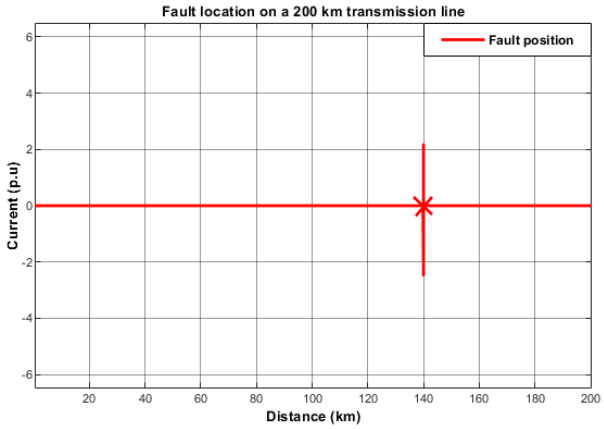


Figure 24: L-G Fault Pin-pointed at 140.9 km.

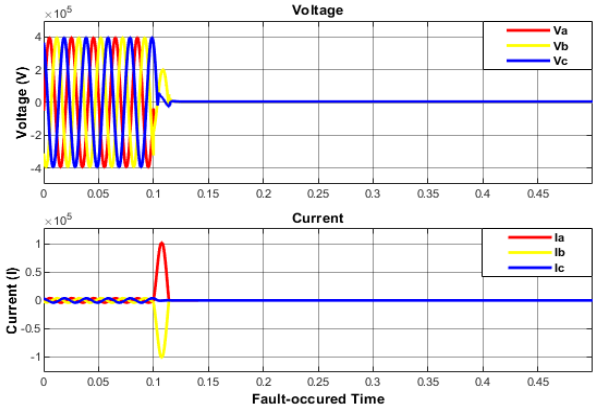


Figure 25: Distorted Voltage & Current Oscillations Induced by L-L Fault.

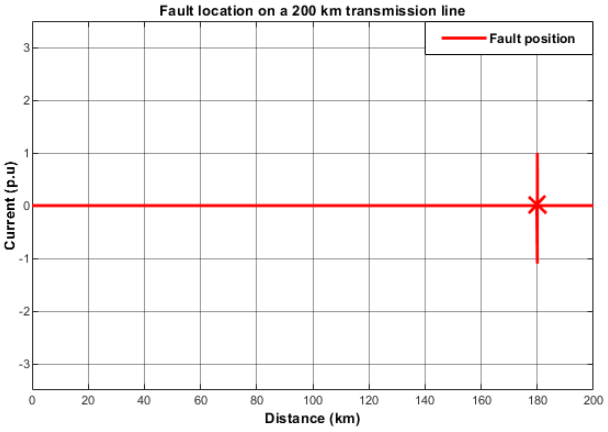


Figure 26: L-L Fault Pin-pointed at 179.2 km.

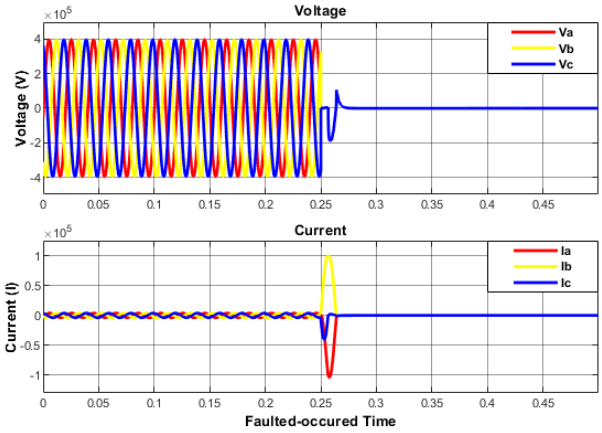


Figure 27: Distorted Voltage & Current Oscillations Induced by L-L-L Fault.

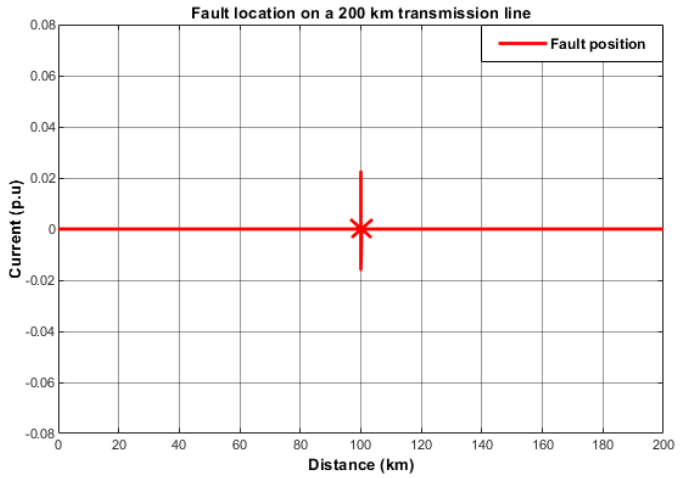


Figure 28: L-L-L Fault Pin-pointed at 98.9 km.

9. Practical Results

In this section, the calculated fault distances using the impedance-based technique are presented in Table 2. The results were obtained from test fault cases involving different faults on a 200 km transmission line. The accuracy of the fault distances was calculated using Equation (32).

$$\%error = \frac{|Actual\ distance - Calculated\ distance|}{Total\ length\ of\ the\ line} \times 100 \qquad (32)$$

Table 2: Error of Estimation: Impedance-based Technique.

Length of the Line (km)	Fault Type	Distance of the Fault (km)	Calculated Fault Distance (km)	Error (%)
200	L-G	80	78.7	1.65
200	L-L	160	146.8	6.6
200	L-L-G	10	6.4	1.8
200	L-L-L	120	115.2	2.4
200	L-L-L-G	40	32.1	3.95
200	L-L	180	176.5	1.75
200	L-G	140	136.7	1.65
200	L-L-L	100	106.4	-3.2

It is observed that the impedance-based technique achieves accurate estimations that are within 1-15% of the designed transmission line.

10. Results Comparison

This section presents the percentage error of the simulated results utilising the proposed algorithm against the calculated results of the impedance-based technique for the designed transmission line. Table 3 illustrates that the proposed algorithm provides a more accurate fault location than the impedance-based technique in all scenarios.

It is shown in Table 3 that the proposed algorithm gives precise predictions with an approximation of less than 1% from the simulated transmission line. As an outcome, it is evident that the proposed algorithm provides more precision.

Table 3: Error of Estimation: Impedance-based Technique against the Proposed Algorithm.

Length of the Line (km)	Fault Type	Distance of the Fault (km)	Calculated Fault Distance (km)	Error (%)
200	L-G	80	1.65%	0.05%
200	L-L	160	6.6%	-0.25%
200	L-L-G	10	1.8%	-0.3%
200	L-L-L	120	2.4%	0.15%
200	L-L-L-G	40	3.95%	-0.5%
200	L-L	180	1.75%	0.4%
200	L-G	140	1.65%	-0.45%
200	L-L-L	100	-3.2%	0.55%

11. Conclusion

This work presents the results of an impedance-based technique to the results of the proposed algorithm on a 200-km transmission line. It was observed that the impedance-based techniques are susceptible to producing errors or incorrect predictions. The presence of faults induced from high impedance sources produces an extra impedance to the ground, which negates the impedance calculation and produces errors in the distance to the fault. This results in inaccuracies that can affect a distance-to-fault estimation by 1-15 % of the overall length. Comparison research was undertaken against the impedance-based techniques to validate the proposed algorithm. The simple reactance technique is considered to be the most basic technique. However, the precision of such a technique suffers in a non-homogenous network due to fault resistance, load current, and distant infeed. The Takagi technique, for instance, is load resistant; however, insensitive to distant infeed. For most scenarios, double-ended impedance-based fault location techniques achieve better outcomes. Studying the fault location usage situation can help determine what extra devices are required to improve the performance of fault-location techniques.

In the present work, an expansion of preceding work [1] introduced a novel algorithm for estimating faults on transmission lines. Extensive trials on a simulated transmission line led to the conclusion of this work. It was observed that the impedance-based technique achieves accurate estimations within 1-15% of the designed transmission line, and the proposed algorithm gives precise predictions with an approximation of less than 1% from the simulated transmission line, reaching a precision of 98.9%. As an outcome, it was evident that the proposed algorithm provides more precision.

In future work, the authors will test the proposed algorithm on a distributed system due to the availability of multiple incomers and feeders connected to the system. Furthermore, the authors will perform a study by simultaneously integrating the proposed algorithm and an artificial neural network (ANN) technique on a complex transmission line and comparing the two algorithms' accuracy.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgement

The author wishes to convey his gratitude to the university for offering him the possibility to work with such a stellar supervisor as Dr. PF Le Roux and offer impactful commentary.

References

[1] M.K Ngwenyama, P.F Le Roux, L.J. Ngoma, "Conventional Method for Electrical Transmission System Fault Location Detection," in 2021 6th Asia Conference on Power and Electrical Engineering (ACPEE), 69-76, 2021, doi: 10.1109/ACPEE51499.2021.9437010.

[2] S.M. Hakimi, A. Hasankhani, M. Shafie-khah, M. Lotfi, J.P. Catalão, "Optimal sizing of renewable energy systems in a Microgrid considering electricity market interaction and reliability analysis," Electric Power Systems Research, **10**, 7678, 2022, doi: 10.1016/j.epsr.2021.107678.

[3] P. Bunnoon, "Fault detection approaches to power system: state-of-the-art article reviews for searching a new approach in the future" International Journal of Electrical and Computer Engineering, **3**(4), 553, 2013.

[4] S.A. Aleem, N. Shahid, I.H. Naqvi, "Methodologies in power systems fault detection and diagnosis," Energy Systems, **6**(1), 85-108, 2015, doi: 10.1007/s12667-014-0129-1.

[5] S.S. Gururajapathy, H. Mokhlis, H.A. Illias, "Fault location and detection techniques in power distribution systems with distributed generation: A review," Renewable and sustainable energy reviews, **74**, 949-958, 2017.

[6] A. Keshavarz, R. Dashti, M. Deljoo, H.R. Shaker, "Fault location in distribution networks based on SVM and impedance-based method using online databank generation," Neural Computing and Applications, **34**(3), 1-17, 2021.

[7] J. Wang, Y. Zhang, T. Li, "Equivalent characteristic impedance based hybrid-HVDC transmission line fault location," Electric Power Systems Research, **10**, 7055, 2021, doi: 10.1016/j.epsr.2021.107055.

[8] X. Tong, H. Wen, "A novel transmission line fault detection algorithm based on pilot impedance," Electric Power Systems Research, **10**, 6062, 2020, doi: 10.1016/j.epsr.2019.106062.

[9] J. Van Gompel, D. Spina, C. Develder, "Satellite based fault diagnosis of photovoltaic systems using recurrent neural networks," Applied Energy, **11**, 7874, 2022, doi: 10.1016/j.apenergy.2021.117874.

[10] A.H. Hassanabad, D. Nazeipur, "Design and Simulation of a Control System for Investors in Wind Turbines.", **6**, 6.

[11] X. Wang, Y. Shen, "Fault-tolerant control strategy of a wind energy conversion system considering multiple fault reconstruction," Applied Sciences, **7**, 94, 2018.

[12] M. Nemati, M. Bigdeli, A. Ghorbani, "Impedance-based fault location algorithm for double-circuit transmission lines using single-end data," Journal of Control, Automation and Electrical Systems, **31**(5), 1267-1277, 2020, doi: 10.1007/s40313-020-00620-w.

[13] J. Doria-Garcia, C. Orozco-Henao, L. Iurinic, J.D. Pulgarín-Rivera, "High impedance fault location: Generalized extension for ground faults," International Journal of Electrical Power & Energy Systems, 105387, 2020, doi: 10.1016/j.ijepes.2019.105387.

[14] L. De Andrade, T.P. de Leão, "Impedance-based fault location analysis for transmission lines," in PES T&D 2012, 1-6, 2012, doi: 10.1109/TDC.2012.6281527.

[15] D.W. Thomas, C. Christopoulos, R.J.d.O. Carvalho, E.T. Pereira, "Single and double ended travelling-wave fault location on a MV system," 2004, doi: 10.1049/cp_20040098.

[16] M. Ngwenyama, P. Le Roux, L. Ngoma, "Traveling Wave fault location detection technique for high voltage transmission lines," in 2021 2nd International Conference for Emerging Technology (INCET), 1-7, 2021, doi: 10.1109/INCET51464.2021.9456334.

[17] M.T. Hagh, K. Razi, H. Taghizadeh, "Fault classification and location of power transmission lines using artificial neural network," in 2007 International Power Engineering Conference (IPEC 2007), 1109-1114, 2007.

[18] S. Ekici, "Support Vector Machines for classification and locating faults on transmission lines," Applied soft computing, **12**(6), 1650-1658, 2012, doi: 10.1016/j.asoc.2012.02.011.

[19] A. Prasad, J.B. Edward, C.S. Roy, G. Divyansh, A. Kumar, "Classification of faults in power transmission lines using fuzzy-logic technique," Indian Journal of Science and Technology, 1-6, 2015, doi: 10.17485/ijst/2015/v8i30/77065.

[20] L. Wei, W. Guo, F. Wen, G. Ledwich, Z. Liao, J. Xin, "Waveform matching approach for fault diagnosis of a high-voltage transmission line employing harmony search algorithm," IET generation, transmission & distribution, **4**(7), 801-809, 2010, doi: 10.1049/iet-gtd.2010.0104.

[21] L. Ji, X. Tao, Y. Fu, Y. Fu, Y. Mi, Z. Li, "A new single ended fault location method for transmission line based on positive sequence superimposed network during auto-reclosing," IEEE Transactions on Power Delivery, **34**(3), 1019-1029, 2019, doi: 10.1109/TPWRD.2019.2901835.

[22] A. Di Tomasso, G. Invernizzi, G. Vielmini, "Accurate single-end and double-end fault location by traveling waves: a review with some real applications," in 2019 AEIT International Annual Conference (AEIT), 1-6, 2019.

[23] F. Aboshady, D. Thomas, M. Sumner, "A new single end wideband impedance based fault location scheme for distribution systems," Electric Power Systems Research, 263-270, 2019, doi: 10.1016/j.epsr.2019.04.034.

[24] J. Barati, A. Doroudi, "Novel modified impedance-based methods for fault location in the presence of a fault current limiter," Turkish Journal of Electrical Engineering & Computer Sciences, 1881-1893, 2018, doi: 10.3906/elk-1711-127.

[25] F. Wang, X. Feng, L. Zhang, Y. Du, J. Su, "Impedance-based analysis of grid harmonic interactions between aggregated flyback micro-inverters and the grid," IET Power Electronics, **11**(3), 453-459, 2018, doi: 10.1049/iet-pel.2017.0356.

[26] A. Abu-Siada, M.I. Mosaad, S. Mir, "Voltage–current technique to identify fault location within long transmission lines," IET Generation, Transmission & Distribution, **14**(23), 5588-5596, 2020, doi: 10.1049/iet-gtd.2020.1012.

[27] K. Morgan, W. Gamal, K. Samuel, S.D. Morley, P.C. Hayes, P. Bagnaninchi, J.N. Plevris, "Application of impedance-based techniques in hepatology research," Journal of clinical medicine, **50**, 2020, doi: 10.3390/jcm9010050.

[28] D. Guillen, C. Salas, L. Fernando Sanchez-Gomez, L.M. Castro, "Enhancement of dynamic phasor estimation-based fault location algorithms for AC transmission lines," IET Generation, Transmission & Distribution, **14**(6),1091-1103, 2020, doi: 10.1049/iet-gtd.2019.0051.

[29] C. Zhang, Y. Yu, Y. Wang, M. Zhou, "Takagi–Sugeno fuzzy neural network hysteresis modeling for magnetic shape memory alloy actuator based on modified bacteria foraging algorithm," International Journal of Fuzzy Systems, **22**(4), 1314-1329, 2020, doi: 10.1007/s40815-020-00826-9.

[30] A. Macioł, P. Macioł, B. Mrzygłód, "Prediction of forging dies wear with the modified Takagi–Sugeno fuzzy identification method," Materials and Manufacturing Processes, **35**(6), 700-713, 2020, doi: 10.1080/10426914.2020.1747627.

[31] S. Roostaee, M.S. Thomas, S. Mehfuz, "Experimental studies on impedance based fault location for long transmission lines," Protection and Control of Modern Power Systems, 1-9, 2017.

[32] M.N. Hashim, M.K. Osman, M.N. Ibrahim, A.F. Abidin, "Investigation of features extraction condition for impedance-based fault location in transmission lines," in 2017 7th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), 325-330, 2017, doi: 10.1109/ICCSCE.2017.8284428

[33] A. Dalcastagne, S. Zimath, "A study about the sources of error of impedance-based fault location methods," in 2008 IEEE/PES transmission and distribution conference and exposition: Latin America, 1-6, 2008, doi: 10.1109/TDC-LA.2008.4641697

[34] K. Ramar, H. Low, E.E. Ngu, "One-end impedance based fault location in double-circuit transmission lines with different configurations," International Journal of Electrical Power & Energy Systems, **64**, 1159-1165, 2015, doi: 10.1016/j.ijepes.2014.09.006.

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 2, 94-102 (2022)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Online Support for Tertiary Mathematics Students in a Blended Learning Environment

Mary Ruth Freislich*,1, Alan Bowen-James[2]

[1] School of Mathematics and Statistics, University of New South Wales, Sydney, 2052, Australia

[2]Le Cordon Bleu Business School, Sydney, 2112, Australia

ARTICLE INFO

Article history:
Received: 14 November, 2021
Accepted: 20 February, 2022
Online: 18 March, 2022

Keywords:
Blended learning environments
Scaffolding
SOLO taxonomy
Tertiary mathematics

ABSTRACT

The context for the study was a naturally occurring quasi-experiment in the core mathematics program in a large Australian university. Delivery of teaching was changed in a sequence of two initial core mathematics subjects taken by engineering and science students. The change replaced one of two face-to-face tutorial classes per week by an online tutorial. Tasks in the online tutorial were designed to lead the students through the week's topics, using initially simpler tasks as scaffolding for more complex tasks. This was the only change: syllabus and written materials were the same, as was students' access to help from staff and discussion with peers. The study compared learning outcomes among students in two adjacent years: Cohort 1, the last before the change, and Cohort 2, in the first implementation of the change to a blended learning environment. Learning outcomes were assessed by a method derived from the SOLO taxonomy, which used a common scale for scoring written answers to examination questions in the two cohorts. In the first mathematics subject students doing online tutorials had significantly higher scores than those studying before the change. In the second mathematics subject there were no significant differences. The conclusion was that the online tutorials gave an advantage to students beginning university study and gave adequate support to those in the subject taken a little later. It can be concluded that the use of an online teaching component in the delivery of university mathematics programs is not only justifiable but desirable, subject to careful design of the teaching material offered.

1. Introduction

This paper is an extension of work originally presented at the *IEEE 2020 International Conference on Computer Science and Computational Intelligence* [1]. Extended content is mainly in the research background section, dealing with scaffolding in blended learning environments and the importance of the design of online teaching material. It includes discussion of the unsuitability for mathematics learning of existing instruments used to evaluate students' approaches to studying. Results contain effect sizes. The discussion includes more detail, and a conclusion section has been added.

The reported project stems from a change in first year mathematics teaching in a large Australian university, which replaced one out of two face-to-face tutorials with on online tutorial consisting of a set of tasks designed to lead the student through the required material step by step from simpler to more complex tasks. The change was made in the core sequence of first year mathematics subjects, Mathematics 1A and Mathematics 1B, which is taken by science and engineering students. Before the change, teaching was entirely face-to-face, consisting of lectures to large groups, plus two tutorials given to problem solving and answering students' questions, in much smaller groups.

The only change in the organization of teaching was the replacement of one face-to-face tutorial by an online tutorial. In the online tutorial, immediate feedback identified errors without giving solutions. Completing the tutorial tasks earned a small contribution to the student's final mark.

The study was carried out before higher education was disrupted by the Covid19 epidemic, and apart from the change in one tutorial, there was no other change in the organization and material in the core sequence. That is, the syllabus and the written

*Corresponding Author: Mary Ruth Freislich, m.freislich@unsw.edu.au

https://dx.doi.org/10.25046/aj070208

teaching material did not change, and the online tasks were very similar to those used in face-to-face tutorials. All students had access to help from staff and discussion with peers was still available in the face-to-face teaching groups. The study dealt with two cohorts of students in adjacent years, the last year before the change, and the first year of its introduction. Admission criteria had not changed, and the secondary school mathematics course taken by local students had not changed. It is therefore reasonable to conclude that the two cohorts were comparable in background and level of selection. In the year of the change, survey results indicated that students were satisfied with the teaching delivery.

It seems, therefore, that the natural quasi experiment afforded by the change offered good control for a study of any effect that the change might have on the quality of learning outcomes among students in the two cohorts. Making a comparison requires a valid method for evaluating outcomes, and validation requires the use of observable evidence. The course evaluation procedures advocated in improvement programs by the Accreditation Board for Engineering and Technology [2] emphasize the importance of direct assessments.

The present study uses observable evidence from students' final examination scripts. The comparison is based on a method described below in Section 2. It is emphasized here that the method has potential importance because it is direct and criterion-based. Before giving detail about the method, one needs to examine existing research that suggests directional predictions that can be tested.

2. Background

2.1. Students' use of resources

The newly implemented program studied in the present work represents provision of an online learning resource, rather than online instruction, given that the majority of instruction was face-to-face, and the online segment involved students' work rather than online instruction. This means that the program is not comparable with projects involve online instruction such as that reported in [3].

In [4] the authors make the important point that evaluation of any new learning resource can be invalidated if there is no evidence about whether students have used the resource. Findings in [5] were that mathematics students who were offered several optional leaning resources, they tended to use only one of the set of resources. It is worth clarifying that the present study deals with only one new resource, so that the problem of choice does not arise. There is also no uncertainty about whether the new resource was used because students' work online leaves an audit trail.

2.2. Student learning and the transition to university mathematics

There are three research strands that are important to the purpose of the present work.

First, in the review of research on student learning made in [6], it is noted continuing importance is found for an approach to learning that contains a continuing purpose of understanding material and attempting to link and compare different ideas. Understanding and high quality of learning are unlikely when students' approach is atomistic, focused on the accumulation of unrelated detail. For mathematics students, understanding is achieved and tested by active problem-solving. The instruments used in the British and Australian research reviewed in [6] measure the search for understanding using items describing wide reading and venturing beyond the syllabus. Among undergraduate mathematics students, such items are irrelevant for all but the most highly gifted students, but at all levels of talent the intention and achievement of understanding relate to activity in doing mathematical tasks appropriate to the level of study.

Active problem solving requires effort and persistence, which relates to North American research underlying the *National; Survey of Student Engagement (NSSE)*, [7]. This research strand found that self-regulation of study is very important to students' learning, and this is an obvious requirement for activity in working on mathematical tasks.

Directly relevant Australian work on student engagement in mathematics learning is supported by the theoretical outline given in [8] The authors note that student engagement in mathematics is multidimensional, with fuzzy boundaries between categories. The principal components that have been identified as affecting student engagement are consistent with components of the student learning research. *Expectancy-value theory* is defined by the value given to a learning goal and the learner's expectation of achieving it. This entails interest factors and confidence, as well as practical reasons for valuing achievement, and research has shown that it relates to quality of learning outcome. Similar importance is attached to the factor of self-regulated study identified in the North American research.

University students' choice of mathematics requires some previous success, and the choice implies that they attach some value to the subject. But there is no lack of evidence that many students find the transition to university mathematics very difficult, and the evidence comes from a wide variety of settings. Examples are afforded by the work reported in [9], for Britain, in [10] for Sweden and in [11] for Australia. Beginning university students can find self-regulation difficult. For mathematics students, previous levels of motivation, goal setting and self-regulation will not be sustained at university level if successful mathematical activity is not sustained.

Experience of difficulty may lead to discouraged and anxious avoidance of attempted engagement with mathematical tasks. The work reported in [12], done in an Australian setting indicates, that beginning mathematics students can benefit from learning support

that facilitates engagement with mathematical tasks. The next requirement is for evidence relevant to beneficial types of support.

2.3. Scaffolding and transfer of responsibility

The material for the online tutorials was specially designed to lead the student through the week's mathematical topic using a sequence of tasks that progressed from simple to more complex. Immediate feedback was given for each response, informing the student only whether the response was correct or not, without giving a solution. In addition, the sequences of tasks were designed so that solutions to earlier tasks could help with the later more difficult tasks. The online work could be done in multiple sessions within a specified time period, so the student could temporarily leave a task to look up material, ask for help, or discuss it with peers. Such a design has the potential to function as *scaffolding* for the extension of students' understanding.

Scaffolding is defined as intervention by a teacher to support students in achieving a learning goal that they would be unlikely to achieve without support [13]. There has been considerable discussion of the method of intervention and the design of the teacher's intervention, so that it extends the student's own reasoning without imposing or supplying a solution. The original idea rests on Vygotsky's thesis, described in [14], that the most valuable instruction is that which leads a learner into a development defined as being in the *zone of proximal development*. That is, the learner is already on the border of extended capability, and hence can reach extension with minimal appropriate help. The idea of scaffolding is defined by interaction between teacher and student. The authors in [15] found that interactive scaffolding led by the teacher in relatively small community college mathematics classes was very much more successful than previously used approaches. For large enrolment groups, limits of resourcing make the original form of scaffolding impossible. But it is argued here that the design of the online tutorials affords an approximation to scaffolding, because the gradient of task difficulty and the immediate feedback provide indirect assistance in the extension of understanding, with the limitation of the feedback also implying that assistance is not too intrusive.

Transfer of responsibility to the learner is also an important underlying goal of providing scaffolding [16]. It is pointed out in [17] that, for scaffolding to make its widest contribution, it needs a definition that empowers the learner, so that the student becomes independent of the presence of an insightful teacher as agent. In the context of mathematics, they propose problem solving as the means of creating self-scaffolding. In contrast to face-to-face tutorials, online tutorials give all responsibility for work on the given tasks to the student, with the minimal assistance designed to foster effort and persistence. Organizational responsibility in scheduling time is also required, but the important factor is the design of the tasks facilitating active engagement in the tasks, which serves to build the understanding, independence and self-regulated study found important in the studies described here and in Section 2B.

2.4. Blended mathematics teaching and the importance of design

Evidence is available that well-designed online materials can function in this way. Studies of statistics programs [18], [19] indicate that achievement gains follow careful adjustment of materials, designed to integrate the learning environment consistently, and to foster understanding. The results of [18] are particularly important, because the material provided to students was revised from year to year, and benefits to students' achievement appeared only in later years. These results are compatible with the established distinction between medium as a means of delivery and the designed study program as the goods delivered [20] Rapid and flexible delivery can give an advantage only if goods of value are delivered.

The study described in [21] is also highly relevant to the idea of scaffolding afforded by suitably designed material. It deals with a very large group of statistics students of variable academic and national background, who were offered an online tutorial system that proceeded from diagnostic testing to select tasks best adapted to each student's stage of learning. The study found that the time spent using the online program was positively related to achievement, with the strongest effect among students whose scores on Vermunt's *Inventory of Learning Styles* [22] indicated that they were less well adapted to university study.

2.5. Assessing learning outcomes

In Australian work on learning outcomes, [23] the researchers developed a classification of the quality of learning outcomes based on actual responses to a variety of educational tasks. The classification used criteria defined by the complexity, adequacy of coverage, and consistency of observable responses to set tasks. They defined a system of levels of outcome called the *Structure of Observed Learning Outcomes (SOLO) Taxonomy.* The value of the reference to the observable is clear. The researchers claimed that the classification was invariant across disciplines and justified the claim by giving illustrations from the work done in the principal areas of school study, across the middle years of schooling, from upper primary level to junior secondary. The SOLO levels, as defined in [23] are listed in Table 1 below.

The SOLO split between the *Multistructural* and *Relational* levels is based on consistency in reasoning, and so reflects the dichotomy between understanding relationships and atomistic display of facts which is of obvious importance in mathematics, with achievement of the relational level providing evidence of understanding. The wide applicability of the SOLO taxonomy is not relevant to the present study, but the issue of consistent reasoning is central to it. The applicability of SOLO to

mathematics was based on research that identified patterns of errors and misconceptions in students' mathematics learning.

Table 1: The SOLO taxonomy

Level	Definition
Prestructural	No valid response
Unistructural	One aspect of the problem correctly identified, but no diversity of aspects presented, so that questions of consistency cannot arise.
Multistructural	Multiple relevant information presented and used, but without considering relationships between different parts, so that inconsistency appears.
Relational	Multiple relevant information presented and used in a way that recognizes relationships and achieves consistency within the given task.
Extended abstract	Multiplicity recognized and consistency achieved over a context beyond that of the given task.

The SOLO taxonomy has been used at tertiary level as a framework for defining intended learning outcomes for programs in mathematics and computer science [24] and its application in other science disciplines at tertiary level has been found to be a valuable diagnostic tool [25]. The SOLO levels were adapted for the work reported in [26] to define a method of evaluating levels of learning outcomes in tertiary students' mathematics.

The focus was on examination performance in early undergraduate years, so the highest SOLO level was not considered relevant. The other four levels were used to construct a scoring system intended to provide a common scale usable across tasks involving the same mathematical material, examined at a similar level of difficulty.

The criteria used were, first, logical consistency, and second, adequate coverage of the task. A student's response to an examination question was assigned to one of six levels, labelled from 0 to 5. Levels 4 and 5 required the logical consistency of the SOLO relational level, with 5 given for a completely correct solution, and 4 given if there was a small error that did not affect consistency, like a minor slip in arithmetic or a copying error. Levels 0 and 1 correspond to SOLO Prestructural and Unistructural levels: nothing right or only one relevant aspect of the problem identified. Solutions with an error of logic at the Multistructural SOLO level, with more than one good step presented, were classified as level 2 or 3, depending on how much of a satisfactory solution was present. Examination questions were split into self-contained tasks, and each was scored independently. A composite score was obtained by summing the task scores, weighted using the proportion of the examination marks assigned to each.

The method does not attempt the generality claimed for the SOLO taxonomy. Validity is claimed only for the close relationship between tasks, depending on the stability of syllabus, staffing, student intake, teaching materials and most of the implementation of teaching in the two adjacent year groups. The SOLO taxonomy is well adapted to mathematics because its criteria fit the requirements of mathematical tasks. But its most important characteristic is its being defined in terms of the observable. The North American Accreditation Board for Engineering and Technology [2]) argues that a teaching program cannot be adequately evaluated without a direct method for examining students' learning outcomes, one which is closely fitted to the actual study program, both of which requirements apply to the method described. Applying the scoring method is similar to examination marking, and scores correlate at over 0.9 with examination marks, which implies similar ranking. What the method is intended to achieve is a common ranking for the two year-groups' performance on similar tasks. It is worth noting also that a direct method of examining learning outcomes has advantages over the use of questionnaires to assess approaches to studying. Two reasons are important. The first is intrinsic: direct assessment avoids problems associated with the reliability of self-reported data about behaviour and attitudes. The second reason is the mismatch between the existing instruments used to assess approaches and the study of mathematics, at least at undergraduate level. This has already been mentioned in connection with the approach instruments described in [6]. But one should also note that similar remarks apply to the North American *NSSE,* and the Australian Survey of Student Engagement (AUSSE) derived from it [27] derived from it.

The point here is that the *AUSSE* measure higher level thinking by items dealing with extended essay- style writing and multiple revision of drafts. In the development work for the AUSSE, it was found [27] that science students had low scores of higher-level thinking, but it is probable that such results are contaminated by the inadequacy of the instrument.

3. Method

3.1. Sample

The target population was the set of students enrolled for Mathematics 1A and 1B, in adjacent years, taking the groups from the first time in each year that the unit was offered. Simple random samples were drawn from those students who sat the final examination. This means that those who did not survive to the final examination could not be considered, but this restriction applies to all the groups being compared. Questions involving students' gender were not part of the study, but gender information was available, and was recorded, because any gender-related patterns that might emerge would be of interest. Sample numbers are in Table 1. The proportions of females and males in the sample are very similar to proportions in the total groups.

Table 2: Sample

Cohort	Mathematics1A		Mathematics 1B	
	Female	Male	Female	Male
1 (from the last year before the change)	53	152	38	142
2 (from the first year when the change was introduced)	49	154	44	153

3.2. Analyses

The two cohorts were compared within each of the two mathematics subjects. In each subject, four groups defined by cohort and gender were compared using analysis of variance. In the case of a significant overall result, differences between groups were examined using least significant differences. For cases where there were significant results, effect sizes were calculated. Analyses were done using the open-source package Rstudio [28].

4. Results

4.1. Mathematics 1A

Descriptive statistics are in Table 3, and the analysis of variance data are in Table 4.

Table 3: Mathematics 1A Descriptive statistics

Cohort		Female	Male
1 All teaching face-to-face	Mean	10.66	10.07
	St. dev.	3.58	3.62
	n	63	152
2 Blended teaching	Mean	12.87	12.24
	St. dev.	3.93	3.73
	n	49	154

Table 4: Mathematics 1A Analysis of variance

Analysis of variance				
Source	**Sums of squares**	**df**	**Mean squares**	**F**
Between groups	521.29	3	173.76	12.74***
Residual	5510.47	404	13.64	
Total	6031.44	407		

*** $p < 0.001$

The means for Cohort 2 are higher than those for Cohort 1, and the analysis of variance gives a high level of significance to differences between groups. Results for comparisons between pairs of groups using least significant differences are in Table 5

Table 5: Least significant differences

Groups in order of means				
Cohort		1 female	2 male	2 female
1 male		0.90	5.14***	4.62***
1female	t		2.79**	3.10**
2 male				1.04

** $p < 0.01$; *** $p < 0.001$

The differences are significant for all comparisons of Cohort 1 groups with Cohort 2 groups, and no within-cohort comparisons between females and males were significant. The purpose of the study did not include gender comparisons but grouping by gender was in the analysis because it was possible that different delivery of teaching might have different effects for females and males.

Effect sizes for the four significant comparisons are in Table 6. The interpretations use the classification described in [29] and are high or very high in all cases.

Table 6: Effect sizes

Group	Cohort 2 male		Cohort 2 female	
	Effect size		Effect size	
Cohort 1 male	0.60	Very high	0.77	Very high
Cohort 1 female	0.46	High	0.46	High

4.2. Mathematics 1B

Descriptive statistics are in Table 7 and the analysis of variance results are in Table 8. There were no significant differences between groups in Mathematics 1B.

Table 7: Mathematics 1B: Descriptive statistics

Cohort		Female	Male
1 All teaching face-to-face	Mean	11.48	11.46
	St. dev.	3.77	3.74
	n	38	142
2 Blended teaching	Mean	11.82	11.32
	St. dev.	3.46	3.44
	n	44	154

Table 8: Mathematics 1B Analysis of variance

Analysis of variance				
Source	**Sums of squares**	**df**	**Mean squares**	**F**
Between groups	8.52	3	2.84	0.82 ns
Residual	4819.82	373	12.92	
Total	4828.34	378		

5. Discussion

It is worth noting here again that no gender differences were found. Marginally higher mean scores for females probably only reflect the higher selection of the female groups, given that tertiary mathematics groups still contain considerably more males.

It is clear that the results for Mathematics 1A show advantages in the online component of delivery of teaching. The advantage is in the direction predicted from the research background, subject to the importance of the design of the online teaching material. The digital audit trail afforded by the technology gives assurance that the online learning resource was used by the students, which functions as an additional control factor. Because Mathematic 1A is the first core mathematics subject taken by engineering and science students, one can

conclude that the online program facilitated students' transition to university study.

The lack of significant differences between the two cohorts in Mathematics 1B can be explained by combining evidence from the literature with the conclusion given for Mathematics 1A. Mathematics 1B is the second subject in first-year core mathematics, its students are at least one semester further into university study than most students in Mathematics 1A and are more highly selected because they have already passed Mathematics 1A. In [21] the findings indicated that online resources were most helpful to students who were initially less well adapted to university study. The mathematics 1B groups, therefore, are likely to have less need of help than students who are mostly new to university study.

But the finding for Mathematics 1B is still useful evidence because it indicates that the online program shows no disadvantage compared with fully face-to face teaching. This means that, if one regards the medium as a delivery vehicle, the results indicate delivery of adequate goods. The speed and flexibility of delivery therefore become relevant. The audit trail permitted by the technology also enables improvement of the online material through tracking areas where students have most difficulty. The method of comparison of outcomes used in the present study can be used to compare different sets of online material, serving as the direct Students' written assessments can be scanned into digital records, which opens the way to a cyclic use of technology to provide research material for evaluation of what the technology delivers. Such material would also permit research on changes in students' learning over some years.

In a review of research on fully online teaching of undergraduate mathematics, [30] it is reported that the results are mostly unfavorable to online teaching. It should be emphasized, however, that the present study represents a different field, because the blended learning environment involved retained easy contact with staff and peers. That is to say, the learning environment was not exclusively online and, indeed, assumed a degree of offline interpersonal engagement.

It should also be noted that the direct assessment of quality of learning outcomes in the present study has advantages over alternative methods. It clearly is unreasonable to judge online teaching using correlations between results of assessments of different teaching components, but the use of grades alone also does not provide a clear determination of efficacy. Hence, results of the study [31], which used grades, cannot be considered as corroboration for the present study.

It was noted In Section 2B that instruments used to assess students' approaches to studying are not well adapted to mathematics learning. The underlying concept of the value of a search for understanding is clearly important in all fields, so that, even after some decades of stabilization of existing instruments, adaptation to mathematics would be useful. Records from online tutorial tasks and written examinations could be combined with initial qualitative investigation of students' approaches to and experience of studying mathematics.

6. Conclusion

The results indicate that the use of an online component in the delivery of first year tertiary mathematics can be justified as producing enhanced learning outcomes among beginning students, and no disadvantage to those at a slightly later stage, provided that the online teaching material is carefully designed to lead the students from simpler to more complex tasks. Hence any recommendation for the extended use of online teaching material, and any future research of online mathematics teaching, must focus primarily on the quality of the design of that material.

The present study is limited to first-year mathematics. It follows that investigation in other contexts and later stages of university study would be a necessary supplement. The increasing use of online teaching delivery affords the opportunity for such work. In addition, one should note that the technology furnishes detailed records of students' use of materials and performance on assessment tasks that provide valuable data for study.

The present study did not address students' experience of studying. In the background section it was noted that existing self-report questionnaires on students' approaches to studying are unsuitable for mathematics learning. The development of suitable instruments with a similar purpose, but targeting more appropriate approaches, is an open field. The development of such instruments would be facilitated by initial exploratory work using qualitative methods to elucidate salient aspects of students' experience of studying mathematics.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors wish to acknowledge the support of the University of New South Wales and Le Cordon Bleu Australia.

References

[1] M.R. Freislich, A. Bowen-James, "Observable learning outcomes among tertiary mathematics students in a newly implemented blended learning environment". In 2020 IEEE International Conference on Computer Science and Computational Intelligence *(CSCI),* 976-980, 2020, doi:10.1109/CSCI151800.2020.00181.

[2] Accreditation Board for Engineering and Technology, 2020,*ABET accreditation*. https://www.abet.org.

[3] S. Lambert, "Reluctant mathematicians: skills-based MOOC scaffolds a wide range of learners", Journal of Interactive Media in Education*,* **2015**(1), 1-11, 2015 doi.org/w5334/jime.bb.

[4] P. Sharma, M.J. Hannafin2007, "Scaffolding in technology enhanced learning environments",. Interactive Learning Environments, **15**(1), 27-46, 2007, doi:101080/10494820600996972.

[5] M. Inglis, A. Palipura, S. Trenholm, J. Ward, 2011, "Individual differences in students' uses of optional learning resources'", Journal of Computer Assisted Learning*,* **27**(6), 490-502, doi:10.1111/j.1365-2729.2011.00417.x.

[6] J.T.E.Richardson, 'Student learning in higher education: A commentary', *Educational Psychology Review,* vol. 29, pp. 353-362, 2017, doi: 1007/s10648-017-9410-x.

[7] Indiana University School of Education, National Survey of Student Engagement, University of Indiana, 2017.

[8] H.M. Watt, M. Goos, M, "Theoretical foundations of engagement in mathematics'" Mathematics Education Research Journal, **29**(2), 133-142, 2017, doi:10.1007/s13394-017-0206-6.

[9] M. McAlinden, A. Noyes, "Mathematics in the disciplines at the transition to university", Teaching Mathematics and its Applications, **38**(2), 61-73, .2019, doi.org10.1093/teamat/hry004.

[10] S.H. Bengmark, H. Thunberg, T.M. Winberg, 2017, "Success-factors in transition to university mathematics", International Journal of Mathematical Education in Science and Technology, **48**(7), 988-1001, 2017,doi: 10.1080/0020739X.2017.1310311.

[11] P.W. Hillock PW, R.N. Khan, 2019, "A support learning programme for first-year mathematics", International Journal of Mathematical Education in Science and Technology, **50**(7), 1073-1086, 2019, doi:10.1080/0020739X.2019.16569026.

[12] L.J. Rylands, D. Shearman, "Mathematics learning support and engagement in first year engineering", International Journal of Mathematical Education in Science and Technology., **49**(8), 1133-1147, 2017, doi: 10.1080/0020739X.2018.1447699.

[13] C. Quintana, 2021, 'Scaffolding inquiry', in R.G. Duncan, & C.A. Chinn, (Eds.), International handbook of inquiry and learning, Routledge, 176-188.

[14] B.H. Johnsen, BH, 2020, 'Vygotsky's Legacy Regarding Teaching-Learning Interaction and Development', In B.H. Johnsen (ed.), Theory and Methodology in International Comparative Classroom Studies, Cappelen Damm Akademisk, 82-98.

[15] T. Gula, C. Hoesler, W. Majciejewski, W, "Seeking mathematics success for college students: A randomised field trial of an adapted approach", International Journal of Mathematical Education in Science and Technology*,* **48**(8),127-143, 2021, doi: 10.1080/0020739X2015.1029026.

[16] B.R. Belland, A.E. Walker, M.W. Olsen H. Leary, H., "A pilot meta-analysis of computer-based scaffolding in STEM education", *Educational Technology and Society,* 18(1), 183-197, 2015, https://www.jstor.org/stable/jedtechsoci.18.1.18.

[17] D. Holton, D. Clarke, D., "Scaffolding and metacognition"; International Journal of Mathematical Education in Science and Technology*,* **37**(2), 127-143, 2006, DOI:10.1080/0020730500285818.

[18] L. Zetterqvist, "Applied problems and use of technology in an aligned way in basic courses in probability and statistics for engineering students: a way to enhance understanding and increase motivation,", Teaching Mathematics and its Applications*,* **36**(2), 108-122, 2017, doi.org/ 10.1093/teamat/ hrx004

[19] A.H. Jonsdottir, A.A. Bjornsdottir, G. Stefansson, G., "Difference in learning among students doing pen-and-paper homework compared to web-based homework in an introductory statistics course", Journal of Statistics Education*,* **25**(1), 12-20, 2017, doi: 10.1080/10691898/2017.1291289.

[20] R.E. Mayer, "Thirty years of research on online learning", Applied Cognitive Psychology, **33**(2), 152-159, 2019, doi: 10.1002/acp.3482.

[21] D.J. Tempelaar, B. Rienties, B. Giesbers, "Who profits most from blended learning?'" *Industry and Higher Education*, **23**(4), 285-292, 2009, doi.org/10.1145/3170358.3170385.

[22] J.D. Vermunt, Y.J. Vermetten, YJ, 2004, 'Patterns in student learning: relationships between learning strategies, conceptions of learning and learning outcomes', Educational Psychology Review, **16**(4), 359-384, 2004, doi:org/10.1007/s10648-004-0005-y.

[23] J.B. Biggs, K.F. Collis, KF (2014). Evaluating the quality of learning: The SOLO taxonomy. Academic Press, 2014.

[24] C. Braband, B. Dahl, B, "Using the Solo taxonomy to analyze competence progression in science curricula", Higher Education, **58**(4),.531-549, 2009, doi: 10.1007/s10734-009-9216-4.

[25] L.C. Hodges, L. Harvey, "Evaluation of student learning in organic chemistry using the SOLO taxonomy", Journal of Chemical Education*,* **80**(7) 785-787, 2003, doi:org/10.1021/ed080 p785.

[26] M.R. Freislich, A. Bowen-James, 2019 "Effects of a change to more formative assessment among tertiary mathematics students", *Anziam Journal Electronic Supplement*, **61**, C255-C271, doi:10.21914/anziamj.v6110.15166.

[27] A. Radloff, H. Coates, 2009, Doing more for learning: Enhancing engagement and outcomes. Australian Council for Educational Research, 2009.

[28] RStudio, Open source and professional software for data science, <https://rstudio.com>,, 2020

[29] S. Higgins, M. Katsipataki, 2016, "Communicating comparative findings from meta-analysis in educational research: some examples and suggestions", International journal of Research and Method in Education, **30**(3), 237-254, doi: 10/1080/ 1743727X. 2016.1166486.

[30] S. Trenholm, J. Peschke, "Teaching undergraduate mathematics fully online: a review from the perspective of communities of practice", International Journal of Educational Technology in Higher Education*,* **17**, Article 37, 2020, doi:org/10.1186/s41239-020-00215-0.

[31] B. Loch, R. Borland, N. Sukhurovka, 'Implementing blended learning in tertiary mathematics teaching', The Australian Mathematical Society Gazette, **46**(2), 90-102, 2019.

Appendix: Examples of the scoring method

A. Algebra 1

Find conditions on $b_1, b_{2,,} b_3$ to ensure that the following system of equations has a solution.

B. Algebra 2

a) Find all roots in the complex numbers of

$$z^5 + 1 = 0$$

$$\begin{array}{rcrcrcl} x & + & 2y & . & . & = & b_1 \\ x & + & y & - & z & = & b_2 \\ 2x & + & y & - & 3z & = & b_3 \end{array}$$

Solution

$$\begin{pmatrix} 1 & 2 & 0 & b_1 \\ 1 & 2 & -1 & b_2 \\ 2 & 2 & -3 & b_3 \end{pmatrix} Step\ 1 \rightarrow$$

$$\begin{pmatrix} 1 & 2 & 0 & b_1 \\ 0 & -1 & -1 & b_2 - b_1 \\ 0 & -3 & -3 & b_3 - 2b_1 \end{pmatrix} Step\ 2 \rightarrow$$

$$\begin{pmatrix} 1 & 2 & 0 & b_1 \\ 0 & -1 & -1 & b_2 - b_1 \\ 0 & 0 & 0 & b_3 - 3b_2 + b_1 \end{pmatrix}$$

Conclusion

Solutions exist if and only if

$$b_3 - 3b_2 + b_1 = 0$$

Table 9 Scoring examples for Algebra 1

Score	Example
5	All correct
4	Step 1 correct. Step 2; $\begin{pmatrix} 1 & 2 & 0 & b_1 \\ 0 & -1 & -1 & b_2 - b_1 \\ 0 & 0 & 0 & b_3 - 5b_2\ b_1 \end{pmatrix}$ [Mistake in arithmetic.] Conclusion: solutions exist if and only if $b_3 - 5b_2 + b_1 = 0$
3	Row operations correct to the end of Step 2. But conclusion given as: $b_1 \neq 0,\ b_1 \neq b_2, b_3 - 3b_2 + b_1 \neq 0$
2	Row operations correct to the end of Step2. No conclusion.
1	Step 1 correct. Then replace Row 2 by Row 2 + (1/2) Row 1, giving $Row\ 2 = (0 \quad 0 \quad -1 \quad b_2)$ [This shows row operations are not understood.]

Factorise $z^5 + 1$ over the complex numbers.

Factorise $z^5 + 1$ over the real numbers.

Solution

Put $z = re^{i\theta}$. Then

$$r^5 e^{5i\theta} = 1\, e^{(\pi + 2k\pi i)}$$

So $r = 1$ and

$$5\theta = \pi + 2k\pi. \quad \theta = \frac{(2k+1)\pi}{5}$$

Distinct solutions occur for

$$k = 0, 1, -1, 2, -2.$$

So the solutions to the equation are:

$$e^{\frac{\pi i}{5}}, \quad e^{\frac{3\pi i}{5}}, \quad e^{\frac{-\pi i}{5}}, \quad e^{\frac{5\pi i}{5}} = -1, \quad e^{\frac{-3\pi i}{5\pi}}$$

$$z^5 + 1 =$$

$$(z+1)\left(z - e^{\frac{\pi i}{5}}\right)\left(z - e^{\frac{-\pi i}{5}}\right)(z - e^{\frac{3\pi i}{5}})(z - e^{\frac{-3i}{5\pi}})$$

$$z^5 + 1 =$$

$$(z+1)\left(z^2 - 2z\cos\left(\frac{\pi}{5}\right)1\right)\left(z^2 - 2z\cos\left(\frac{3\pi}{5}\right) + 1\right)$$

Table 10 Scoring examples for Algebra 2

Score	Example
5	All correct
4	Correct (a), (b), then (c) $(z+1)(z^2 - 4\,z + 1)(z^2 - 2\cos\frac{3i\pi}{5}z + 1)$
3	Correct (a), (b), then (c) $(z+1)(z - e^{\frac{i\pi}{5}})(z - e^{\frac{i\pi}{5}})(z - e^{\frac{3i\pi}{5}})(z - e^{\frac{3i\pi}{5}})$.
2	Roots given as $e^{\frac{2ikp}{5}}$ then (b) corresponding to this, no (c)
1	$z = e^{\frac{2kip}{5}}$ and no more

C. Calculus 1

a) State the Mean Value Theorem
b) Use the theorem to prove sinhx> x for x> 0

Solution

a) If $f(x)$ is continuous on $[a, b]$ and differentiable on (a, b), then there exists c in (a, b) such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

b) $f(x) = \sinh x$ is continuous and differentiable everywhere and $f'(x) = \cosh x$. So, there is c ε $(0, x)$ such that

$$\frac{\sinh x - \sinh 0}{x - 0} = \frac{\sinh x}{x}$$

$$= \cosh c = \frac{e^c + e^{-c}}{2} > 1$$

It follows that $\sinh x > x$ for x > 0

Table 11 Scoring examples for Calculus 1

Score	Example
5	All correct
4	Correct up to $\frac{\sinh x}{x} = \cosh c$, then a sketch showing $\cosh x > 1$, but conclusion stated as $\frac{\sinh x}{x} > 0$, and hence $\sinh x > x$
3	Correct up to $\frac{\sinh x}{x} = \cosh c$, then " $\cosh c < 1$ so $x < \sinh x$ as required."
2	Correct statement of the theorem, no more
1	Ratio formula for the theorem stated, no conditions, no more

D. Calculus2

Determine whether the following improper integral converges. (Give reasons for your answer.)

$$\int_1^\infty \frac{1}{\sqrt{(1 + x^6)}}\, dx$$

Solution

$$\frac{1}{\sqrt{(1 + x^6)}} < \frac{1}{\sqrt{x^6}} = \frac{1}{x^3}$$

$$\int_1^\infty \frac{1}{x^3}\, dx = \lim_{R\to\infty} \int_1^R \frac{1}{x^3}\, dx.$$

$$= \lim_{R\to\infty} \left(-\frac{1}{2R^2} + \frac{1}{2}\right) = \frac{1}{2}$$

So, the original integral converges, by the comparison test.

Table 12 Scoring examples for Calculus 2

Score	Example
5	All correct
4	Chosen comparison right, but integral evaluated as $-1/4x^2$, with consistent valid conclusion
3	Evaluation of $\int_1^\infty \frac{1}{X^3} dx$ correct, but no comparison made.
2	Wrote $\frac{1}{\sqrt{(1+x^6)}} < \frac{1}{x^3}$ but no more
1	Stated $\int_1^\infty f(x)\, dx = \lim_{R\to\infty} \int_1^R f(x) dx$ but no more

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 3, 95-105 (2022)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

A Secure Trust Aware ACO-Based WSN Routing Protocol for IoT

Afsah Sharmin*, Farhat Anwar, S M A Motakabber, Aisha Hassan Abdalla Hashim

Faculty of Engineering, Department of Electrical and Computer Engineering, International Islamic University Malaysia, Kuala Lumpur, 53100, Malaysia

ARTICLE INFO

Article history:
Received: 14 April, 2022
Accepted: 16 May, 2022
Online: 27 May, 2022

Keywords:
Internet of Things
Wireless Sensor Network
Routing algorithms
Ant Colony Optimization
Security
Trust Value
Energy consumption

ABSTRACT

The Internet of Things (IoT) is the evolving paradigm of interconnectedness of objects with varied architectures and resources to provide ubiquitous and desired services. The popularization of IoT-connected devices facilitating evolution of IoT applications does come with security challenges. The IoT with the integration of wireless sensor networks possess a number of unique characteristics, so the implementation of security in such a restrictive environment is a challenging task. Due to the perception that security is expensive in terms of computation, power and user-interface components, and as sensor nodes or low-power IoT objects have limited resources, it is desired to design security mechanisms especially routing protocols that are light weighted. Bio-inspired mechanisms are shown to be adaptive to environmental variations, robust and scalable, and require less computational and energy resources for designing secure routing algorithms for distributed optimization. In IoT network, the malicious intruders can exploit the routing system of the standardized routing protocol, e.g., RPL (The Routing Protocol for Low-Power and Lossy Networks), that does not observe the node's routing behavior prior to data forwarding, and can launch various forms of routing attacks. To secure IoT networks from routing attacks, a secure trust aware ACO-based WSN routing protocol for IoT is proposed here that establishes secure routing with trustworthy nodes. The trust evaluation system, is enhanced to evaluate the node trust value, identify sensor node misbehavior, and maximize energy conservation. The performance of the proposed routing algorithm is demonstrated through MATLAB. Based on the proposed system, to find the secure and optimal path while aiming at providing trust in IoT environment, the average energy consumption is minimized by nearly 50% even as the number of nodes has increased, as compared with the conventional ACO algorithm, a current ant-based routing algorithm for IoT-communication, and a present routing protocol RPL for IoT.

1. Introduction

The IoT (Internet of Things) is an evolving technology that performs a significant role in interconnecting intelligent devices or objects that surround us into a network. Integration of wireless sensor networks (WSNs) and IoT, offer a wide variety of applications domains that contour human life and also have influence on economic benefits. The IoT applications have touched its presence in many spaces, such as smart homes, smart cities, smart grid systems, banking, healthcare, environmental monitoring, transportations, data management and analysis and agriculture etc. The evolution of novel applications, systems, and technologies are intensifying attention from the research perception as well. Fueled by the extensive use of systems of interconnected intelligent objects or things enabled by wireless technology such as radio frequency identification (RFID), Wi-Fi, Bluetooth, embedded sensor and actuator nodes, cell phones, IoT is transforming into a fully integrated future internet from the static internet that would provide autonomous, smart behavior and pervasive communication networks for smart connectivity and context-aware computation. This paper is an extension of work originally presented in ICCCE'21 [1].

The present network protocols for wireless communications become inadequate when it comes to the IoT because of the large upsurge of IoT objects, diversity of continually emerging devices and possessions, and heterogeneity among objects' architectures.

*Corresponding Author: Afsah Sharmin, Faculty of Engineering, Department of Electrical and Computer Engineering, International Islamic University Malaysia, 53100 Kuala Lumpur, Malaysia Tel: +60142217380
Email: afsahsharmin@gmail.com

Especially, when the objects or nodes in the IoT system have limited resources in term of energy, memory and processor and the topology changes due to the mobility of the nodes. There are numerous applications, systems, and services from diverse manufacturers, as well as a wide range of hardware and software requirements, making a comprehensive compliance process for efficient and secure IoT-communication difficult to achieve. Also, the adaptation process of conventional communication protocol to take into account the structural and logical characteristics of the IoT modules is also very challenging.

During IoT routing, which influences and aids the interconnectivity of devices, a crucial deliberation focuses on energy efficiency, secure communication, scalability, computational complexity, autonomy, changed environmental issues, node mobility, resource constraints, and QoS (quality of service) requirements for specific applications. The sensors deployed in an IoT system, are energy-constrained and characterized by their self-organization; they sense, monitor and collect data, and perform computational functions while communicating over wireless networks and lossy channels. Because of the distinctive features of IoT networks, the system is subject to a variety of attacks, and many IoT devices are low-powered and computationally weak, and are not built to address security and privacy issues, a security breach could occur in such a system. Despite the fact that various IoT-specific routing protocols have been designed for providing routing decisions, within satisfying resource consumption, they have not been exhaustively verified for trustworthiness [2]. A secure routing protocol is essential for secure exchange of data with the intended parties rather an attacker and a mechanism is required for the stipulation of predetermined participants or discovering trustworthy nodes to collaborate with. Wireless sensor network which is constructed on autonomous nodes collaboration, plays a vital role in providing ubiquitous computing facilities to the diversification of IoT. There are numerous threats and challenges in the area of communication security, and wireless communication is particularly vulnerable to data exposure. The importance of route security is likewise high because the nodes are spatially scattered over a large area and the base station may be located distant from the information-carrying sensor node or device, requiring multi-hop communication to cooperatively send data over the network to a main location or the sink node for which routing path is necessary [3].

To protect the information in IoT devices, a significant range of secure routing algorithms and security mechanisms, including cryptographic techniques for message integrity, have been proposed by the scientific community. When malicious nodes or internal adversarial nodes or internal compromised nodes are present, the keys exchanged for interactions with the other nodes in the network are compromised as well. Most of the secret keys distribution algorithms are computationally expensive and take additional resources such as large memory space and CPU cycles, and that would result in performance degradation, while making it difficult to distinguish between malicious and non-malicious nodes using solely cryptographic measures. As a result, they are inappropriate for resource-constrained network systems. The notion of providing security is pricy with regard to compute, electrical energy, and user-interface components due to low-powered IoT objects, sensor and actuator nodes. If the encryption keys are accessed by the attackers, the whole network's data could be susceptible to exposure. If the protocol does not take the node's behavior into account throughout the routing process, security attacks like Rank attacks and Sybil attacks can be carried out without difficulty, paving the way for further insider attackers. These attacks can be mitigated by employing trust-aware secure routing protocols.

Existing WSN and IoT routing protocols are unable to adequately set of scales security and energy consumption, resulting in routes that are not globally optimum and might fail to function in the face of malicious attacks, threats and vulnerabilities. Bio-inspired processes offer low-cost options for developing secure routing algorithms that find the optimal path. Furthermore, finding trustworthy neighbors is a critical responsibility. Thus, an accompanying security solution known as trust management has been applied and enhanced [4]. In order to manage the network's highly dynamic topology while preserving energy efficiency during data transfer, various intelligent systems and biological systems, as well as the techniques by which they solve their everyday challenges, are used in the construction of secure routing algorithms. The ant colony optimization (ACO) system is a bio-inspired algorithm that uses the notion of self-organization to aid ants' coordination for solving problems. This technique is notably inspiring for addressing security issues in IoT network routing, as ants create paths that satisfy precise constraints in a graph. Bio-inspired processes are robust, adaptive, and scalable, and they aid in the design of optimal algorithms and distributed systems. The probability formula is utilized for route selection in ACO, which is a probabilistic process, while the pheromone update formula is used for pheromone trail updating [5].

EICAntS (Efficient IoT communications based on ant system) is an ant-inspired routing strategy for optimizing IoT communications that was proposed in [6]. The energy parameter is used in the calculation of the global efficiency factor, which represents the ant colony system's pheromone estimations. This approach extends network lifetime while reducing energy usage. The energy impact concentrates the data class that the node manipulates. The various difficulties afflicting the energy factor, for instance small-scale multi-path fading and large-scale fading, free-space path loss in wireless communications are not indicated here. Furthermore, no precise details for calculating the energy level of the nodes are provided. Three routing metrics, ETX (Expected Transmission Count), load or content, and residual energy, are utilized separately and in combination in [7] to enhance the design of the proactive routing protocol RPL (Routing Protocol for Low Power Lossy Network) objective function (OF), that is used to automate the route development method, for IoT applications. Residual energy (RE) in conjunction with ETX (EE) and an upgraded timer setup is effective for energy consumption. On the other hand, unlike the ant colony based approach, which employs the mechanism at work in ant-colony foraging, there is no optimization model used. Using the principles of rank threshold limitations and hash chain authentication, a secure-RPL (SRPL) protocol is suggested in [8] to minimize the influence of rank manipulation. This technique is seemed to be computationally expensive as it combines cryptography with hash chain authentication. In addition, nodes are vulnerable to insider attacks. In [9], the authors suggested a trust-based threshold method for the selection of a parent node to provide security countermeasures to Rank attacks amid RPL routing. The scheme's benefit is that the attacking node is recognized in the course of selection process of

the parent node, which mitigates Rank attacks. The scheme's downside is that additional susceptible attacks, such as blackhole and Sybil attacks, cannot be identified and alleviated well. The authors proposed approaches for the detection and mitigation of Rank attacks that are inconsistent with RPL-supported IoT in [10]. The node's trustworthiness is not considered by the technique, which leads to further security issues, targeting network traffic and resources.

In RFSN [4] framework, the sensor nodes keep reputation about other nodes in the system. Within this framework, a beta reputation system that uses Bayesian formulation has been employed. Using a watchdog mechanism, a node observers the behavior of other nodes. In this way, their reputation is built over a period which help to evaluate their trustworthiness to collaborate. Also, their future behavior is predicted. Direct and indirect reputation are built up using direct observations and second hand information respectively. The statistical expectation of the probability distribution signifies the reputation, which is used to calculate trust. However, this schema does not include a provision for distributing information about a bad reputation. As a result, it is unable to cope with uncertainty. Secure alternate path routing in sensor network (SeRINS) detects and isolates the compromised nodes by providing key management system along with the neighbor report system where the inconsistent routing information have been injected by those nodes [11]. Here, the compromised node is found out using neighbor report technique. The base station then broadcasts the compromised node's ID, key ring to the entire network and that malicious node is excluded using revocation of its cryptographic keys network-wide. However, the proposed technique needs huge changes to apply in the network as it is majorly embedded in the routing arrangement and neighboring nodes can eavesdrop.

A hybrid tree-based search approach called ANT-BFS is presented in [12] to determine the best and shortest information transmission route in order to enhance network performance. ACO is used in conjunction with breadth first search algorithm to investigate the neighbors so as to identify the solution or the requisite node. The amount of energy used is reduced with this strategy. The execution of BFS in ACO, on the other hand, necessitates more memory and computation time. In [13], quantum computation method is introduced and a new WSN routing algorithm, named the Quantum Ant Colony Multi-Objective Routing (QACMOR), is proposed to monitor in complex manufacturing environments. The node pheromone is characterized by quantum bits and to update the pheromone concentration of the path, the quantum gates are rotated. This method improves convergence performance and saves energy consumption. However, the computational complexity of the algorithms and effects on QoS are not addressed in this technique and need to be considered as QoS is posed by real-time applications. In [14], a routing protocol REL for IoT based on residual energy and wireless link quality estimate is suggested to improve reliability and energy efficiency. It enhances the quality of service (QoS) of IoT applications. To improve protocol reliability, received signal strength indication (RSSI) and signal to noise ratio (SNR) are used to generate the link quality estimate for wireless links. To reduce protocol overhead, an opportunistic piggyback technique is implemented, and the residual energy is transmitted to adjacent nodes to increase energy consumption. Despite this, no better approach is used, as opposed to the ant-inspired routing algorithm, which makes use of the ant-based system.

The proposed system of ours [15] has been analyzed more here to efficiently balance security and energy consumption. The proposed routing algorithm has considered important communication parameters for data transmission in an IoT network, such as mobility and energy parameters. This paper extends the work reported in [1], where a secure bio-inspired routing protocol based on ant colony optimization (ACO) systems is proposed, with the intension of providing trust in WSNs while improvising efficient IoT communications.

The relevant work is introduced in Section 1 and the rest of this paper is organized as follows. The system model and energy consumption model are discussed in Section 2. Section 3 depicts the proposed scheme in detail, including the proposed secure ACO algorithm and its design concept, the trust model used as a security mechanism, and trust assessment. The performance of the proposed system is evaluated in Section 4. Finally, some concluding remarks are provided in Section 5.

2. System Model and the Energy consumption model

The chosen network system is based on an IoT sensor organization in which the deployed nodes, sensors and actuators M_i are dispersed throughout the monitored area at random. The graph G linked with the nodes and symmetrical communication links makes up the system model for an IoT communication network. The energy and computational resources available to the sensor nodes in the region are the same. The received signal strength indication (RSSI) can be used by the nodes to calculate the estimated distance of the transmitters, where the transmission power must be acknowledged. The nodes can adjust transmission power and keep records of their neighbors' information updates. The presented system uses the radio energy model of wireless communications [16] and is implemented utilizing (1) to analyze the energy consumption. The quantity of energy consumed is determined by the distance between transmitting and receiving nodes, the size of the packets, and a distance-threshold value, d_0. The two types of energy consumption models applied here are free-space (the transmission power attenuates inversely proportional to d^2) and multi-path fading (the received power is falling off inversely with d^4) models. The following equations are used to calculate the energy consumption (E_{tran}) by the sensors during the transmission of an m-bit data:

$$E_{tran}(m,d) = \begin{cases} mE_{elec} + m\epsilon_{fs}d^2 & if\ d < d_0 \\ mE_{elec} + m\epsilon_{mp}d^4 & if\ d \geq d_0 \end{cases} \quad (1)$$

where ϵ_{fs} and ϵ_{mp} are the amplifying radio's energy consumption in the free-space and multi-path fading models, respectively. The distance is denoted by d, while the threshold value for the distance is denoted by d_0. Electronic devices' circuitry is powered by energy dissipation, E_{elec}. Now, $E_{Rx}(m)$, the reception energy for an m-bit data for a node, can be calculated as follows:

$$E_{Rx}(m) = mE_{elec} \quad (2)$$

The following equation is used to compute the residual energy (E_{res_i}) of a node n_i :

$$E_{res_i} = E_{tot_i} - E_{tran_i} \quad (3)$$

where E_{res_i} denotes the residual energy, E_{tot_i} represents the total initial energy and E_{tran_i} represents the transmission energy.

3. Proposed Secure Routing Protocol based on ACO

3.1. Proposed improved ACO Algorithm

a) The state-transition formula:

By examining the nodes' stable energy consumption while keeping in consideration the security issues for next-hop routing to determine the most trustable route getting to the node that delivers the certain required provision, an enhanced network routing algorithm based on ACO is proposed here. Hence, the next hop selection by the ants depends upon residual energy level of the neighbor nodes, i.e., the node with greater energy level possessing more probability of being selected higher, and their trust value, i.e., regarding the nodes' high trust value as probabilistic next-hop for routing. Assume if an ant *m* is located at node *i* at time *t*, it will comply to the following probability formula to choose the subsequent node *j* as the information forwarding node of the ensuing route for the proposed improvement of our ant colony optimization based routing algorithm [15]:

$$P_{ij}^{m} = \begin{cases} \frac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}(t)]^{\beta}[\vartheta_{ij}(t)]^{\gamma}[T_{ij}(t)]^{\psi}E_j}{\sum_{s\subset allowed_m}[\tau_{is}(t)]^{\alpha}[\eta_{is}(t)]^{\beta}[\vartheta_{is}(t)]^{\gamma}[T_{is}(t)]^{\psi}E_s} & , j \subset allowed_m \\ 0, \quad others & \end{cases} \quad (4)$$

$$\text{here,} \quad \eta_{ij}(t) = \frac{1}{d_{ij}} \quad (5)$$

where $\tau_{ij}(t)$ is the amount of pheromone deposited on edge (i, j) and $\eta_{ij}(t)$ is the state transition desirability of edge (i, j). A priori knowledge, typically the heuristic value $\eta_{ij}(t)$ is $1/d_{ij}$ and d_{ij} is the distance between *i* and *j*. There are two impact factors, α and β, that control the influence of the pheromone intensity and heuristic value respectively. In accordance with the average node mobility or speed, the stability factor, $\vartheta_{ij}(t)$, is determined, where γ is the mobility constant. $T_{ij}(t)$ is the high trust value of nodes at time *t* or the trust metric and ψ is the impact factor that control the influence of trust level among the nodes to further communication. The calculation of the trust metric will be provided below. E_j represents the node residual energy that ant *m* would visit.

a) The local update:

After an ant finish mapping a node *i* to node *j*, the corresponding pheromone intensity $[\tau_{ij}(t)]$ is updated by a local pheromone updating rule according to (6). Besides enhancing diversity of the algorithm, local pheromone update is augmented here in case a large quantity of pheromone value is accumulated down the pathways, while preventing faster local convergence. As a result, the pheromone measure is restored and controlled through the use of a threshold rating, and the updating rule which is assessed by following equation:

$$\tau_{ij}(t+1) = \begin{cases} T, & \tau_{ij}(t+1) > T \\ (1\text{-}\rho)\tau_{ij}(t) + \Delta\tau_{ij}(t) & else \end{cases} \quad (6)$$

$$\Delta\tau_{ij}(t) = \sum_{k=1}^{m} \Delta\tau_{ij}^{k} \quad (7)$$

where ρ signifies the local pheromone decay parameter, $\rho \in (0,1)$, a threshold value T is provided to restrict excessive pheromone accumulation, $\Delta\tau_{ij}(t)$ is the appended pheromone deposition of link (i, j), which is typically specified as below:

$$\Delta\tau_{ij}^{k} = \begin{cases} \frac{S}{L_k} & if\ kth\ ant\ travels\ on\ the\ edge\ (i,j) \\ 0 & otherwise \end{cases} \quad (8)$$

where *S* is a constant represents the strength of pheromone, L_k is the cost of the k^{th} ant's tour known as length, and *m* is the number of ants.

3.2. Trust Model used as a Security Mechanism

Trust framework involves two participants: trustors and trustees, who work associatively to accomplish a particular job based on a node's estimated trust value to determine its trustworthiness. A weighed value index, denoted as the node's trust rating, is assessed depending on the judgement of a node's prior behaviour, which also establishes the node's reputation [4]. The trust management approach identifies malicious and other attackers and compromised nodes in the communication network by evaluating their trust value in order to deal with unpredictability about the nodes' future activities. This reflects a node's trustworthiness while engaging with its neighbors, either directly or indirectly, to complete a set of specified activities. Consequently, the behaviour of a node is observed and that defines the positive or negative interactions over time, which is demonstrated as expressions to represent the calculable range of values, such as the trust value rating, while also indicating a node's reputation metric. For secure routing, neighbors or adjacent nodes with greater trust metric values are chosen, whereas nodes having lower values of trust or if trust values of these suspect nodes do not increment with time, then these nodes are identified as malicious. The security mechanism described in [4], manipulating a watchdog mechanism, is utilized and expanded upon here, where a beta reputation system based on Bayesian formulation is applied via direct/indirect observations to signify reputation metric.

The direct trust (DT_{ij}) is the rating of the latest activities of node *j* by node *i*. and it is calculated through the expected value of the probability distribution function. While allowing for consideration of the beta distribution and beta function, as a previous distribution property in the communications between the nodes, yields the direct trust value that node *i* has on node *j*. The direct trust is represented as follows:

$$DT_{ij} = \frac{\acute{\alpha}_j + 1}{\acute{\alpha}_j + \acute{\beta}_j + 2} \quad (9)$$

where $\acute{\alpha}_j$ indicates the successful or cooperative interactions and $\acute{\beta}_j$ indicates the unsuccessful or non-cooperative interactions or interactive behaviors between node *i* and *j* accordingly from the perspective of node *i*.

While an entity can make a precise direct trust judgement based on direct observation without a third party involvement for its adjacent nodes, it relies on the recommendations of trusted nodes to assess trust for packet transmission to nodes that are not directly connected. In case of uncertainty, presume that the evaluating node *i* needs the recommendation from a third entity and acquires reputation rating of node *j*, via their commonly adjoining nodes *k*. According to the principle of trust transfer decline, the recommended trust metric is calculated by following equation:

$$RT_{ij}^{k} = DT_{ik} * DT_{kj} \tag{10}$$

here, RT_{ij}^{k} gives the recommended trust value that node *i* possesses about node *j* offered by the common neighbor nodes *k*, and is derived by the product of the direct trust values, DT_{ik} and DT_{kj}. Accordingly, DT_{ik} represents direct trust rating between nodes *i* and *k*, and DT_{kj} represents direct trust rating between nodes *k* and *j*.

The trust metric $T_{ij} \in [0,1]$ of node *i* holds for *j* is the operational trust rating that is computed by collecting interactive records from third parties through direct observation or indirect observation. The weighted average, an associatory trust aggregation function, is computed by combining the estimates RT_{ij}^{k}, where distinct recommenders' provisions are brought into consideration for computing RT_{ij}^{k} particularly from each trusted edge. The trust metric is given as follows:

$$T_{ij} = \sum_{k \in N_i} \left(RT_{ij}^{k} * w_k \right) \tag{11}$$

$$\text{here, } w_k = \frac{DT_{ik}}{\sum_{k \in N_i} DT_{ik}}, \; k = 1,2, \ldots, N_i. \tag{12}$$

$$DT_{ik} = \frac{\acute{\alpha}_k + 1}{\acute{\alpha}_k + \acute{\theta}_k + 2} \tag{13}$$

where the weight w_k is assigned depending on recommenders' trust levels to lessen the influence of personal choice. In the above equation, $w_k (0 \le w_k \le 1, \sum_{k=1}^{N_i} w_k = 1)$ is the weight of RT_{ij}^{k}. The direct or indirect recommendations for node *j* received by node *i* from a set of trusted nodes denoted as N_i. It also indicates the number of received recommendations that is utilized. DT_{ik} represents the direct trust values between nodes *i* and *k*, while $\acute{\alpha}_k$ and $\acute{\theta}_k$ represent the prior recommendation or reputation metric, successful and unsuccessful interactive records accordingly that node *i* already possesses about node *k*.

In the proposed system, every sensor node manages and controls its own pheromone traces while not adding too much overload to the network, and maintaining the lightness of the model. Moreover, to gather ratings, there is not any central or supervising entity and each transmitted ant carries the sensors' identifications along with the pheromone traces.

On the contrary, a reputation rating depending upon pheromone value, τ_{ij}, of a communication route can be established here where the higher is the pheromone trace, the quality of the path, the higher is the security. Every node saves its own pheromone traces and the pheromone traces for its neighbors. In this sense, a more secured route is with more pheromone deposits, implying that a linked node holds greater packet forwarding or collaborative capabilities. The deterministic factor as well as this pheromone measure, $\tau_{ij} \in [0, 1]$, will determine the probability of ants selecting one path or another and the trust value in association with reputation metric provided by an entity to another node specifies the deterministic factor. If the reputation of a node at time *t* is denoted by $\varphi_{ij}(t)$, then the following equation can be applied for the detection of a malicious node:

$$\tau_{minimum} = \frac{\sum_{i=1}^{n_k} \tau_{ij}(t)}{n_k} \tag{14}$$

where τ_{ij} represents the pheromone quantity in between nodes *i* and *j*, and the number of *i*'s neighbors is n_k. If $\varphi_{ij}(t) < \tau_{minimum}$, which indicates the node's reputation falls below the minimum reputation conditions, $\tau_{minimum}$, then security threat or node's misbehavior is detected, and this node is identified for its malicious tasks, and will have fewer forwarding capabilities.

3.3. Trust Assessment

High-trust level nodes are used for routing decisions or secure communications by the proposed method. During the trust calculation process when the trust values have been determined, a trust assessment system is further adopted for ranking the highest to the lowest trust values, T ([0, 1]). This will further help to detect and eliminate the misbehaving node, where nodes with lower trust values are categorized as malicious. The membership degree and fuzzy classification of nodes' trust are implemented here. Three grades or level of trust have been provided for trust evaluation of a node by using fuzzy judgment as: distrust, uncertain and completely trust level or state which is represented in Table 1. Three fuzzy subsets T1, T2 and T3, as shown in Fig. 1, and the corresponding membership functions are defined as m1 (t), m2 (t) and m3 (t) and m1 (t) + m2 (t) + m3 (t) = 1.

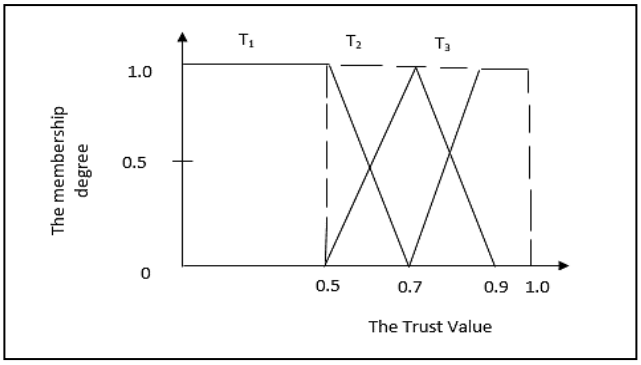


Figure 1: The membership function of node's trust

Table 1: Trust States

Three fuzzy subsets (T)	Trust level
$\boldsymbol{T_1}$	Distrust

T_2	Uncertain
T_3	Completely trust

3.4. The global update and the fitness-function

The global updating of pheromone concentration is worked out after all of the ants have constructed their solutions, finishing their search and have appeared at the target node. In addition, after the search is completed, each ant corresponds to a routing path. To begin, the path estimation rating, which is a route assessment function, is provided as (15) using the existing node energy, route length [m^{th} ant's route length L_m^k in k^{th} iteration], and trust metric. Some nodes will die prematurely if the residual energy [E_{res_i} for a sensor node n_i] is not analyzed as it is in the typical ACO method, reducing the network's overall lifespan. The path's fitness value can be calculated as follows:

$$f_{(fitness)_m^k} = \frac{E_{res_i}}{L_m^k} * T_{ij} \tag{15}$$

here T_{ij} is the trust metric of node i holds for j. Then the global pheromone updating applies on the optimum path which is the best-so-far solution, offering the largest fitness value. The pheromone intensity is updated globally according to the following equation:

$$\tau_{ij}(t+1)=(1-\delta)\tau_{ij}(t)+\sum_{m=1}^{n}\Delta\tau_{ij}^m \tag{16}$$

$$\Delta\tau_{ij}^m = \begin{cases} R * f_{(fitness_{best})_m^k}, & if\ mth\ ant\ visits\ the\ edge\ i,j \\ 0, & otherwise \end{cases} \tag{17}$$

where $0< \delta <1$ is the global pheromone decay parameter, R is the constant for recompensing the pheromone, n denotes the total number of ants, and $\Delta\tau_{ij}^m$ is the increase of pheromone concentration of the edge (i,j) utilized by m^{th} ant, which is proportionate to the maximal cost of fitness equation, $f_{(fitness_{best})_m^k}$, if edge (i,j) is associated with the global best route.

3.5. Proposed Improvement

Clustering is contemplated on attaining scalability while maintaining security. As a result, the routing protocol presented in [15] can be used in conjunction with a clustering based routing approach, such as LEACH [16], a hierarchical clustering protocol. It takes into account the data forwarding probability, nodes' current residual energy, the trust metric, and nodes distance from the base station (BS) and improves the optimal cluster head (CH) selection technique. The flow diagram representing the proposed enhancement is shown in Figure 2.

The probability ($Prob_i$) of a node being selected as a cluster head, an ant m can apply the probability calculation equation given in (18). Node i is presumed to be the present cluster head node then the next node j to be selected as the subsequent cluster head, where the trust metric (T_{ij}), P_{ij}^m , the node distance ($dist_i$), as well as two control parameters (α and β) are used, and the following probability equation is applied:

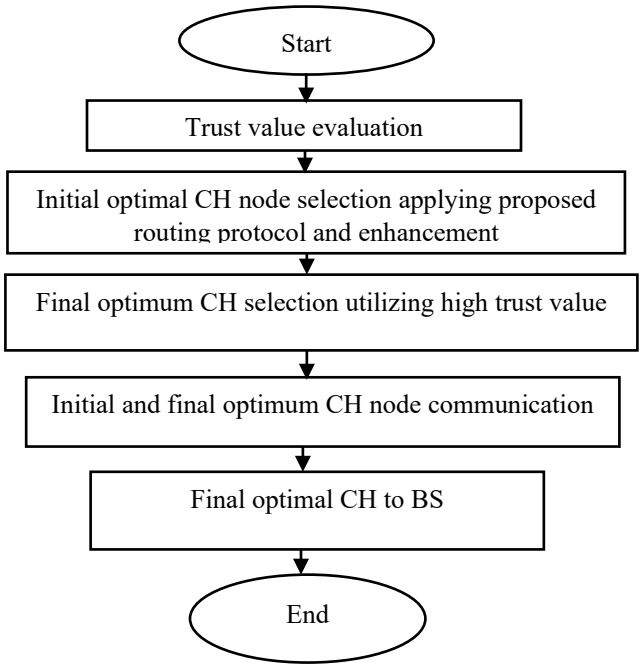


Figure 2: The flow diagram of the proposed improvement

$$Prob_i(t)=\frac{dist_i*\alpha+[P_{ij}^m(t)]*\beta}{\sum_1^{N_i} dist_i*\alpha+[P_{ij}^m(t)]*\beta} * T_{ij} \tag{18}$$

here T_{ij} signifies the trust metric, P_{ij}^m is computed from (4), and N_i is the set of nodes in the cluster.

3.6. Trusted Parent Selection

Algorithm 1: Trust Calculation and selection of trusted parent

Let $M_1 \leftarrow$ any obtainable entity in the Neighbour_List[ ]
Let $M_2 \leftarrow$ another entity next to M_1 in the Neighbour_List[ ]
Calculate

$$T_{ij} = \sum_{k \in M_i} \left(RT_{ij}^k * w_k\right)$$

while node is not found in Malicious_Class_List do
If (M_1.*ETX_metric*<= *ETX_metric-limit*) & (M_2.*ETX_metric*<= *ETX_metric-limit*)
If (M_1.*Rank* <= *Self_Rank*) & (M_2.*Rank* <= *Self_Rank*)
Selected_Parent = $M_1.T_{ij} > M_2.T_{ij}$? M_1:M_2;
else
if (M_1.*Rank* <= *Self_Rank*) || (M_2.*Rank* <= *Self_Rank*)
Selected_Parent = M_1.*Rank* < M_2.*Rank* ? M_1: M_2
else
Selected_Parent = NULL;

end if
else
If (M_1.*ETX_metric* <= *ETX_metric-limit*) ||
(M_2.*ETX_metric* <= *ETX_metric-limit*)
Selected_Parent = M_1.*ETX_metric* <= M_2.*ETX_metric* ?
M_1: M_2;
else
Selected_Parent = NULL;
end if
end while
return Selected_Parent
End. //of program.

The algorithmic procedure implemented here has been given above for selecting the trusted-parents. It includes calculation of the trust values of the nodes and a trust-based method for the selection of parents. The algorithm utilizes the ETX metric as specified in [17]. For the initiation for the optimum parent swap, the minimum required variation of the computed trust value for a node is denoted as $M_1.T_{ij}$. The node having the maximal trust rating along the node's routing path is searched for by the algorithm among all the routes, while the path would also have minimum ETX values, given in (19). The ETX limit represents the maximum ETX rating assessed to be the optimal prospective parent, whereas a node will not select its neighbours that have superior rank as its possible chosen parents. It will also ensure that there is no loop. The trust threshold (Trust assessment Table 1) is utilized for a trusted parent preference, during trust calculation for selecting the node as the chosen parent. Moreover, the rank order is maintained as specified in [18]. Upon identification of a malicious node as a parent, the child node reassigns itself with a different parent from the offered list for selecting a parent node.

The ETX metric, or expected transmission count is calculated as:

$$ETX_{(i,j)} = \frac{1}{D_f * D_r} \tag{19}$$

where D_f defines the forward data delivery and D_r is the reverse data delivery or acknowledgement from the receiver.

4. Result and Discussion

MATLAB is used to accomplish the performance evaluation and simulation. The routing protocol proposed here is compared to the benchmark protocols, where the conventional ACO algorithm, EICAntS algorithm [6], a current ant-based routing method for IoT communication, and a present proactive routing protocol for low power lossy network (RPL) [7] for IoT have been considered as benchmark protocols. There are 100 nodes dispersed in a $100m \times 100m$ area. Some malicious nodes are also deployed across the network at random. The initial trust value is calculated which is set as 0.6, observing the number of interactions, and for that taking reasonable value is crucial. The simulation parameters are set as: $\alpha = 1, \beta = 1, \gamma = 1, \rho = 0.05, \delta = 0.05$. More parameters are presented in the Table 2.

Table 2: Simulation Parameters

Parameters	Values
T	100
Initial trust value	0.6
Initial energy per node	0.5 joule
Node-speed	2 m/s ~ 5 m/s
Transmitted message bits	4000 bits
Distance of transmission	50 m

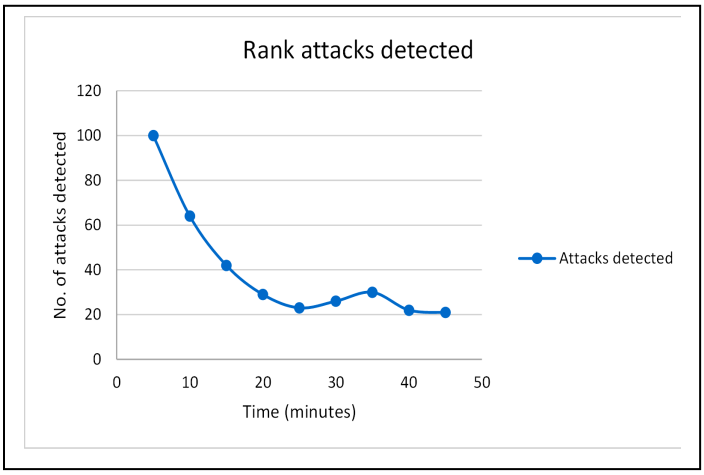


Figure 3: Rank attack detected by proposed secure ACO algorithm

In Trust calculation process, by using the computed trust value, a trustor node assesses a trustee node. It employs the trust metric value to evaluate whether the trustee node is adequately reliable enough of fulfilling an allotted task, and the trust threshold system (Trust assessment Table 1) is utilized for the assessment. Each node collects the direct trust value of directly connected neighbours and recommended trust value of indirectly connected neighbours. The focus of this research is on detecting and isolating internal attacks, particularly Rank and Sybil attacks. A malicious node modifies its rank in a rank attack for disrupting the network route topology, whereas a Sybil node, by using fake identities, tries to subvert the network process. The malicious nodes, taking part in internal attacks of the network, are more challenging to detect as they are aware of the system information of the network. By using node overhearing and monitoring methods, this secure trust-based system perceives unusual route transmission towards a node, and that might be an indication of a rank attack. By assigning a greater weight to a node's existing trust value, a Sybil attack node is detected and isolated. It is also needed not to attribute its observed prior behavior too much weight while defending against a Sybil node as its initial behaviour might be well. So, if it does not have any worthy packet sending behavior that can be observed, its trust value will remain below the threshold, which is necessary for secured communication.

For the simulation study, in the phase of implementing Rank attack, a malicious node initially keeps up with a fine prior behaviour for roughly 5 to 10 seconds. After that during every cycle, it broadcasts spuriously low Rank values and initiates its attack.

From Figure 3, it can be seen that the proposed secure routing protocol is effective at detecting and isolating the Rank attacks. In the course of routing operations, about 100 attacks have been

detected in the first five minutes. Although with the simulation progress, the number of attacks detected has steadily decreased.

In RPL routing operation, a node examines potential parents that have lower rank values than itself and then selects as its chosen parent. In this way, the rank of a node changes and realignment takes place for a child node to another selected parent node that has a smaller rank value. A Rank attack proceeds where the attacker takes advantage of this attribute in RPL routing. It presents itself with a superior rank value to its adjacent nodes and the neighbours are attracted and deceived by this.

The frequency of node rank changes is shown in Figure 4. From the comparison in between MRHOF-RPL (Minimum Rank with Hysteresis Objective Function-RPL) [18] and the secure system presented here, it is observed that the benchmark protocol has notably higher vulnerability to node rank changes than the proposed algorithm, demonstrating a vulnerability to Rank attacks. However, this proposed scheme, persistently has maintained low frequency of node rank changes during all of the simulation period.

From Figure 5, it can be observed that the proposed secure routing protocol is effective at detecting and isolating the Sybil attacks. During the routing procedures, about 272 attacks have been detected in the first five minutes but the number of attacks detected have decreased with time.

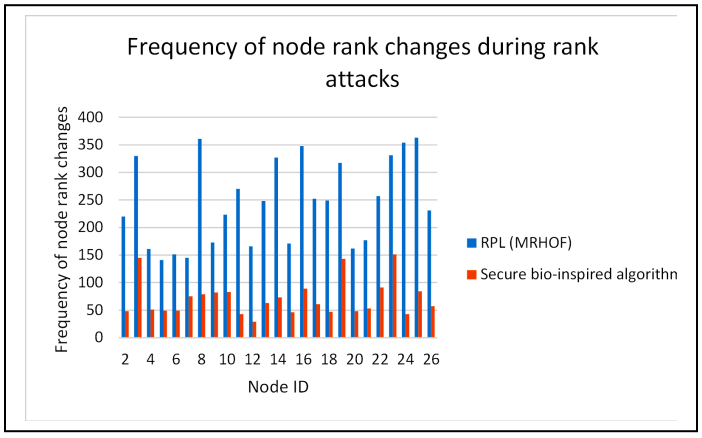


Figure 4: Frequency of node rank changes comparison

The offered ant colony metaheuristic-based routing method is used to discover the optimum pathway for routing packets from the originating node to the target node with the least amount of energy consumption and the highest level of security. The trustworthy nodes are chosen for data transfer in order to build a secure routing path. The graph in Figure 6 shows the relationship between the detection times of a malicious node and the number of nodes in the network. The detection time is defined as the number of malicious nodes found in relation to the simulation time. The percentage of malicious nodes has been retained fixed in this graph, and the value 1 for detection time indicates that no malicious nodes have been found.

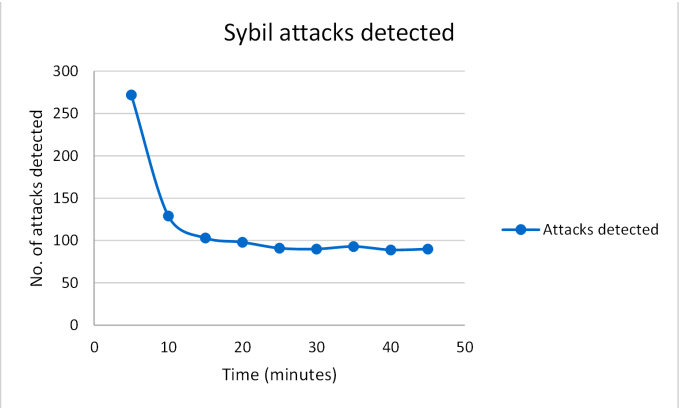


Figure 5: Sybil attack detected by proposed secure ACO algorithm

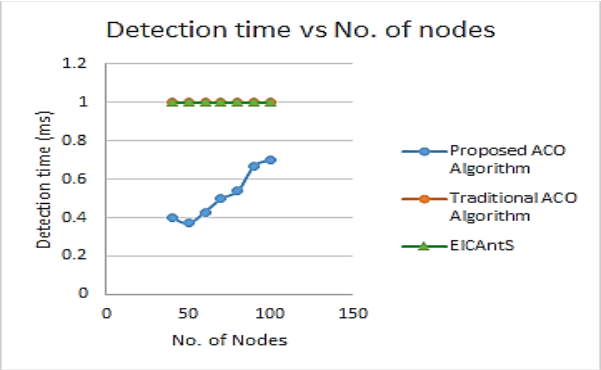


Figure 6: Detection times comparison of a malicious node to the No. of nodes.

The detection of malicious nodes in the network is not considered by the traditional ACO algorithm and EICAntS protocol. As a result, the systems fail to detect any malicious nodes, resulting in lower security and performance. However, the proposed scheme not only converges into the best-so-far path but also the most secure route by taking into account essential transmission factors along with the use of trust to improve security. With the detection and isolation of malicious nodes, it outperforms benchmark protocols while discovering and collaborating with trustworthy nodes via utilizing a trust assessment system. Figure 7 shows the comparison of a malicious node's detection times represented on the ordinate to the percentage of malicious nodes represented on the abscissa accordingly.

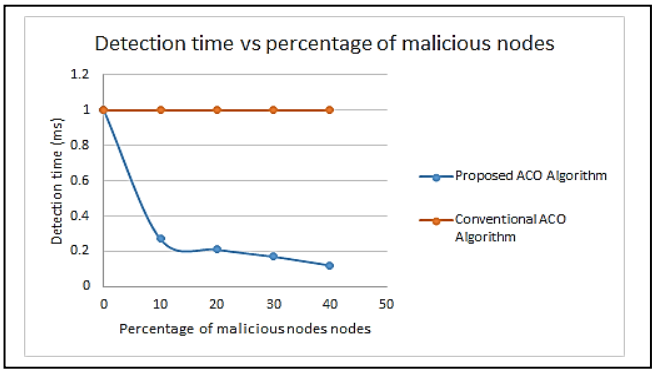


Figure 7: Detection times comparison of a malicious node to the percentage of malicious nodes.

Figure 8 shows a contrast of the consumed energy per transmission of the nodes for the protocol proposed here, the ant-based routing method for IoT communication, i.e., EICAntS, the standard ACO algorithm, and the RPL protocol, demonstrating that the suggested ACO algorithm is better in terms of consuming a lesser amount of energy. The cumulative energy consumption of each node is displayed here every transmission for each individual search operation. In comparison to the previous benchmark protocols, the suggested calculation has clearly achieved refinement, resulting in a substantially lower energy consumption, nearly 50% less for the majority of nodes.

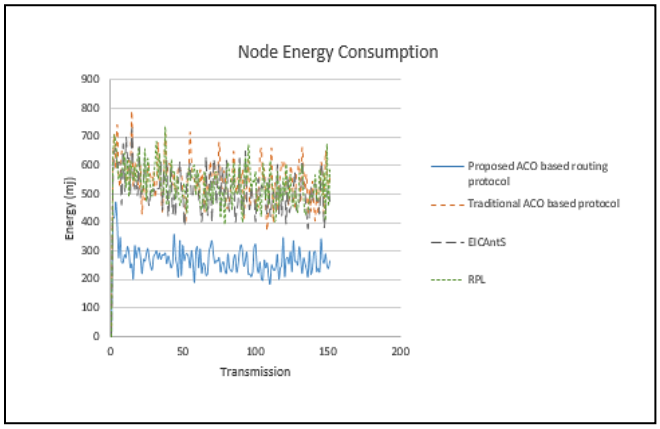


Figure 8: The energy consumption per transmission of the nodes comparison

The presented routing technique based on ACO consumes less energy compared to the traditional ant colony optimization metaheuristic algorithm, proactive routing protocol for low power lossy network (RPL), and efficient IoT communications based on ant system (EICAntS) routing protocol, even in the situations where the number of nodes increases, as shown in Figure 9, where the average energy consumption is lower. When contrasted to the benchmark routing techniques, it is clear that using the suggested ACO-based routing algorithm as an explication reduces average energy consumption, by approximately 50% less and makes the algorithm lightweight. Because the proposed approach enables the optimum packet forwarding path for transmission to be determined, and retransmissions are avoided, providing reliable communication. This method likewise reduces the number of updating phases while optimizing the route selection strategy. According to the outcomes, the more nodes there are, the higher the energy consumption. Another result is that the more malicious nodes there are, the more energy is consumed. The presented framework retains scalability by using less power than the standard protocols taken as the benchmark, even as the number of nodes in the network expands.

As demonstrated in Figure 10, the average End-to-end delay performance metric rises when the number of nodes grows. The proposed routing protocol lessens the repetition issue while also enhancing the procedure for selecting a route because numerous packets have to be sent again to the intended destination if the optimal path is not found and utilized to deliver the packets. Compared to the mentioned benchmark algorithms here, the suggested technique performed well with regard to average End-to-end delay, achieving a nearly 40% decrease in end-to-end delay. Figure 11 shows the throughput results. The network throughput is measured by calculating the total number of packets sent over the complete simulation time, or the measure of digital data transmitted per time unit via a communication link. It is usually expressed in bits per second (bps), although it can also be expressed as data packets delivered per-second or per-time-slot. From the contrast, it is clearly shown that the results obtained by the method proposed here are better than the results recorded by the other network.

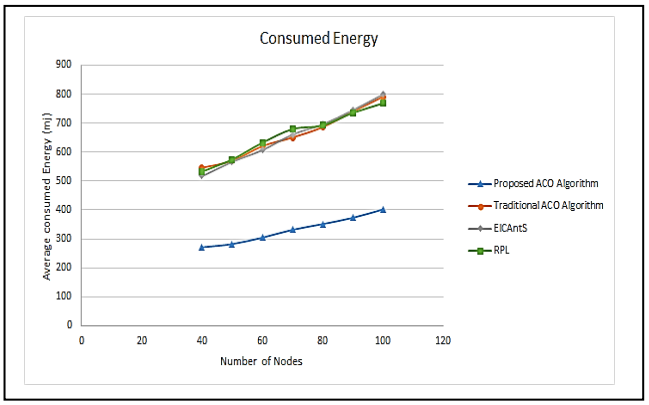


Figure 9: The average energy consumption comparison.

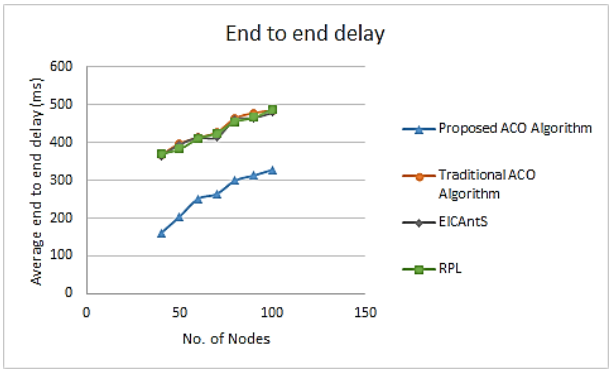


Figure 10: Comparison of the average End-to-end delay with regard to the number of nodes using fitness function

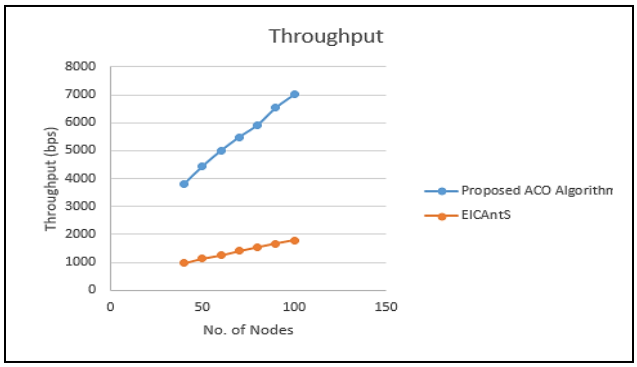


Figure 11: Comparison of Throughput

The findings of the proposed algorithm's calculation of packet delivery rates are shown in Figure 12. Since the protocol devises the ability to deliver numerous packets shortly while also reaching the destination, the proposed approach achieved satisfactory outcomes despite the crucial node quantity. It offers a system for determining the most secure information-transmission path among the network's several routes. Many packets are diverted or

dropped out when there are malicious nodes utilizing other strategies. However, due to security mechanisms, the proposed technique is used to deliver most of the data.

The packet loss ratio is also seen in Figure 13 when malicious nodes are present. The packet delivery ratio diminishes when the percentage of nodes that are malicious rises or as other attackers and compromised nodes exist in the pathways amid communication nodes. When attackers and compromised nodes cannot be detected and packets outreaching the target node successfully decline, the ratio of packet loss or misdirection is significant in the case of the specified benchmark methodologies.

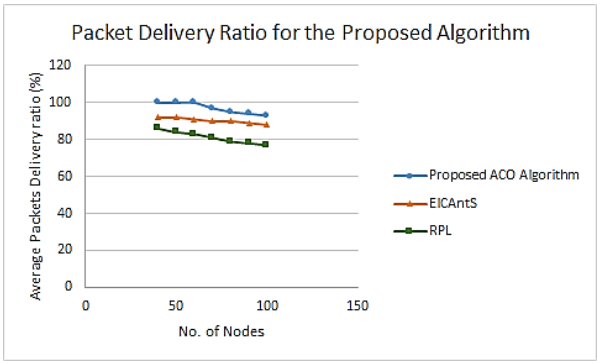


Figure 12: Packet Delivery Ratio (PDR) with the number of nodes

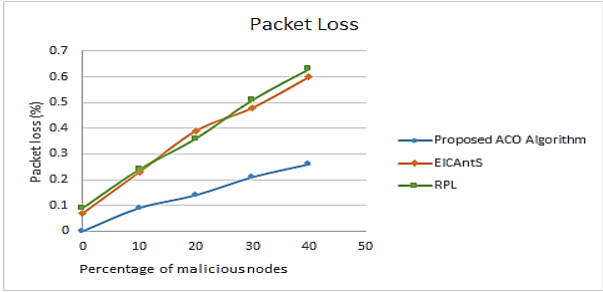


Figure 13: Packet Loss in the existence of malicious nodes

5. Conclusions

Though the IoT technology would be evolving in the coming decade, its multiple and complex aspects need to be considered in the process of developing effective communication protocols. An ACO-based WSN routing algorithm for IoT is proposed in this paper, which uses a trust-based security system while considering the limited resources restraints in sensors or low-power IoT objects, as well as the special necessity of security in the data forwarding process. The trust value is worked out to determine a node's trustworthiness for packet transfer. In order to evaluate route performance, the proposed enhancement and the route assessing function include the trust metric, as well as the existing energy of the nodes and the route length. The energy factor, the trust metric, and the average mobility of the nodes are all included in the ACO algorithm's probability formula as well. When compared to the benchmark methods, the presented ACO-based routing algorithm lowered energy consumption by almost 50% even as the number of nodes rose, making the algorithm lightweight and scalable. It also showed a nearly 40% reduction in end-to-end delay. The routing protocol generates a secure and globally optimal route based on the related information, which includes the neighboring nodes' trust value and residual energy, as well as the path cost from the adjacent node to the sink node. The proposed technique can retain a higher packet delivery ratio due to the security mechanism, which ensures the system's efficacy in addition to the global optimization. Furthermore, by providing trustworthy routing paths, the proposed routing protocol can efficiently balance energy consumption and security.

As future work, presented secure routing protocol would be improved to implement in a real-world setting to estimate the algorithm's performance. Moreover, it will be elaborated to deal with additional conspiring attacks such as a Rank attacking node colluding with Selective Forwarding attacks or having collusion with a Blackhole or a Sybil attack. Finally, previously trusted nodes will be re-assimilated based on their trust levels after having recouped their battery power. These nodes will be deployed administratively to ensure network's balanced secure communication.

References

[1] A. Sharmin, F. Anwar, S.M.A. Motakabber, A.H.A. Hashim, "Secure ACO-Based Wireless Sensor Network Routing Algorithm for IoT," in Proceedings of the 8th International Conference on Computer and Communication Engineering, ICCCE 2021, 190-195, 2021, doi:10.1109/ICCCE50029.2021.9467223.

[2] J. Granjal, E. Monteiro, J. Sa Silva, "Security for the internet of things: A survey of existing protocols and open research issues," IEEE Communications Surveys and Tutorials, **17**(3), 1294– 1312, 2015, doi:10.1109/COMST.2015.2388550.

[3] W. Dargie, C. Poellabauer, Fundamentals of Wireless Sensor Networks: Theory and Practice, 2011, doi:10.1002/9780470666388.

[4] S. Ganeriwal, M.B. Srivastava, "Reputation-based framework for high integrity sensor networks," in Proceedings of the 2004 ACM Workshop on Security of Ad Hoc and Sensor Networks, SASN'04, 66–77, 2004, doi:10.1145/1029102.1029115.

[5] L. Bianchi, M. Dorigo, L.M. Gambardella, W.J. Gutjahr, "A survey on metaheuristics for stochastic combinatorial optimization," Natural Computing, 8(2), 239-287, 2009, doi:10.1007/s11047-008-9098-4.

[6] S. Hamrioui, P. Lorenz, "Bio inspired routing algorithm and efficient communications within IoT," IEEE Network, **31**(5), 74-79, 2017, doi:10.1109/MNET.2017.1600282.

[7] S.S. Solapure, H.H. Kenchannavar, "Design and analysis of RPL objective functions using variant routing metrics for IoT applications," Wireless Networks, **26**(6), 4637–4656, 2020, doi:10.1007/s11276-020-02348-6.

[8] G. Glissa, A. Rachedi, A. Meddeb, "A secure routing protocol based on RPL for internet of things," in 2016 IEEE Global Communications Conference, GLOBECOM 2016-Proceedings, 1-7, 2016, doi:10.1109/GLOCOM.2016.7841543.

[9] I. Kenji, T. Matsunaga, K. Toyoda, I. Sasase, "Secure parent node selection scheme in route construction to exclude attacking nodes from RPL network," IEICE Communications Express, 299–303, 2015, doi:10.1587/comex.4.340.

[10] R. Stephen, L. Arockiam, "E2V: Techniques for Detecting and Mitigating Rank Inconsistency Attack (RInA) in RPL based Internet of Things," in Journal of Physics: Conference Series, **1142**(1), 012009, 2018, doi:10.1088/1742-6596/1142/1/012009.

[11] S.B. Lee, Y.H. Choi, "A secure alternate path routing in sensor networks," Computer Communications, **30**(1), 153–165, 2006, doi:10.1016/j.comcom.2006.08.006.

[12] R. Khoshkangini, S. Zaboli, "Efficient Routing Protocol via Ant Colony Optimization (ACO) and Breadth First Search (BFS)," International Conference on Internet of Things (IThings 2014), (March), 375–381, 2014, doi:10.1109/iThings.2014.69.

[13] F. Li, M. Liu, G. Xu, "A quantum ant colony multi-objective routing algorithm in WSN and its application in a manufacturing environment," Sensors (Switzerland), **19**(15), 3334, 2019, doi:10.3390/s19153334.

[14] K. Machado, D. Rosário, E. Cerqueira, A.A.F. Loureiro, A. Neto, J.N. de Souza, "A routing protocol based on energy and link quality for internet of things applications," Sensors (Switzerland), **13**(2), 1942–1964, 2013, doi:10.3390/s130201942.

[15] A. Sharmin, F. Anwar, S.M.A. Motakabber, "Energy-Efficient Scalable Routing Protocol Based on ACO for WSNs," 2019 7th International

Conference on Mechatronics Engineering, ICOM 2019, 1-6, 2019, doi:10.1109/ICOM47790.2019.8952053.

[16] W.B. Heinzelman, A.P. Chandrakasan, H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," IEEE Transactions on Wireless Communications, **1**(4), 660-670, 2002, doi:10.1109/TWC.2002.804190.

[17] J. Vasseur, M. Kim, K. Pister, N. Dejean, D. Barthel, "Routing metrics used for path calculation in low power and lossy networks," Draft-Ietf-Roll-Routing-Metrics, 2011.

[18] T. Winter and P. Thubert "RPL: IPv6 Routing Protocol for Low power and Lossy Networks," IETF Internet-Draft, 2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 2, 103-111 (2022)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Solar Energy Assessment, Estimation, and Modelling using Climate Data and Local Environmental Conditions

Clement Matasane[1,*], Mohamed Tariq Kahn [2]

[1]*Cape Peninsula University of Technology (CPUT), Electrical and Electronic Engineering Department, Symphony Way, Bellville Campus, Bellville, 7925, South Africa*

[2]*Cape Peninsula University of Technology (CPUT), Research Chair: Energy, Director: Energy Institute, Head: Centre for Distributed Power and Electronic Systems, Head: Centre for Research in Power Systems, Symphony Way, Bellville Campus, Bellville, 7925, South Africa*

ARTICLE INFO

Article history:
Received: 17 July, 2021
Accepted: 11 February, 2022
Online: 18 March, 2022

Keywords:
Solar Energy
Radiation
Insolation
Climate Data
Potential energy
Geographical Parameters

ABSTRACT

On Renewable Energy (RE), this field covers the most significant share of the world energy demand and challenges on the expensive measurement and maintenance equipment to be used. In all studies and designs, global solar radiation (GSR) measurements require assessment, estimation, and models to be applied together with the environment and meteorological data on installing stations at the specific location. These meteorology stations provide measured data throughout the year/ annually or at specified periods, depending on the site of interest. This study includes assessment and estimations of the solar radiation at the Vhembe District using the geographical data measured daily, monthly, and throughout a year in the area. It provides variables such as the geographical maps of the solar availability at a minimum and maximum temperatures obtained during the annual analyses. Determining the solar radiation at a specific location for energy generation involves several procedures, estimations, and calculations using the climatological weather data measurements through MATLAB simulations. In addition, the Geographical Remote Sensing (RS) and Mappings, and Spreadsheet Graph Analytics, were applied to the measured data from the nine installed Weather Stations (WS) in the Vhembe District area was used. The analysis determines the minimum and maximum solar radiation equations associated with the local climate patterns in accommodating the theoretical bases and period changes. The paper contributes to the main project objectives on renewable energy assessment for potentials and generation at a micro/small scale in the district. These parameters are fundamental in estimating and determining the potential solar energy radiation using its extraterrestrial solar radiation per day/ weekly/ monthly. Annual periods towards methods to develop micro/small energy projects for rural and urban communities for domestic and commercial use. As a result, the meteorology analysis is being presented in this study.

1. Introduction

This paper is an extension of work initially presented at the 2019 IEEE PES/IAS Power Africa Conference held in Abuja, Nigeria [1]. This article provides an extension and detailed result to determine the daily, monthly, annual, and solar potential and radiation within the Vhembe District. This demonstrates the estimating of the energy potential as part of the sub-energy potentials obtained from the wind, biomass/biogas, and hydro energy for the optimal energy generation in the area.

Solar energy applications play a significant role in health, agriculture, civil engineering, and the environment for their execution to support the energy demands within the domain [1], [3]. Hence, evaluating the solar energy potential at any specified location requires accurate solar radiation information. The sunshine duration from the most common variable for predicting

*Clement Matasane, CPUT, Symphony Way, Bellville, 7925, South Africa, +27(021) 4603383 & matasanec@cput.ac.za

global solar radiation (GSR), so sunshine duration can be easily calculated, reliable, and widely used. In increase, solar radiation is the primary root of energy and varies per amount of energy received at different locations [2], [28], [30]. The current developments towards sustainable energy savings and generation using the solar photovoltaic (PV) units have accelerated the maturation process and investment in the area [3]-[5], [11], [29]. In the Limpopo Province (coordinates as 22°50 "22. 08" S and 30°18 "36" E), there is adequate sunlight, which can be more utilized for solar energy applications as shown in Figure 1 to Figure 3. The images were obtained from the Global Solar Atlas developed by the Energy Sector Management Assistance Program (ESMAP) and SOLARIS supported by the World Bank. It is thus essential to harvest and store this natural resource to find a solution to energy shortages and environmental degradation at the state. It is of the view that solar energy systems are considered the most cost-effective and economic power systems in providing off-grid electricity generation in rural areas in the province.

Estimating the renewable energy (RE) potentials using geographical and climatological data requires thorough calculations per specified location. Hence the quantity of solar energy per location is essential, as shown aside from the direct average irradiation, global horizontal irradiation, and potential photovoltaic power for the region from Figure 1 to Figure 3.

More geographical, climatic, and analysis data were added to present conditions and their placement in this paper. Figure 1 to Figure 3 illustrates the available solar map available in the provinces, demonstrating that in the region of Limpopo Province, especially in upper streams, there is enough radiation in considerations for the development of solar energy projects from a small scale to large scale. The accessibility of the radiation in those fields can be demonstrated to last around twelve hours every day as it is one of the hottest areas in the state.

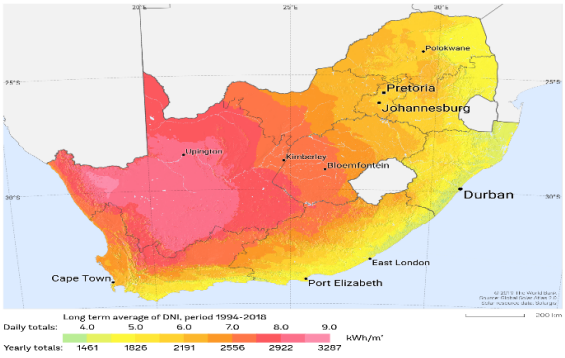


Figure 1. The direct average irradiation through the region (© Global Solar Atlas, 2020)

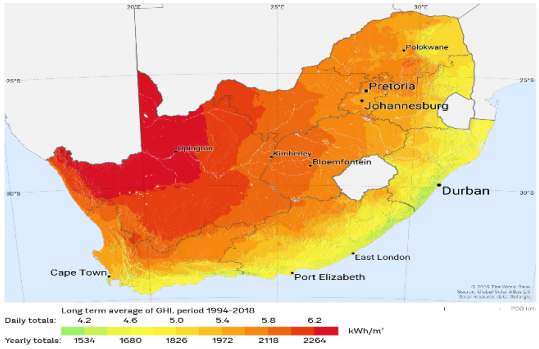


Figure 2. The global horizontal irradiation through the region (© Global Solar Atlas, 2020)

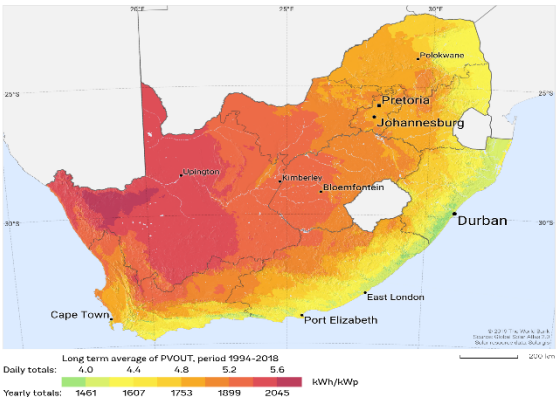


Figure 3. The potential photovoltaic energy for the region (© Global solar Atlas, 2020)

Every country has different solar radiation data but needs other techniques and measurements to determine a specific location. It can be touched on that there are various models produced to determine available solar radiation based on the sunshine hours. In South Africa, mainly, it plays a significant part in renewable energy systems and applications such as in health systems, agriculture, and farming, construction, and housing for domestic and industrial use [1], [28] as it is regarded as the most efficient and economical alternative resource and unused abundant sunshine available throughout the year.

It is essential to know where renewable energy sources come from, as almost all sources originate entirely from the sun [5]. Therefore, the sun's arrays that get into the atmosphere are subjected to several elements, including array absorption, scattering, reflection, and transmission through the atmosphere, before arriving at the earth's surface or location of interest. This array is separated into three categories: - diffused, reflected, and direct solar radiation representing the solar energy from the sunlight. Solar radiation data that arrives at the ground level is essential for many wide ranges of applications for meteorology, applied science, environmental sciences, agricultural hydrology and sciences, soil sciences, and physics. In summation, these include the modeling and estimation of crops and crop evapotranspiration, particular health sectors and medicine, and other research in the natural sciences [6], [7], [24].

Many projects are currently being developed, focusing on South Africa and looking into energy systems and their habits. The importance of using solar energy schemes in most parts of Limpopo Province has increased. Independently, improving and raising awareness on issues towards climate change and factors regarding the residential area is one of the significant economic challenges [8]–[10], [27]. This has been understood as the potential importance and opportunity in addressing energy security and carrying out of one the research niche area within the environmental Millennium Development Goals (MDG) on energy [11], [26].

The solar radiation data measured provide information on how much solar energy is reflected along the airfoil during the specific period (i.e., hourly, daily, monthly, and yearly) at the particular position. These measured data are of importance for efficient solar

energy research for utilization and providing energy through natural resources such as photosynthesis [3]-[5]. In modern energy engineering, many cases are calling for equipment subjected to solar radiation [6], [7], [11]. This includes technology equipment such as street lights, traffic lights, remote control gates, public messages and notices, and other solar energy-supplied units installed in the residential district without proper grid-connected electricity supplies. The sunlight power source furnishes this equipment through radiation at defined energy absorption, which causes different significance. Using the radiation data within the Vhembe District allows determining the solar radiation parameters connected to the power needed for solar energy use.

Solar energy forms part of the ultraviolet intensity spectrum, determined by the physical wave solar constant (K). This electromagnetic constant flows through a unit area (A) by the solar array directly to the earth's surface distance from the sun. As a result, to measure the necessary solar energy, it is essential to know the duration time (T) of the solar array path through the aura to the specified destination location [12]. In addition, the amount of incoming solar radiation on the Earths' surface is the measurable amount during the minimum (Tmin). The maximum (Tmax) temperature (that is, in Degree Celsius) at relatively on a daily average of sunshine per hour (hrs); this requires a validated model to be used [12], [26]. Hence the data predictions at the specified location are of importance for the estimation and design in energy conversion for domestic employment industrial and commercial applications [13], [27].

This study introduces a mathematical and estimation radiation model and calculations developed for the Vhembe District to design solar energy strategies as per the radiation data found through the installed meteorological stations. The solar energy is specified and analysed in different geographical locations that enable the parameters in calculations for any positioning on an hourly basis or day by day. As a result, a specific amount of radiation will be employed to determine the possibilities with radiation measurement in planning solar energy systems.

Thus, the mathematical solar energy model used for data prediction provides a vast solar energy potential in the Vhembe District in hourly, daily, and monthly solar radiation on the location determined. Once more, the Vhembe District has a very complex topology, hill mountainous zone area due to its position, and extensive territory with plantations and vegetation, as shown in Figure 4 to Figure 8. As a result, climate data are critical throughout the year due to the location's land cover and weather patterns.

Figure 4. The Thathe Forestry and plantation

Figure 5. The Tshakuma Mango plantation

Figure 6. The Levubu banana plantation

Figure 7. The Elim litchis plantation

Figure 8. The Phiphidi Falls and river

These plantations (including corn, wheat, and sugar cane) furnish the wood wastes through organic materials that can be

applied directly or converted into biofuels or bugs to be burned as fuels to generate energy. Besides, with the availability of the upstream and downstream rivers, there are opportunities that micro/small hydropower systems can be developed to assist the small farmers within the area for the irrigation and plantations, as shown in Figure 8.

In all four districts, people live under a very disadvantageous eco-system and environment, such as limited access to grid-connected electricity and using most natural resources such as traditional biomass, encroachments, paraffin, etc., and gas their energy resources. Also, the district has 14 primary commodities, as shown in Table 1. The majority of the community depends on farming as their economic sustainability, improving their livelihoods, and creation of employment for the rural communities [3], [14], [24].

Table 1. Type of agricultural farmers in the district [2]

Commodities	No. of Smallholder farmers
a) Backyard gardens	644
b) Banana	409
c) Citrus	16
d) Fish	81
e) Garlic	39
f) Guava	128
g) Litchi	4
h) Livestock's	15 652
i) Macadamia	512
j) Mango	758
k) Poultry	992
l) Tomato	2015
m) Vegetable gardens	2300
TOTAL	**23 636**

With such commodities, an estimation of waste to bio-energy plants could be manufactured to enable the communities to use their waste materials and turn them into a valuable product that can gain them. This will be utilized in determining their energy needs or a marketable product as the source of income in supplying waste to the bio-energy plant that could be got within their area. This will enable the communities to have a whiter, healthier environment and potential job creation and admission to improved energy that improves their living standards. The same uses with biomass, wind and hydro energy resources fail to be taxed.

The measurements received from the solar radiation and humidity climate weather data through the instruments utilized are much subject to stability error function as exposed to heat transport within the aura. This heat transfer is of the drift by 10% of the determined values. In summation, there is a relatively 1% humidity loss of the instruments per month. According to [24], many available studies refer to the global solar radiation models. This includes available models in estimating the daily, weekly, monthly, and annual radiation used for solar energy estimation purposes.

2. Materials and Methods

Throughout the study, the following materials, methods, and analyses were carried out during the estimations and modeling, namely: - Weather data measured throughout the installed metrology weather stations, the GIS maps obtained through Remote Sensing and Mapping downloads for the Vhembe District and using the Photovoltaic graphical modeling through the Matlab software.

Basic Solar Irradiance Measurement

Solar radiation depends on the sunshine that arrives on the earth during the daytime, with its specific latitude location and the atmospheric transmittance (K). Besides the net solar radiation reaching the earth's surface, some are lost and be used for other heating methods, which are turned into additional energy that can be measured using specific instruments [16], [19].

The radiometer is an instrument for measuring irradiance in equal quantities of solar energy at a specified wavelength range measured. The most significant concern was thought in choosing a site placement regarding the determined climatological measurements within the country of interest during the day or in the year. Besides, it was understandable that new models and techniques exist and are being developed in improving the measurement techniques for estimating solar radiation energy with accurate, readable available meteorological parameters. Hence, considering solar radiation on horizontal and tilted surfaces forms part of the estimations [20]. Furthermore, in computing the global radiation, one should take the daily solar radiation absorption (Rs) on the ground, together with the extraterrestrial insolation (Q) and the mean daily solar through the sky transmittance (K) according to equation 1:

$$Rs = Q \times K \tag{1}$$

The constant, K, has a variation of K_C or K_O when there is a clear forecast and K_i on the intermediate days during the year [18], [21]. In summation, in estimating the direct solar radiation (I), one must recognize that it depends on the actual length (r) between the ground and sunlight during the incident measurement. As such, direct solar radiation (I) is known as the dower of the so-called extraterrestrial solar radiation (Io), which arrives at the earth's surface directly from the atmosphere [20], [24].

Basic Solar Radiation Intensity

The parameters affecting the solar radiation intensity within the atmospheric region are important in solar energy as other arrays are reflected throughout the air. That is, the spectrum of the radiation emitted by the sun is about the power in the ultra-violet region as the solar radiation beam (i.e., constant (I_0)) passes through the atmosphere when the sun is at its mean distance from the earth [21], [22], [25]. This value is

$$I_0 = 1.37 \pm 0.02 kW/m^2 \tag{2}$$

This constant varies as the light travels through the clouds, absorbed or scattered, reflection and based on the climate latitude and longitude of the location area for the solar energy. This value diverges by 3% as the earth's orbit is elliptical, and the distance from the sun varies all year round. The variations distance between the sun and the world is due to the earth's orbit caused by the actual intensity of solar radiation outside the atmosphere to differentiate

from I_0 by a few per cents to strike into account these variations by a mean factor, F in Degree Celsius.

$$F = 1 - 0.0335 \sin 360(n_d - 94)/365 \quad (3)$$

Where n_d is a specific day of the year (i.e., $n_d = 1$ for January and $n_d = 365$ for December), the argument of the sine function is in degrees. All the values of solar radiation intensity given below, which are in the sun at its average distance from the earth, must be multiplied by F to obtain the actual values on a day. During January, as the weather is clear, the sun is closer to the world, the solar radiation is 3% larger than the average, and in July, when the earth is furthest from the sunshine, the solar radiation is 3% less than the norm.

Prediction Model and Determination of Solar Energy Radiation at Specific Area and Particular Time

The site selection directly impacts the potential renewable energy systems (RES) projects in many different ways, including technical, economic, and environmental aspects. However, one of the critical roles in the PV power plants is the inconsistency and variability of solar irradiation, which can be geographically dissimilar from one location to another [23], [28]. To measure the specific amount of solar radiation at a particular area and a specific time, it is essential to define the angle of inclination by cos(δ) of solar arrays to the perpendicular of the earth by considering the area of interest, cos(θ) and expressed by the total amount of watts per meters square (W/m^2) and the joules per meter squares (J/m^2) [12], [25]. Also, the RE subject field requires three parameters, namely, global horizontal irradiance (GHI), direct normal irradiance (DNI), and diffuse horizontal irradiance (DHI) [13], [24], [29]. Furthermore, many methods have been introduced to measure global solar irradiation in that respect. These methods have been modified in several ways to suit different models. The simplest example to calculate the global solar radiation is shown in equation 4 [3], [25], 30

$$H = H_O \left[A + B\left(\frac{n}{L_d}\right)\right] \quad (4)$$

The H and H_O are the daily solar radiation and the daily extraterrestrial radiation in $MJm^{-2}d^{-1}$; A and B are constant-coefficient; n and L_d are the sunshine hours per day and location day length in Hours (hrs). The constant-coefficient values are subject to the location of the study and its weather conditions throughout the year [25], [28]. Using the captured weather data, several states are shown using the equation to estimate solar and weather conditions.

In measuring the solar radiation, territorial solar irradiation (E_{ss}) and global solar irradiation (EET) are considered. Hence, the total solar energy (E_s) above the atmospheric level is equal to the absolute atmospheric solar power at sea level multiplied by the length of day (N) per change of temperature (T) throughout the year as determined by equation 5. The latitude (L) and angle of declination (δ) by the sunlight must also be consider parenting the surface of absorption [31].

$$E_s = (Io + 1) + 0.34 \cos\frac{2\pi N}{365} \times L \quad (5)$$

Where the Io is the solar constant = $1{,}367 W/m^2$ (is the extraterrestrial radiation as the earth orbits around the sun) and N is the number of the day for solar absorption, 0.34 being the constant coefficient of solar irradiance at the atmospheric level and L is the length of the day being calculated by equation. The $\cos\frac{2\pi N}{365}$ It is the calculated angle of declination of the sun during the day through per year during the earth's orbit.

The length of the day was calculated by: -

$$L = \frac{2}{15}\cos^{-1}(-\tan L \times \tan \delta) \quad (6)$$

This was determined by the length (L) of days per the solar decline angle (δ) as calculated by equation 7.

$$\delta = 23.45 \sin\frac{(284+N)}{365} \quad (7)$$

In normal circumstances for the Vhembe Region weather measurements, the minimum length was 11.2hrs, and the maximum size of the day was 13.9hrs [3], [25]. Therefore, the solar energy (E_s) for the minimum and maximum was determined by equations 8 and 9.

$$E_{s(\min)} = (I_0 + 1) + 3{,}808 \cos\frac{2N}{365} \quad (8)$$

$$E_{s(\max)} = (I_0 + 1) + 4{,}726 \cos\frac{2N}{365} \quad (9)$$

As a result, equations 1 to 7 were acknowledged in the patterns, estimations, measurements, and computations of the solar radiation energy and demonstrated by the Matlab graphical responses in Figure 17 to Figure 19 for the radiance measurements.

3. Analysis and Discussions

3.1. Meteorology Data Acquisition Analysis

Measurements were remotely captured throughout one year, from January to December 2018. The data acquisition (DAQ) system was used to obtain data from the nine Weather Stations (WS) used during the data collection (i.e., Hanglip, Shefeera, Tsianda, Thohoyandou WO, Dzanani Biaba Agric, Mphefu, Joubertstroom Plantation, Vondo - Bos and Tshivhasie Tea Venda). Table 2 gives the locational longitude and latitude coordinates of the installed weather stations used during the data collection for the study.

Table 2. The location of the weather stations installed in the Vhembe District

Weather Station Name	Longitude (°, E)	Latitude (°, N)
Hanglip	101.07	41.95
Shefeera	94.68	40.15
Tsianda	98.48	39.77
Thohoyandou WO	103.08	38.63
Dzanani Biaba Agric	100.13	37.33
Mphephu	30.03	22.89
Joubertstroom Plantation	22.57	29.19
Vondo - Bos	30.33	23.93
Tshivhasie Tea Venda	22.96	30.35

It was challenging to measure solar radiation in many locations due to the cost of equipment to be used, maintenance, and calibrations to obtain accurate values. Hence, the South African Weather Stations (SAWS) meteorological weather data were used in all the nine weather stations (WS) been installed, in concert with the Agricultural Research Council for the Institute for Soil, Climate and Water (ARC-ISCW) in providing the data. As the results, within reference to the measurements obtained, the data were used to define the solar energy potential for the Vhembe District at specified locations to evaluate its amount for power generation. This data was remotely captured and used to calculate the potential solar energy per location for power generated by photovoltaic system modules. Table 3 shows essential parameters in measuring solar irradiation at different positions. These were applied to evaluate the accessibility of solar irradiation at the designated place.

Table 3: The measurements units used for solar radiation evaluations

Parameters	Units
Global radiation	G (W/m^2)
Diffuse radiation	G_d (W/m^2)
Beam radiation	G_n (W/m^2)
Sunshine hours	σ (hrs)
Maximum and minimum temperature	T_{min} and T_{max} (°C)
Humidity	H (%)
Pressure	P (Pa)
Visibility	F (m)
Wind speed and directions	V(m/s); N,W,E,S
Air mass	ρ(kg/s)

Besides, the efficiency and error calculations were considered difficult to measure the solar radiation at the geographical location due to other factors, including absorbed or reflected by the atmosphere. Following the data obtained, available renewable energy resources within the Vhembe District were of importance and the peoples' quality of life as purpose in using the solar energy to meet their energy demands.

Figure 9 and Figure 10 show the monthly graphical meteorological analysis of the area for minimum and maximum temperature and the length of the day (sunshine hours) as per the yield throughout the year on the direct solar radiation measurement. These measurements were acquired through the day's duration in the district to get a micro solar system for the community expectation and energy demands.

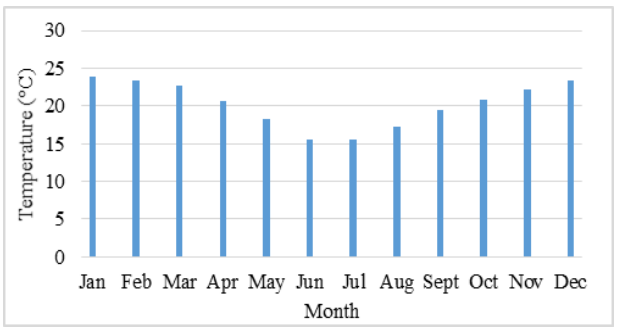


Figure 9: The monthly temperature measurement (Average - 20 °C)

It was noticed that the highest values of the solar insolation are during the summer months (Jan to Apr and Sept to Dec), and the lowest values are during the winter months (May, Jun, Jul, and Aug) as applicable per the day during that time for the season.

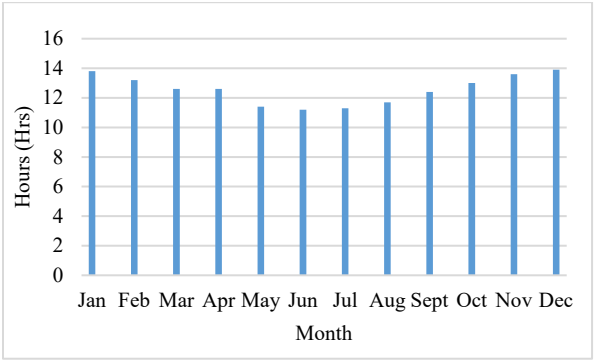


Figure 10: The monthly length of the day measured (Average - 12.5Hrs)

Furthermore, the desirable amount of solar energy was measured during the allowable day time duration (from 06:00 – 19:00) for the whole year per location during the 24hr time interval. The measurements were used to determine the minimum and maximum coefficients variations through the solar radiation calculations using the Matlab software analysis.

3.2. Remote Sensing and GIS Mapping Analysis

The Vhembe District is situated at 22.7696° S Latitude, 29.9741° E 25 Longitude of the Limpopo Province. At an altitude of 250m above mean sea level, a study was conducted in estimating the monthly and annual solar radiation, using the climate and geographical parameters. These areas and the outcomes obtained will assist the researchers and public entities interested in working on solar energy developments to have reference and locations' conditions that they can use for solar energy estimation. Figure 11 shows the potential solar available within the district and per municipality used for domestic use during the solar energy estimation. This map provides an overview of available heat energy to be used.

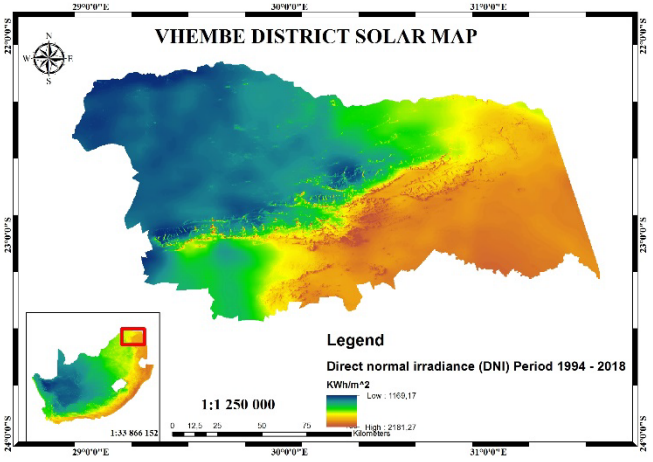


Figure 11: The Solar Map of the Vhembe Strict Area

The analysis shown includes the daily solar radiation assessment from all the locations and using other parameters to see

the solar potential available, as shown by Figure 10 to Figure 13. Thither are many solar radiation databases available for most sections of the countries around the globe.

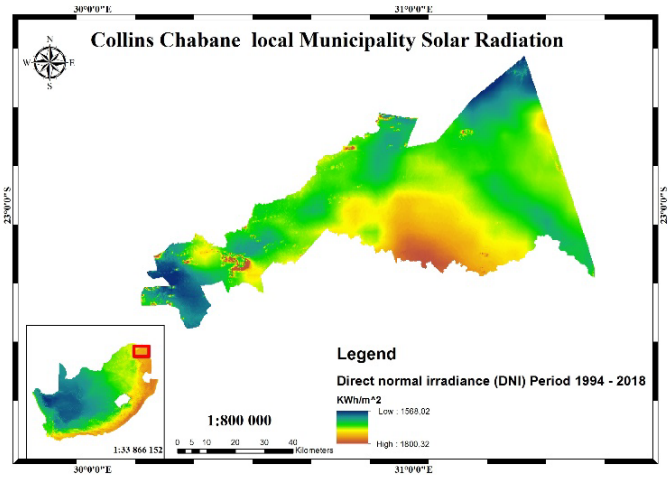


Figure 12: The Areal Solar for the Collins Chabane Municipality

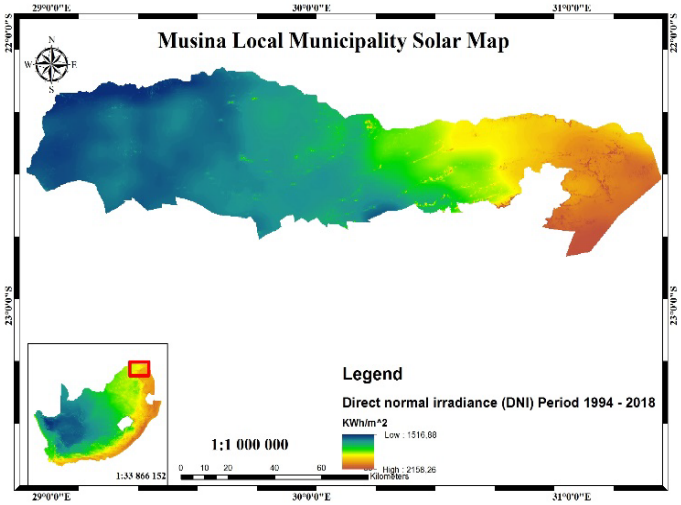


Figure 13: The Musina Municipality Solar Map

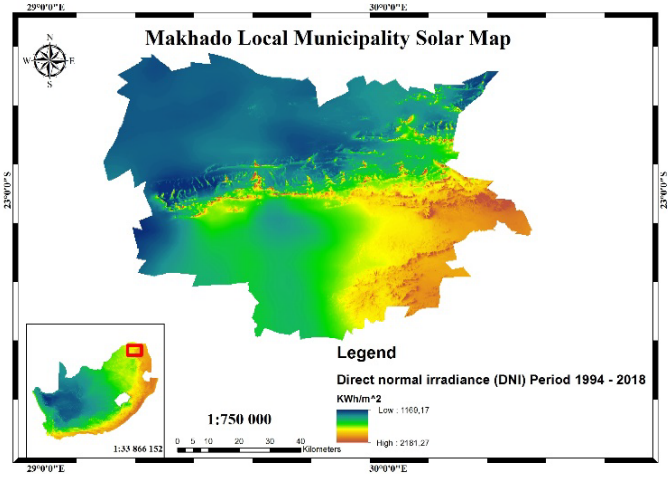


Figure 14: The Thulamela Municipality Solar Map

These were set for the annual weather changes to determine the solar maps of the Vhembe strict and its municipalities (Musina, Collins Chabane, Thulamela, and Makhado Municipality). As a result, the Remote Sensing (RS) for GIS was employed to settle the territorial dominion's solar maps.

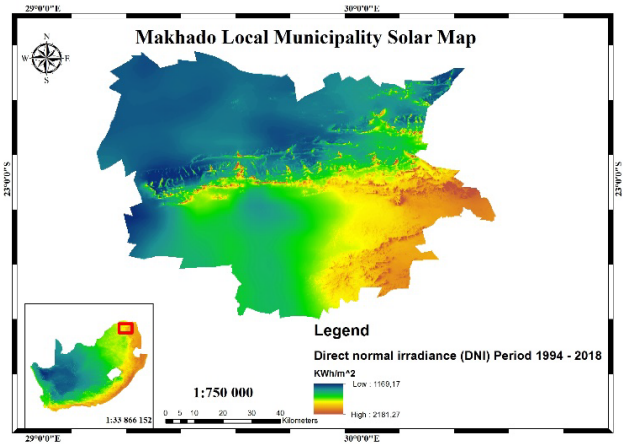


Figure 15: The Makhado Municipality Solar Map

3.3. Computational Solar Analysis using Matlab

During the data analysis, the daily solar radiation, the intensity of direct radiation (W/m²) through an average sun hour of the solar insolation, as shown in Figures 17 to 19. It was noted that high radiation is received during the summertime. There are low irradiance and temperatures in wintertime, which is demonstrated by the lowest and highest measurements for power potential and public presentation.

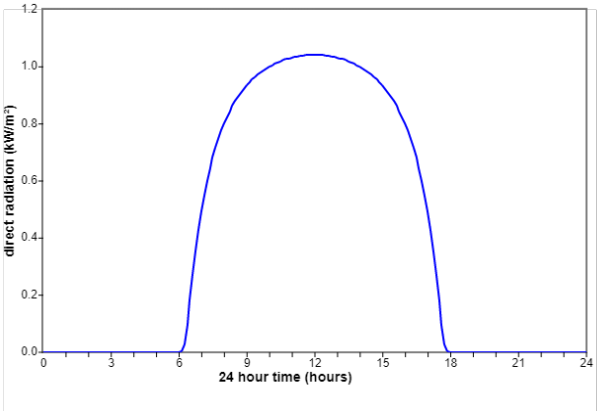


Figure 16: The daily solar radiation for the district

Figure 16 shows the daily solar irradiance curve during the number of hours during sun hours of the day. This is the direct radiation per hour that is generated.

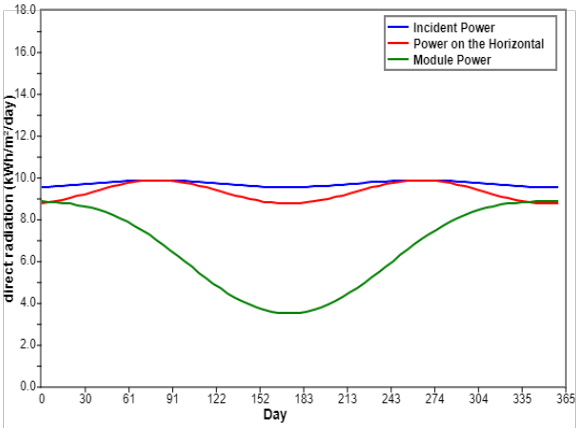


Figure 17: The daily incident power generation per direct intensity

Figure 16 to Figure 18 shows the maximum amount of power directly received without any clouds during the regular sun hours. This amount is set at different to determine how much the specific location's radiation is per period defined. As a result, the power required is generated.

The area is known for its abundant radiation and available solar resources, which significantly influence the design, configuration, and cost of power systems produced. It was observed that the highest values of the solar insolation are during the summer months (Jan to Apr and Sept to Dec), and the lowest values are during the winter months (May, Jun, Jul, and Aug), as demonstrated by Figure 14 to Figure 19. These estimates and weather patterns have been obtained and analyzed through weather stations installed in the Vhembe District and the Matlab software analysis as part of the computations estimations.

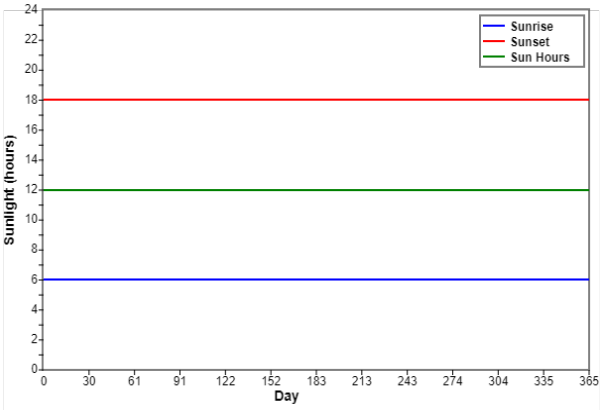


Figure 18: The average sun hours of the solar insolation

Figure 18 shows the mean daily solar availability based on the three curves corresponding to the incoming solar insolation. The daily insolation shown is the number of sun hours rising and sunset. The limited results are helpful for the conception and estimation of the power module needed to take out the solar energy schemes to be set up. This will provide the estimates of the available irradiation generation concept.

The approximation of solar radiation energy is vital in designing the solar energy system or devices. The estimation and size are not easily determined due to the cost and techniques required to practice. As a result, there is a demand for building theoretical methods for estimating solar radiation, such as the empirical relationships using commonly measured climatological data measured at the specific area [17], [18]. Due to its position and extensive territory, the Vhembe District has a complex topology, hill, and mountainous zone area; hence, the nine installed weather stations found different climate data.

4. Conclusions

The estimation model uses the most recent data measured throughout (January to December 2018) of the meteorological data obtained in the nine installed weather stations. This analysis demonstrates that the required solar radiation values for the potential use in the field are accepted by the temperatures commonly measured by the installed weather stations around the Vhembe District. The obtained results can be the best exemplar for solar estimation at different geographical and climatic locations.

In estimating PV radiation's optimal and potential role, it is essential to consider the location suitability to maximize the solar energy received and the power generated at the selected position. In this paper, the meteorological estimations, graphics, and formulae were applied to influence the behaviour of the Vhembe District climatic conditions as one of the rural regions of involvement in deploying renewable energy technologies such as solar schemes. Consequently, the analysis demonstrates that the required solar radiation values for the potential use in the field are estimated by the temperatures commonly measured by the installed weather stations around the Vhembe District. Furthermore, intensive research studies within specific locations should be carried out to identify and find out the environmental matters linked with the placement and its natural resources to see potential energy sources available for community use.

It is noted that the active use of solar energy schemes has an environmental impact compared to other authors. These solar energy technologies should be increasingly introduced within the rural areas taking into account the suitability and energy potential of the region.

In summary, solar energy provides many advantages over other alternative energy sources. As presented in the paper, a simple principle of solar heat energy can be utilized in various applications. Nonetheless, it is mentioned that solar energy has its drawbacks or limitations like high initial price, depending on the weather, and challenges in energy storage. As a result, the South African Government is increasingly introducing initiatives with plans in providing subsidized programs to increase an effort in encouraging solar energy use in the rural regions. With the application of the solar assessment, the local community will use these findings to assist in determining the potential locations to deploy and install the solar systems for their local use, agriculture, and community use.

Disclosing a conflict of interest

The authors have no conflict of interest to declare.

Acknowledgment

The acknowledgment is towards the support of the Energy Institute (EI) Members and the Centre for Distributed Power & Electronic Systems (DEECE), Dr. K. Aboalez, Dr. M. Adonis, Dr. A. Raji, and Dr. Ali-Mustafa-Ali Almaktoof on their expertise and

supervision in developing, compiling and writing this publication. I want to thank the Research Directorate (RD) Unit under the Office of Deputy-Vice Chancellor Research Innovation and Technology and Partnership (DVC-RITP) for financial support. They appreciate the South African Weather Stations (SAWS) and the Agricultural Research Council for the Institute for Soil, Climate and Water (ARC-ISCW) of data provided. I would likewise like to recognize the 2019 IEEE Power Africa Conference as this extended paper forms part of a conference paper presented at the Abuja, Nigeria conference.

References

[1] C. Matasane, M.T.E Kahn, "Solar Radiation Estimations Using the Territorial Climatological Measurements in Vhembe District, Limpopo Province for Solar Energy Potential Estimation and Use," 2019 IEEE Power Africa Conf.: Abuja, Nigeria, 2019, doi:10.1109/PowerAfrica.2019.8928806.

[2] C. Matasane, C. Dwarika, R. Naidoo, "Modelling the Photovoltaic Pump Output Using Empirical Data from Local Conditions in the Vhembe District," 2014 International Conference on Social Education and Community Conf, doi:10.5281/zenodo.1096759.

[3] M. S. Gadiwala[1,2], A. Usman[2], M. Akhtar[2], K. Jamil[2], "Empirical Models for the Estimation of Global Solar Radiation with Sunshine Hours on Horizontal Surface in Various Cities of Pakistan," Pakistan Journal of Meteorology, **9**(18), 2013.

[4] A. E. Lawin[1,*], M. Niyongendako[2], C. Manirakiza[2], "Solar Irradiance and Temperature Variability and Projected Trends Analysis in Burundi," Climate 2019, **7**(6), 83, 2019, doi:10.3390/cli7060083.

[5] S. Zekai, "Solar Energy Fundamentals and Modeling Techniques," Atmosphere, Environment, Climate Change and Renewable Energy, **22**, 2008.

[6] T. A. McMahon[1], M. C. Peel[1], L. Lowe[2], R. Srikanthan[3], T. R. McVicar[4], "Estimating actual, potential, reference crop and pan evaporation using standard meteorological data: a pragmatic synthesis," Journal of Hydrol. Earth Syst. Sci., **17**, 1331–1363, 2013, doi:10.5194/hess-17-1331-2013.

[7] M. Paulescu, E. Paulescu, P. Gravila, V. Badescu, "Weather Modeling and Forecasting of PV Systems Operation, Green Energy and Technology," Springer-Verlag London, 17–42, 2013, doi:10.1016/j.rser.2016.11.222.

[8] P. Jayakumar, Solar Energy Resource Assessment Handbook: APCTT Asian and Pacific Centre for Transfer of Technology of the United Nations – Economic and Social Commission for Asia and the Pacific (ESCAP), 2009.

[9] L. Mary[1], A. E. Majule[2], "Impacts of climate change, variability and adaptation strategies on agriculture in semi-arid areas of Tanzania: The case of Manyoni District in Singida Region, Tanzania," African Journal of Environmental Science and Technology, **3**(8), 206–218, 2009, doi:10.5897/AJEST09.099.

[10] D. R. Brooks, Monitoring Solar Radiation and Its Transmission through the Atmosphere, Department of Mechanical Engineering and Mechanics, Drexel University, Philadelphia, PA, USA, **2**, 2006.

[11] A Guide to Energy's Role in Reducing Poverty, Energizing the Millennium Development Goals, UNDP, 2005.

[12] S. A. Kalogirou, "Solar thermal collectors and applications," Progress in Energy and Combustion Science, **30**, 231–295. 2004.

[13] O. I. Kordun, The influence of solar radiation on sheet steel structures temperature increment, Achieves of Civil Engineering, **LXI** (1), 2015.

[14] S. S. Ndwakhulu, An evaluation of the performance of the Department of Agriculture in Limpopo Province in improving the livelihood of smallholder farmers during the period 1994–2004, with special reference to the Vhembe District, MSc Thesis, University of Stellenbosch, 2007.

[15] K. Bakirci, "Models of solar radiation with hours of bright sunshine: a review," Renewable Sustainable Energy Review, **13**, 2580–2588, 2009, doi:10.1016/j.rser.2009.07.011.

[16] K. Sukarno[1], Ag. S. Abd. Hamid[1], J. Dayou[1], M. Z. H. Makmud[2], M. S. Sarjadi[2], "Measurement of Global Solar Radiation in Kota Kinabalu Malaysia," ARPN Journal of Engineering and Applied Sciences, **10**(15), 2015.

[17] A. A. El-Sebaii, F. S. Al-Hazmi, A. A. Al-Ghamdi, and S. J. Yaghmour, "Global, direct and diffuse solar radiation on horizontal and tilted surfaces in Jeddah, Saudi Arabia," Applied Energy, **87**(2), 568–576, 2010.

[18] X. Li u, X. Mei, Y. Li, "Evaluation of temperature-based global solar radiation models in China," Agricultural and Forest Meteorology, **149**(9), 1433–1446, 2009, doi: 10.1016/j.agrformet.2009.03.012.

[19] S. Becker, "Calculation of direct solar and diffuse radiation in Israel," International Journal of Climatology, **21**, 1561–1576, 2001.

[20] C. K. Pandey and A. K. Katiyar, "Solar Radiation: Models and Measurements Techniques," Journal of Energy, **2013**, ID 305207, doi:10.1155/2013/305207.

[21] S. Becker; "Calculation of Direct Solar and Diffuse Radiation in Israel," International Journal of Climatology, **21**, 1561–1576. 2001, doi:10.1002/joc.650.

[22] Z. Samani, "A General Solar Radiation Estimation Model Using Ground Measured Meteorological Data in Sarawak, Malaysia," Journal of Telecommunication, Electronic, and Computer Engineering, **10**(1), 99–105, 2015.

[23] Z. Hassan, Optimal Design and Analysis of Grid-Connected Solar Photovoltaic Systems, Ph.D Thesis, Concordia University, 2018.

[24] J. Singh, A. Kruger, "Is the summer season losing the potential for solar energy applications in South Africa?" Journal of Energy South Africa, **28**(2), 2017, doi:10.17159/2413-3051/2017/v28i2a1673.

[25] S. Bandy, W.A. Zainal, "A General Solar Radiation Estimation Model Using Ground Measured Meteorological Data in Sarawak, Malaysia," Journal of Telecommunication, Electronic, and Computer Engineering, Universiti Teknikal Malaysia Melaka, **10**(1), 99–105, 2018.

[26] T. R. Govindasamy, N. Chetty, "Quantifying the global solar radiation received in Pietermaritzburg, KwaZulu-Natal to motivate the consumption of solar technologies," Open Physics, **16**(1), 2018, doi:10.1515/phys-2018-0098.

[27] E.V Tikyaa, A. Akinbolati, M. Shehu, "Assessment of empirical models for estimating mean monthly global solar radiation in Katsina," FUDMA Journal of Sciences (FJS), **3**(1), 333–344, 2019, doi:10.4314/bajopas.v4i2.5.

[28] Z. E. Mohamed*#, R. M. Farouk**, H. H. Saleh***, "The Performance Evaluation of Mathematical Models for Predicting MAD-GSR in Egypt," Global Journal of Pure and Applied Mathematics, **14** (7), 897–918, 2018.

[29] T. S. Mulaudzi, N. E. Maluta, and V. Sankaran, "Evaluation of the global solar irradiance in the Vhembe district of Limpopo Province, South Africa, using different theoretical models," Turkish Journal of Physics, **39**, 264–271, 2015, doi:10.3906/fiz-1505-9.

[30] Ş. Ozan, T. Kuleli. "Estimation of SR over Turkey using artificial neural network and satellite data." Applied Energy **86**, 7–8 (2009): 1222-1228, doi:10.1016/j.apenergy.2008.06.003.

[31] A. A. El-Sebaii, F. S. Al-Hazmi, A. A. Al-Ghamdi, S. J. Yaghmour, "Global, direct and diffuse solar radiation on horizontal and tilted surfaces in Jeddah, Saudi Arabia," Applied Energy, **87**(2), 568–576, 2010, doi:10.1016/j.apenergy.2009.06.032.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 1, 107-124 (2022)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Efficient Publicly Verifiable Proofs of Data Replication and Retrievability Applicable for Cloud Storage

Clémentine Gritti*, Hao Li

Computer Science and Software Engineering Department, University of Canterbury, Christchurch, 8011 New Zealand

ARTICLE INFO

Article history:
Received: 13 December, 2021
Accepted: 09 February, 2022
Online: 28 February, 2022

Keywords:
Proofs of Retrievability and Reliability
Verifiable Delay Functions
Cloud Storage Services

ABSTRACT

Using Proofs of Retrievability (PORs), a file owner is able to check that a cloud server correctly stores her files. Using Proofs of Retrievability and Reliability (PORRs), she can even verify at the same time that the cloud server correctly stores both her original files and their replicas. In 2020, a new PORR combined with Verifiable Delay Functions (VDFs) was presented by Gritti. VDFs are special functions whose evaluation is slow while verification is fast. Therefore, those functions help guarantee that the original files and their replicas are stored at rest. Moreover, an important feature of the 2020 PORR solution is that anyone can verify the cloud provider's behaviour, not only the file owner. This paper extends Gritti's version. In particular, a realistic cloud framework is defined in order to implement and evaluate accurately. Results show that this PORR solution is well suitable for services provided for cloud storage.

1 Introduction

Cloud data storage has exploded over the last decade, for both business and personal purposes. The latter have offered the opportunity to delegate the storage of individuals' files, through various services that have been developed and diversified with competitive fees for data owners (e.g. copies of files stored in different storage locations). However, due to its complex structure, many unfortunate events have happened over the last few years. In 2015, lightning strikes engendered data loss on Google centers[1]. The startup *Front Edge CNC'* used to keep its online production data in Tencent Cloud Storage but realised in 2018 that it was completely lost[2]. More recently, according to McAfee, 99% of misconfiguration incidents occurring in a public cloud environment are not detected, thus exposing enterprises and organisations to a huge risk of undetected data breaches[3]. Therefore, it has become urgent to carefully design cloud storage solutions to overcome all possible unfortunate events such as the aforementioned ones.

An ideal solution will enable a simple deal between a client, who owns some files, and a cloud provider (also called cloud server), who offers safe and attractive cloud storage services. For instance, the cloud provider may propose to store copies of a file in addition to the file itself, and spread them across multiple servers, avoiding single point of loss. The cloud provider may also offer to decrease fees per copy, to encourage the client to adopt the replication process as much as possible. More precisely, both the client and cloud provider are financially incentivized: the client wants to save fees as much as possible while uploading files as many as possible, while the cloud provider wants to earn as much as possible while saving storage as much as possible. Those assumptions clearly motivate their behaviours. The rational cloud provider may claim to store the client's files while actually not, and thus offering the free storage space to other clients. The malicious client may claim to upload copies of the files, with lower fees, but rather uploading different files where fees would have higher if behaving honestly.

This work is an extension of [1], in which a new solution was proposed for secure and efficient cloud storage for a client who owns some files stored on a cloud provider. Security is enhanced by allowing detection of a misbehaving cloud provider that does not store the client's files correctly and prevention against a malicious client who tries to save fees. Informally, the cloud provider is asked to create copies of the client's files stored across various locations and can be challenged to prove the client that all files and their copies are entirely stored. The client uploads encrypted files

*Corresponding Author: Clémentine Gritti, CSSE Department, University of Canterbury, clementine.gritti@canterbury.ac.nz
[1] https://www.bbc.com/news/technology-33989384
[2] https://medium.com/genaro-network/tencent-was-claimed-ten-million-for-data-loss-due-to-cloud-hard-drive-glitch-344a26449fe2
[3] https://www.computerweekly.com/news/252471175/Enterprises-exposed-to-data-loss-by-cloud-configuration-errors

to the provider but does not have the responsibility of creating the copies, avoiding the client to upload fake copies (i.e. files that are not at all linked to the original ones). Efficiency is conserved even with improved security. In particular, experimental results presented for the first time in this paper show that the mechanisms used to preclude malicious behaviours incur low costs on the computational and communication processes.

1.1 Problem Statement

Let a client subscribe to data storage services offered by a cloud provider. Proof of Data Possession (PDP) and Proofs of Retrievability (POR) are cryptographic protocols proposed to enable the client to verify that the cloud provider actually proceeds as agreed, that is correctly stores the client's data.

The latter protocol, POR [2], guarantees the client that her data is available in its entirety. The former protocol, PDP [3, 4], allows the client to verify that the stored data has not undergone any modifications. Recent PDP works [5, 6, 7] offer a replication feature: the cloud provider creates copies of the data and the client can check in a single instance that both the original and copied data are all correctly stored.

However, the aforementioned solutions do not reflect the actual framework. Limited bandwidth is offered for free and costs grow quickly with the increasing bandwidth. In most of existing works, the replication and uploading processes rely on the client, imposing huge burden and fees to the latter. In addition, fees for storage of extra replicas of a given file are less consequent than fees for storage of different files. Hence, a *malicious* client may attempt to upload different files while claiming that they are replicas of the same file. The cloud provider cannot observe such a trick since data is assumed to encrypted before being uploaded, and thus are not readable.

Limitations also raise on the cloud provider's side. The latter may appear to be financially motivated and consequently act maliciously. For instance, it attempts to not store all file replicas and offers the resulting unoccupied storage space to other potential customers. Then, such a *rational* cloud provider creates the missing replicas on the fly when the client who owns the original file asks for storage verification. In addition, the client may put tacit trust on its cloud provider and neglect to carefully read the service level agreements [8]. Likely, the client does not take the time and effort to verify how the cloud provider deals in storing her data.

The design of an appropriate solution for secure data storage in the cloud is clearly encouraged by the aforementioned financial concerns. Clients are motivated by reducing the costs due to bandwidth and storage while the cloud provider is stimulated by partitioning data to make profit. Those concerns have been carefully studied and overcome in [9]. The authors presented an extension of POR, called Proof of Retrievability and Reliability (PORR), to encompass the correct storage verification of both file and its replicas at once. However, the main obstacle of the PORR instantiating in [9] is the private feature of verification: only the client is able to check that the cloud provider correctly stores her file and its replicas.

In [1], we proposed a new PORR protocol overcoming all the aforementioned challenges at once. Along with getting over malicious clients and rational providers, we overcame the likely laziness: the client can delegate the task of checking that the cloud provider has been acting honestly on her data.

1.2 Idea

Our solution is designed to enable a client to upload her file and the cloud provider to generate the replicas of this file. By doing so, we prevent malicious clients to upload different files and claim that there are all replicas of the same file.

Moreover, we offer the client the guarantee that the cloud provider correctly stores both the original file and its replicas at rest. By using slow functions, we prevent rational cloud providers to compute replicas of an original file on the fly when challenged to prove correct storage. Indeed, evaluating a slow function is noticeably slow while verifying its unique evaluation output is fast and easy. Therefore, if the cloud provider tries to generate a replica on demand, it will take a noticeable time to evaluate the slow function attached to the replica and output the unique solution, rather than just storing the replica along with the output from the evaluated slow function.

The cloud provider is challenged by a *verifier*, that can be anyone on behalf of the client. The verification allows a client to ensure that the cloud provider stores the original file and all replicas in their entirety.

Our PORR solution with public verification is a combination of the publicly verifiable POR protocol with RSA signatures from Shacham and Waters [10, 11] and of the slow exponentiation-based Verifiable Delay Functions (VDFs) in finite groups from [12].

We suggest an extension of the POR scheme built by Shacham and Waters [10, 11], where replicas of the challenged file are contained in the verification mechanism. Public verification allows anyone, on behalf of the client, to request the cloud provider a proof of correct storage. Such a feature overcomes a possible lack of data integrity awareness from the client. In order to stop the cloud provider to generate the file replicas on demand when being challenged by the verifier, we incorporate slow functions from [12], namely VDFs. Of course, storing VDF outputs must result into storage costs that are at least as high as storing replicas as required. In fact, rational cloud providers are expected to commit minimal computation resources when generating a correct response. Therefore, the cost of replying to a challenge on the fly should be more than the cost of storing the replicas correctly.

To our knowledge in [1], the author presented the first PORR protocol that both delegates the construction of the replicas to the cloud provider and that permits anyone to verify that the latter correctly stores the files and their replicas.

We give an overview of the PORR protocol in Figure 1. A client encodes and processes a file M as M^*, and outsources it to the cloud provider. The latter commits to store M^* entirely across a set of r storage nodes. This means that M^* is updated to contain the original copy of M and r replicas. On inputs the public parameters, the verifier (possibly the client as in the figure) can efficiently launch challenge requests and verify the responses from the cloud provider. The cloud provider must store both the original file and its replicas at rest, at r storage nodes.

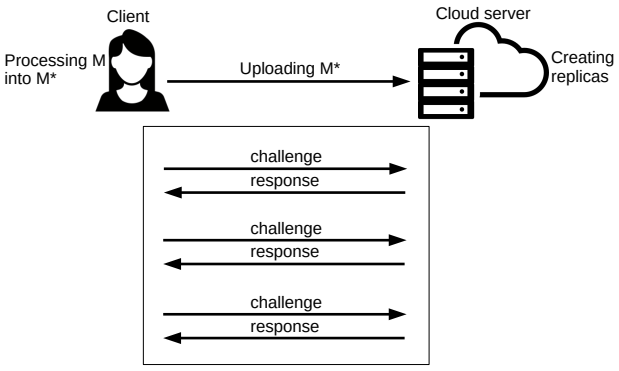


Figure 1: **PORR protocol:** First the client prepares her file M by encrypting and encoding it, resulting into M^*. Second, she uploads M^* to the cloud server along with some additional elements, such as a tag T. The cloud server generates the replicas based on the additional information. Later, and multiple times, the client and cloud server enter into a challenge-response protocol to allow the former to check whether the latter correctly stores her file and its replicas.

1.3 Related Work

We first present existing works on Provable Data Possession and Proof of Retrievability. We then move on Proof of Reliability, where several variants have been proposed based on different mechanisms, namely data replication, erasure codes, proof of location and proof of space. Finally, we review slow functions that allow to generate puzzles, and in particular Verifiable Delay Functions.

1.3.1 Provable Data Possession and Proof of Retrievability.

Provable Data Possession: The authors in [3] introduce the concept of Provable Data Possession (PDP), which enables a client to privately verify the integrity of her files stored at an untrusted cloud provider without the need to retrieve the entire files. Thereafter, the authors in [13] improve the efficiency of the previous PDP scheme [3] by using symmetric encryption. Subsequently, PDP constructions for static data have been proposed in the literature [14]–[17]. PDP schemes with additional properties, such as data dynamicity, have been suggested in [18]–[23]. Several works have been published recently [24]-[30], proposing PDP schemes preserving data privacy and publicly verifying data integrity.

Proof of Retrievability: In [2], the author defined Proof of Retrievability (POR), that is a similar concept to PDP. In POR, the client can correct existing erros and recover the whole files, in addition to check that the cloud provider stores her files in their entirety. In [10], the author proposed two POR schemes: a first one with public verification built on BLS signatures, and a second one with private verification based on pseudo-random functions. In [31], the author improved Shacham-Waters POR schemes by achieving only a constant number of communication bits per verification rather than linear in the number of sectors. Subsequent works on distributed systems follow in [32]–[34], as well as POR protocols with data dynamicity in [35, 36].

1.3.2 Proof of Reliability.

Based on Data Replication: The authors in [5] extend the scheme from [3] by enabling the file owner to verify that at least k replicas are stored by the cloud provider. In [37, 38], the author propose a mechanism with a tunable masking feature in order to let the cloud provider be responsible of the repair operations and the client manage the process of repair. In [39], cross-client deduplication is enabled at the file level by extending [38]. In [6, 7], clients are able to modify and add some pieces of a given file, while to check the correctness of replications of this file. In [40], the replication process is made transparent in a distributed system for cloud storage by extending [18]. The aforementioned solutions, a similar framework is encountered: replicas are prepared by the client, and further sent to the cloud provider with the original files.

More recently, the Proof of Retrievability and Reliability (PORR) protocol is defined for the first time in [9] to illustrate the case where the cloud provider prepares the replicas rather than the client. Tunable puzzles are used for replicating the file, guaranteeing the the replicas are stored at rest. Nevertheless, only a private verifying process is available.

Based on Erasure Codes: In [33], the author provides a high availability and integrity layer for cloud storage. By using erasure codes, data retrievability and reliability are guaranteed among distributed storage servers, and clients can detect and repair file corruption. The authors in [41] achieve a more efficient mechanism for repairs by extending HAIL [33] and moving bulk computations to the cloud provider. The authors in [42] present a scheme for remote data checking in network coding-based distributed storage systems. The scheme minimizes the communication overhead of the repair component compared to the erasure code-based approaches. The repair mechanism in [42] is improved in [22], such that the computation cost for the client is reduced. Following the idea of adding a third party auditor [22], in [43] the author offer a network coding-based scheme such that the repair mechanism is executed between the auditor and the cloud provider, and the client is let apart.

A proof of fault tolerance, as a protocol based on erasure codes, is proposed in [44]. The scheme uses technical characteristics of rotational hard drives to build a time-based challenge, such that the client can check that her encoded files are stored at multiple storage nodes within the same cloud provider. Thereafter, the authors in [45] construct POROS for erasure code-based cloud storage systems by combining POR with time-constrained operations. In order to set a time threshold for the response generation and thus to force the cloud provider to store replicas at rest, rotational hard drive-based ideas are leveraged similar to the ones in [44].

More recently, in [46] the author uses erasure codes to guarantee data reliability, and ensures that the burden of replica generation as well as the repair operations are shifted to the cloud provider. Nevertheless, contrary to [44, 45], the solution in [46] does not rely on the technicity of rotational hard drive technology. Indeed, such method is inevitably dependent on the parameters of the underlying erasure code system. However, in [46], only the client can check the behaviour of the cloud provider in storing her files and replicas.

Proof of Location: Proofs of Location (PoL) [47, 48] have been introduced to prove the geographic position of data, such that whether certain servers in a given country store such data. A PoL scheme is presented in [48] as a combination of geo-location mechanisms and the Shacham-Waters POR schemes [10]. The PoL model looks similar to the PORR one, such that files are uploaded only once by the client to the cloud provider. Then, the cloud provider generates the file tags, prepares the replicas and scatters the latter across servers that are geo-located at different places. Individual POR instances are invoked by the client with all servers to verify whether the latter correctly store the files. On the contrary, PORR schemes must enable clients to launch a single instance to check that all file replicas are correctly stored instead of one instance per replica.

Proof of Space: Proofs of Space (PoS) [49] allow the prover to reply correctly as long as a given amount of space or time per execution is invested. In our case, we require that the prover invests a minimum amount of space. In addition, we need to get a single instance to check that all file replicas are correctly stored, that is not made possible with PoS. Hence, the notion of PoS does not suit our requirements.

1.3.3 Slow Functions.

Time-lock puzzles [50], benchmarking [51], timed commitments [52] and client puzzles [53, 54] rely on the sequential nature of large exponentiation in a finite group.

In [53], the author design a slow function by extracting modular square roots. No algorithm is known for computing modular exponentiation which is sub-linear in the bit-length of the exponent. Nevertheless, the puzzle difficulty is limited to $O(\log p)$, meaning that a very large prime p (being the order of the group) must be chosen to produce a difficult puzzle. In [54], the author consider the sequential nature of Dwork-Naor solution to suggest chaining a series of such puzzles together (e.g. for lotteries). However, this construction does not provide asymptotically efficient verification, hence it can rather be seen as a pseudo-Verifiable Delay Function (VDF).

Time-lock puzzles [50] involve computing an inherently sequential function, by repeating squaring in a RSA group. Used in the context of POR with data reliability guarantees, the cloud provider requires an apparent amount of time to solve those puzzles while the latter are efficiently solved by the client using a trapdoor. Moreover, storing the solution of a time-lock puzzle incurs extra storage costs that are at least as large as the required storage for replicas. The difficulty of a RSA-based puzzle can be easily adjusted by the client (who creates the puzzle) to cater for variable length and different cost metrics. However, such puzzles are not guaranteed to be publicly verifiable: the client (or a dedicated verifier) uses a secret element to prepare each puzzle and to verify the solution.

A given number of sequential steps are required for VDFs, resulting into a unique output verified in an efficient and public way. The construction of a VDF with a trusted setup was proposed [55]. The authors in [12] observe that the trusted setup can be discarded by choosing a sufficiently large random number N, such that $N = pq$ with high probability, for p, q two large prime factors. Nevertheless, the adversary would benefit for parallelizing the arithmetic due to the large size of N, and the running time of the verifier would then increase. The authors in [12] rather suggest to build a VDF from exponentiation, based on the assumption that the adversary cannot run a long pre-computation from the publication of the public parameters to the evaluation of VDF. Hence, this solution achieves proofs that are shorter than the proofs in [55] at a similar level of security.

1.3.4 Work Comparison.

In Table 1, we compare our work with main other ones, based on the specific features that enable secure cloud storage against malicious servers and clients. We omit the Proof-of-Space (PoS) mechanism for Proof of Reliability since it is too far-off our work.

Table 1: Comparing existing protocols similar to PORR

Type	Work	Proof	Replicas? Where?	Public? Private?
Original PDP	[3]	PDP	no replicas	private
	[13]	PDP	no replicas	private
Original POR	[2]	POR	no replicas	private
	[10]	POR	no replicas	private (pseudo)
	[10]	POR	no replicas	public (BLS)
Proof of Reliability (data replication)	[5]	POR	replicas at client's side	private
	[39]	POR	replicas at client's side	private
	[9]	POR	replicas at server's side	private
	ours	POR	replicas at server's side	public
Proof of Reliability (erasure codes)	[33]	other	replicas at client's side	private
	[46]	POR	replicas at server's side	private
Proof of Reliability (proof of location)	[48]	POR	replicas at server's side but individual POR instances	private

Most of the compared works are based on PDP and POR proofs. When replica services are offered, we consider proof of correct data storage in a single instance for both original files and their copies (otherwise specified). Replica options are an important service to be offered by cloud providers, in order to avoid unfortunate data loss due to technical issues (e.g. single point of failure).

In almost all works, only private verification is offered. Private verification means that only the client can challenge the server for correct storage. We think that such an assumption is too strong as clients may be too lazy to do so. We suggest that delegating such a checking process to a third party will guarantee a better storage experience for both clients and cloud providers.

The closest work to ours is definitely the work from [9]. Indeed, the authors aim to develop a cloud storage solution to prevent both rational cloud providers and malicious clients. To do so, they combine POR instances with RSA-based puzzles [50].

One limitation in [9] is from the private aspect of POR instance verification. Indeed, either the client or someone in possession of the secret key can perform the verification of integrity of stored data. We aim to develop a publicly-verifiable PORR scheme, that allows everyone, using only public elements, to check that the cloud provider correctly stores original files and their replicas.

2 Contributions and Road Map

In [1], the first Proof of Retrievability and Reliability (PORR) with public verification was proposed. The original file of the client is first uploaded to a cloud provider. The replica plan has been agreed by both the client and cloud provider. Following that plan, replicas of the original file are prepared by the cloud provider. Anyone (and possibly the client) can challenge the cloud provider to check the integrity of the stored files and replicas.

In [1], the publicly verifiable RSA-based POR scheme from Shacham and Waters [10, 11] was combined with the exponentiation-based VDF construction from [12]. This allows us to define a common parameter setting, roughly a RSA-based one, while benefiting all the properties offered by the two schemes. Namely, the verifier can efficiently launch a large number of challenge requests and checks the cloud provider's responses using only public elements, while the cloud provider is forced to store the client's original file along with all the replicas at rest (otherwise being timely noticed).

In this paper, we provide the experimental analysis of our publicly verifiable PORR protocol. Implementation and evaluation results reveal realistic applications. Communication overhead between the client, the cloud provider and the verifier remains fair. Computation costs are affordable for the client and verifier, while they are made such that the cloud provider does not gain in computing the replicas on the fly from challenge requests. We also compare our solution with MIRROR [9], a PORR prototype with private verification, built upon Waters and Shacham's POR [10] and RSA-based puzzles [50].

Therefore, our solution fits cloud-related bandwidth and storage requirements, while preventing malicious clients and rational cloud providers from being successful.

In the following Sections 3 and 4, we recall the definition and construction of Shacham-Waters POR [10, 11] and of VDF from [12]. The reader can directly jump to Section 5 if being familiar with concepts of POR and VDF. In Section 5, our publicly verifiable PORR protocol presented in [1] is described. In Section 6, we depict the implementation framework of our PORR and analyse the experimental results. We conclude this paper in the last section.

3 RSA-based POR with Public Verification

In this section, we recall the publicly verifiable RSA-based POR construction from [10, 11]. Since our choice of slow functions [12] is based on RSA and also publicly verifiable, such POR construction is welcomed. The RSA-based POR construction in [11] is an extension of the RSA-based PDP construction in [4].

3.1 Definition

A Proof Of Retrievability (POR) comprises five algorithms. The client runs the two first algorithms. A public and secret key pair (pk, sk) are generated by running the key generation algorithm. For each file M, the client computes a corresponding tag T. Using the secret key sk, the file and tag generation algorithm processes the file M, resulting into M^*, and computes the tag T.

The client challenges the cloud provider when she wants to verify her files are stored in their entirety. A challenge set Q, selected when running the challenge generation algorithm, defines the file blocks that the cloud provider must prove to store correctly. From this challenge, the latter replies by running the response generation algorithm and shares the response $resp$ with the client. The latter can then verifies $resp$ by running the verification algorithm. The output of this algorithm tells the client whether or the cloud provider stores her files in their entirety.

In the above description, we suggest that the client verifies the cloud provider's response. However, in a publicly verifiable case, anyone can verify $resp$ since only public elements are required to run the verification algorithm. In a privately verifiable case, only someone with a secret element can check $resp$, the client in general would have such competency, but a private verifier can also be given a secret key. We only consider the first case in our paper.

POR protocols must be proved correct and sound. Correctness requires that for all key pairs (pk, sk) output by the key generation algorithm, for all file $M \in \{0,1\}^*$, and for all pair (M^*, T) output by the file and tag generation algorithm, the verification algorithm accepts with challenge and response respectively output by the challenge generation algorithm and the response generation algorithm. Soundness states that if a cheating cloud provider convinces the verifier that it is storing the file M, then it is actually storing the file M. Shacham and Waters formalize the notion of an extractor algorithm that interacts with the cheating cloud provider using the POR protocol.

3.2 Notations

The processed file M' (obtained from applying an erasure code on the file M) is split into n blocks, and each block is then split into s sectors. Each sector $m_{i,j}$, for $i \in [1, n]$ and $j \in [1, s]$, is an element of $\mathbb{Z}_N$. For instance, for a processed file M' that is b-bit long, there are $n = \lceil b/s \log(N) \rceil$ blocks.

A challenge is an l-element set $Q = \{(i, v_i)\}$. The size l of the set Q is a system parameter. Each tuple $(i, v_i) \in Q$ can be described as follows:

- The value i is a block index in $[1, n]$.

- The value v_i is an element in $E \subseteq \mathbb{Z}_N$. Let $E \subseteq \mathbb{Z}_N$ be the set of coefficients v_i chosen for verification requests. For instance, $E = \mathbb{Z}_N$, such that coefficients v_i are randomly chosen from $\mathbb{Z}_N$.

3.3 Construction

Let κ be the security parameter. Let κ_1 be a bit length such that the difficulty of factoring a $(2\kappa_1 - 1)$-bit modulus fits the security parameter κ. Let $max(E)$ be the largest element in E. Let κ_2 be the bit length equal to $\lceil \log(l \cdot max(E)) \rceil + 1$ [10, 11].

The construction of publicly verifiable POR with RSA signatures given in [11] is as follows:

Key Generation: On input a security parameter κ, the randomized key generation algorithm outputs the public key and secret key pair (pk, sk) as follows. Let $(spk, ssk) \leftarrow \mathsf{S.KeyGen}(\kappa)$ be a random signing key pair. Choose two primes p and q at random such that $p, q \in [2^{\kappa_1 - 1}, 2^{\kappa_1} - 1]$. Let $N = pq$ be the RSA modulus such that $2^{\kappa_1 - 2} < N < 2^{\kappa_1}$. Let $G : \{0,1\}^* \rightarrow \mathbb{Z}_N^*$ be a full-domain hash function, seen as a random oracle. Pick at random a prime e of length $2\kappa_1 + \kappa_2$ bits, and set $d = e^{-1} \mod \phi(N)$.

Set the public key $pk = (N, e, G, spk)$ and the secret key $sk = (N, d, G, ssk)$.

File and Tag Generation: On inputs the secret key sk and the file M, the file and tag generation algorithm outputs a processed file M^* and the tag T as follows:

1. Apply the erasure code on M to obtain M'.
2. Split M' into n blocks for some integer n, each of them being s sectors long, resulting into a $n \times s$ matrix $\{m_{i,j}\}_{i \in [1,n], j \in [1,s]}$. Each sector $m_{i,j}$ is an element of $\mathbb{Z}_N$.
3. Choose a random file identifier $id \in \mathbb{Z}_N$.
4. Pick at random s random numbers $u_1, u_2, \cdots, u_s \in \mathbb{Z}_N^*$.
5. Let $T_0 = id \| n \| u_1 \| u_2 \| \cdots \| u_s$. Compute the file tag $T = T_0 \| \mathsf{S.Sig}_{ssk}(t_0)$.
6. For each $i \in [1, n]$, compute $\sigma_i = (G(id \| i) \cdot \prod_{j=1}^{s} u_j^{m_{i,j}})^d \mod N$.
7. The processed file is $M^* = (\{m_{i,j}\}_{i \in [1,n], j \in [1,s]}, \{\sigma_i\}_{i \in [1,n]})$.

Challenge Generation: On inputs the secret key sk and the tag T, the randomized challenge generation algorithm outputs the challenge set Q as follows:

1. Use the key spk to verify the signature on T. If the signature is invalid, then output 0 and halt.
2. Otherwise, recover $id, n, u_1, u_2, \cdots, u_s$.
3. Pick at random an l-element subset I from the set $[1, n]$. For each $i \in I$, choose a random element $v_i \in E$. Let $Q = \{(i, v_i)\}_{i \in I}$.

Response Generation: On inputs the processed file M^* and the challenge set $Q = \{(i, v_i)\}_{i \in I}$, the deterministic response generation algorithm outputs a response $resp$ as follows:

1. Compute $\mu_j = \sum_{(i,v_i) \in Q} v_i m_{i,j} \in \mathbb{Z}$ for $j \in [1, s]$ (note that there is no modular reduction).
2. Compute $\sigma = \prod_{(i,v_i) \in Q} \sigma_i^{v_i} \mod N$.
3. Set the response $resp = (\{\mu_j\}_{j \in [1,s]}, \sigma)$.

Verification: On inputs the response $resp$, the deterministic verification algorithm checks first whether each μ_j is in the range $[0, l \cdot N \cdot max(E)]$. If some values are not in the range, then halt and output 0. Otherwise, check whether the equation $\sigma^e = \prod_{(i,v_i) \in Q} G(id \| i)^{v_i} \times \prod_{j=1}^{s} u_j^{\mu_j} \mod N$. If so, then it outputs 1; otherwise, it outputs 0.

3.4 Security

The security for POR is linked to *unforgeability*, *extractability* and *retrievability*. Informally, the security proof is split in three parts such that:

- The first part shows that the verifier can not receive an illegitimate (i.e. forged) response from the adversary (i.e. the cheating cloud provider).
- The second part shows that there exists an extractor extracting a constant fraction δ of file blocks (previously encoded with an erasure code) as soon as there exists an adversary that succeeds the verification process a noticeable number of times. Indeed, all responses that have been checked should be legitimate.
- The last part shows that one can use an erasure code to rebuild the whole file M if the constant fraction δ of file blocks has been recovered.

The RSA-based POR construction has been proved correct and sound. We let the readers refer to [10, 11] for the formal security proof.

4 Verifiable Delay Functions From Exponentiation in a Finite Group

In order to prevent rational behaviour from cloud providers, we aim to use slow functions (also called puzzles). Slow functions are constructed such that:

- Puzzles and file blocks are combined, resulting into r correct replicas of the file M. Homomorphic properties required for compact proofs are preserved.
- The client can adjust the difficulty of the puzzles supply multiple cost metrics.
- The costs from the storage of the solution of the puzzle are at least as high as the storage needed for the replicas of the file M.
- The cloud provider needs more time to solve the puzzles than the client, since the latter has access to a trapdoor.

In this section, we recall the construction of VDFs [12] that we use in our PORR. In order to successfully combine Shacham-Waters POR with VDFs, we opt for the exponentiation-based version with bounded pre-computations in a RSA group. This version is secure against attackers with bounded pre-computations, from a generalization of exponentiation-based time-lock puzzles in groups of unknown order.

4.1 Definition

An algorithm is said to run with p processors in parallel time t if one can implement it on a PRAM machine with p parallel processors that run in time t. The total sequential time refers to the time required for computation on a single processor.

A Verifiable Delay Function (VDF) is composed of three algorithms, to set up the system, evaluate the slow function and verify a solution. On inputs a security parameter κ and a delay parameter t, the randomized setup algorithm outputs the public parameters pp containing an evaluation key ek and a verification key vk. This algorithm must be polynomial-time in κ. The delay parameter t must be sub-exponentially sized in κ.

On inputs the evaluation key ek and a puzzle x from some known sampleable set, the evaluation algorithm outputs the solution y and a (possibly empty) proof π. For all public parameters pp and for all puzzles x, the evaluation algorithm must run with $poly(\log(t), \kappa)$ processors in parallel time t. For a given puzzle x, there must be a unique output y whose verification will be correct.

On inputs the verification key vk, the puzzle x, the solution y and its proof π, the deterministic verification algorithm outputs either 1 if y is a valid solution for the puzzle x, or 0 otherwise. The verification algorithm must run in total time polynomial in $\log(t)$ and κ. This algorithm is much faster than the evaluation one.

The VDF is expected to be *sequential* where honest parties can compute (y, π) by running the evaluation algorithm in t sequential steps, while no parallel-machine adversary with a polynomial number of processors can make the distinction between the output y from a random value in many less steps. Moreover, the VDF is expected to be *efficiently computable*, where honest parties run the verification algorithm as fast as possible such that the total time should be $O(polylog(t))$. The VDF is finally expected to be *unique* where for all puzzles x, the value y is difficult to calculate such that the verification algorithm outputs 1 while y is not an output of the evaluation algorithm on inputs pp and x.

4.2 Notations

Given an integer n, let $[1, n]$ denote the set of integers $\{1, 2, \cdots, n\}$. Let $L = \{l_1 = 3, l_2 = 5, \cdots, l_t\}$ be the first t odd primes. The parameter t is the provided delay parameter. Let P be the product of the primes in L, i.e. $P = l_1 \cdot l_2 \cdots l_t$. The parameter P is a large integer with about $t \log t$ bits.

4.3 Construction

Here, we describe in details the exponentiation-based solution for VDF in [12].

Setup: The trusted setup process is the following:

1. Set a RSA modulus $N = pq$ (for instance, 4096 bits long) such that the prime factors p, q are strong primes. The factorization of N is only known by the trusted setup algorithm. Let $H : \mathbb{Z} \rightarrow \mathbb{Z}_N^*$ be a random hash function.
2. For a given pre-processing security parameter B (for instance, $B = 2^{30}$), do the following:
   - Compute $H(i) = h_i \in \mathbb{Z}_N^*$ and $g_i = h_i^{1/P} \in \mathbb{Z}_N^*$ for $i \in [1, B]$.
   - Set $ek = (\mathbb{Z}_N^*, H, g_1, g_2, \cdots, g_B)$ and $vk = (\mathbb{Z}_N^*, H)$.

While the parameters of the verifier are short, the ones of the evaluator are not.

Evaluation: Solving a puzzle x works as follows:

1. Map the puzzle x to a random subset $L_x \subseteq L$ of size κ and a random subset S_x of κ values in $[1, B]$, using a random hash function.
2. Let P_x be the product of all the primes in L_x and let $g = \prod_{i \in S_x} g_i$.
3. The puzzle solution is $y = g^{P/P_x}$.

The computation of the solution takes $O(t \log t)$ multiplications in $\mathbb{Z}_N^*$.

Verification: Verifying a solution y works as follows:

1. Compute P_x and S_x as in the evaluation algorithm on inputs ek and x.
2. Compute $h = \prod_{i \in S_x} H(i)$.
3. Output 1 if and only if $y^{P_x} = h$.

We observe that exactly one element $y \in \mathbb{Z}_N^*$ will be accepted as a solution for a puzzle x. The verification process takes only $\tilde{O}(\kappa)$ group operations.

4.4 Security

Security is defined in face of an attacker able to perform polynomially bounded pre-computations. A VDF scheme must satisfy:

Correctness and Soundness. Every output of the evaluation algorithm must be accepted by the verification algorithm. The solution y for a puzzle x is guaranteed to be unique because the evaluation algorithm evaluates a deterministic function on the sampleable set of puzzles. The proof π does not require to be unique but should be sound and a verifier cannot be convinced that some different output is the correct VDF outcome.

τ-Sequentiality. No adversary should be able to compute an output for the evaluation algorithm on a random puzzle in parallel time $\tau(t) < t$, even with up to many parallel processors, and after a potentially large amount of pre-computations.

We let the readers refer to [12] for more details on the security models for VDFs.

The above construction does not satisfy the definition of a secure VDF presented in [12]. More precisely, an adversary who is able to run a large pre-computation once the parameters pp are known can break the above construction. There are various possible pre-computation attacks requiring tB group operations in $\mathbb{Z}_N^*$ [12].

New parameters must be generated after B challenges; otherwise, the scheme is not secure. This is sufficient for our application of a VDF, for instance by choosing $B = 2^{30}$. Regarding experiments of other solutions [9, 45, 46], storage challenges never excess such B.

5 RSA-based PORR with Public Verification

In this section, we describe our RSA-based solution for Proof Of Retrievability and Reliability (PORR) with public verification, using exponentiation-based VDFs to prevent the cloud provider to generate replicas on the fly when being challenged.

A client encodes and then processes a file M into M^*, and outsources the latter to the cloud provider. The cloud provider then commits to store M^* entirely across a set of r storage nodes with reliability guarantee R. This means that M^* contain the original copy of M along with replicas.

A PORR protocol is executed between a client and a cloud provider provided by its k storage nodes. The goal of such protocol is to enable either the client or anyone else to check the integrity and reliability of the processed file M^*.

5.1 Definition

Informally, our PORR protocol combines the RSA-based POR scheme of Shacham and Waters [10] and the exponentiation-based VDF scheme from [12].

The *Setup* phase initiates the protocol. It corresponds to the key generation and file and tag generation algorithms of Shacham-Waters POR scheme and to the setup algorithm of the VDF scheme. The client generates the parameters of the protocol, corresponding to the ones found in the two underlying schemes. She prepares her to-be-stored M by encrypting and processing it, and by generating the tag T and the authenticators σ_i for $i \in [1, n]$ (where n is the number of blocks) as in the POR scheme [10]. According to the agreed number r of replicas that the cloud provider must store, the client also prepares the VDF puzzles $x_{i,j}^{(k)}$, for $i \in [1, n]$, $j \in [1, s]$ and $k \in [1, r]$. In other words, there is one challenge per sector per replica (we recall that there are s sectors per block).

Once the cloud provider receives the file-related elements to be stored, the Setup phase is over. It can then start the second phase, namely the *Replica Generation* phase, in order to create the r replicas of the original file M. To do so, the cloud provider evaluates the VDF puzzles $x_{i,j}^{(k)}$ by running the evaluation algorithm. It appends each solution $y_{i,j}^{(k)}$ with the corresponding sector replica $m_{i,j}^{(k)}$, for $i \in [1, n]$, $j \in [1, s]$ and $k \in [1, r]$.

The next three phases can be requested multiple times. There is an interaction between a verifier (the client herself or someone on her behalf) and the cloud provider. During the *Challenge Generation* phase, the verifier generates the challenge *chal* and sends it to the cloud provider, by running the challenge generation algorithm of Shacham-Waters POR scheme. During the *Response Generation* phase, the latter replies back to the client with a response *resp* by running the response generation algorithm of POR scheme. Finally, during the *Verification* phase, the verifier then checks *resp* using only public elements, using the verification algorithms of Shacham-Waters POR scheme and of the VDF scheme. Indeed, verification exactly contains two steps: one check for POR and one check for VDF. If the output is 1, then the verifier is guaranteed that the cloud provider stores the file in its entirety along with its r replicas.

5.2 Construction

Here, we describe in details our publicly verifiable RSA-based PORR solution.

Setup: This phase includes the POR-based key, file and tag generations along with the VDF-based setup process.

Let κ be the security parameter. Let $(\mathsf{S.KeyGen}, \mathsf{S.Sign}, \mathsf{S.Verify})$ be a digital signature scheme. Choose two primes p and q at random such that $p, q \in [2^{\kappa_1 - 1}, 2^{\kappa_1} - 1]$. Let $N = pq$ be the RSA modulus such that $2^{\kappa_1 - 2} < N < 2^{\kappa_1}$. Let $G : \{0,1\}^* \to \mathbb{Z}_N^*$ be a full-domain hash function, seen as a random oracle. Pick at random a prime e of length $2\kappa_1 + \kappa_2$ bits, and set $d = e^{-1} \mod \phi(N)$. Let t be the delay parameter and B be the security parameter for VDFs. Let $L = \{l_1 = 3, l_2 = 5, \cdots, l_t\}$ be the first t odd primes and $P = l_1 \times l_2 \times \cdots \times l_t$. Let $H : \mathbb{Z} \to \mathbb{Z}_N$ be a hash function, seen as a random oracle.

The client wishes to store a file $M \in \{0,1\}^*$ at the cloud. Without loss of generality, the file M is assumed to be encrypted and encoded (using the specific erasure code). Encryption guarantees confidentiality and encoding guarantees extractability and retrievability.

As in [10], the file M is first split into n blocks, and then split into s sectors. Let us denote a sector as $m_{i,j} \in \mathbb{Z}_N$, for $i \in [1, n]$ and $j \in [1, s]$. Bit representation of each sector $m_{i,j}$ includes a characteristic pattern (e.g. a sequence of zero bits), in order to guarantee extractability [9]. Pattern length and file size are dependent such that the former should be larger than $\log_2(n \cdot s)$.

The client runs the algorithm $\mathsf{S.KeyGen}(\kappa)$ and gets the signing and verification key pair (ssk, spk). She also chooses an identifier $id \in \mathbb{Z}_N$ for the processed file M^*. She then picks at random s non-zero elements $u_1, u_2, \cdots, u_s \in \mathbb{Z}_N$. The client computes $T_0 = id \| n \| u_1 \| u_2 \| \cdots \| u_s$ and then $T = T_0 \| \mathsf{S.Sign}_{ssk}(T_0)$. Moreover, the client calculates $\sigma_i = (G(id \| i) \cdot \prod_{j=1}^{s} u_j^{m_{i,j}})^d \mod N$ for $i \in [1, n]$. We notice that all operations are done in the multiplcative group $\mathbb{Z}_N^*$ of invertible integers modulo N.

Both the cloud provider and client have agreed to create r replicas of the file M and store all of them at rest. She com-

putes $h_i = H(i)$ and $g_i = h_i^{1/P}$ for $i \in [1, B]$. She also chooses the values $x_{i,j}^{(k)}$ for $i \in [1, n]$, $j \in [1, s]$ and $k \in [1, r]$.

Finally, the client uploads the processed file $M^* = (\{m_{i,j}\}_{i\in[1,n],j\in[1,s]}, \{\sigma_i\}_{i\in[1,n]})$ to the cloud provider. She also forwards the public parameters $params = (N, e, G, H, spk, L, P, T, \{g_w\}_{w\in[1,B]}, \{x_{i,j}^{(k)}\}_{i\in[1,n],j\in[1,s],k\in[1,r]})$ to the cloud provider and anyone interested in playing the role of the verifier.

The client keeps secret the tuple (N, d, G, ssk).

Replica Generation: This phase includes the evaluation process of the underlying VDF.

The cloud provider calculates the solution $y_{i,j}^{(k)}$ for each $x_{i,j}^{(k)}$, and then build the replica $m_{i,j}^{(k)}$ of the original sector $m_{i,j}$, for $k \in [1, r]$.

First, the cloud provider maps $x_{i,j}^{(k)}$ to $L_{i,j}^{(k)} \subseteq L$ of size κ and the random subset $S_{i,j}^{(k)}$ of κ values in $[1, B]$, using a random hash function. Second, it sets $P_{i,j}^{(k)}$ as the product of all primes in $L_{i,j}^{(k)}$ and computes $g_{i,j}^{(k)} = \prod_{w\in S_{i,j}^{(k)}} g_w$. Third, the cloud provider computes the solution $y_{i,j}^{(k)} = (g_{i,j}^{(k)})^{P/P_{i,j}^{(k)}} \in \mathbb{Z}_N$.

Finally, it computes $m_{i,j}^{(k)} = m_{i,j} + y_{i,j}^{(k)}$ as the k-th replica of the sector $m_{i,j}$.

Challenge Generation: This phase corresponds to the challenge generation of POR.

First, the verifier (possibly the client) generates the challenge $chal$. Given $T = T_0 \| \mathsf{S.Sig}_{ssk}(T_0)$, check that $\mathsf{S.Sign}_{ssk}(T_0)$ is a valid signature by running the algorithm $\mathsf{S.Verify}_{spk}$. If the signature is invalid then halt.

Thereafter, elements $id, n, u_1, u_2, \cdots, u_s$ are recovered the verifier. The latter then sets $I \subset [1, n]$ of l elements and randomly selects l elements $v_i \in \mathbb{Z}_N$, for $i \in I$. Then, let $Q = \{(i, v_i)\}_{i\in I}$ where i is defined as the index of the block m_i. A set $R \subset [1, r]$ is also set by the verifier. Finally, let $chal = (Q, R)$ be forwarded to the cloud provider.

Response Generation: This phase corresponds to the response generation of POR.

Upon reception of the challenge $chal$, the cloud provider creates its response $resp$. First, it computes $\mu_j = \sum_{(i,v_i)\in Q} v_i m_{i,j} \in \mathbb{Z}$ and $\sigma = \prod_{(i,v_i)\in Q} (\sigma_i \cdot \prod_{j=1}^{s} \prod_{k\in R} u_j^{m_{i,j}^{(k)}})^{v_i} \mod N$. It sets $resp = (\{\mu_j\}_{j\in[1,s]}, \sigma)$ and forwards it to the verifier.

Verification: This phase includes both POR- and VDF-based verification steps.

Upon reception of the response $resp$, the verifier verifies that whether each μ_j is in the range $[0, l \cdot N \cdot max(\mathbb{Z}_N)]$. The verifier halts and outputs) as soon as some values are not in the range. Otherwise, the verifier checks whether the following equation holds:

$$\sigma^e = \prod_{(i,v_i)\in Q} G(id\|i)^{v_i} \times \prod_{j=1}^{s} u_j^{\mu_j(1+|R|e)} \times \Big(\prod_{(i,v_i)\in Q} \prod_{j=1}^{s} \prod_{k\in R} u_j^{v_i (h_{i,j}^{(k)})^{1/P_{i,j}^{(k)}}} \Big)^e \mod N \quad (1)$$

The verifier outputs 1 if the equation holds. She outputs 0 otherwise.

5.3 Security

In this section, we first describe the possible misbehaving entities in PORR, w.r.t. the cloud provider and client. We then present the security goals of a PORR scheme and define the correctness requirements, according to the ones in [9]. Our PORR solution must guarantee three security goals, namely:

- **Extractability:** The client can recover the original file M in its entirety.
- **Soundness of replica generation:** The replicas of the original file M must be correctly generated.
- **Storage allocation commitment:** The cloud provider utilizes at least as much storage as required to store the original file M and its replicas.

Traditional security notions are either embedded in those security notions (e.g. integrity), or assumed to be guaranteed by default (e.g. confidentiality).

We also sketch the security proofs of our scheme according to their respective security goals. We let the reader reach [9, 10, 11, 45, 46] for more details about the PORR security models and proofs. Indeed, the extractability proof is based on the security of the original POR scheme [10, 11]. Proofs for soundness of replica generation and storage allocation commitment follow the same arguments than in MIRROR [9]. Those security goals are however made guaranteed based on the security of the VDF puzzles [45, 46] in our case rather than of the RSA-based puzzles [50] as in MIRROR.

5.3.1 Misbehaving Entities.

Two entities, namely a cloud provider and a client, participate in our PORR protocol. Both are assumed to attempt malicious behaviours.

Rational Cloud Provider: If the cloud provider cannot save any costs by misbehaving, then it is likely to simply behave honestly. The advantage of an adversarial cloud provider depends on the ratio between costs for storage and accessibility to various resources (e.g. computing) and their availability. Hence, a cloud provider is restricted to a bounded number of concurrent threads of execution.

Therefore, we say that a cloud provider is *rational* when it can achieve cost savings by cheating. For instance, the cloud provider may attempt to get some storage space saved while

the overall cost for operations has not increased. Such cost is limited to the number of storage servers along with limited computational and storage capacities of each server. If supplying computational resources incurs additional costs, then the cloud provider invests in extra computing resources if such a strategy would result in lower total costs (including the underlying costs of storage).

Let us assume that a rational cloud provider achieves to generate a valid response while not reliably storing the client's data. If in order to reach such a behaviour, the cloud provider either provides more resources for storage or provides more resources for computing than resources when it follows the protocol honestly, then the cloud provider likely decides to not behave maliciously.

The client is protected from a misbehaving cloud provider that is not storing the file in its entirety by considering the *extractability* notion. The client is also protected from a misbehaving cloud provider that is not committing enough storage to keep all the file replicas by considering the *storage allocation* notion. Both properties guarantee both the integrity of the original file and its replicas. This means that the cloud provider invests enough redundancy to keep the client's data safe.

Malicious client: We say that a client is *malicious* when she can encode additional data in the replicas by cheating. This additional data cannot be found in the original file.

A client may attempt to abuse on storing more data in the file replicas than what has been approved between the client and cloud provider. In particular, replicas may have a lower cost than their original files (e.g. Amazon S3). Therefore, additional data may be inserted into replicas by a malicious client.

The cloud provider is protected from such a misbehaving client by considering the *correct replication* notion.

The security model used to build our security proofs comes from [9, 45, 46]. We do not consider the confidentiality of the outsourced file: we simply assume that the client encrypts it before the start of the PORR protocol.

We first show the correctness of our PORR scheme. We then move forward to the security properties with relation to cheating cloud provider and client, namely extractability, storage allocation and correct replication.

Informally, our PORR must achieve the following properties:

- *Extractability:* The file can be recovered in its entirety if and only if at least a fraction δ is stored at the cloud provider.
- *Storage allocation:* Misbehavior will be detected with overwhelming probability if less than a fraction δ is stored at the cloud provider.
- *Correct replication:* The client does not participate in the replica generation. Moreover, the size of the file is independent of the size of the parameters needed to create the replicas.

5.3.2 Correctness

If both the cloud provider and the verifier are honest, then on input the challenge *chal* sent by the verifier (output by the challenge generation algorithm), the response generation algorithm (run by the cloud provider) generates a response *resp* such that the verification algorithm outputs "1" with probability 1. This means that an honest cloud provider should always be able to pass the verification of proof of data reliability. From it, the PORR scheme is said to be correct.

During the verification phase, if both the verifier and cloud provider are honest, then on input $Q = \{(i, v_i)\}_{i\in I}$ generated by the verifier, the cloud provider should output a response *resp* such that the Equation (1) holds with probability 1.

Proof. Let N be the modulus, e be the public exponent and d the private exponent. The elements $P_{i,j}^{(k)}$ and $S_{i,j}^{(k)}$ are calculated as in the Replication Generation phase. We recall that $h_{i,j}^{(k)} = \prod_{w\in S_{i,j}^{(k)}} H(w)$ and so:

$$(y_{i,j}^{(k)})^{P_{i,j}^{(k)}} = ((g_{i,j}^{(k)})^{P/P_{i,j}^{(k)}})^{P_{i,j}^{(k)}} = ((h_{i,j}^{(k)})^{1/P})^P = h_{i,j}^{(k)} \mod N$$

From a challenge set $Q = \{(i, v_i)\}_{i\in I}$, with $\mu_j = \sum_{(i,v_i)\in Q} v_i m_{i,j}$ and $\sigma = \prod_{(i,v_i)\in Q} \sigma_i^{v_i}$, we get the following mod N:

$$\begin{aligned}
\sigma &= \prod_{(i,v_i)\in Q} (\sigma_i \cdot \prod_{j=1}^{s}\prod_{k\in R} u_j^{m_{i,j}^{(k)}})^{v_i} \\
&= \prod_{(i,v_i)\in Q} \sigma_i^{v_i} \times \prod_{(i,v_i)\in Q}\prod_{j=1}^{s}\prod_{k\in R} (u_j^{m_{i,j}^{(k)}})^{v_i} \\
&= \prod_{(i,v_i)\in Q} (G(id||i) \cdot \prod_{j=1}^{s} u_j^{m_{i,j}})^{v_i d} \times \prod_{(i,v_i)\in Q}\prod_{j=1}^{s}\prod_{k\in R} (u_j^{m_{i,j}+y_{i,j}^{(k)}})^{v_i} \\
&= \prod_{(i,v_i)\in Q} (G(id||i)^{v_i} \cdot \prod_{j=1}^{s} u_j^{v_i m_{i,j}})^{d} \times \prod_{(i,v_i)\in Q}\prod_{j=1}^{s}\prod_{k\in R} u_j^{m_{i,j} v_i} \\
&\quad \times \prod_{(i,v_i)\in Q}\prod_{j=1}^{s}\prod_{k\in R} u_j^{y_{i,j}^{(k)} v_i} \\
&= \Big(\prod_{(i,v_i)\in Q} G(id||i)^{v_i} \times \prod_{j=1}^{s} u_j^{(\sum_{(i,v_i)\in Q} v_i m_{i,j})}\Big)^d \\
&\quad \times \prod_{j=1}^{s}\prod_{k\in R} u_j^{(\sum_{(i,v_i)\in Q} v_i m_{i,j})} \times \prod_{(i,v_i)\in Q}\prod_{j=1}^{s}\prod_{k\in R} u_j^{v_i (h_{i,j}^{(k)})^{1/P_{i,j}^{(k)}}} \\
&= \Big(\prod_{(i,v_i)\in Q} G(id||i)^{v_i} \times \prod_{j=1}^{s} u_j^{\mu_j}\Big)^d \times \prod_{j=1}^{s} u_j^{|R|\mu_j} \\
&\quad \times \prod_{(i,v_i)\in Q}\prod_{j=1}^{s}\prod_{k\in R} u_j^{v_i (h_{i,j}^{(k)})^{1/P_{i,j}^{(k)}}}
\end{aligned}$$

and so

$$\begin{aligned}\sigma^e &= \prod_{(i,v_i)\in Q} G(id\|i)^{v_i} \times \prod_{j=1}^{s} u_j^{\mu_j} \times \Big(\prod_{j=1}^{s} u_j^{|R|\mu_j} \\ &\times \prod_{(i,v_i)\in Q} \prod_{j=1}^{s} \prod_{k\in R} u_j^{v_i (h_{i,j}^{(k)})^{1/P_{i,j}^{(k)}}}\Big)^e \\ &= \prod_{(i,v_i)\in Q} G(id\|i)^{v_i} \times \prod_{j=1}^{s} u_j^{\mu_j(1+|R|e)} \\ &\times \Big(\prod_{(i,v_i)\in Q} \prod_{j=1}^{s} \prod_{k\in R} u_j^{v_i (h_{i,j}^{(k)})^{1/P_{i,j}^{(k)}}}\Big)^e \mod N\end{aligned}$$

5.3.3 *Extractability*

The client is protected against a malicious cloud provider that is not storing the file in its entirety, through the notion of extractability.

An honest client should recover her file M with high probability. Following the notion of extractability in [10, 11], if a cloud provider can convince a honest client with high probability that it stores the file M^* (i.e. the processed version of the original file M), then there exists an extractor algorithm that, given enough interaction with the cloud provider, can extract the file M.

Sketch of Proof. Let us define a game between an adversary and an environment. The environment is a simulation of honest clients and honest verifiers. The adversary is allowed to submit requests to the environment to create new clients, along with their public and secret parameters, to let them prepare files and their replicas, and to verify correct storage. More precisely, the environment simulates honest client and verifier, and it further provides the adversary with oracles for the algorithms to set up the system, upload the file and replicas, generate the challenge and verify the response.

At the end, the adversary chooses a client with its file M and outputs a cloud provider who can execute the verification process with this client and the file M. A cloud provider is said to be ϵ-admissible if the probability that the verifier does not abort is at least ϵ. The adversary thus picks the client along with the client's secret tuple (N, d, G, ssk), the file M and the public parameters *params*, and simulates a cheating cloud provider. Let the latter succeed in making the verification algorithm yield "1" in an non-negligible ϵ fraction of PORR executions. We say that the PORR scheme meets the extractability guarantee, if there exists an extractor algorithm such that given sufficient interactions with the cheating cloud provider, it recovers M.

The computations executed to upload the file M and to verify correct storage are similar operations to the ones in the publicly verifiable POR scheme in [10, 11]. Therefore, the extractability arguments given in [10, 11] apply to our PORR solution directly. An additional assumption is made on the existence of an erasure coding mechanism applied to the file, to guarantee the entire recovery of the file M from any file fraction δ. We let the reader refer to [10, 11] to obtain additional information on the choice of the erasure codes.

5.3.4 *Soundness of Replica Generation.*

Contrary to extractability notion, soundness of replica generation aims to protect the cloud provider against a malicious client who tries to encode additional data in the replicas.

Sketch of Proof. The replica generation is said to be sound such that if the client is involved in the replica generation, then the cloud provider can get the assurance that the additional uploaded data represents indeed correctly built replicas that do not encode extra data. In our PORR, this situation is solved by not letting the client be fully involved in the replica generation. Indeed, while the client generates the puzzles whom solutions are created by the cloud provider, the latter is responsible of defining each replica with its puzzle solution.

The replica generation does not allow to encode a significant amount of extra data in the replicas. Indeed, the replication process takes as inputs elements whose size is independent of the file size. The replica generation is also said to be correct if replicas represent indeed copies of the uploaded file M. Indeed, the file M can be efficiently recovered from any replica M_k. There exists an efficient algorithm which given the file tags, the public parameters and any replica M_k outputs M.

5.3.5 *Storage Allocation Commitment.*

Similarly to the extractability notion, the storage allocation commitment property aims to protect a client against a cloud provider who does not commit enough storage to store all replicas.

The storage allocation commitment notion forces the cloud provider to store the outsourced file and its replicas at rest. Therefore, a cheating cloud provider that participates in the above extractability game [10, 11] and devotes only a fraction of its storage space to store the file and the replicas entirely, cannot convince the verifier to accept its response with overwhelming probability.

Sketch of Proof. The storage allocation ratio of the cloud provider is defined as follows:

$$\rho = \frac{|st|}{|M| + |M_1| + \cdots + |M_r|}$$

where st corresponds to the storage of the cloud provider that has been allocated for storing the original file M and its replicas $M_1, \cdots, M_r$. The file has been first encrypted and the replicas are copies of the processed file, thus they cannot be further compressed. We can assume that the cloud provider aims to save storage, thus it holds that $0 \leq \rho \leq 1$. The storage allocation commitment ensures that $\delta \leq \rho$ for a threshold $0 \leq \delta \leq 1$ agreed with the client (see the above extractability notion).

First, we want the cloud provider to use at least as much storage as needed to keep the file and its replicas. Second, we see our scheme as a POR protocol applied to both the original file and all the replicas. Let a challenge contain sectors that are not correctly stored. Then, our scheme guarantees that the cloud provider will fail with overwhelming probability unless the correct reconstruction of those sectors is possible.

Lastly, a malicious cloud provider should give a noticeable effort in reconstructing missing sectors. We can easily investigate such

an effort as follows. Let the cloud provider store the whole file but only parts of the replicas. The cheating cloud provider will require a significantly higher effort in recomputing missing replicas, compared to an honest service provider that has all the replicas stored in their entirety. By using slow functions (here the VDFs), the time in getting back the missing parts of replicas is noticeable from the verifier's point of view.

More precisely, the misbehaving cloud provider must compute the puzzle solutions $y_{i,j}^{(k)}$ in order to recompute the missing sectors. Since those elements $y_{i,j}^{(k)}$ are different for each replica, knowing (or reconstructing) one element $y_{i,j}^{(k)}$ from one replica sector does not help the cloud provider in deriving elements from other replica sectors. A rational cloud provider should thus require a significantly higher effort compared to an honest cloud provider in recomputing missing replicas.

Given that the VDF evaluation function requires a noticeable amount of time and effort compared to the associated VDF verification function, this incurs additional (significant) computational overhead on the cloud provider to compute the puzzle solutions on the fly rather than storing them at rest.

Let δ be the threshold selected by the client (see Extractability). Let us assume that less than a fraction δ of all sectors of a replica is stored at the cloud provider. Thus, for each element $y_{i,j}^{(k)}$ in the challenge, the probability to recompute it is at least $1 - \delta$. In addition, $l \cdot s$ sectors are contained in a challenge. Hence, the number of values to recompute is $l \cdot s \cdot (1 - \delta)$. Moreover, the costs from the time effort for recomputing these elements $y_{i,j}^{(k)}$ exceed the costs from the storage of those elements. This is made possible by having a number of challenges linear in the security parameter κ. If so, then the overall probability to avoid these computations is negligible in κ.

6 Implementation and Evaluation

We are interested in implementing and evaluating our PORR solution in a realistic cloud framework. We also wish to compare our results with the ones from [9], since their solution is the closest one to ours.

In this section, we first describe MIRROR, the PORR scheme presented in [9]. We then describe our implementation setting, and we discuss our results and compare them with the ones from [9]. We choose to compare our results with MIRROR's ones since, to our best of knowledge, MIRROR is the closest prototype to ours. Unlike existing schemes, the cloud provider replicates the data by itself in both MIRROR and our PORR. Therefore, expensive bandwidth resources are traded with cheaper computing resources.

As a summary, we evaluate an implementation of our PORR prototype within a realistic cloud setting and we compare the performance of it with MIRROR [9].

6.1 MIRROR

MIRROR [9] is the first scheme to enable the cloud provider to generate the replicas of the client's files by itself. Such move permits to trade expensive bandwidth resources with cheaper computing resources. Cloud providers and clients are likely to adopt it easily since storage services are improved while financial costs are lowered.

The authors in [9] present new PORR definition and security model. Their definition extends Shacham-Waters POR from [10]. Their security model encompasses security risks that have not been covered in previous multi-replica POR schemes, namely security against malicious clients and rational cloud providers. We have sketched the security proofs of our PORR based on their model.

The authors give a solution for PORR, called MIRROR, and prove it secure according to their new security model. Their motivations mostly rely on business matters for cloud providers and financial incentives for clients. They propose a tunable replication scheme by combining Linear Feedback Shift Registers and RSA-based puzzles [50]. By doing so, the burden incurred by the construction of the replicas is shifted to the cloud provider. The latter swaps higher resources for bandwidth with lower resources for computation. In addition, a realistic cloud framework has been defined for the implementation and evaluation of MIRROR. The results show that MIRROR is applicable to real cloud environments.

The main difference between MIRROR and our PORR solution is the nature of the verification process, private and public respectively. The authors in [9] propose to use privately verifiable puzzles to prevent rational behaviour from cloud providers. Therefore, such feature forces the puzzle creator, namely the client, to check herself the solution generated by the cloud provider. We claim that such requirement is too strict on the client's side. If we aim to design a solution for the general public, then we think about individuals that have limited knowledge on cloud storage security, thus that would not follow the process to prevent and/or detect malicious cloud provider behaviours. Instead, we suggest to use VDFs that offer public verification. This implies that anyone else on behalf of the client can verify that the cloud provider has been acting honestly.

6.2 Implementation

We implement our PORR scheme in order to analyze the computational efforts from honest and rational cloud providers, along with the ones from client and verifier.

We show that our protocol produces fair computational and communication overheads on the client, cloud provider and verifier, meaning that our solution is realistically applicable in cloud storage environments. Indeed, computation costs are affordable for the client and verifier, while they are made such that the cloud provider does not gain any advantage in computing the replicas on the fly from the verifier's challenge requests. Therefore, our solution fits cloud-related bandwidth and storage expectations, while preventing malicious clients and rational cloud providers from being successful.

Our implementation setting follows the one in [9] for a more accurate and legitimate comparison. Our code is written in Python 3.8. The selected hash function is SHA-256. The whole test environment is deployed on a PC running an Intel Core i7-9700 with 32GB of memory. We create four Virtual Machines (VMs) to design our test environment, namely:

- One VM represents the client that owns files to be stored on a cloud storage server.

- One VM represents the cloud provider (server) that offers cloud storage services.
- Two VMs represent storage nodes, linked to the cloud provider, that guarantee data replication.

We depict the test environment topology in Figure 2.

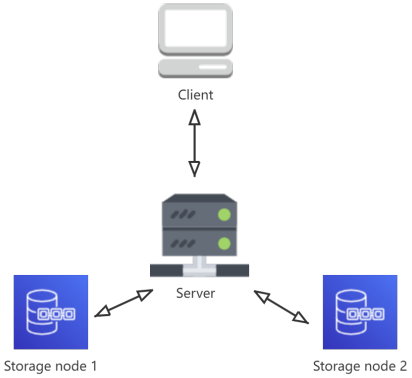


Figure 2: Test environment topology

To emulate a realistic Wide Area Network (WAN), the communication between the various VMs is bridged using a 100 Mbps switch. All traffic exchanged in the environment is shaped with NetEm [56]. The setting parameters are the following:

- We set the packet loss rate at 0.1%, and the correlation rate of lost packets at 0.001% [57, 58, 59].
- We set the delay rate to fit normal distribution with a mean of 20ms and a variance of 4ms [60].
- We set the packet corruption rate at 0.1% [61].
- We set the rate of the package being out of order at 0.2% [62].

Each data point is averaged in our plots by five independent measurements. Where appropriate, we include the corresponding 95% confidence intervals.

Parameter selection. We select our parameters similarly to [9] in order to achieve a better comparison. Let the modulus N be 2048-bit long (hence, 1024-bit primes p and q). We choose a block size equal to 8KB as such value offers the most balanced performance in average. This means that files of 64MB are split into n blocks where $n = 8,000$. Each block is then split into s sectors where $s = 32$. Then, r replicas are created for each file where $r = 2$. The cloud provider is thus expected to keep $8000 \times 32 \times 2 < 52 \cdot 10^4$ file elements, namely the original file sectors and the replica sectors. The number of blocks per challenge will be set to $|I| = 40$.

We consider three network storage protocols, namely File Transfer Protocol (FTP), Server Message Block (SMB) and Network File System (NFS), to enable the client to access her files stored in the cloud, through a computer network. NFS is used on Unix operating systems and SMB on Windows operating systems. The deployment of FTP is wider, as long as the communication port is open. In order to select one of those three network storage protocols, we test their uploading speed. Figure 3 shows the uploading speed values (in MB per second) for network storage protocols FTP, SMB and NFS. Due to the instability of network transmission, the uploading speed results show some ups and downs. The average FTP speed is the fastest, at 17.4 MB per second, which is around 1 second faster than NFS and 3 seconds faster than SMB. Hence, FTP was chosen as the default setting.

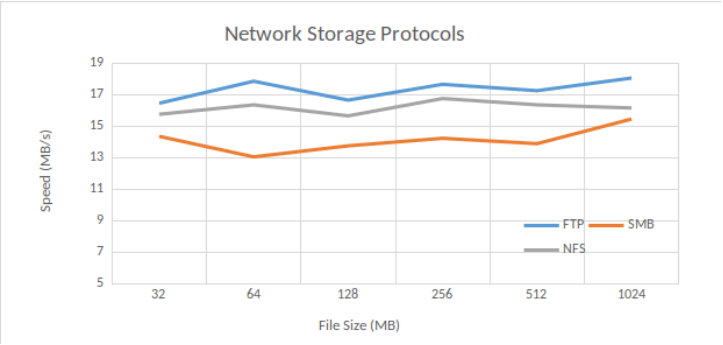


Figure 3: Network Storage Protocols

We summarize our parameter selection in Table **??**. When not specifically mentioned in Section 6.3, the default parameter values are kept as in Table **??** for our protocol evaluation.

Table 2: Parameter selection

Parameter	Default value
RSA modulus size $\|N\|$	2048 bits
RSA prime size $\|p\|$	1024 bits
RSA prime size $\|q\|$	1024 bits
Pre-processing parameter B	2^{30}
Delay parameter t	2^{16}
File size	64MB
Block size	8192 Bytes
Sector size	256 Bytes
Number of replicas r	2
Number of challenges $\|I\|$	40
Network Storage Protocol	FTP

VDF implementation. We recall that our PORR scheme is based on the exponentiation-based VDF scheme [12]. The latter has the following characteristics.

The VDF evaluation algorithm runs with $poly(\log(t), \kappa)$ processors in parallel time t, where κ is the security parameter and t is the delay parameter. We say that an algorithm runs with p processors in parallel time t if one can implement it on a PRAM machine with p parallel processors running in time t.

The VDF verification algorithm runs in total time polynomial in $\log(t)$ and κ. This algorithm should be much faster than the evaluation one.

Our PORR scheme uses VDFs in order to allow the verifier to detect a cheating cloud provider computing the puzzle solutions on the fly rather than keeping them at rest. Therefore, we aim to notice a difference between the response time of a honest cloud provider

and of a rational one. More precisely, a verifier should be able to observe a delay in responding from a cheating cloud provider. We choose a delay parameter t equal to 2^{16}.

The security of the VDF scheme based on exponentiation [12] relies on the assumption that the adversary cannot run a long pre-computation from the publication of the public parameters until the evaluation of the puzzles. Given the pre-processing parameter B, the VDF scheme is secure up to B puzzles. New public parameters must be generated once B puzzles have been used.

The authors in [12] suggest to set $B = 2^{30}$ for a reasonable trade-off between usability and security. Let us consider $B = 2^{30}$ and 64MB files. Up to 2000 files can be stored at the cloud with only one VDF instance. 2000 files roughly correspond to 128GB of data. Another VDF instance will be necessary if more files need to be stored. We recall that the setup algorithm runs in polynomial time in the security parameter κ, meaning that this algorithm is run easily and relatively fast.

We examine whether such condition on the pre-processing parameter B restricts the applicability of our PORR solution. In addition, we verify the costs incurred from computing and storing. Indeed, our results should convince us that our PORR solution fits real world requirements. We summarize our VDF implementation characteristics in Table **??**.

6.3 Evaluation and Discussion

Response generation. We measure the time that the cloud provider takes in order to generate its response, according to different numbers of challenges. Results are shown in Figure 4.

We consider two types of cloud providers: a rational cloud provider and a honest cloud provider. The former needs more time to generate its response since it requires to compute puzzle solutions on the fly. On the other side, since the honest cloud provider has already computed the puzzle solutions and stores them at rest, its response generation is noticeably faster. Therefore, we are guaranteed that the verifier (and thus the client) will notice a rational cloud server from a honest cloud provider since the time difference for response generation is consequent.

The average response time of an honest cloud provider is less than 2 seconds when $r = 2$ in [9], and a delay of almost 1 second is observed from a rational cloud provider. In our case, if the cloud provider tries to compute the 2 replicas on the fly, then it needs to compute the solutions y of 2 puzzles x, meaning that it runs in parallel time t with $poly(\log(t), \kappa)$ processors twice.

From Figure 4, we notice that a rational server needs almost twice the time than a honest server to generate its response, for a given number of challenges. For instance, for 40 challenges, a rational server generates its response in 0.63 seconds while a honest server in 0.31 seconds. Moreover, the response generation time increases with the number of challenges.

MIRROR's response generation depends on the bit size of the factors (that are elements needed to prepare the replicas) [9]. As long as the blind factors are greater or equal to 70 bits, the rational server should not gain any reasonable advantage in misbehaving. For instance, when the coefficients' size is equal to 70 bits, a rational server in MIRROR needs around 1 second to generate its response based on 40 challenges. Since we do not use blind factors, a rational server in our PORR only requires 0.63 seconds.

For a higher security level, that is a bigger blind factors' bit size (say 200 bits), a MIRROR rational server generates its response in roughly 2.8 seconds. In our case, in order to reach a higher security level, we can increase the number of challenges. For instance, for a number of challenges equal to 200, a rational server needs 4.86 seconds to generate its response. Such time cost happens since the rational server must calculate puzzle solutions on the fly.

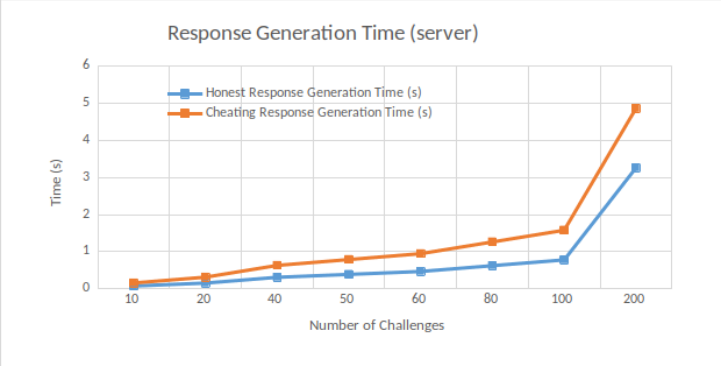


Figure 4: Response Generation Time

File preparation. We measure the time required to prepare a file before storing it on the cloud. Specifically, we measure the time for setup (i.e. key generation and tag generation) and replica generation (i.e. puzzle evaluation). Results are shown in Figure 5.

The time spent for preparing the file and its replicas increases exponentially with the size of the file. For instance, for a 16MB file and its 2 replicas, we need around 50 seconds while for a 256MB file, we require more than 700 seconds.

Moreover, file preparation takes around 180 seconds for a 64MB file and its 2 replicas in our case. We notice that setup and replica generation take around 500 and 700 seconds respectively in MIRROR [9], given the same file setting. Therefore, file preparation is almost 7 times faster in our PORR than in MIRROR for a 64MB file.

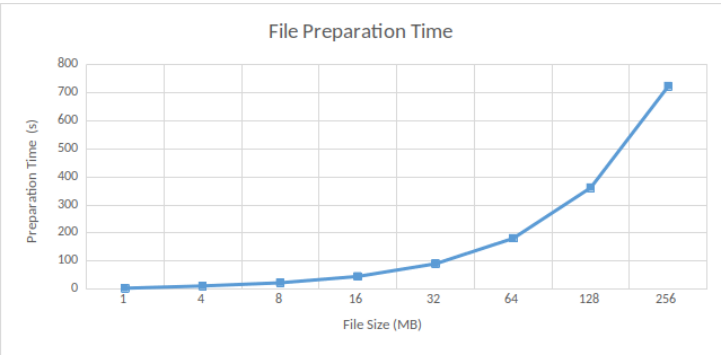


Figure 5: File Preparation Time

Verification. We measure the time required to verify correct storage at rest. Specifically, we measure the time for response generation and response verification. We consider multiple file sizes, and thus multiple numbers of challenges, since the latter depend

on the former. In particular, a large file means a bigger number of challenges. The number of challenges is calculated as the half of the file size (in MB). For example, 32 challenges are requested for a 64MB file. According to the default parameter settings, the verifier thus checks 128KB of data for each MB file data. Results are shown in Figure 6.

The verification time experienced by the client is 0.57 seconds in our case, while being 0.8 seconds in MIRROR [9]. The number of challenges increases with the file size, thus impacting the verification time. Given a file size of 128MB, the number of challenges is 64 and the verification time reaches 1.14 seconds. Note that the number of challenges in MIRROR is fixed at a value of 40, whatever the file size. Therefore, the verification time is constant, equal to 0.8 seconds.

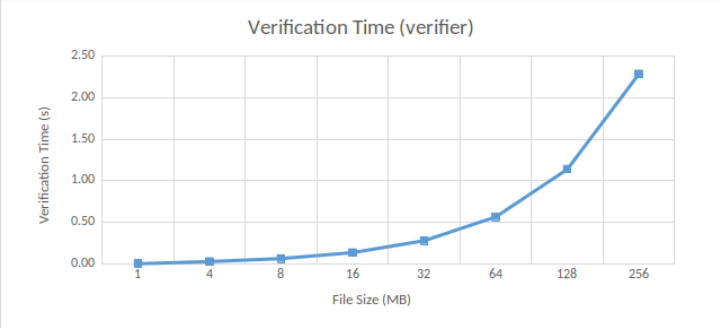


Figure 6: Verification Time

Puzzle evaluation. We measure the time spent for puzzle evaluation only. We are interested in such measurement since this part takes the longest time in file preparation. Results are shown in Figure 7.

Puzzle evaluations contribute to almost 3/4 of the file preparation. In the default configuration, that is given a 64MB file, file preparation time reaches 180 seconds, where around 130 seconds are spent on puzzle evaluations. Since the number of sectors increases with the file size, this leads to the linear growth of puzzle evaluation time.

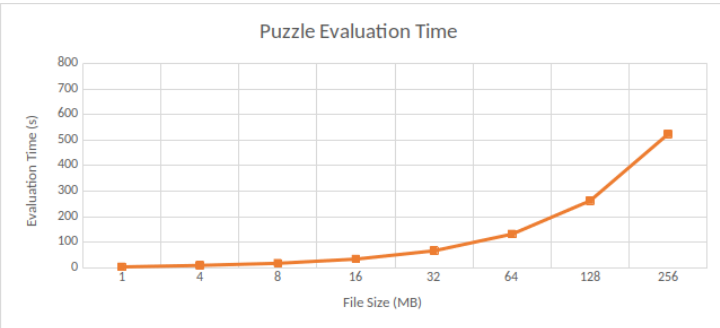


Figure 7: Puzzle Evaluation Time

Latency w.r.t. block sizes. We measure the time required for response generation and verification, according to the block sizes. We choose a sector size of 256 Bytes. Results are shown in Figure 8. We notice that a smaller block size gives a shorter time. Indeed, block size and challenge number are related, meaning that response generation and verification are faster as the block size decreases.

MIRROR [9] gets a similar trend to ours, but less obvious. In MIRROR, the verification time is around 1 second for a 1MB block size, and 1.1 seconds for a 2MB block size. In our case, the verification time is around 0.1 seconds for a 1MB block size, and 0.2 for a 2MB block size, thus an increase by 50%.

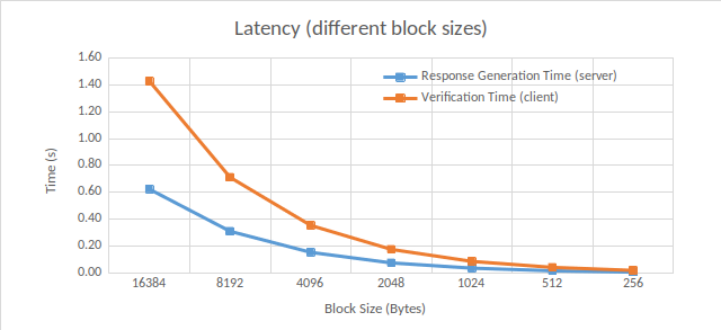


Figure 8: Latency (blocks)

Latency w.r.t. sector sizes. We measure the time required for response generation and verification, according to the sector sizes. We choose a block size of 8192 Bytes. For a given block size, a small sector size means a bigger number of sectors in that block. Therefore, the verification time is impacted since more data must be checked while the number of blocks and thus of challenges is fixed. Results are shown in Figure 9.

The authors in [9] do not explicitly analyze the effect of sector size of their protocol MIRROR. We expect that, for a given file, when the sector size decreases, its number increases, and thus impacts, with a rise, the server's response time and verifier's verification time.

Our default sector size is 256 Bytes, as a trade-off between efficiency and security. Verification is done in 0.36 seconds with such value. On the other size, a sector size equal to 32 Bytes results in getting a verification time 15 times slower.

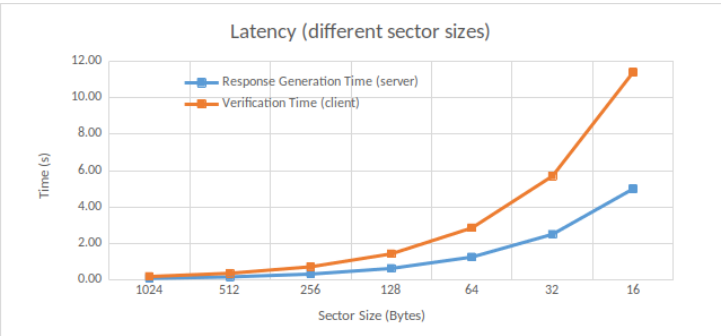


Figure 9: Latency (sectors)

Financial costs. We compare the financial costs between our PORR and MIRROR [9], w.r.t. file preparation. The replication process (as named in MIRROR) roughly corresponds to our file preparation. Results are shown in Figure 10.

The graph records those cost differences between the two protocols. The Amazon EC2 processor rental price is US$ 0.404 per hour. Such price is thus US$ 0.000112 per second, and has been multiplied with our file preparation time.

Given a file of size 384MB, our PORR and MIRROR will both incur a cost equal to 0.18 USD. Our PORR protocol seems more interesting from a financial point of view with file sizes smaller than 384 MB, saving from 2 to 15 times compared to MIRROR. On the other side, MIRROR seems more attractive with larger file sizes.

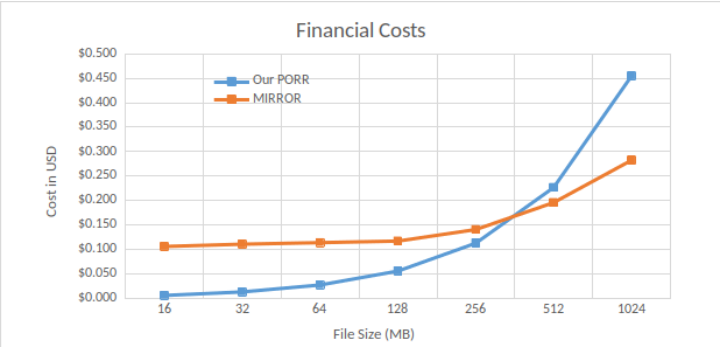


Figure 10: Financial Costs

Additional remarks. In MIRROR [9], replicas are computed as the product of the original file sector and two blind factors. However, the cloud provider may just keep the factors in reserve, and multiply by the sector when requested for verification. The gain may not be substantial and a cloud provider may just store the replicas.

Our replicas are computed as the addition of the original file sector and one blind factor. Similarly to [9], the cloud provider may or may not just keep the factor and add the sector when challenged. Future task will investigate whether a better design can prevent such behaviour from the cloud provider.

Summary. Our results show that our PORR prototype manages to trade expensive bandwidth resources with cheaper computing resources. Those results are likely to be well accepted by cloud providers and clients with the promise of better storage services and lower financial costs. Moreover, some of those results, especially on both technical and financial aspects, show that our PORR is more competitive with small file sizes, compared to MIRROR [9]. Therefore, our PORR solution may become one of the economically-viable, applicable systems that offer verifiable replicated cloud storage. In particular, our PORR fits the demands from indivudals and small businesses with lower file sizes to store on a cloud.

7 Conclusion

In this paper, we presented a recent PORR protocol, first proposed in [1], where one can check in a single instance that the cloud provider correctly stores an original file and its replicas. We combine our PORR with the slow function VDF to enable anyone to verify the cloud provider's behaviour, not only the file owner.

Implementation and evaluation of our solution have been carried out. Results show that such design is well applicable in realistic cloud environments. Multiple results, on technical and financial aspects, show that our PORR is more competitive with small file sizes, compared to MIRROR [9] and with the noticeable advantage of offering public verification.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment We thank the reviewers for their valuable comments.

References

[1] C. Gritti, "Publicly Verifiable Proofs of Data Replication and Retrievability for Cloud Storage," in 2020 International Computer Symposium (ICS), 431–436, 2020, doi:10.1109/ICS51289.2020.00091.

[2] A. Juels, B. S. Kaliski, Jr., "Pors: Proofs of Retrievability for Large Files," in Proceedings of the 14th ACM Conference on Computer and Communications Security, CCS '07, 584–597, ACM, New York, NY, USA, 2007, doi:10.1145/1315245.1315317.

[3] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, D. Song, "Provable Data Possession at Untrusted Stores," in Proceedings of the 14th ACM Conference on Computer and Communications Security, CCS '07, 598–609, 2007, doi:10.1145/1315245.1315318.

[4] G. Ateniese, R. Burns, R. Curtmola, J. Herring, O. Khan, L. Kissner, Z. Peterson, D. Song, "Remote Data Checking Using Provable Data Possession," ACM Trans. Inf. Syst. Secur., **14**(1), 12:1–12:34, 2011, doi:10.1145/1952982.1952994.

[5] R. Curtmola, O. Khan, R. Burns, G. Ateniese, "MR-PDP: Multiple-Replica Provable Data Possession," in Proceedings of the 2008 The 28th International Conference on Distributed Computing Systems, ICDCS '08, 411–420, IEEE Computer Society, Washington, DC, USA, 2008, doi:10.1109/ICDCS.2008.68.

[6] A. F. Barsoum, M. A. Hasan, "Integrity Verification of Multiple Data Copies over Untrusted Cloud Servers," in Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (Ccgrid 2012), CCGRID '12, 829–834, IEEE Computer Society, USA, 2012, doi:10.1109/CCGrid.2012.55.

[7] A. Barsoum, M. Hasan, "Provable Multicopy Dynamic Data Possession in Cloud Computing Systems," IEEE Transactions on Information Forensics and Security, **10**, 485–497, 2015, doi:10.1109/TIFS.2014.2384391.

[8] G. S. Prasad, V. S. Gaikwad, "A Survey on User Awareness of Cloud Security," Int. J. of Engineering and Technology, **7**(2.32), 131–135, 2018, doi:10.14419/ijet.v7i2.32.15386.

[9] F. Armknecht, L. Barman, J.-M. Bohli, G. O. Karame, "Mirror: Enabling Proofs of Data Replication and Retrievability in the Cloud," in Proceedings of the 25th USENIX Conference on Security Symposium, SEC'16, 1051–1068, USENIX Association, USA, 2016.

[10] H. Shacham, B. Waters, "Compact Proofs of Retrievability," in Proceedings of the 14th International Conference on the Theory and Application of Cryptology and Information Security: Advances in Cryptology, ASIACRYPT '08, 90–107, Springer-Verlag, Berlin, Heidelberg, 2008, doi:10.1007/978-3-540-89255-7_7.

[11] H. Shacham, B. Waters, "Compact Proofs of Retrievability," Full version, 2008, https://hovav.net/ucsd/dist/verstore.pdf.

[12] D. Boneh, J. Bonneau, B. Bünz, B. Fisch, "Verifiable Delay Functions," CRYPTO 2018, 757–788, Springer-Verlag, Berlin, Heidelberg, 2018.

[13] G. Ateniese, R. Di Pietro, L. V. Mancini, G. Tsudik, "Scalable and Efficient Provable Data Possession," in Proceedings of the 4th International Conference on Security and Privacy in Communication Netowrks, SecureComm '08, 9:1–9:10, ACM, New York, NY, USA, 2008, doi:10.1145/1460877.1460889.

[14] C. Wang, Q. Wang, K. Ren, W. Lou, "Privacy-preserving Public Auditing for Data Storage Security in Cloud Computing," in Proceedings of the 29th Conference on Information Communications, INFOCOM'10, 525–533, IEEE Press, Piscataway, NJ, USA, 2010.

[15] S. Yu, C. Wang, K. Ren, W. Lou, "Achieving Secure, Scalable, and Fine-grained Data Access Control in Cloud Computing," in Proceedings of the 29th Conference on Information Communications, INFOCOM'10, 534–542, IEEE Press, Piscataway, NJ, USA, 2010.

[16] Z. Hao, S. Zhong, N. Yu, "A Privacy-Preserving Remote Data Integrity Checking Protocol with Data Dynamics and Public Verifiability," IEEE Trans. on Knowl. and Data Eng., **23**(9), 1432–1437, 2011, doi:10.1109/TKDE.2011.62.

[17] Y. Yu, M. H. Au, Y. Mu, S. Tang, J. Ren, W. Susilo, L. Dong, "Enhanced privacy of a remote data integrity-checking protocol for secure cloud storage," International Journal of Information Security, 1–12, 2014, doi:10.1007/s10207-014-0263-8.

[18] C. Erway, A. Küpçü, C. Papamanthou, R. Tamassia, "Dynamic Provable Data Possession," in Proceedings of the 16th ACM Conference on Computer and Communications Security, CCS '09, 213–222, ACM, New York, NY, USA, 2009, doi:10.1145/1653662.1653688.

[19] Y. Zhu, H. Wang, Z. Hu, G.-J. Ahn, H. Hu, S. S. Yau, "Dynamic Audit Services for Integrity Verification of Outsourced Storages in Clouds," in Proceedings of the 2011 ACM Symposium on Applied Computing, SAC '11, 1550–1557, ACM, New York, NY, USA, 2011, doi:10.1145/1982185.1982514.

[20] Y. Zhu, G.-J. Ahn, H. Hu, S. S. Yau, H. G. An, C.-J. Hu, "Dynamic Audit Services for Outsourced Storages in Clouds," IEEE Transactions on Services Computing, **6**(2), 227–238, 2013, doi:http://doi.ieeecomputersociety.org/10.1109/TSC.2011.51.

[21] C. Wang, Q. Wang, K. Ren, W. Lou, "Ensuring data storage security in cloud computing," in in Proc. of IWQoS'09, 2009.

[22] A. Le, A. Markopoulou, "NC-Audit: Auditing for Network Coding Storage," CoRR, **abs/1203.1730**, 2012.

[23] C. Wang, Q. Wang, K. Ren, N. Cao, W. Lou, "Toward Secure and Dependable Storage Services in Cloud Computing," IEEE Trans. Serv. Comput., **5**(2), 220–232, 2012, doi:10.1109/TSC.2011.24.

[24] B. Wang, B. Li, H. Li, "Oruta: privacy-preserving public auditing for shared data in the cloud," IEEE Transactions on Cloud Computing, **2**(1), 43–56, 2012, doi:http://doi.ieeecomputersociety.org/10.1109/TCC.2014.2299807.

[25] B. Wang, B. Li, H. Li, "Knox: Privacy-preserving Auditing for Shared Data with Large Groups in the Cloud," in Proceedings of the 10th International Conference on Applied Cryptography and Network Security, ACNS'12, 507–525, Springer-Verlag, Berlin, Heidelberg, 2012, doi:10.1007/978-3-642-31284-7_30.

[26] C. Wang, S. S. Chow, Q. Wang, K. Ren, W. Lou, "Privacy-Preserving Public Auditing for Secure Cloud Storage," IEEE Transactions on Computers, **62**(2), 362–375, 2013, doi:http://doi.ieeecomputersociety.org/10.1109/TC.2011.245.

[27] X. Fan, G. Yang, Y. Mu, Y. Yu, "On Indistinguishability in Remote Data Integrity Checking," **58**(4), 823–830, 2015, doi:http://dx.doi.org/10.1093/comjnl/bxt137.

[28] B. Wang, B. Li, H. Li, "Panda: Public Auditing for Shared Data with Efficient User Revocation in the Cloud," IEEE T. Services Computing, **8**(1), 92–106, 2015, doi:10.1109/TSC.2013.2295611.

[29] C. Gritti, W. Susilo, T. Plantard, "Efficient Dynamic Provable Data Possession with Public Verifiability and Data Privacy," in E. Foo, D. Stebila, editors, Information Security and Privacy, 395–412, Springer International Publishing, Cham, 2015.

[30] C. Gritti, R. Chen, W. Susilo, T. Plantard, "Dynamic Provable Data Possession Protocols with Public Verifiability and Data Privacy," in J. K. Liu, P. Samarati, editors, Information Security Practice and Experience, 485–505, Springer International Publishing, Cham, 2017.

[31] J. Xu, E.-C. Chang, "Towards Efficient Proofs of Retrievability," in Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security, ASIACCS '12, 79–80, Association for Computing Machinery, New York, NY, USA, 2012, doi:10.1145/2414456.2414503.

[32] K. D. Bowers, A. Juels, A. Oprea, "Proofs of Retrievability: Theory and Implementation," in Proceedings of the 2009 ACM Workshop on Cloud Computing Security, CCSW '09, 43–54, ACM, New York, NY, USA, 2009, doi:10.1145/1655008.1655015.

[33] K. D. Bowers, A. Juels, A. Oprea, "HAIL: A High-availability and Integrity Layer for Cloud Storage," in Proceedings of the 16th ACM Conference on Computer and Communications Security, CCS '09, 187–198, ACM, New York, NY, USA, 2009, doi:10.1145/1653662.1653686.

[34] Y. Dodis, S. Vadhan, D. Wichs, "Proofs of Retrievability via Hardness Amplification," in Proceedings of the 6th Theory of Cryptography Conference on Theory of Cryptography, TCC '09, 109–127, Springer-Verlag, Berlin, Heidelberg, 2009, doi:10.1007/978-3-642-00457-5_8.

[35] Q. Wang, C. Wang, J. Li, K. Ren, W. Lou, "Enabling Public Verifiability and Data Dynamics for Storage Security in Cloud Computing," in Proceedings of the 14th European Conference on Research in Computer Security, ESORICS'09, 355–370, Springer-Verlag, Berlin, Heidelberg, 2009.

[36] E. Shi, E. Stefanov, C. Papamanthou, "Practical Dynamic Proofs of Retrievability," in Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security, CCS '13, 325–336, ACM, New York, NY, USA, 2013, doi:10.1145/2508859.2516669.

[37] B. Chen, R. Curtmola, "Towards Self-Repairing Replication-Based Storage Systems Using Untrusted Clouds," in Proceedings of the Third ACM Conference on Data and Application Security and Privacy, CODASPY '13, 377–388, Association for Computing Machinery, New York, NY, USA, 2013, doi:10.1145/2435349.2435402.

[38] B. Chen, R. Curtmola, "Remote data integrity checking with server-side repair," Journal of Computer Security, **25**(6), 537–584, 2017, doi:10.3233/JCS-16868.

[39] I. Leontiadis, R. Curtmola, "Secure Storage with Replication and Transparent Deduplication," in Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy, CODASPY '18, 13–23, Association for Computing Machinery, New York, NY, USA, 2018, doi:10.1145/3176258.3176315.

[40] M. Etemad, A. Küpçü, "Transparent, Distributed, and Replicated Dynamic Provable Data Possession," in M. J. J. Jr., M. E. Locasto, P. Mohassel, R. Safavi-Naini, editors, Applied Cryptography and Network Security - 11th International Conference, ACNS 2013, Banff, AB, Canada, June 25-28, 2013. Proceedings, volume 7954 of *Lecture Notes in Computer Science*, 1–18, Springer, 2013, doi:10.1007/978-3-642-38980-1_1.

[41] B. Chen, A. K. Ammula, R. Curtmola, "Towards Server-Side Repair for Erasure Coding-Based Distributed Storage Systems," in Proceedings of the 5th ACM Conference on Data and Application Security and Privacy, CODASPY '15, 281–288, Association for Computing Machinery, New York, NY, USA, 2015, doi:10.1145/2699026.2699122.

[42] B. Chen, R. Curtmola, G. Ateniese, R. Burns, "Remote Data Checking for Network Coding-Based Distributed Storage Systems," in Proceedings of the 2010 ACM Workshop on Cloud Computing Security Workshop, CCSW '10, 31–42, Association for Computing Machinery, New York, NY, USA, 2010, doi:10.1145/1866835.1866842.

[43] T. P. Thao, K. Omote, "ELAR: Extremely Lightweight Auditing and Repairing for Cloud Security," in Proceedings of the 32nd Annual Conference on Computer Security Applications, ACSAC '16, 40–51, Association for Computing Machinery, New York, NY, USA, 2016, doi:10.1145/2991079.2991082.

[44] K. D. Bowers, M. van Dijk, A. Juels, A. Oprea, R. L. Rivest, "How to tell if your cloud files are vulnerable to drive crashes," in Y. Chen, G. Danezis, V. Shmatikov, editors, Proceedings of the 18th ACM Conference on Computer and Communications Security, CCS 2011, Chicago, Illinois, USA, October 17-21, 2011, 501–514, ACM, 2011, doi:10.1145/2046707.2046766.

[45] D. Vasilopoulos, K. Elkhiyaoui, R. Molva, M. Onen, "POROS: Proof of Data Reliability for Outsourced Storage," in Proceedings of the 6th International Workshop on Security in Cloud Computing, SCC '18, 27–37, Association for Computing Machinery, New York, NY, USA, 2018, doi: 10.1145/3201595.3201600.

[46] D. Vasilopoulos., M. Önen., R. Molva., "PORTOS: Proof of Data Reliability for Real-World Distributed Outsourced Storage," in Proceedings of the 16th International Joint Conference on e-Business and Telecommunications - Volume 2: SECRYPT,, 173–186, INSTICC, SciTePress, 2019, doi: 10.5220/0007927301730186.

[47] Z. N. J. Peterson, M. Gondree, R. Beverly, "A Position Paper on Data Sovereignty: The Importance of Geolocating Data in the Cloud," in Proceedings of the 3rd USENIX Conference on Hot Topics in Cloud Computing, HotCloud'11, 9, USENIX Association, USA, 2011.

[48] G. J. Watson, R. Safavi-Naini, M. Alimomeni, M. E. Locasto, S. Narayan, "LoSt: Location Based Storage," in Proceedings of the 2012 ACM Workshop on Cloud Computing Security Workshop, CCSW '12, 59–70, Association for Computing Machinery, New York, NY, USA, 2012, doi:10.1145/2381913.2381926.

[49] S. Dziembowski, S. Faust, V. Kolmogorov, K. Pietrzak, "Proofs of Space," in R. Gennaro, M. Robshaw, editors, Advances in Cryptology – CRYPTO 2015, 585–605, Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.

[50] R. L. Rivest, A. Shamir, D. A. Wagner, "Time-Lock Puzzles and Timed-Release Crypto," Technical report, USA, 1996.

[51] J. . Cai, R. J. Lipton, R. Sedgewick, A. C. . Yao, "Towards uncheatable benchmarks," in [1993] Proceedings of the Eigth Annual Structure in Complexity Theory Conference, 2–11, 1993.

[52] D. Boneh, M. Naor, "Timed Commitments," in M. Bellare, editor, Advances in Cryptology — CRYPTO 2000, 236–254, Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.

[53] C. Dwork, M. Naor, "Pricing via Processing or Combatting Junk Mail," in Proceedings of the 12th Annual International Cryptology Conference on Advances in Cryptology, CRYPTO '92, 139–147, Springer-Verlag, Berlin, Heidelberg, 1992.

[54] A. K. Lenstra, B. Wesolowski, "A random zoo: sloth, unicorn, and trx," Cryptology ePrint Archive, Report 2015/366, 2015, https://eprint.iacr.org/2015/366.

[55] B. Cohen, K. Pietrzak, "Simple Proofs of Sequential Work," Cryptology ePrint Archive, Report 2018/183, 2018, https://eprint.iacr.org/2018/183.

[56] NetEm, "NetEm, the Linux Foundation," http://www.linuxfoundation.org/collaborate/workgroups/networking/netem, 2021, accessed on 11/04/2021.

[57] T. Flach, N. Dukkipati, A. Terzis, B. Raghavan, N. Cardwell, Y. Cheng, A. Jain, S. Hao, E. Katz-Bassett, R. Govindan, "Reducing Web Latency: The Virtue of Gentle Aggression," in Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM, SIGCOMM '13, 159–170, Association for Computing Machinery, New York, NY, USA, 2013, doi:10.1145/2486001.2486014.

[58] M. Dahlin, B. B. V. Chandra, L. Gao, A. Nayate, "End-to-End WAN Service Availability," IEEE/ACM Trans. Netw., **11**(2), 300–313, 2003, doi: 10.1109/TNET.2003.810312.

[59] S. Sundaresan, W. de Donato, N. Feamster, R. Teixeira, S. Crawford, A. Pescapè, "Broadband Internet Performance: A View from the Gateway," SIGCOMM Comput. Commun. Rev., **41**(4), 134–145, 2011, doi: 10.1145/2043164.2018452.

[60] D. Dobre, G. Karame, W. Li, M. Majuntke, N. Suri, M. Vukolić, "PoWerStore: Proofs of Writing for Efficient and Robust Storage," in Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security, CCS '13, 285–298, Association for Computing Machinery, New York, NY, USA, 2013, doi:10.1145/2508859.2516750.

[61] D. Zhuo, M. Ghobadi, R. Mahajan, K.-T. Förster, A. Krishnamurthy, T. Anderson, "Understanding and Mitigating Packet Corruption in Data Center Networks," SIGCOMM '17, 362–375, Association for Computing Machinery, New York, NY, USA, 2017, doi:10.1145/3098822.3098849.

[62] V. Paxson, "End-to-End Internet Packet Dynamics," SIGCOMM Comput. Commun. Rev., **27**(4), 139–152, 1997, doi:10.1145/263109.263155.

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 2, 133-142 (2022)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

An Interdisciplinary Approach to Fracture of Solids from the Standpoint of Condensed Matter Physics

Mark Petrov*

Department of Strength and Durability of Materials and Structural Components, Aeronautical Research Institute named after S. A. Chaplygin, Novosibirsk, 630051, Russia

ARTICLE INFO

Article history:
Received: 27 January, 2022
Accepted: 14 March, 2022
Online: 12 April, 2022

Keywords:
Strength
Creep
Fatigue
Inelasticity
Rheology
Damages

ABSTRACT

Instead of approaches of solid mechanics or a formal description of experimental data an interdisciplinary approach is proposed to consider failure and deformation as thermodynamic processes. Mathematical modeling of the processes is carried out using rheological models of the material. One fracture criterion is used, that formally corresponds to the achievement of a threshold concentration of micro-damage in any volume of the material. The prediction of the durability of materials under constant or variable temperature and force conditions is performed by time steps, including situations with changes in the material structure. Calculations of durability of structural components are based on the relationship of plastic flow and failure processes distributed over the volume of the material.

1. Introduction

In our article we have shown the possibilities and necessity of applying an interdisciplinary approach to solving the problem of flow and fracture of materials [1]. Kauzmann was one of the first scientists who applied the theory of reaction rates to the yielding of solids, examining creep as a process of directional diffusion under the effect of applied stresses [2]. Assuming that the applied stress reduces the energy barrier in one direction and increases the barrier to approximately the same extent in the opposite direction, he derived an equation for the excess number of transfer acts per unit time in the direction of applied stresses:

$$n_* = 2A\exp\left(-\frac{H}{kT}\right)\sinh\left(\frac{\Delta H}{kT}\right), \qquad (1)$$

where H is the initial height of the energy barrier, ΔH is the variation of this height under the stress effect, k is the Boltzmann constant, T is the absolute temperature, and A is the reaction constant. At high values of ΔH, the reverse flow through the barrier is usually ignored, and Eq. (1) takes the form

$$n_* = A\exp\left(-\frac{H-\Delta H}{kT}\right). \qquad (2)$$

Kauzmann assumed that ΔH depends in a linear manner on stresses, and (2) was confirmed by some experiments.

If some process in the material is caused by thermal activation, the dependence of the process rate on the stress σ and temperature is described by the Arrhenius equation in which the pre-exponential factor depends in the general case on the stress and temperature. Specific types of the expressions $V_0(\sigma, T)$ and activation energy $U(\sigma)$ are determined by the range of temperature–force conditions of loading. Each such region is characterized by the dominance of some deformation mechanism or mass transfer mechanism. The physical interpretation of this equation is based on the theory of overcoming potential barriers. The exponent in (3) is interpreted as the probability of the transition through the barrier or as the fraction of atoms that are in the activated state at each time instant [3].

$$V = V_0(\sigma, T)\exp\left[-\frac{U(\sigma)}{kT}\right], \qquad (3)$$

Based on this concept, the analysis of strength and deformation characteristics of any material should be started from the analysis of results of simple experiments on fracture at constant stress and temperature to identify the basic features of these processes corresponding, for example, to the form (2). Tests performed under monotonic loading provide additional data, which may ensure a more accurate description of the material behavior [4, 5]. In this case, the main research method is the thermally activation analysis.

* Corresponding Author: Mark Petrov1, post box 166, 630089, Novosibirsk, Russia, markp@risp.ru

With cyclic loading with small amplitudes leading to fatigue failure, thermal activation analysis cannot be performed. The process of failure is localized and distributed in the volume of the material according to the field of internal stresses. The problem can be solved only by mathematical modeling of local processes of fracture, based on the same patterns of fracture that were revealed during the study of the creep of the material.

There are objective reasons for the lack of reliable methods of durability calculations. Because of the large variety of operation conditions, investigations were separated into individual fields, and the bearing capacity of particular structural elements was studied only for particular conditions of their operation. Internal processes in the material under fracture are rather complicated and versatile; the lack of information about their relationship with macroscopic properties of solids gave rise to many approaches both to understanding the fracture phenomenon and to developing engineering methods of estimating the bearing capacity of various structures.

Despite comprehensive investigations, there is no unified concept, which would allow successful evaluation of strength and durability of structures under hostile conditions of their operation. There are many publications dealing with physical and metal science aspects of strength and durability. These studies assist in understanding what happens in the material and explain experimentally observed specific features of the material behavior. However, such studies are not directly related to calculations of strength and durability in practice. There are many approaches and methods for determining the bearing capacity of structures depending on the loading character and temperature, though each of these approaches and methods is applicable only in a limited range of operation conditions. If the range of operation conditions is extended, there arises a problem of matching these approaches. The problem is difficult because the basis of the problem solution, i.e., the material itself, is ignored. It is sufficient to say that even different units of durability measurement are different for different loading types: these may be the time, or the number of cycles, or even the sum of loads. It is necessary to revise the traditional methods used for estimating strength and durability of structures from the viewpoint of physics of material properties.

The solution of the problem of assessing the durability of materials in structures under arbitrary thermal-force loading is impossible without constructing new models of continuous media on the basis of physics and thermodynamics of internal processes that occur in loaded solids. It is only an adequate presentation of a solid as a physical medium that offers a possibility of considering the entire multitude of interrelated processes of deformation and fracture, structural transformations, and physical and chemical effects. The analysis of experimental data from this viewpoint leads to qualitatively new ideas of material properties and allows determining the optimal volume of the experiment and the sequence of obtaining the characteristics of new alloys, thus, reducing the cost and time of structural design.

2. Basic laws of failure and deformation of materials

Examination of the kinetics of fracture of polymers, pure metals, alloys, and other materials showed that the following dependence of durability as the inverse of the average rate of failure $\dot{\omega}$ on the temperature and stress is satisfied in many cases (for a mole of a substance by replacing the Boltzmann constant k with the universal gas constant R):

$$\tau = \tau_0 \exp\left(\frac{U_0 - \gamma\sigma}{RT}\right), \tag{4}$$

or in the general case when temperature, stress and parameter γ depend on time t,

$$\dot{\omega} = \nu_0 \exp\left(-\frac{U_0 - \gamma(t,\sigma,T)\sigma(t)}{RT(t)}\right), \tag{5}$$

where U_0 is initial activation energy of fracture, γ is the structure-sensitive coefficient (activation volume), $\nu_0 = 1/\tau_0$ – characteristic Debye frequency [6, 7]. The expression for the plastic strain rate at a constant stress (steady creep stage) obtained in the same experiments has a similar form

$$\dot{\varepsilon}_p = \dot{\varepsilon}_0 \exp\left(-\frac{Q_0 - \alpha\sigma}{RT}\right), \tag{6}$$

indicating a close relationship of the fracture processes with the processes of plastic deformation. A comparison of the parameters of (4), (5) and (6) for many materials in fact shows the equality (within the limits of the error of experimental data processing) of U_0 and Q_0, γ and α, and the product $\tau_0\dot{\varepsilon}_0$ is equal to the residual strain ε_* accumulated at the steady creep stage [6]. The residual strain changes only slightly (approximately by an order of magnitude) with a large change in the duration of fracture (9–10 orders) [7]. The values of the pre-exponential factors in (4) and (6) determined in processing of experimental data for different materials were in the range 10^{-11}–10^{-14} s for τ_0 and 10^{12}–10^{13} s^{-1} for $\dot{\varepsilon}_0$.

In many cases, the activation characteristics of atomic rearrangement processes are reflected in the macroscopic characteristics of the solid under loading. Therefore, natural attempts have been made to examine the mechanism of these processes by means of the thermal activation analysis. For this purpose, in accordance with (2), (3), (4) or (5), we plot the logarithm of the process rate on the stress at different temperatures and on the reciprocal value of temperature and different stresses.

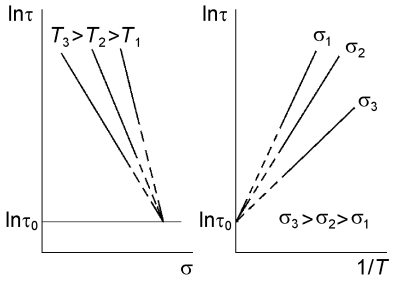


Figure 1: Temperature–force dependences of durability for determining the activation parameters of the fracture process

If, for example, (4) is valid and its parameters are constant, then we obtain a series of straight lines in the corresponding coordinates, with the lines converging in a band (Fig. 1 and 2). At the same time, it is evident that the process itself may lead to changes in the state

of the medium in which it takes place. This results in changes of the parameters and in their dependence on both the external conditions (σ, *T*) and the stage of the process, i.e., internal conditions. For materials, these are structural changes, being the result of the combined effect of different atomic mechanisms in different configurations at each scale level.

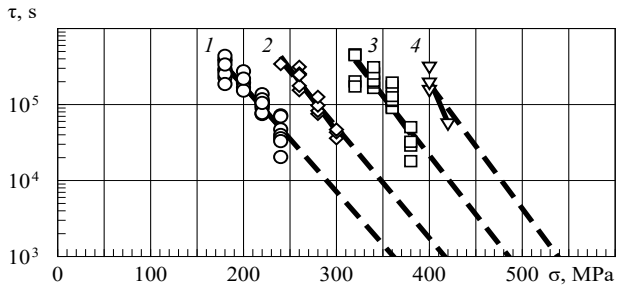


Figure 2: Temperature–force dependences of durability for determining the activation parameters of the fracture process

In the article [1], the force dependences of the activation energy of fracture and deformation for alloy 1201 T1 were shown in Figure 1 (Al-Cu-Mn system). The same refers to duralumin (the durability of its specimens is shown in Figure 2, Al-Cu-Mg system). The force dependence of the activation energy of fracture obtained in this experiment is shown in Figure 3. In the range of stresses and temperatures in which they were tested, the coefficient γ also has a constant (minimum) value, and the specimens demonstrate structurally stable "long-term strength," which can be recalculated from one temperature-force mode to another. Similar dependences were previously given earlier for this and other aluminum alloys, including quasi-stable states at $\gamma = \gamma_{max}$ [1, 4, 5, 8, 9]. All this can be seen only through the thermal activation analysis, taking into account, among other things, the quantum effects of low-temperature fracture of materials [9].

For mechanical engineers who are used to terms "strength" or "long-term strength," we can offer a more stringent strength characteristic – the strength parameter. Let us define it as $P_b = 1/\gamma$. Then the above-mentioned specimens in a certain temperature-time interval, regardless of the loading speed and temperature, will have $P_b = const$, i.e., exactly the same "strength" determined only by the material structure (activation volume γ). The dimension of P_b is MPa·mol/kJ, and it is also independent of the loading trajectory if the loading rate is variable. Thus, the comparison of the "strength" tests results is justified.

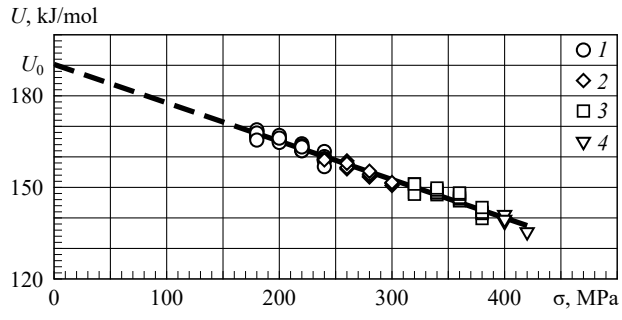


Figure 3: Force dependences of the activation energy of fracture of 80 duralumin specimens (Fig. 2) in the range of stresses of 180÷420 MPa and temperatures of 398÷473 K; temperature, K: *1* – 473, *2* – 448, *3* – 423, *4* – 398

For mechanical engineers who are used to terms "strength" or "long-term strength," we can offer a more stringent strength characteristic – the strength parameter. Let us define it as $P_b = 1/\gamma$. Then the above-mentioned specimens in a certain temperature-time interval, regardless of the loading speed and temperature, will have $P_b = const$, i.e., exactly the same "strength" determined only by the material structure (activation volume γ). The dimension of P_b is MPa·mol/kJ, and it is also independent of the loading trajectory if the loading rate is variable. Thus, the comparison of the "strength" tests results is justified.

An extensive experiment analyzed by the methods described here revealed the influence of creep of the binder on the strength properties of fiberglass plastic [10]. Figure 4 shows the force dependences of the activation energy of fiberglass plastic fracture under longitudinal bending (a) and tensile loading (b). For longitudinal bending, the mean values for 20 or 40 specimens tested in each mode are shown. The diamonds denote modes of monotonic loading with different rates, and the circles show loading by a constant bending moment. For tensile loading, the data for each specimen are provided. The stress scatter corresponds to monotonic loading, and the scatter of the activation energy corresponds to a constant load.

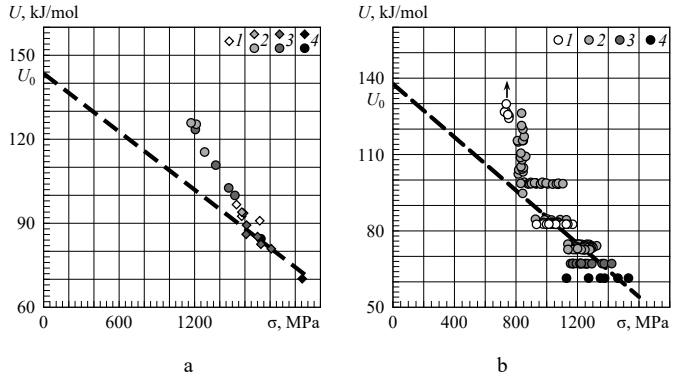


Figure 4: Force dependences of the activation energy of fiberglass fracture (rods with a diameter of 5.5 mm) under longitudinal bending (a) and tensile loading (b) [10]; temperature, *T*, °C:
a) *1* – +60, *2* – +50, *3* – +20, *4* – −30; b) *1* – +50, *2* – +9–20, *3* – −8, *4* – −30

The lines drawn through those values of $U(\sigma)$ that satisfy the equation of a straight line with the minimum slope showed approximately the same values of the initial activation energy U_0. The deviations from these lines illustrate the role of binder creep in the distribution of forces over the fibers in the composite, which is equivalent to changing the material structure. Thus, the straight lines correspond apparently to some stable state of the material structure, when the initial stage of creep has already ended. With a decrease in stresses and an increase in the duration of the failure process, the experimental values of $U(\sigma)$ deviate from these lines in the direction of increasing durability. Rapid loading or temperature reduction leads to more uneven loading of the fibers, and they begin to break down sequentially. The rod failure process is finalized even at lower loads [10] in contrast to metals in which the rupture stresses of the specimens increase in proportion to the growth of the logarithm of the loading rate (or to the decrease in the logarithm of the fracture time).

Under monotonous loading, the values of $U(\sigma)$ are calculated by the equivalent failure time τ_{eq} in accordance with (5) reduction

to maximum stresses, based on the same damage in accordance with the Bailey criterion [11], according to the formula

$$\tau_{eq}\nu_0 \exp\left(-\frac{U_0-\gamma\sigma_{\max}}{RT}\right)=\int_0^{t_*}\dot{\omega}dt=1, \qquad (7)$$

where t_* is the loading time along a particular trajectory to a stress σ_*, often less $\sigma_{\max}$, at which specimen fracture occurs. The error in determining the activation volume γ turns out to be small, since the entire process of failure is short-lived and concentrated in the range of action of high stresses.

The change in the activation volume γ is associated with a change in the structure of the material and each material has its own reasons and characteristics. These can be, for example, relaxation processes of internal stresses, diffusion of alloying elements in the alloy matrix or creep of the binder in the composite material. And each of them requires separate study and modeling.

3. Mathematical modeling of the rheological properties of the material

In accordance with the patterns of fracture and deformation of materials (1) and (2) observed in the experiment, new bodies were introduced into rheology, called Zhurkov (Zh) and Kauzmann (Km) bodies [12]. Denoting $A=\varepsilon_*\nu_0\exp(-Q_0/RT)$ and $B=\alpha/RT$ in (6), we obtain the rheological equations of the Zh and the Km solid:

$$\dot{\varepsilon}_p = A\exp(B\sigma) \text{ and } \dot{\varepsilon}_p = 2A\,\mathrm{sh}(B\sigma). \qquad (8)$$

The sequential and parallel connections of these bodies with the Hooke body (H) having an elastic modulus M form bodies similar to the Maxwell and Kelvin (Voigt) bodies, which describe the general and local plastic flow in materials. They are indicated by the symbols PM_1 and PM_2 (with Zh bodies) or PM_5 and PM_6 (with Km bodies) [1, 12]. A set of such elements is a structural model of the material that shows in Figure 5. The difference from the similar mechanical structural model of the material based on the Saint-Venant body [13] is the replacement of dry friction elements with elements that describe plastic flow kinetics (8). An element of the general plastic flow of the material has also been added.

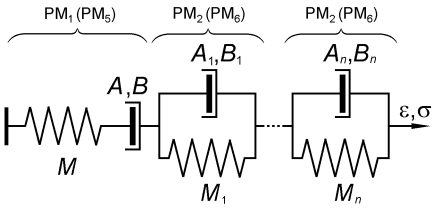


Figure 5: Structural model of a material describing the elasticity, creep, and hysteresis-type inelasticity by a set of elements with a parallel connection of an elastic Hooke's body and a plastic flow body (Zhurkov's or Kauzmann's body) [1, 4]

The plastic flow and the process of material failure are both occurring and interrelated [14]. Therefore, describing the rheological properties of materials, it is possible to associate them with the accumulation of damage and to assess the durability of structural elements under arbitrary external effects [5].

The rheological equation, for example, of the PM_1 solid (Fig. 5) as the equality of the total strain rate to the sum of the elastic strain rate of the H solid and the plastic strain of the Zh solid (8) has the following form [1, 12]:

$$\dot{\varepsilon}=\frac{\dot{\sigma}}{M}+A\exp(B\sigma). \qquad (9)$$

This equation has the following solutions [12]:

- at a constant strain ($\dot{\varepsilon}=0$), we obtain the equation of stress relaxation

$$\sigma=-\frac{1}{B}\ln[\exp(-B\sigma_0)+ABMt]; \qquad (10)$$

- at a constant strain rate ($\dot{\varepsilon}=C$), substituting the integration constant in (10) by the function and deriving the linear equation, we obtain the expression

$$\sigma=-\frac{1}{B}\ln\left\{\exp[-B(\sigma_0+MCt)]+\frac{A}{C}[1-\exp(-BMCt)]\right\} \qquad (11)$$

As the time progresses ($t\to\infty$), this expression yields the flow stress (yield stress)

$$\sigma=-\frac{1}{B}\ln\left(\frac{A}{C}\right), \qquad (12)$$

which depends on the strain rate and temperature;

- at loading with a constant rate $\dot{\sigma}=D$, we obtain the dependence of strain on time

$$\varepsilon=\varepsilon_0+\frac{Dt}{M}+A\exp(B\sigma_0)\frac{\exp(BDt)-1}{BD}. \qquad (13)$$

At $D=0$, this dependence transforms to the steady creep equation

$$\varepsilon=\varepsilon_0+A\exp(B\sigma_0)t$$

Here σ_0 and ε_0 are the stress and strain at the time instant $t=0$.

Solutions of (9) for a constant strain rate C (11) or for a constant loading rate D (13) give two different relationships between stresses and strains. As a result, we get several "theories of plasticity" [5]. If the material structure changes during plastic flow, the parameters of (9) should be replaced with functions describing the transition of the material from one state to another. This can be done by analyzing the experimental deformation curves by time steps [1, 12]. No new "theories of plasticity" are required. And stress relaxation according to solution (10), in which the material also is also fractured, does not require any energy expenditure. Everything happens due to the internal energy of a solid, the measure of which is temperature.

For the PM_2 solid (Fig. 5), on the basis of solving the equilibrium and strain compatibility equations, we can write the following rheological equation:

$$\dot{\varepsilon}\exp(BM\varepsilon) = A\exp(B\sigma)\,. \qquad (14)$$

Integration of this equation for $\sigma = \sigma_0 + Dt$ yields the solution

$$\varepsilon = \frac{1}{M}\left[\sigma_0 + Dt + \frac{1}{B}\ln\left\{\begin{matrix}\exp[-B(\sigma_0 - M\varepsilon_0 + Dt)] \\ +\dfrac{AM}{D}[1-\exp(-BDt)]\end{matrix}\right\}\right]. \qquad (15)$$

For the loading rate $D = 0$, we obtain

$$\varepsilon = \frac{1}{M}\left[\sigma_0 + \frac{1}{B}\ln\left\{\exp[-B(\sigma_0 - M\varepsilon_0)] + ABMt\right\}\right]. \qquad (16)$$

When stress relaxation (10) or strain relaxation (16) occurs, and the stresses reach a small value, one should use similar solutions of rheological equations with Kauzmann bodies (at $\sigma \to 0$ the rate of plastic strain $\dot{\varepsilon}_p \to 0$) [12]. This should be especially borne in mind at elevated temperatures.

For a PM_5 body with constant strain, instead of (10) we have

$$\sigma = \frac{2}{B}\operatorname{artanh}\left[\tanh\left(\frac{B\sigma_0}{2}\right)\exp(-2ABMt)\right], \qquad (17)$$

or, for the calculation procedures [15],

$$\sigma = \frac{1}{B}\ln\frac{1+X}{1-X};\ \left(X = \frac{\exp(B\sigma_0)-1}{\exp(B\sigma_0)+1}\exp(-2ABMt)\right).$$

For the PM_6 body at constant stresses, instead of (16) we obtain

$$\varepsilon = \frac{1}{M}\left[\sigma_0 - \frac{2}{B}\operatorname{artanh}\left\{\begin{matrix}\tanh\left[\dfrac{B(\sigma_0 - M\varepsilon_0)}{2}\right] \\ \times\exp(-2ABMt)\end{matrix}\right\}\right], \qquad (18)$$

Solutions (10), (16) and (17), (18) provide completely identical results in the range of high stresses. Therefore, when using the Km solid in the models, it is more efficient to use the solutions for models with the Zh solid in appropriate sections of the loading program, because this is a simpler procedure. This also refers to the algorithms of processing the experimental data for determining the parameters of the rheological models. Figure 6 shows the stresses in the PM_1 and PM_5 solids with their rapid deformation to the establishment of constant flow stresses and subsequent curing at a fixed strain. Calculations were carried out for duralumin: Q_0 = 192.6 kJ/mol, α = 0.142 kJ/(mol·MPa).

If the material is characterized by the same behaviour in tensile and compressive loading, it is only necessary to change the signs of the stresses, strains, and their rates to the opposite signs when passing to the compression region. Otherwise, the parameters A and B should also differ. The algorithm of calculations in the transition through zero should be also accurate. The time step in unloading should be selected in such a manner that the stresses in the flow elements should approach zero prior to "reversing" of the equations. Otherwise, the strains would be determined with errors.

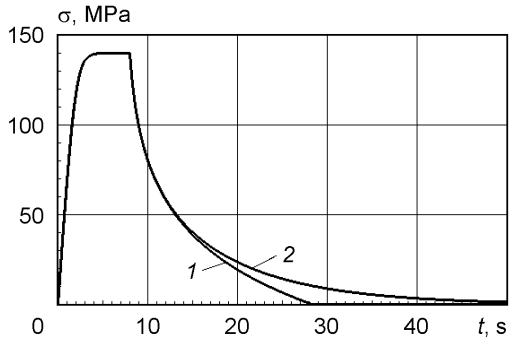


Figure 6: Deformation of the PM_1 (*1*) and PM_5 (*2*) solids with a constant strain rate to the "yield limit" and subsequent stress relaxation at a constant strain (the parameters of the rheological solids were taken for the D16 T material at 573 K)

When the stresses are greater than $-\ln(AM/D)/B$ in (15), there is something like a functional relationship between the stresses and strains in subsequent loading. In this case, we have "plasticity with hardening." If loading is terminated, we obtain equations of the so-called logarithmic creep [16], which were interpreted analytically in (16). In processing experimental data for an actual material, it is necessary to separate the plastic flow with actual hardening accompanied by changes in the material structure and by a decrease in the activation volume α from the local flow. The latter takes place in a set of local volumes; in each volume, it is characterized by its own activation parameters.

As a test problem, we study with a stress jump. Experimental results of such tests have long been known [16]: in the unsteady stage of creep, the initial stress increased or decreased by a jump, and after some time it returned to the same level.

Figure 7 gives results of calculations of the deformation process in the D16 T alloy performed using its model. The flow characteristics of the material predicted by the model are exactly the same as those of the real flow. No additional conditions apart from specification of the loading program and the temperature are required [12].

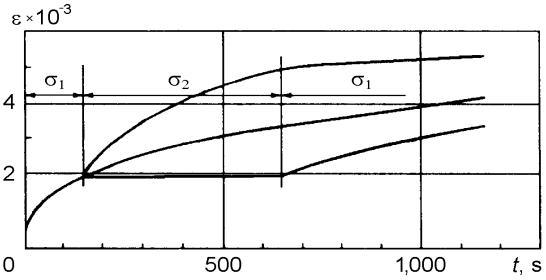


Figure 7: Creep with a stress jump: calculations using the model of the material D16 T (423 K); σ_1 = 300 MPa, σ_2 = 270 and 310 MPa

Other examples of the calculation of strains of specimens during loading and unloading are given in the article [5]. Experimental data, with which the results of calculations are compared, are contained in the article [17] or obtained by the author himself.

For such calculations based on the amplitude dependence of inelasticity, parametric identification of the structural model of the material (PM_2 or PM_6 bodies) is performed by dividing it into components that characterize each structural element. The typical amplitude dependence of the inelastic deformation of the material in the form of the width of the inelasticity loop is shown in Figure 8.

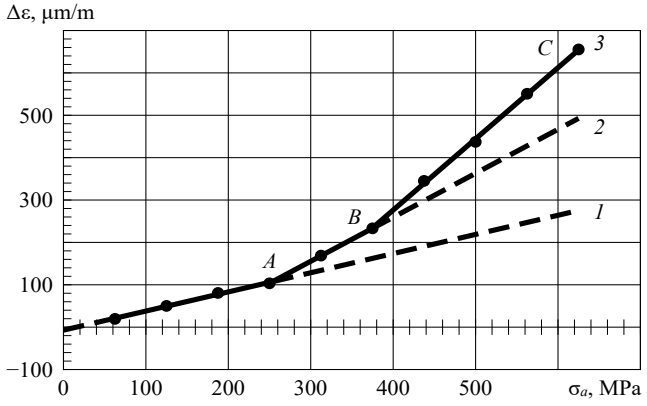


Figure 8: An example of a typical amplitude dependence of the inelasticity loop width of an eight-layer unidirectional carbon fiber reinforced plastic

The broken line *3* in the figure shows the value of the loop width: the maximum distance between the loading and unloading curve $\varepsilon = f(\sigma)$, calculated at a constant mean value of the cycle stresses. The data are taken from an experiment performed on unidirectional carbon fiber reinforced plastic. Up to point *A* (line *1*), there is always relaxation-type inelasticity in any material [18, 19]. As the loading amplitude increases, hysteresis-type inelasticity additionally appears (segment *AB* on line *2*). This is followed by a new increase in inelasticity (segment *BC*). Each loop width increment is ascribed to one structural element of the material model, which will determine its durability in the corresponding range of amplitudes.

The dependence of durability on mean cyclic stresses is taken into account in the rheological model of the material by changing the loop width through the change of the parameter $\dot{\varepsilon}_0$ in (6) for each structural element of the material model. For this purpose, in each amplitude range it is necessary to test with a different asymmetry index [4], and the endurance value *N* (the number of cycles passed during the specimen fracture) will be inversely proportional to the increment of the loop width in this range.

After parametric identification of the mathematical model carried out using experimental data for a specific frequency and temperature of tests, it is possible to proceed to calculations of the durability of the material under arbitrary changes in temperature and stress within the studied range of temperature-force dependences of the deformation activation energy and the fracture activation energy. When the material structure changes, the parameters *A* and *B* in (8) should be replaced by functions describing the accompanying thermally activated processes or the results of some other external effects leading to these changes.

An example of this is the fracture of duralumin at various combinations of temperature and stress. A precipitation aged alloy, which has reached the first maximum of hardness, undergoes an phase aging stage in the process of failure (intermetallic precipitation). Its hardness, having reached the second maximum, begins to decrease. This also happens in the absence of stresses, and the process accelerates under load, which affects the residual strain of the specimens. Figure 9 shows the dependence of the residual strain of duralumin specimens on the tensile test mode. The observed minimum of residual strain during the period of steady creep is associated with the achievement of the maximum hardness of the material.

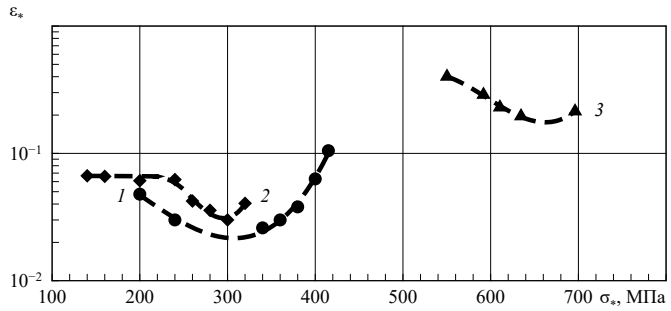


Figure 9: Dependence of the residual strain of duralumin specimens on the stresses and temperature of the tests (creep under constant and monotonically increasing loads); *T*, K: *1* – 423, *2* – 448, *3* – 293÷523; *1*, *2* – average values for several specimens, *3* – true residual strains in the neck of specimens under monotonic loading; σ_* – initial stress values at constant load (*1*, *2*) and the highest stress values at specimen rupture (*3*).

The change in hardness over time of alloys similar in composition was studied in [20–23] and others. The time to reach a certain state of the alloy in the process of transformation, which occurs due to thermal activation, is determined by a typical expression of the theory of the rates of processes [24]

$$\tau_p = \tau_{p0} \exp\left(\frac{Q_p}{RT}\right), \tag{19}$$

where Q_p is the activation energy of precipitate growth, the value of which is equal to the effective value of the activation energy of diffusion of alloying elements in the alloy matrix. The activation energy of diffusion, similar to what we observe during failure, depends approximately linearly on stresses, and the diffusion process is accelerated as a result of plastic deformation [18]. Then, taking into account the simultaneity of the process of precipitations in the centers, which additionally arise due to the accumulated plastic strain ε_{res}, expression (19) takes the form

$$\tau_p = \tau_{p0} \exp\left(\frac{Q_p - \beta\sigma}{RT} - m\varepsilon_{res}\right). \tag{20}$$

In the case of an additional increase in the number of precipitation centers due to a greater concentration of vacancies, the pre-exponential multiplier in (20) should include a multiplier expression that takes into account the temperature of the cooling medium during quenching [18]. At this temperature, the equilibrium concentration of vacancies has time to be established [20], and in this form this expression can participate in the calculations of the aging process at low temperatures and in the description of the recovery process. The characteristic maximum hardness of isolate-aged alloys and the corresponding minimum of

plasticity make it possible to estimate the activation parameters in the expression (20).

By equating the failure time (4) and the time to reach the minimum of the residual creep strain (20) obtained under certain loading conditions (Fig. 9), it is possible to estimate the parameters included in (20). Having the parameters of (20), we obtain an expression for the conditional rate of the precipitation process $V_p = 1/\tau_p$, the integral of which in time will give one when the second hardness maximum is reached. The next task is to link the aging rate with the strength and deformation properties of the material. This can be done through the hardness of the alloy, since indentation of the indenter is the same process of failure associated with large plastic strains [18].

In the absence of a load for triple alloys (2.5% Cu, 1.14% Mg and 3.0% Cu, 1.36% Mg), similar in composition to D16 T and AK4-1 T1 alloys of the same system, Hardy obtained a Q_p value of 32 and 33 kcal/mol, respectively [22]. This is approximately equal to the value of the activation energy of diffusion of copper in aluminum (32.6 kcal/mol [25], 1.4 ± 0.1 eV or 31.8 ± 2.3 kcal/mol [26]). After processing the reference data on changes in the deformation characteristics of AK4-1 T1 alloy specimens under monotonous loading after different aging modes, we obtained an estimate of $Q_p = 32.4$ kcal/mol and $\tau_{p0} = 10^{-10}$ s. That is, all estimates of the activation energy of the decay process of a solid solution turn out to be quite close.

After the introduction of the hardness function into the equation of the failure rate (5), with which the activation parameters of the fracture process are associated, and then calculations of the thermo-cyclic loading of structural elements were performed, taking into account the decay of a supersaturated metallic solid solution in these precipitation hardening alloys. Calculations of the fatigue failure process at low temperatures do not require taking into account such structural transformations, and it is quite acceptable to assume the structure of the material corresponding to the initial state of the alloy [18].

4. Examples of applying an interdisciplinary approach to practical tasks

Having the activation parameters U_0 and γ (which correspond to parameters A and B in Figure 5), it is possible to perform calculations for those loading conditions when the material flows throughout the entire volume, regardless of how the stresses and temperature change. The internal stresses in the so-called "fracture centers" naturally change, and this requires special modeling. Figure 10 shows the comparison of experimental data with the calculation for different temperature-time and temperature-force loading conditions of structural specimens made of AK-1 T1 alloy. Vertical lines correspond to the actual scatter of durability in the experiment, horizontal lines to the range of calculated estimates made taking into account the basic errors of the test program by load and temperature.

The figure shows that the calculated estimates of durability fall within a twofold range of deviations from the experimental data, which is usually observed when testing the same material of different batches. The calculations are made taking into account the decay of supersaturated metallic solid solution in a given alloy, aged to the second maximum hardness (T1 state), representing the parameter $\dot{\varepsilon}_0$ in (6) as the product of the residual strain by the frequency multiplier $\varepsilon_* \nu_0$ and relating it to the change in the hardness of the alloy. In all cases, fracture occurs as a result of creep, regardless of how the stresses in material or the dangerous places of structural components change [4, 18].

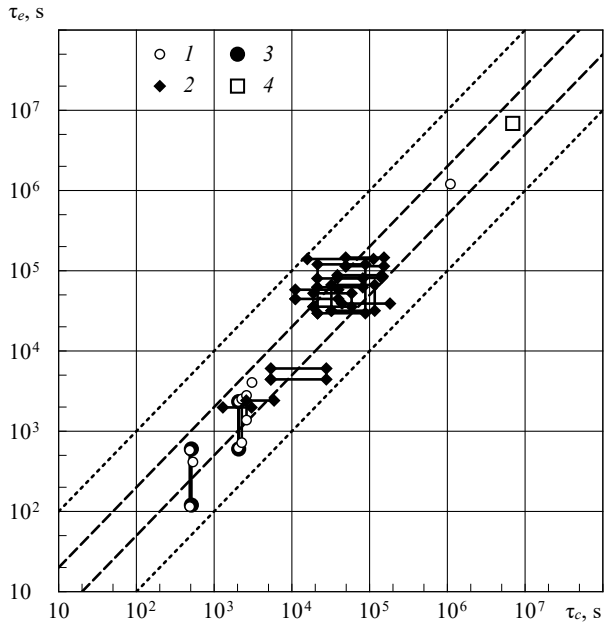


Figure 10: Comparison of calculated (τ_c) and experimental (τ_e) values of the durability of specimens and structural components made of AK4-1 T1 alloy tested at various loads and temperatures: *1*, *2* – strip with a central hole and a longitudinal stringer under thermo-cyclic loading [4]; *3* – rod: constant and variable stresses at 543 K at 10 Hz [18]; *4* – full-scale structure fracture, tested under the specified temperature–force program (with addition of crack propagation period)

The determination of the remaining parameters of the structural model of the material (Fig. 5) requires cyclic loading at a constant mean stress component of the cycle σ_m. The values of temperature, frequency and shape of the loading cycle must be set. Stepwise increasing the amplitude of loading, we obtain the amplitude dependence of inelasticity (Fig. 8), which is used to select amplitude values for fatigue tests according to characteristic points. That is, for example, for the *AB* and *BC* ranges, two amplitude values must be selected each. Then, these modes must be tested with two mean load components. For each amplitude range, it is sufficient to know for any one mode the inelastic strain in the loading cycle. After parametric identification of the model, it is possible to perform calculations at a different temperature, frequency, cycle shape and generally at arbitrary changes in them, if one assumes that no changes in the structure occur in the material. Otherwise, this requires a separate study, and the material model parameters must be replaced by functions that represent these changes.

Using the relationship between inelastic strains and damage accumulation, the mathematical model makes it possible to calculate the durability for various spectra of external effects, be it stress or temperature, representing their implementation by piecewise linear approximation. Having solutions of differential equations, for example, (9) and for other structural elements of the

model at constant stresses or strains and linearly varying, for example, (15) and (16), it is possible to calculate any arbitrary process of temperature-force loading [4].

In Figure 11 shows a comparison of the calculated estimates of durability of structural components with experimental data for various loading cases. All tests were carried out at a temperature of 293 ±2 K. Calculations were performed for a temperature of 293 K, assuming the structure of the material unchanged, corresponding to its initial state.

As in the previous example (Fig. 10), the calculated estimates of durability were made using a model of a design element that transforms in time the nominal stresses or loads into strains in the places of their concentration [4, 9]. For the specimens whose durability is marked by points *3*, *4* and *6*, the calculations were performed for two different quality batches of this material.

As in the previous example (Fig. 10), the calculated estimates of durability were made using a model of a design element that transforms in time the nominal stresses or loads into strains in the places of their concentration [4, 9]. For the specimens whose durability is marked by points *3*, *4* and *6*, the calculations were performed for two different quality batches of this material.

The time step of calculations at wide-band load spectrum is chosen minimum 0.25 or 0.5 period of the highest-frequency component of the spectrum. All load spectra were represented by equivalent polyharmonic pseudo-random processes (PRP) having the same spectral density, or by a real loading process recorded in operation [27]. The degree of discreteness of the spectrum depends on the material and type of the design element.

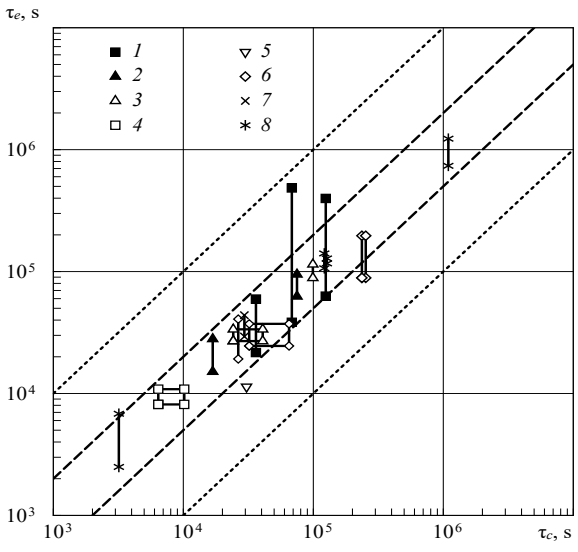


Figure 11: Comparison of calculated (τ_c) and experimental (τ_e) values of durability of structural specimens and structures made of 1201 T1 alloy tested under different loading programs (T = 293 ±2 K): *1* – plate-bar without notch, constant spectral density value in the interval 0.5÷10.5 Hz, 6 harmonics; *2* – plate-bar without notch, narrowband random noise in the interval 0÷5.5 Hz, 13 harmonics; *3* – plate-bar with notch, narrowband random noise in the interval 0÷5.5 Hz, 13 harmonics; *4* – plate-bar with notch, block 87-step program, triangular cycle shape at 10 Hz; *5* – acoustic tests of panels in the interval 0÷200 Hz; *6* – notched plate-bar, forced flight cycle 1200 s [23], compiled from records of bending moments on the wing of an airplane-laboratory; *7* – notched plate-bar, forced flight cycle GAG at 0.025 Hz; *8*–30 mm wide plate-bar with a central hole 20 mm, cyclic tests in the frequency range 0.1÷40 Hz with different cycle shape of loading

As in the case of variable temperatures (Fig. 10), the calculated estimates of durability are located in the range of twofold deviations from their experimental values. The calculations were performed based on the average statistical data of the durability of one of the semi-finished products of this material. To do this, two values of the mean cycle stresses are taken for each amplitude which selected by the inelastic characteristics of the material.

In each case, the structures are loaded in a different way; therefore, obtaining estimates requires statistical data on the typical operation conditions. Calculations are performed with averaged statistical data on loading, i.e., the averaged spectral density of the processes, which is then transformed to a discrete spectrum by the method of summation of elementary random functions [18]. As a result, one obtains a PRP, which is statistically equivalent to a real random process.

Examples of calculations for various PRPs, compared with the experiment, are given in the article [5]. The same real loading spectrum was modeled by a different number of harmonics distributed in several ways by frequency. This shows the significant effect of changes in the dispersion of the process in the high-frequency part of the spectrum on the durability of structural components.

To distinguish creep from fatigue, the units of measurement of durability must be uniform. Any unit of measurement always has a physical justification and a reference value [28]. The unit of measurement "cycle" does not exist in any system of units of measurement and cannot exist, since in each case it has a different content. Therefore, it is possible to distinguish cyclic creep from fatigue only if the durability is expressed in units of time, that is, the way the failure process actually occurs. In Figure 12 shows the dependences of durability on tensile stresses at their constant value and at cyclic tension with different frequencies and constant σ_{min} = 40 MPa. The abscissa shows the equivalent stresses σ_{eq} corresponding to the constants at which the durability has the same value in accordance with expression similar to (7), –

$$\int_0^{\tau} \dot{\omega}(t,\sigma)dt = \tau\nu_0 \exp\left(-\frac{U_0 - \gamma\sigma_{eq}}{RT}\right).$$

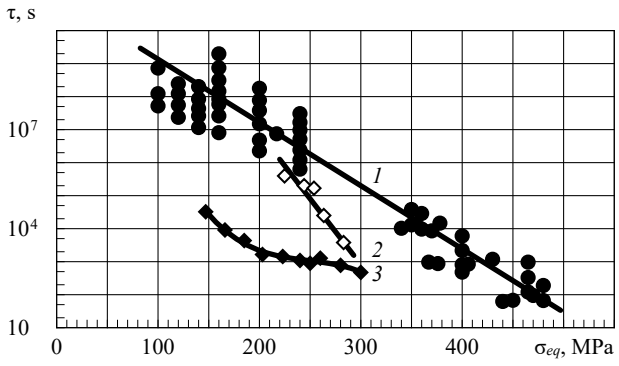


Figure 12: Durability of smooth specimens of AK4-1 T1 alloy tested at constant (*1*) and cyclically varying with a frequency of 0.05 (*2*) and 30 Hz (*3*) stresses (temperature 423 K)

The figure shows that at constant stresses, the logarithm of the durability linearly depends on the stresses, thereby illustrating the

main regularity of fracture (line *1*). With alternating stresses varying with a low frequency (straight line *2*), a decrease in the stress swing brings the value of the cyclic durability closer to the static one. It is clear that here we are dealing with cyclic creep, in which a decrease in durability occurs as a result of a concomitant relaxation of internal stresses, which decrease with a decrease in the loading rate [4, 6]. If the frequency is high (curve *3*), a decrease in the stress swing increases the discrepancy between the durability under static and cyclic loading, which first increases and then becomes smaller, approaching the static durability at $\sigma_a \to 0$. With an increase in the stress swing, fatigue failure will be replaced by fracture from cyclic creep, and curve *3* intersects with straight line *2*.

If the use of expression (5) is quite justified for a frequency of 0.05 Hz, then for fatigue failure at a frequency of 30 Hz, the approximating curve should be considered conditional (the lines in both cases are drawn according to the average logarithmic values of the durability). The process of failure during fatigue occurs in local volumes, the stresses in which are not known. In polymers, they can be evaluated by indirect methods, for example, by infrared spectroscopy [29]. In metal alloys and composite materials-structures, this can be done by inelasticity using mathematical models based on thermodynamic laws of fracture [4, 5, and 18].

Studies have targeted an interdisciplinary approach to the problem of the fracture of solids, for example, [30]. They consider in detail various aspects of the processes, depending on certain loading conditions, the materials used and their structures, and the mechanisms of the processes. A number of areas of knowledge related to the fracture of materials are considered and which explain what happens in this case. This is the knowledge necessary to understand the essence of the events taking place. Our approach does not consider the numerous details of the phenomena observed during fracture. This is like a cross-section of the whole problem in a certain plane. It is based on the fundamentals of the thermodynamics, and all types of fracture are considered from the same positions. The formulas given above are used specifically for calculations. Although for particular cases the mechanics hikes give quite satisfactory results, but fundamentally this does not solve the problem. The progress in solving complex strength problems is possible by combining the knowledge and methods of mechanics, physics and physical materials science.

The above and in article [1] examples show how large a body of information is provided by analysis of the rheological properties of materials and how many possibilities arise in modeling these properties if the mechanical models are filled with physical content. The approach outlined here, unlike the others currently in use, relies on a single conceptual framework: the notions of what a solid is and why it failures. Any material is an atomic-molecular system in thermal motion, and fracture occurs as a result of anharmonicity and stochasticity of the process of thermal vibrations of atoms in a solid body [7]. This is confirmed by a numerical experiment performed by the molecular dynamics method [31]. It follows that fracture is due to the internal energy of the solid, the measure of which is temperature [5, 7]. External effects only change the failure rate of a solid if they change its internal energy (thermal energy, electromagnetic radiation, chemical reactions). The end of the process is the achievement of a certain concentration of damage (microcracks, pores, delamination in composites), leading to its transition to the next dimensional level [4, 5, 7, 9, and 14]. And the prediction of further fracture, such as crack propagation, is reduced to modeling the process of material failure at its tip by the same methods [7, 18, and 32].

5. Conclusion

The proposed methodology for predicting the durability of materials in structures shows that an interdisciplinary approach and reproduction in mathematical models of the processes of their failure and deformation as thermodynamic, allows us to solve those problems that have not yet been solved by methods of fracture mechanics. It is shown that it is based primarily on the study of the physical laws of fracture, which are revealed in experiments to determine durability at constant stresses and temperatures. Then these laws should be applied to other loading cases: variable loads and temperatures. And by involving the knowledge of materials science, it becomes possible to take into account in the calculation procedures other thermally activated processes accompanying the flow and failure of a solid, which changes its structure.

The above approach shows the practical sequence of actions required when investigating the strength properties of any new material. In order to study the strength, it is necessary, as a minimum, to test its specimens with different loading rates and at different temperatures, and then perform thermally activation analysis of the obtained data. To study the resource characteristics of a material, it is necessary to investigate their inelastic properties, comparing the results obtained with the endurance data, while taking into account the temporary nature of fatigue failure. This is followed by mathematical modeling of processes for solving problems of arbitrary temperature-force loading.

Conflict of Interest

The author did not have a conflict of interest with his colleagues using other approaches to the problem under consideration. The author experienced only a benevolent attitude, support and understanding.

Acknowledgment

The author is deeply grateful to H. D. Gringauz and N. A. Moshkin for a number of joint works that gave a lot of new, useful information and initiated the described approach, and with whose permission their results are presented here. He is also grateful to many other colleagues who, to one degree or another, helped him in numerous and laborious experiments.

References

[1] M.G. Petrov, "Interdisciplinary approach to solving problems on the flow and fracture of materials," in XX International Conference on Methods of Aerophysical Research (ICMAR 2020), AIP Conference Proceedings 2351, doi: 10.1063/12.0004278.

[2] W. Kauzmann, "Flow of solid metals from the standpoint of the chemical-rate theory," Transactions of the AIME, **143**, 57–83, 1941.

[3] S. Glasstone, K.J. Laidler, H. Eyring, The theory of rate processes, McGraw-Hil, 1941.

[4] M.G. Petrov, "Mathematical modeling of failure and deformation processes in metal alloys and composites," American Journal of Physics and Applications, **8** (4), 46–55, 2020, doi: 10.11648/j.ajpa.20200804.11.

[5] M.G. Petrov, "Investigation of the longevity of materials on the basis of the kinetic concept of fracture," Journal of Applied Mechanics and Technical Physics, **62** (1), 145–156, 2021, doi: 10.1134/S0021894421010181.
[6] V.A. Stepanov, N.N. Peschanskaya, V.V. Shpeizman, G.A. Nikonov, "Longevity of solids at complex loading," International Journal of Fracture, **11**, 851–867, 1975.
[7] V.A. Petrov, A.Ya. Bashkarev, V.I. Vettegren, Fizicheskiye osnovy prognozirovaniya dolgovechnosti konstruktsionnykh materialov, Polytechnika, 1993.
[8] M.G. Petrov, A.I. Ravikovich, "Deformation and failure of aluminum alloys from the standpoint of kinetic concept of strength," Journal of Applied Mechanics and Technical Physics, **45** (1), 124–132, 2004.
[9] M.G. Petrov, "Fundamental studies of strength physics – methodology of longevity prediction of materials under arbitrary thermally and forced effects," International Journal of Environmental and Science Education, **11** (17), 10211–10227, 2016.
[10] M.G. Petrov, "Rol' protsessa polzuchesti svyazuyushchego v kinetike razrusheniya stekloplastika," in XII Mezhdunarodnoy nauchnoy shkoly-konferentsii: Fundamental'noye i Prikladnoye Materialovedeniye, Izdatelstvo AltGTU, 50–66, 2015.
[11] J. Bailey, "An attempt to correlate some tensile strength measurements on glass," Glass Industry, **20**. 21–25, 1939.
[12] M.G. Petrov, "Rheological properties of materials from the point of view of physical kinetics," Journal of Applied Mechanics and Technical Physics, **39** (1), 104–112, 1998.
[13] A.R. Michetti, "Fatigue analysis of structural components through math-model simulation," Experimental Mechanics, **2**, 69–76, 1977.
[14] S.N. Zhurkov, "Dilatonnyy mekhanizm prochnosti tvordykh tel," Fizika Tverdogo Tela, **25** (11), 33198–3323, 1983.
[15] H.B. Dwight, Tables of integrals and other mathematical data, Macmillan Company, 1961.
[16] A.J. Kennedy, Processes creep and fatigue in metals, Oliver and Boyd, 1962.
[17] T-S. Kê, "Experimental evidence of the behavior of grain boundaries in metals," Physical Review, **71** (8), 533–546, 1947.
[18] M.G. Petrov, Prochnost' i dolgovechnost' elementov konstruktsiy: podkhod na osnove modeley materiala kak fizicheskoy sredy, Lambert Academic Publishing, 2015.
[19] A.S. Nowick, B.S. Berry, Anelastic relaxation in crystalline solids, Academic Press, 1972.
[20] M.L.V. Gayler, P. Parkhouse, "The ageing of high-purity 4 percent copper-aluminium alloy," Journal of Institute of Metals, **66**, 67–84, 1940.
[21] M.L.V. Gayler, "The cold working of a high-purity aluminium alloy containing 4% of copper and its relation to age-hardening," Journal of Institute of Metals, **72**, 543–563, 1946.
[22] H.K. Hardy, "The ageing characteristics of some ternary aluminium-copper-magnesium alloys with copper: magnesium weight ratios of 7 : 1 and 2.2 : 1," Journal of Institute of Metals, **83**, 17–33, 1954.
[23] R. Graf, A. Guinier, "Influence de l'ecrouissage apres trempe sur les phenomenes de precipitation dans l'alliage aluminim-cuivre a 4% de cuivre," Comptes rendus hebdomadaires des seances de l'academie des sciences, **238**, 819–821, 1954.
[24] J.W. Christian, The theory of transformations in metals and alloys. Part I. Equilibrium and general kinetic theory. 2nd ed., Pergamon Press, 1975.
[25] L.H. Van Vlack, Materials science for engineers, Addison-Wesley, 1970.
[26] A. Kelly, R.B. Nicholson, "Precipitation hardening," Progress of Material Science, Pergamon Press, **10**, 151–391, 1963.
[27] M.G. Petrov, "On test programs of aircraft structures," in XVI International Conference on the Methods of Aerophysical Research (ICMAR 2012): abstracts, Part I, Kasan Federal University, 2012.
[28] D. Kamke, K. Krämer, Physikalische Grundlagen der Maßeinheiten, B. G. Teubner, 1977.
[29] V.I. Vettegren, I.I. Novak, K.J. Friedland, "Overstressed interatomic bonds in stressed polymers," International Journal of Fracture, **11**, 789–801, 1975.
[30] T. Yokobori, An interdisciplinary approach to fracture and strength of solids, Wolters–Noordhoff Scientific Publications Ltd, 1968.
[31] V.S. Yuschenko, E.D. Schukin, "Molekulyarno-dinamicheskoye modelirovaniye pri issledovanii mekhanicheskikh svoystv," Fiziko-khimicheskaya mekhanika materialov, **4**, 46–59, 1981.
[32] V.R. Regel, A.M. Leksovskii, S.N. Sakiev, "The kinetics of the thermofluctuation – Induced micro- and macrocrack growth in plastic metals," International Journal of Fracture, **11**, 841–850, 1975.

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 5, 141-148 (2021)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Real-time Measurement Method for Fish Surface Area and Volume Based on Stereo Vision

Jotje Rantung*, Frans Palobo Sappu, Yan Tondok

Department of Mechanical Engineering, Faculty of Engineering, Sam Ratulangi University, Manado, 95115, Indonesia

ARTICLE INFO

Article history:
Received: 21 June, 2021
Accepted: 29 August, 2021
Online: 27 September, 2021

Keywords:
Measurement method
Real-time
Stereo-vision
Fish-surface-area
Fish-volume

ABSTRACT

In the automation of the fish processing industry, the measurement surface-area and volume of the fish requires a method that focuses on processing automation. The creation of a stereo-vision based on real-time measurement method is one of the most essential aspects of this work. To do this task, we completed two steps. The first, the acquisition of the image of the fish using a stereo camera and calibrating the image for size using sample of the image acquisition. Second, by applying image processing techniques and vision system, the fish surface area and fish volume is obtained in real-time. The experimental results of the proposed method have good results for fish surface area and fish volume. The measuring process using stereo-vision only takes a short time, making it suitable for the real-time method.

1. Introduction

Knowing the volume size is useful in the fish processing industry for size sorting, quality assessment, and microbial concentration [1]. After calculating the volume, other physical parameters such as mass and density are also easy to investigate. Surface area and volume are important physical parameters in fish[2,3]. Surface-area and fish volume are important parts in processing into fish products that must be known. In the salting process of fish, the area and thickness of the salt are affected by the surface area. Furthermore, understanding the surface area of fish is needed to calculate heat and mass transfer, determine other physical parameters including gas permeability, weight per unit surface area, and respiration rate. Fish-volume and other physical properties are appropriate for calculating water content, heat transfer, and respiration rates. Likewise, the volume of fish affects the rate of cooling and freezing, which is ultimately useful for determining the heat load on the cooling system and calculating cooling costs [4, 5].

Measurement of surface area and volume of fish is currently generally done manually by using the eyes of people or workers using a manual meter. Manual measurements can result in inaccurate, ineffective, and time-consuming measurements, especially when measuring fish in large numbers. The most commonly used for measuring surface area is the tape method, while the volume measurement used the water displacement method [6]. The tape method has drawbacks, such as timing problems, and may cause measurement errors due to human error. Similarly, because the fish has holes in its gills and mouth, volume measurement using the gas transfer method and the water displacement method is not practical for fish shapes. Analytical estimates based on principal dimensions and weight have been studied [7], but the results of this method are very time-consuming because the sensing process is carried out by human.

In the entire fish processing process, measuring the surface area and volume is critical. To increase production and minimize processing time and costs, the fish processing sector requires technology solutions that focus on processing automation. Innovative advances in image processing have empowered the use of new methods to quantify the surface region and volume of fish correctly, rapidly, and precisely. With the availability of image processing techniques, digital image analysis has begun to be used for simple measurement systems such as line measurements [8]. A partitioning technique utilizing the image processing has been completed to ascertain the surface region and volume of fish [9]. The working principle of this measurement method is offline. Real-time image segmentation to determine the coordinates of image objects has been carried out [10]. The segmentation principle employs stereo-vision as a measuring instrument, which is typically unstable and slow. With the availability of image processing methods and stereo-vision as measuring tools, the automation of the fish processing industry is feasible because to the development of a stereo-vision-based real-time measuring.

*Corresponding Author: Jotje Rantung, Faculty of Engineering, Sam Ratulangi University, Manado, 95115, Indonesia, +6285298392179, jrantung@unsrat.ac.id

https://dx.doi.org/10.25046/aj060516

Based on the previous description, the researcher is interested in investigating and developing new methods for measuring the surface area and volume of fish, specifically real-time measurements, and creating a stereo-vision-based instrumentation system for real-time measurements.

The main objective of this research is to develop a method of measuring the surface area and volume of fish using a stereo camera as a vision instrument so the measurements can be applied in real-time. The development is implemented by combining the ellipsoid approach method and the real-time image segmentation method using a stereo camera as stereo-vision.

2. Related Work

As previously described, the fish handling industry requires mechanical arrangements that attention on mechanization of preparing to expand efficiency and lessen handling time and expenses. Estimating the surface region and volume of fish is a significant stage in the entire process of the fish handling industry. In order to automate the process of measuring the surface area and volume of fish, new methods need to be developed. Due to technological advances in image processing methods and the principle of real-time image segmentation using stereo-vision, research on image processing and stereo-vision applications for measurement is increasingly being carried out.

The first step in real-time measurement using stereo-vision is segmenting the image of the object to be measured. Studies on image segmentation methods are still being carried out at present, all of which aim to improve the quality of the image segmentation of objects. The selection of automatic thresholding for image segmentation based on genetic algorithms was done in [11]. Research on image segmentation of fish objects using the K-means cluster enhancement algorithm to obtain fish body contours by separating the fish image from the background in complex background conditions has been carried out [12]. All the studies mentioned above showed satisfactory results. However, the proposed studies were limited to static imagery.

Researchers have conducted several preliminary investigations on fish image segmentation in addition to the studies described above. Research to determine the amount of fish skin injury has been proposed [13]. The proposed research was to realize the form of injury to fish based on the L*A*B* color space and the HSV color space. Subsequent research proposed a new approach to measuring 2D injury rates in fish with a modified K-means cluster algorithm based on the L*A*B* color space [14]. The experimental results of this study indicate that the proposed new approach is closer to the level of injury and actual injury to fish than the results of the manual threshold method on L*A*B* color images. The two studies above resulted in fish image segmentation, which effectively measures the level of injury to fish skin. However, in practice, it was still carried out offline. An offline image segmentation development is real-time image segmentation. The real-time image segmentation method for online measurement purposes was developed [10]. This real-time image segmentation method uses a stereo camera as stereo-vision to determine the coordinates of the image object. The emphasis was on determining the coordinate image of the object.

Image processing techniques are still being used to develop methods for measuring surface area and volume. Some of them was the determination of the volume and surface area of bubbles [15], estimating the volume and weight of apples using the 3D reconstruction method [16], and the determination of the surface area and volume of the fish using the ellipsoid approach method applied to image processing [9]. However, the determination and estimation of volume and surface area of image objects were still done offline in these studies. The strategy to decide surface region and volume of axisymmetric farming items was created [17]. In this study, the image captured by the CCD camera was processed using adobe photoshop. The use of an inexpensive 3D scanner to measure the surface area and volume of sweet potatoes was proposed [18]. The results of the method were satisfied for the identification of features related to shape using 3D scanner-based measurement.

In this study, we will develop a method of measuring the surface-area and volume of fish in real-time using a stereo camera as stereo-vision. Development is done by combining the ellipsoid approach method that researchers have reported in [9] and the real-time image segmentation method using a stereo-camera as stereo-vision reported in [10]. In the elliptical approach, the fish object is considered an ellipsoid shape by making partitions on three sides of the view according to the analytical model. Partitions on three sides are analyzed as image pixel shape are estimated as surface area and volume. The real-time segmentation process is carried out on a stereo-vision video frame that will capture the 3D coordinate value of the fish object to be measured. Real-time 3D coordinate values in the stereo-vision video frame corresponding to the image pixel values on the partitions of each viewpoint of the fish. The pixel value will be converted to millimeter size according to stereo-vision calibration.

3. Proposed Real Time Measurement Method

This section presents a real-time estimating technique for the surface region and volume of fish by using a vision system. The stereo camera mounted on the highest point of the fish object is used to capture the images. Image calibration for fish size was carried out using one image sample taken with a stereo camera. The stereo camera performs the segmentation procedure in real time on the video frame. The calibration of the stereo camera yielded intrinsic and extrinsic parameters. Open-source computer vision libraries (OpenCV, C++) were used to create an image segmentation algorithm that works in real-time. To extract objects in a “graph-based image,” it utilizes a combination “HSV color space, threshold value, mathematical morphological transformations, and contour detection techniques.” This research provides a method for performing real-time image segmentation. This process uses open-source computer vision (OpenCV, C++) to determine the surface area and volume of the fish. The utilization of a stereo camera to quantify the surface region and volume of fish in 3D directions is introduced as well in this paper. To test the validity of the proposed real-time measurement methodology, the experimental findings of the proposed real-time measurement technique are compared to those of the analytical measurement approach.

3.1. Stereo Camera Image Acquisition

The experimental arrangement for this proposed measuremet is shown in Figure 1. A stereo camera was utilized to capture the images in this study. Figure 2 depicts the target measurement experimental setup. The cross-line serves as a guide for precisely positioning the target fish. A stereo camera can automatically determine the fish's location.

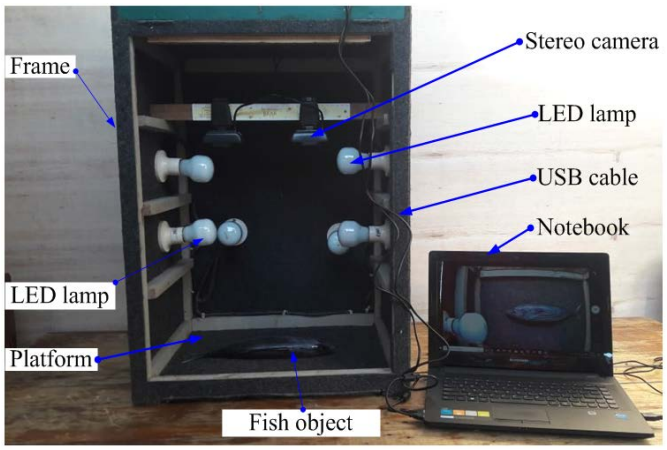


Figure 1: Experimental arrangement of proposed measurement

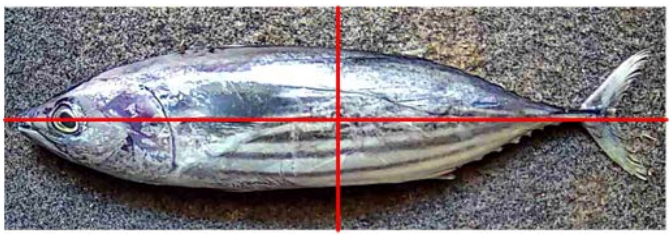


Figure 2: Experimental setup for target measurement

3.2. Image Calibration

A sample image taken with a stereo camera is used to calibrate the size of the fish image. A caliper was used to measure the physical dimensions of the fish in millimeters, as shown in Figure 3. The value of the calibration constant for the measurement of surface area and fish volume was based on the findings of image calibration of one fish object. The number of pixels from the outermost distance of the fish pictures is the unit of measurement in millimeters

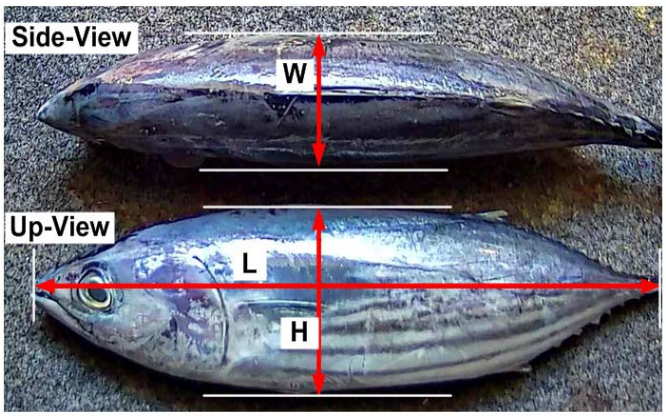


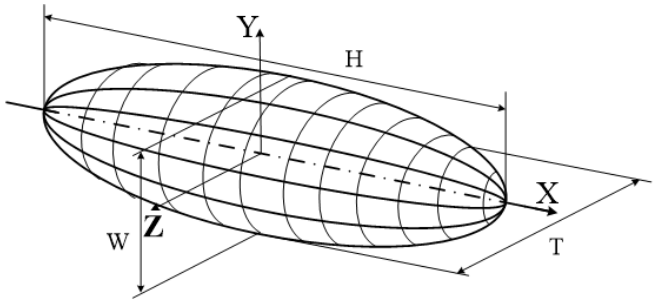
Figure 3: Fish measured by using micrometer

3.3. Partition for image processing method

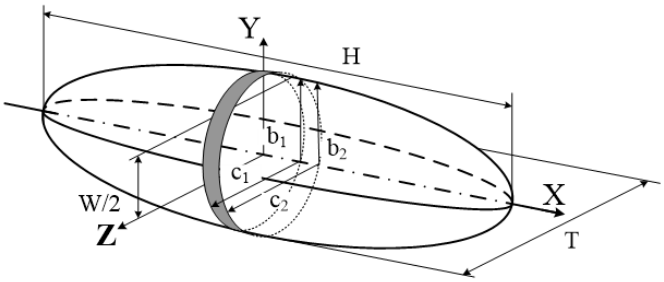
Figure 4 depicts a fish surface region and volume picture preparation technique. Each circle area and volume may be calculated as follows:

$$S_i = \pi(b_i + c_i)\Delta t_i \tag{1}$$

$$S = \sum_{i}^{n/2} S_i \quad \text{for} \quad i = 2,3,\cdots,n/2 \tag{2}$$



(b) Prolate spheroid orientation



(b) General ellipsoid

Figure 4: Model used for image processing method

Area of each disc is calculated as average area of left and right planes, and dt_i is very thin and set as a pixel. The volume of each disc can be calculated as follows:

$$V_i = \pi\Delta t_i (b_i c_i) \tag{3}$$

$$V = \sum_{i}^{n/2} V_i \quad \text{for} \quad i = 2,3,\cdots,n/2 \tag{4}$$

where $b_{1i} = W_{1i}/2$ and $b_{2i} = W_{2i}/2$ denote heights of the discs of the left and right planes in top view, $c_{1i} = H_{1i}/2$ and $c_{2i} = H_{2i}/2$ denote widths of the discs of left and right planes in side view, respectively, dt_i denotes as the thickness between disc of the left and right planes, and n denotes the number of boundary point of the fish contours.

3.4. Stereo Camera Calibration

The focal length of a stereo camera is an important parameter in a measurement algorithm. This value is obtained by calibrating a stereo camera. This parameter determines whether a camera

lens's ability to focus on an object through distorted images is strong or weak. There are four intrinsic parameters: f_x and f_y as the focal lengths of the camera in terms of pixel dimensions in the x and y direction, and (u_0,v_0) is the principal point. The camera usually represents lens distortion that is a radial distortion given as follows:

$$\mathbf{u}_d = \begin{cases} u_d = (u_u - u_0)(1 + k_1 r_u^2 + k_1 r_u^4 \\ v_d = (v_u - v_0)(1 + k_1 r_u^2 + k_1 r_u^4 \end{cases} \quad (5)$$

distortion-free is expressed as $p(u_u,v_u)$ and distortion-normalized image coordinates is expressed as $p(u_d,v_d)$. The radial distortion coefficients are expressed as k_1 and k_2, and $r_u^2 = u_u^2 + v_u^2$. The focal length is determined by the camera model.

$$f = \frac{1}{2}(f_x / m_x + f_y / m_y) \quad (6)$$

In terms of pixel size in the x and y directions, the focal length of the camera are denoted by f_x and f_y .

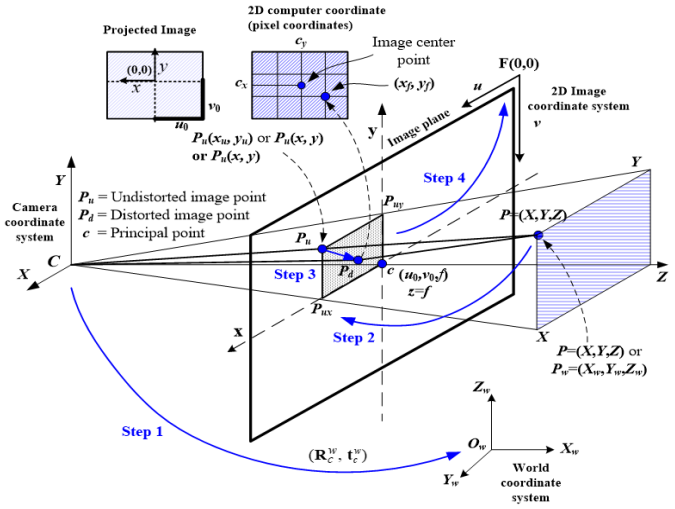


Figure 5: Steps of camera calibration

The most significant role of camera calibration is to identify the four intrinsic parameters and the two distortion coefficients.

The recommended calibration procedure as follows:

- Setup camera, print a pattern and attach it to a planar surface.
- Take a few images of the model plane under different orientations by moving either the plane or the camera.
- Detect the feature points in the images.
- Estimate the intrinsic and extrinsic parameters of the camera.

There are four steps to convert a point from the world coordinates to the computer memory image coordinate, shown in Figure 5.

3.5. Real-Time Image Segmentation

The segmentation process is done in real-time on the video frame by using a stereo camera. In reducing complexity and computation time, hue and value feature spaces are segmented. After that, they combine as a feature image segmentation. A stereo camera is applied to capture the images of the fish. In this work, use contour-based segmentation and mathematical morphological method for real-time segmentation. Image segmentation is done by combining the *HSV* color space, threshold, mathematical morphological transformations, and contour detection techniques to extract objects in graphics-based images. Software design is implemented by using the C++ programming language. A library available in open-source computer vision (OpenCV, C++) is used to implement real-time image segmentation by loading images, creating windows to hold an image in real-time, and saving the image.

3.6. 3D Measurement

The 3D position is obtained from stereo triangulation. First, two images are obtained from a set of three-dimensional test points whose three-dimensional coordinates are known. Secondly, the estimated 3D coordinate of the same points is computed from their projections using the calibrated parameters. Finally, the discrepancy between real and estimated positions is compared. In this case, the accuracy depends on the calibration of both cameras. Figure 6 shows the principle of 3D measurement using a stereo camera. The projective transformation for the basic stereo camera image is sought with the epipolar constraint that the epipolar line is horizontal, and 3D measurements are made from the information about the corresponding point by a stereo camera. The pairs of baseline stereo images are generated from ordinary images with the projective transformation of the axes of X, Y, and Z.

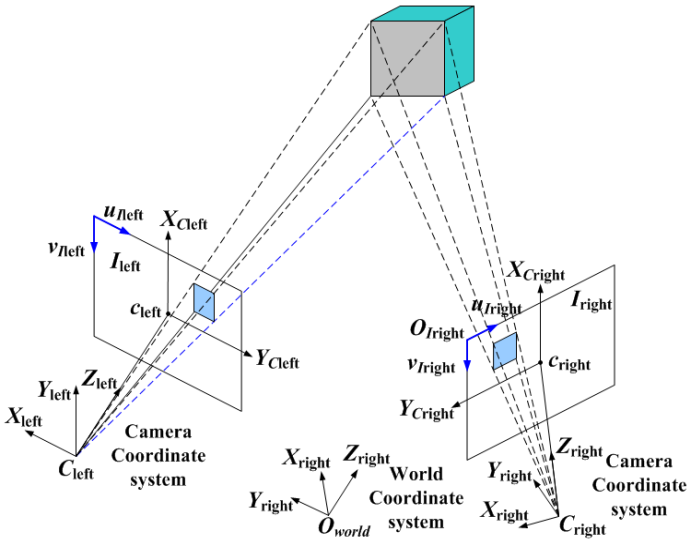


Figure 6: 3D measurement principle using stereo camera

The triangulation of stereo camera model is shown in Figure 7. Generally, stereo systems may have optical axes with the fixation point at a finite distance from the cameras. The left and right image planes are represented by the segments I*left* and I*right*, respectively, and O*l* and O*r* are the centers of projection, or optical centers in the left and right of the camera, respectively. Because the optical axes are parallel, their point of centroid called the fixation point lies intimately far from the cameras.

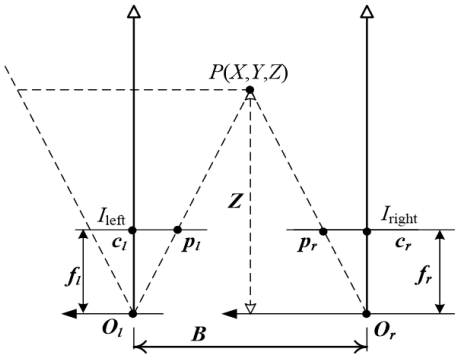


Figure 7: Triangulation of stereo camera mode

By assuming that the 3D coordinate frame has its origin in the optical center of the left camera, the perspective projection from the 3D camera coordinate (X, Y, Z) to the ideal image coordinate (x, y) is as follows:

$$\begin{cases} x_l = \dfrac{X \cdot f}{Z} \\ x_r = \dfrac{(X - B) \cdot f}{Z} - B \\ y_l = y_r = \dfrac{Y \cdot f}{Z} \end{cases} \tag{7}$$

From Equation (7), the Z coordinate can be determined as follows:"

$$Z = f \cdot \frac{B}{B - d} \tag{8}$$

where $x_l = p_l - c_l$, $x_r = p_r - c_r$, and $d = x_r - x_l$ is a disparity that is the difference in image position between corresponding points in the two images. Once Z is determined, it is straightforward to calculate X and Y using similar triangles:

$$X = x_l \cdot \frac{B}{B - d} \tag{9}$$

$$Y = y_l \cdot \frac{B}{B - d} \tag{10}$$

Following the segmentation procedure, the 3D object coordinates are determined. "First, the object is recognized to determine its center, and then the principle of determining 3D coordinates is implemented. The following is how the distance error rate is calculated.

$$e_d = \left(\frac{Z - \text{real distance}}{\text{real distance}} \right) \cdot 100\% \tag{11}$$

3.7. Fish Surface Area and Volume Measurement

Figure 8 shows the flow chart of the software developed to calculate the surface area and volume of the fish. The process begins with obtaining two synchronous images of the cameras and their subsequent correction using the parameters obtained from the calibration. Once the images are corrected, the fish real-time image segmentation algorithm is executed individually for the up-view image and side-view image. It is done for the up-view image to obtain fish length and width in pixels, and the same process is done for the side-view image to obtain fish length and height in pixels.

The first process is done for the up-view image so that length (L) and width (W) are obtained, and these results are stored as a saved model. The length (L) and height (H) is calculated in a side view, and the result is saved as a saved model. For the last process, the values of LWH are used for the calculation process of surface area and volume.

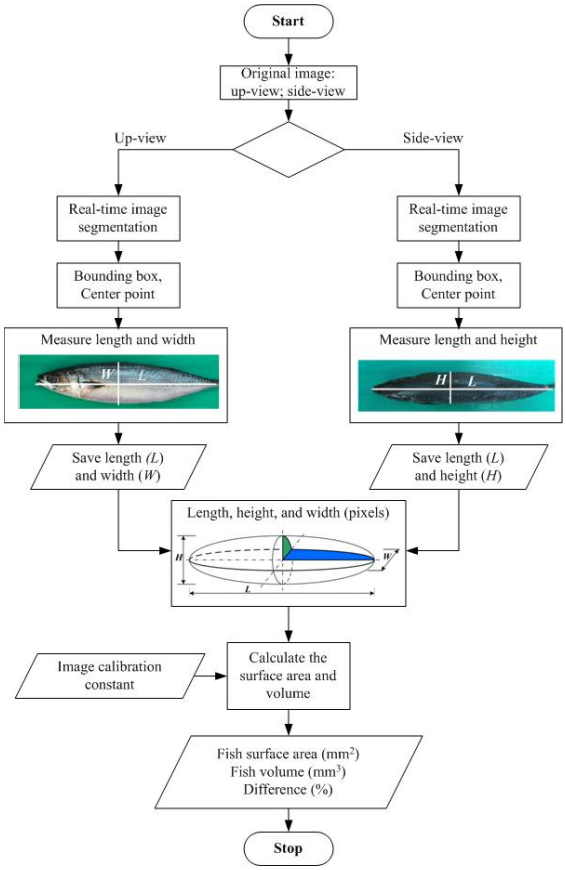


Figure 8: Flowchart to calculate fish surface are and volume

4. Experiment Results

4.1. Experiment Result for Image Calibration and Stereo Camera Calibration

In this experiment, for image calibration, we used a vernier caliper to measure the physical dimension of the fish in millimeters. We use the image calibration result of one fish object as a calibration constant value for measuring fish surface area and volume measurement. The size in millimeters is obtained by measuring the number of pixels from the outermost distance of the fish images. Figure 3 shows the fish image processing results used for image calibration. An example of one fish has a dimension of 290 mm (L) x 51.1 mm (H) x 38.22 mm (W), and the calibration result is obtained as 1 pixel = 0.0251 mm for (L), 1 pixel = 0.00557 mm for (H), and 1 pixel = 0.00475 mm for (W).

Camera intrinsic parameters are calibrated at a working distance of about 1000 mm. The chessboard image is used as a calibration pattern, and a 19" LCD monitor is used to display the image pattern. Table 1 shows the calibration results of the left camera and right camera, respectively.

Table 1: Calibration results of the intrinsic camera parameter

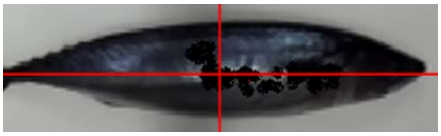
Camera	Focal length (pixel)		Principal point (pixel)	
	f_x	f_y	u_0	v_0
Left	940.49	769.00	674.65	180.35
Right	941.49	769.50	675.65	181.25

Camera	Distortion coefficients		pixel/mm		Focal length (mm)
	k_1	k_2	P_{ux}	P_{uy}	
Left	-98.91	11.34	674	180	2.8289
Right	-99.01	11.84	675	181	2.8311

4.2. *Experiment result for Real-Time Image Segmentation*

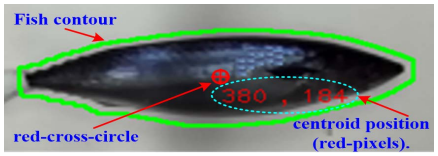
The proposed segmentation method is designed to detect fish objects useful for fish processing. The proposed image segmentation method is applied to different scanning objects before it is applied to the fish. Various common objects, such as simple objects, small and large objects, objects of different shapes, and various objects, are used in this experiment. Once the stereo camera finds the center position of the image fish it wants to measure, the open-source computer vision will segment the fish image, calculating it to prepare for real-time surface area and volume measuring.

The vision targets tested in the experiments are fish as a real detected vision object. Figure 9 shows fish target detection. Figure 9(a) shows the original RGB color image, Figure 9(b) shows the threshold image, and Figure 9(c) shows a segmented image with a centroid position.



(a) Thresholded image

(b) Original RGB color image with the contour image



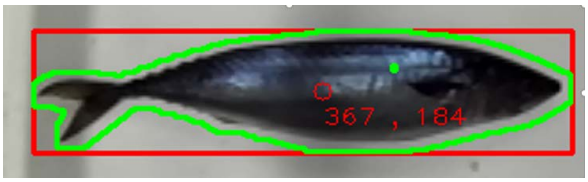
(c) Original RGB color image with the contour image and bounding box

Figure 9: Fish target detection

The image processing stage involves image segmentation for the fish object target, the object centroid, and the bounding box in pixels. The white background used in the experimental setup simplifies the background detection procedure in both RGB color images obtained with a stereo camera.

The results of segmenting fish objects using the lowest threshold (Th = 0) and a bounding box are shown in Figure 10. The best results were found in this research, as shown in Figure 10(a). The feature segmentation findings from this experiment will be used in the next step because of their high accuracy. The results of feature segmentation in Figure 10 are used in the next test. By applying the graph cut algorithm, the results are shown in Figure 11. Finally, image segmentation is obtained from the object's bw image, as shown in Figure 11(a).

(a) Thresholded image

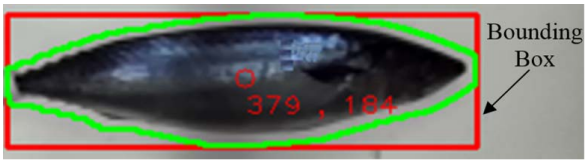


(b) Original RGB color image

Figure 10: Segmentation of the fish object with Th = 0 and bounding box

(a) Thresholded image

(b) Original RGB color image with the contour image



(c) Original RGB color image with the contour image and bounding box

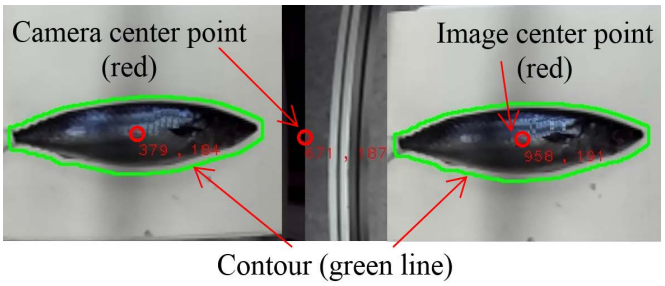
Figure 11: Segmentation of the fish object with Th=0

4.3. *Experimental results of fish surface area and volume measurement*

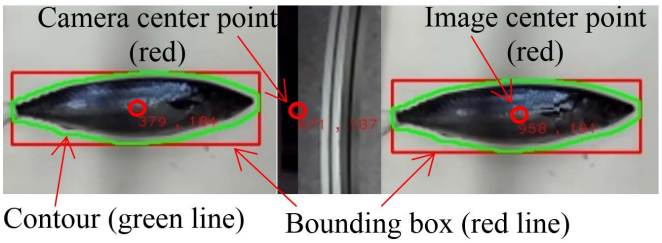
Figure 12 shows the representation of the right camera and the left camera in real-time image segmentation. The white area of the fish threshold image in Figure 12(a) is represented by the inner area of the green line in the original RGB color image of the fish in Figure 12(b). The white area in the threshold image and the green line in the original RGB color image are visible for the left camera and right camera. In other words, image segmentation works well. It is also indicated by the bounding box in the green

line area, the center point of the original left-right image (red), and the center point between the left camera and the right camera.

(a). Left and right thresholded images



(b). Left and right original RGB color images



(c). Left and right original RGB color images with bounding box

Figure 12: Representation of the right camera and left camera

The length, width, and height of the bounding box (L, W, and H) correspond to the fish image's length, width, and height in pixels. Because calculating the surface area and volume can be limited to one image, the centroid will be calculated for both images obtained with the stereo camera system. The thresholded image of the target analyzed in one experiment, as shown in Figure 12(a). Figure 12(b) shows the results of the detection of fish to be analyzed, namely the location of the centroid (red circle) and the contour (green line). Figure 12(c) shows the results of fish detection, which will be used to calculate the surface area and volume of fish. The result obtained in this case shows the centroid location (red circle), contour (green line), and bounding box in the stereo camera image. One example of the object experiment result is shown in Figure 12(b) and Figure 12(c). In this result, the image centroid for the left camera in pixels is (379,184), the center of the image for the right camera in pixels is (958,191), and the image center is for stereo cameras in pixels (671,187). The distance between the planes of the two cameras and the plane of pixels can be determined using this value.

The 2D image coordinates (pixels) of the center of mass in a stereo camera are shown in Table 2. Table 3 shows the results of the 3D coordinates (mm). Based on experimental results for 3D coordinate measurements, the least distance error rate, as listed in Table 3, is at a distance of 550 mm. The distance between the stereo camera and the object will be adjusted to 550 mm to measure the surface area and volume of the fish. The results of using stereo-vision to measure the surface area and volume of four fish in real-time are shown in Table 4. In Table 4, the measurement results using the analytical method are used as a comparison to obtain the effectiveness of the proposed real-time measurement method.

Table 2: Image coordinate of the centroid point

Trial	Centroid in 2D coordinate (pixels)		Real distance Z (mm)
	X	Y	
1	687	203	780
	680	194	660
	671	187	550
2	687	203	780
	680	194	660
	671	187	550
3	687	203	780
	680	194	660
	671	187	550

Table 3: Real-time measurement results of surface area and volume of four fishes

Trial	Centroid in 3D coordinate (pixels)			Real distance Z (mm	Distance error rate e_d (%)
	X	Y	Z		
1	16	20	793.2	780	1.69
	16	20	667.7	660	1.17
	16	20	553.4	550	0.62
2	16	20	792.9	780	1.65
	16	20	667.5	660	1.14
	16	20	553.2	550	0.58
3	16	20	791.6	780	1.49
	16	20	667.1	660	1.08
	16	20	552.6	550	0.47

Table 4: Real-time measurement results of surface area and volume of four fishes

Fishes sample size (mm)				Surface area (mm²)		
	L	H	W	Real-time method	Analytic method	Error e_d (%)
1	290	38.22	51.1	173,159.92	164,428.75	5.31
2	287	37.12	49.0	167,981.93	160,028.51	4.97
3	289	38.20	50.5	170,715.11	163,020.54	4.72
4	287	37.00	48.9	166,746.44	159,963.96	4.24
	L	W	H	Volume (mm³)		
1	290	38.22	51.1	311,256.65	296,406.67	5.31
2	287	37.12	49.0	286,384.78	273,189.71	4.83
3	289	38.20	50.5	305,330.52	291,763.51	4.65
4	287	37.00	48.9	282,648.04	271,750.83	4.01

5. Conclusions

A method for measuring fish surface area and volume in real-time using a stereo camera as a stereo-vision was proposed. The measurement process was done in real-time on the video frame. Image object segmentation is done first before measurement. The hue (*H*), saturation (*S*), and value (*V*) were separately segmented before they are combined. The result of this segmentation was the targeting object.

Experiment results show that real-time image segmentation of the proposed method had a good result. The experiments demonstrated that the calibration process could quickly detect the chessboard corners. After the calibration results, the focal lengths of the left camera and right camera were about 2.8289 mm and 2.8311 mm, respectively. The focal length difference between the two cameras was about 0.00221 mm. For desired object distances of 780 mm, 660 mm, and 550 mm, the distance error rate was less than 2% for both distances of 650 mm and 550 mm in three trial times. However, for 780 mm, the distance error rate was bigger than the two distances of 550 mm and 660 mm and was more than 3%. The 3D coordinate measurement results revealed that increasing the distance increased the distance error in the Z coordinate, which was caused by the camera's vision field of view. As a result, a distance of 550 mm is used in the next analysis, which involves measuring the surface area and volume of the fish.

The surface area and volume of fish were measured by the proposed real-time measurement and compared to the analytic measurement method. Experimental results using a sample of four fishes show that the differences in surface area and volume were 4.24%~5.31% and 4.01%~5.01%, respectively. The process of real-time image segmentation, feature extraction, and measurement of fish surface area and volume takes about 0.0018 milliseconds. The results show that the proposed method can accurately measure the surface area and volume of fish in real-time.

The real-time image segmentation, 3D information, and measurement methods for surface area and volume proposed in this research could be applicable in fish recognition and sorting applications. The 3D information-based stereo camera could be applied to an automated fish processing system to handle fish on a conveyor belt.

The measurement method developed by this research can be utilized to automatically measure the surface area and volume of fish using stereo-vision. Furthermore, by employing this approach, measuring will become easy, faster, more effective, and efficient. The findings of this study can also be accustomed to other fields of research, such as real-time assessment of fruit product dimensions, fish sorting processes on moving conveyors, and moving goods robots.

Acknowledgment

This research was funded by PNBP BLU 2021, Sam Ratulangi University, Manado, Indonesia.

References

[1] E.S. Bridge, R.K. Boughton, R.A. Aldredge, T.J.E. Harrison, R. Bowman, S.J. Schoech, “Measuring egg size using digital photography: Testing Hoyt’s method using Florida Scrub-Jay eggs,” Journal of Field Ornithology, **78**(1), 109–116, 2007, doi:10.1111/J.1557-9263.2006.00092.X.

[2] R. Badonia, A. Ramachandran, T. V Sankar, “Quality Problems in Fish Processing,” Journal of the Indian Fisheries Association, **18**, 283–287, 1988.

[3] A. Getu, K. Misganaw, “Post-harvesting and Major Related Problems of Fish Production,” Fisheries and Aquaculture Journal, **06**(04), 2015, doi:10.4172/2150-3508.1000154.

[4] D. J. Lee, X. Xu, J.D. Eifert, P. Zhan, “Area and volume measurements of objects with irregular shapes using multiple silhouettes,” Optical Engineering, **45**(2), 027202, 2006, doi:10.1117/1.2166847.

[5] A.B. Koc, “Determination of watermelon volume using ellipsoid approximation and image processing,” Postharvest Biology and Technology, **45**(3), 366–371, 2007, doi:10.1016/J.POSTHARVBIO.2007.03.010.

[6] S.M.A. Moustafa, “Theoretical Prediction of Volume, Surface Area, and Center of Gravity for Agricultural Products,” Transactions of the ASAE, **14**(3), 549–0553, 1971, doi:10.13031/2013.38336.

[7] T.Y. Wang, S.K. Nguang, “Low cost sensor for volume and surface area computation of axi-symmetric agricultural products,” Journal of Food Engineering, **79**(3), 870–877, 2007, doi:10.1016/J.JFOODENG.2006.01.084.

[8] F. Pla, J.M. Sanchiz, J.S. Sánchez, “An integral automation of industrial fruit and vegetable sorting by machine vision,” IEEE International Conference on Emerging Technologies and Factory Automation, ETFA, **2**, 541–546, 2001, doi:10.1109/ETFA.2001.997731.

[9] J. Rantung, M.T. Tran, H.Y. Jang, J.W. Lee, H.K. Kim, S.B. Kim, “Determination of the Fish Surface Area and Volume Using Ellipsoid Approximation Method Applied for Image Processing,” Lecture Notes in Electrical Engineering, **465**, 334–347, 2017, doi:10.1007/978-3-319-69814-4_33.

[10] J. Rantung, J.M. Oh, H.K. Kim, S.J. Oh, S.B. Kim, “Real-Time Image Segmentation and Determination of 3D Coordinates for Fish Surface Area and Volume Measurement based on Stereo Vision,” Journal of Institute of Control, Robotics and Systems, **24**(2), 141–148, 2018, doi:10.5302/J.ICROS.2018.17.0213.

[11] 병 룡이, Q. Bao Truong, V. Huy Pham, 김 형 석 Byung-Ryong Lee, H.-S. Kim, “Automatic Thresholding Selection for Image Segmentation Based on Genetic Algorithm,” **17**(6), 587–595, 2011, doi:10.5302/J.ICROS.2011.17.6.587.

[12] H. Yao, Q. Duan, D. Li, J. Wang, “An improved K-means clustering algorithm for fish image segmentation,” Mathematical and Computer Modelling, **58**(3–4), 790–798, 2013, doi:10.1016/J.MCM.2012.12.025.

[13] M.T. Tran, H.H. Nguyen, J. Rantung, H.K. Kim, S.J. Oh, S.B. Kim, “A New Approach of 2D Measurement of Injury Rate on Fish by a Modified K-means Clustering Algorithm Based on L*A*B* Color Space,” Lecture Notes in Electrical Engineering, **465**, 324–333, 2017, doi:10.1007/978-3-319-69814-4_32.

[14] M.T. Tran, J. Rantung, T.H. Nguyen, H.K. Kim, S.B. Kim, “Measurement of injury rate on fish skin and performance comparison based on L*A*B* and HSV color spaces,” Aug. 2021, doi:10.1051/matecconf/201815902010.

[15] J. Wen, Q. Sun, Z. Sun, H. Gu, “An improved image processing technique for determination of volume and surface area of rising bubble,” International Journal of Multiphase Flow, **104**, 294–306, 2018, doi:10.1016/J.IJMULTIPHASEFLOW.2018.02.004.

[16] B. Zhang, N. Guo, J. Huang, B. Gu, J. Zhou, “Computer Vision Estimation of the Volume and Weight of Apples by Using 3D Reconstruction and Noncontact Measuring Methods,” Journal of Sensors, **2020**, 2020, doi:10.1155/2020/5053407.

[17] C.M. Sabliov, D. Boldor, K.M. Keener, B.E. Farkas, “Image Processing Method to Determine Surface Area and Volume of Axi-Symmetric Agricultural Products,” International Journal of Food Properties, **5**(3), 641–653, 2002, doi:10.1081/JFP-120015498.

[18] A. Villordon, J.C. Gregorie, D. LaBonte, “Direct Measurement of Sweetpotato Surface Area and Volume Using a Low-cost 3D Scanner for Identification of Shape Features Related to Processing Product Recovery,” HortScience, **55**(5), 722–728, 2020, doi:10.21273/HORTSCI14964-20.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 2, 143-148 (2022)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Interpretable Rules Using Inductive Logic Programming Explaining Machine Learning Models: Case Study of Subclinical Mastitis Detection for Dairy Cows

Haruka Motohashi[*,1], Hayato Ohwada[2]

[1]*Graduate School of Science and Technology, Department of Industrial Administration, Tokyo University of Science, Noda, Chiba, 278-8510, Japan*

[2]*Faculty of Science and Technology, Department of Industrial Administration, Tokyo University of Science, Noda, Chiba, 278-8510, Japan*

ARTICLE INFO

Article history:
Received: 01 February, 2022
Accepted: 07 April, 2022
Online: 12 April, 2022

Keywords:
Inductive Logic Programming
Model Interpreting
Mastitis Detection

ABSTRACT

With the development of Internet of Things technology and the widespread use of smart devices, artificial intelligence is now being applied as a decision-making tool in a variety of fields. To make machine learning models, including deep neural network models, more interpretable, various techniques have been proposed. In this paper, a method for explaining the outputs of machine learning models using inductive logic programming is described. For an evaluation of this method, diagnostic models of bovine mastitis were trained using a dataset of dairy cows, and interpretable rules were obtained to explain the trained models. As a result, the rules obtained indicate that the trained classifiers detected mastitis cases depending on certain variations in the electrical conductivity (EC) values, and in some of these cases, the EC and lactate dehydrogenase fluctuated in different ways. The interpretable rules help people understand the outputs of machine learning models and encourage a practical introduction of the models as decision-making tools.

1 Introduction

This paper is an extension of a study originally presented at the 2020 IEEE 19th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC) [1].

With the development of Internet of Things technology and the widespread use of smart devices, artificial intelligence is now being used as a decision-making tool in a variety of fields. Moreover, various machine learning models have been proposed to support an efficient and accurate medical diagnosis. Such models are expected to not only detect patients correctly, but also reveal the basis of the diagnosis.

Various techniques have been proposed to make machine learning models, including deep neural network models, more interpretable. Decision-tree-based algorithms (e.g., random forest and lightGBM [2]) provide the feature importance based on the frequency of all features in the trees generated by the algorithms. Some algorithms approximate original complex models (including a deep neural network) locally with simpler interpretable models [3, 4]. For convolutional neural network used to solving image processing tasks, gradient-based highlighting represents important regions in images where the networks focus to detect target objects or track them [5, 6].

Another approach to interpreting machine learning models is to describe their outputs using interpretable rules. Inductive logic programming (ILP) is based on predicate logic and can produce rules using inductive learning. ILP has the advantage of obtaining interpretable classification rules from training data and representing the opinions of domain experts [7, 8].

The interpretability of machine learning models has encouraged their introduction in decision-making applied in fields such as medical, including veterinary, diagnosis. Bovine mastitis, which is an inflammation of the udder or mammary gland owing to physical trauma or infection, is a common disease in dairy cattle, which dairy farmers must control to prevent economic losses.

With the introduction of auto milking systems, it has become easier to measure the indicators needed for cow health management during milking and to detect common diseases in dairy cows, including mastitis. Auto milking systems enable farmers to utilize auto mastitis detection using indicators such as milk yield, electrical conductivity, fat, protein, lactose and blood in the milk, and milk flow rate. In addition, SCC and various systems using statistical

[*]Corresponding Author: Haruka Motohashi, Faculty of Science and Technology, Tokyo University of Science, 2641 Yamazaki, Noda-shi, Chiba Prefecture 278-8510, Japan, Tel: +81-4-7124-1501, & 7420701@ed.tus.ac.jp

models and machine learning, including artificial neural networks, have also been reported [9]–[10].

In this study, we propose a method for explaining the outputs from machine learning models using ILP. For an evaluation of the method, diagnosis models of bovine mastitis were trained using a dataset of dairy cows, and interpretable rules were generated using ILP and the explained outputs of the mastitis detection model.

2 Method Used to Explain Classifiers through Interpretable Rules

An overview of the method used in this study is presented in Figure 1. This method aims to generate logic rules using background knowledge and outputs of a classifier, and to interpret the classification model. Interpretable rules are generated using ILP. In contrast to ordinary machine learning models such as deep neural networks, a resultant set of rules produced using ILP generally represents patterns in the given datasets. In this study, ILP is applied to outputs from classifiers trained using machine learning methods, and the set of rules generated describe how the model classifies the instances.

Although other methods based on a linear local approximation [3, 4] represent the importance of each feature, an explanation of machine learning models using ILP describes the models using nonlinear relationships with multiple features. Thus, methods for explaining classifiers can be applied to models with a complex architecture, such as deep neural networks [11, 12].

In this study, although the architectures of the machine learning models are maintained, the outputs of the models are given to an ILP system. Therefore, the proposed method is available regardless of the machine learning methods used for model training. Moreover, the definition of predicates used in ILP can be distinguished from the features in the classifiers and the predicates take advantage as a way to reflect knowledge of domain experts.

As shown in Figure 1, in the first step, a classification model is trained using machine learning methods and outputs of the classification are obtained. Background knowledge of the dataset and its outputs are then added into an ILP system, called Parallel GKS [13]. Finally, the classifier is interpreted based on the set of rules.

3 Case Study: Explanation of Bovine Mastitis Detection Model

To evaluate the proposed method, a classification model for the subclinical mastitis detection of dairy cows was trained using machine learning. With the introduction of auto milking systems, it has become easier to measure the indicators needed for cow health management during milking and to detect common diseases in dairy cows including mastitis.

Previous studies [9]–[10] used records of veterinary treatments and somatic cell count (SCC) for labeling the data of every milking as clinical or subclinical mastitis, and their models predicted the status of the quarters during each milking. SCC is generally used for the diagnosis of subclinical mastitis, and the most frequently used threshold for defining subclinical mastitis is 200,000 cells/mL [14]. However, it is thought that SCC can be affected by other factors such as the lactation number, stress, season, and breed [15].

As novel mastitis detection approaches, some biomarkers for mastitis detection have been discovered [16]–[17]. In particular, lactate dehydrogenase (LDH), which is related to inflammation, and according to a previous study is the biomarker with the lowest validation [18], is measured practically using a commercial milking machine. The result is applied to calculate the risk of developing mastitis as part of a milk analysis. However, mastitis detection using such biomarkers is still not a common approach for farmers because the equipment required is quite expensive. Therefore, in this study, a dataset in which cows labeled as either healthy or having subclinical mastitis based on the LDH values was prepared, and a common measurement, i.e., the electrical conductivity (EC), was used as a feature through the application of machine learning.

The dataset used in this study was collected between September 2018 and December 2021 at a farm in Hokkaido, Japan. On this farm, cows are milked any time they want, and items other than mastitis risk are measured during every milking. Data from September 2018 to August 2020 collected on the farm were used to train the detection model, and the remaining data were used to evaluate the model.

Datasets measured using an auto milking machine (a DeLaval Voluntary Milking System ™; VMS) and a milk analyzer (a DeLaval Herd Navigator ™; HN) were used. The HN measures the LDH, which is an index of subclinical mastitis, in milk and is used to calculate the risk of contracting mastitis.

The mastitis risk takes a value of zero to 100. On a farm, an HN measurement of greater than 70 allows farmers to suspect that a cow has mastitis. Therefore, in this study, subclinical mastitis cases were determined based on the mastitis risk. If her mastitis risk is above 70, the cow has subclinical mastitis; otherwise, her udder is disease-free.

3.1 Data Preprocessing

In this detection model, two features are calculated based on the EC obtained from VMS. One is the maximum EC values (max_EC) in the udder, and the other is ratio of the maximum to minimum EC values, i.e., the inter-quarter ratio (IQR). The EC is one of the measurements related to mastitis, and its value increases when a cow has mastitis [19]. According to our previous study [1], this mastitis detection model includes a four-day time series of these two features from three days prior to the prediction day, with eight features in total.

As mentioned above, cows on this farm are milked any time they want, and items without mastitis risk are measured during every milking. In addition, LDH in milk (indicating the risk of mastitis) is generally measured once every day to every three days, and the next measurement day is determined depending on the risk value. Therefore, this dataset consists of time series data, and the number of data points of a cow differs each day. Thus, features are calculated using data on the milking with the highest EC.

The preprocessed datasets in this study are shown in Table 1. In the test dataset, labeled samples are used for an evaluation of the detection model, and unlabeled samples are only used for generating rules through ILP.

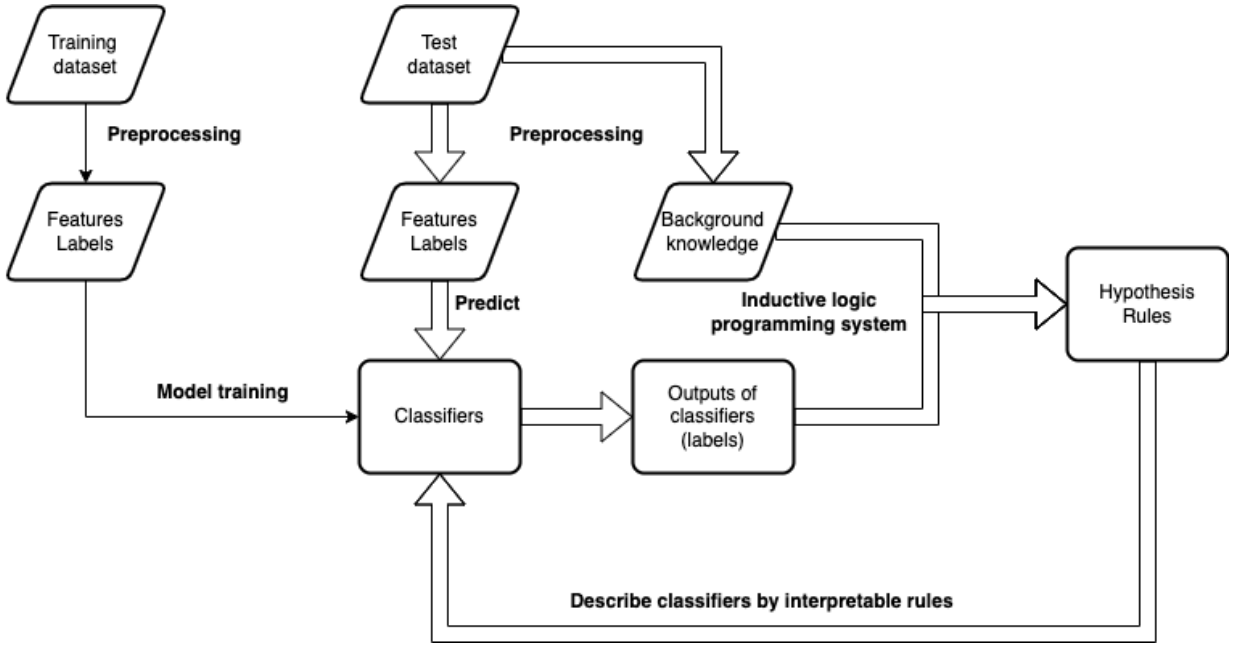


Figure 1: Overview of our method used to explain classifiers through interpretable rules.

Table 1: Number of cows and samples in the training and test datasets.

Dataset	Training	Testing
Data extraction period	2018-9-23 – 2020-11-1	2020-11-2 – 2021-12-2
Samples	98825	122372
Subclinical mastitis	4176	2121
Fine	94649	50771
unlabeled	-	69480

3.2 Learning Classification Models

After data preceding, classifiers for bovine mastitis detection are trained using a support vector machine (SVM). In this case, cows with subclinical mastitis are sparsely present in the dataset, and hence under-sampling (using the repeated edited nearest neighbors algorithm [20, 21]) was applied to the training dataset, and a regularization parameter in SVM (C) was adjusted using class weights, which are inversely proportional to the class frequencies in the input data.

The trained models were evaluated through a 10-fold cross validation using the sensitivity, specificity, and area under a receiver operating characteristic curve, as in previous studies on mastitis detection [9, 22, 23]. To evaluate the small number of false positives, the precision was also used for the evaluation.

Samples in the test datasets were given to the trained dataset and classified as fine or having mastitis. Using these outputs, interpretable rules explaining the detection model were generated using ILP.

3.3 Generation of Rules Using ILP to Explain the Trained Models

Using the outputs from the trained classifiers and background knowledge of samples in the test dataset, interpretable rules were obtained through ILP learning. Like other machine learning methods, ILP algorithms extract patterns in the samples with a certain label. In this study, an ILP system called GKS [24, 13] was used to employ ILP and generate rules.

Background knowledge is used in ILP learning, similar to features in other machine learning methods, and is represented by predicates in terms of the logic program. To describe the numerical features in terms of ILP, features were discretized based on the definition of the predicates for background knowledge and rules. In this case, three arguments were defined, as shown in Table 2. Such background knowledge of samples labeled by the trained classifier was given to the ILP system, and rules consisting of the predicates were generated.

A variable id in Table 2 represents one sample in the datasets. @IQR is a predicate which describes difference of the EC values between cows' quarters on the prediction day and whose variable, IQR, takes four values ($<= mean - std$, $> mean - std$, $> mean$, $> mean + std$ where $mean$ is 1.08 and std is 0.07) defined by discretized values of IQR (features in the detection models).

Table 2: Definition of the predicates in the subclinical mastitis detection model.

Predicate	Definition
@IQR,+id,+IQR @IQR,+id,-IQR @IQR,+id,#IQR	the maximum value of EC / the minimum value of EC in quarters (discretized by three thresholds: $mean - std$, $mean$, $mean + std$)
@delta_maxEC,+id,+day1,+day2,#delta @delta_maxEC,+id,-day1,+day2,#delta @delta_maxEC,+id,#day1,+day2,#delta @delta_maxEC,+id,+day1,-day2,#delta @delta_maxEC,+id,-day1,-day2,#delta @delta_maxEC,+id,#day1,-day2,#delta @delta_maxEC,+id,+day1,#day2,#delta @delta_maxEC,+id,-day1,#day2,#delta @delta_maxEC,+id,#day1,#day2,#delta	the difference of max_EC between two days ($minus$, $flat$, $plus$)
@before_day,+day1,-day2	sequence of days (day_0 is the prediction day and day_n is the day following day_{n+1})

@delta_maxEC represents the difference of max_EC values between two consecutive days (from three days prior to the prediction day) and a variable delta takes three values ($minus$, $flat$, $plus$ which describes the value of max_EC decreases or increases by over 0.4 or remain flat).

@delta_maxEC also has two variables representing targeted days from three days prior to the prediction day and these variables take four values ($day_3, day_2, day_1, day_0$). @before_day is a predicate which expresses an ordinal relation between these four values. This predicate contributes to generate rules flexibly, which consider difference of the values between arbitrary consecutive two days and mention changes of the values between variable prediods before the prediction day, unlike tree-based algorithms.

Outputs from the learned classifier and the background knowledge were given to parallel GKS and ILP learning was employed. Finally, interpretable rules for mastitis detection models were obtained.

Table 3: Subclinical mastitis detection performance in the training dataset (a 10-fold cross validation was used) and the test dataset.

	Sensitivity	Specificity	Precision	AUC
Training	0.668	0.814	0.137	0.809
Test	0.667	0.840	0.148	0.831

4 Result and Discussion

To evaluate our method for explaining machine learning models using interpretable rules, it was applied to the classification problem of bovine mastitis detection. Table 3 lists the evaluation results for the classifier of subclinical mastitis detection using an SVM. In the test dataset, 79 records of veterinary treatment were included, and 66 out of 79 (83.5%) records were detected as subclinical mastitis by the classifier, whereas the target label of the model underestimated the mastitis risk. Therefore, the trained model detected some of the clinical mastitis cases correctly, although the precision of the classifier was 15%.

Using outputs from the classifier, ILP learning was employed, and interpretable rules were obtained. The rules that were the most readable and related to the real conditions of mastitis are listed below. These rules describe the relationship between mastitis and the variation of the EC values.

Rule1 pos(A) :- IQR(A, > $mean+std$), delta_maxEC(A, B, C, $plus$), delta_maxEC(A, day0, day1, $flat$)

Rule2 pos(A) :- delta_maxEC(A, B, C, $minus$), delta_maxEC(A, C, D, $minus$)

Rule3 pos(A) :- IQR(A, > $mean$), delta_maxEC(A, day2, B, $plus$)

Rule4 pos(A) :- delta_maxEC(A, B, C, $minus$), delta_maxEC(A, day2, day3, $plus$)

These rules were described variation of max_EC values before rising mastitis risk by combination of @delta_maxEC. Values A, B and C given to variables day1 or day2 in this predicate represented arbitrary consecutive days before the prediction day, which made genereated rules more scalable.

Among the rules obtained, some clearly explaining the classifier are described in detail.

Rule1 pos(A) :- IQR(A, ¿mean+std), delta_maxEC(A, B, C, plus), delta_maxEC(A, day0, day1, flat)

Rule 1 indicates that cows whose IQR value is extremely high and whose max_EC values continuously increase have a high risk of subclinical mastitis. Figure 2 is a case corresponding to Rule 1. This rule also indicates that the detection model comprehends the typical relationship between bovine mastitis and electrical conductivity of milk. In this case, EC and LDH increased simultaneously, and the classifier detected a high mastitis risk as the mastitis risk alarm based on LDH.

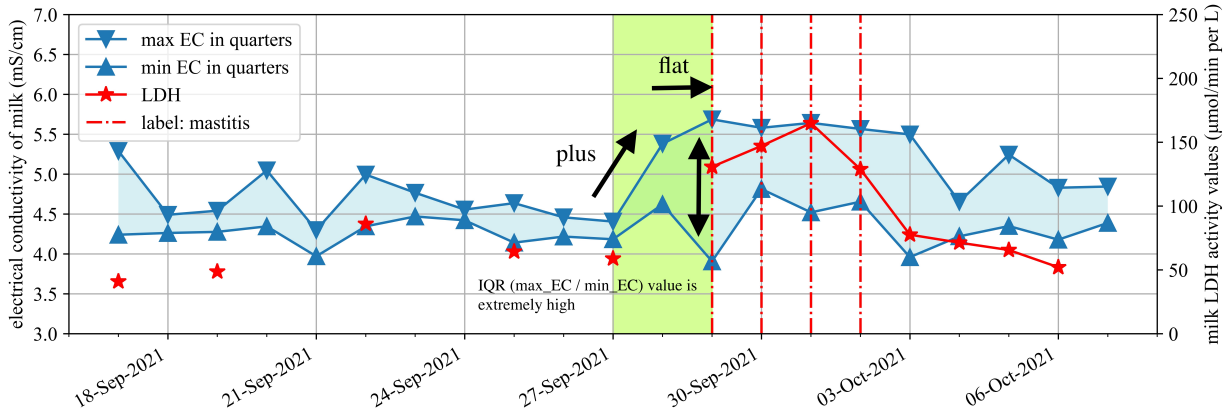


Figure 2: One of the subclinical mastitis cases corresponding to Rule 1.

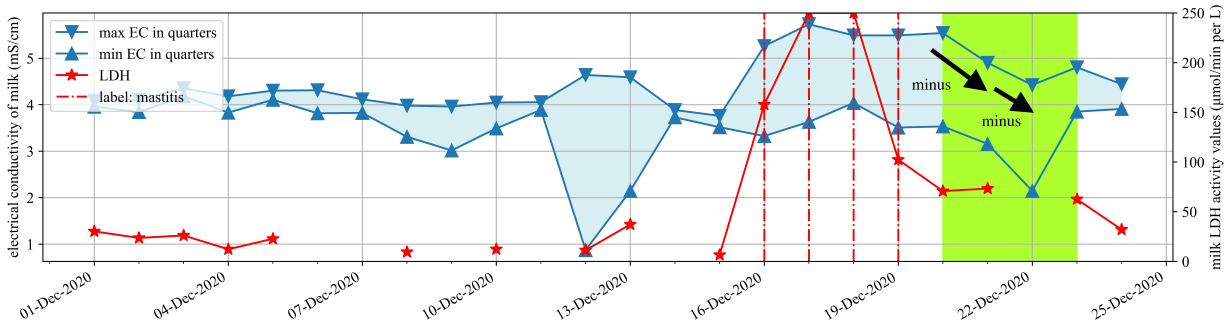


Figure 3: One of the subclinical mastitis cases in which the values of EC and LDH changed differently corresponding to Rule 2.

Rule2 pos(A) :- delta_maxEC(A, B, C, minus), delta_maxEC(A, C, D, minus)

Rule 2 indicates that cows whose max_EC values have been decreasing for two straight days have a risk of subclinical mastitis. Rule 2 apparently conflicts with Rule 1. However, this rule explains some cases of cows with subclinical mastitis. Figure 3 shows an example of cases in which the values of EC and LDH increased at disparate times. In this case, the mastitis risk values based on LDH, i.e., the target label of the classifier used in this study, become lower before the EC values began decreasing. However, the cow was deemed to be disordered by the farm staff and received veterinary treatment on December 23rd. Therefore, Rule 2 suggests that the trained classifier detected clinical cases that continued after the LDH values began decreasing.

The rules genereated by ILP practically interpreted the mastitis detection model in this study and provided explanation of the mastitis detection, which were available for users of this detection system, staff of the farm, to understand how to classify cows as mastitis or fine. Providing reason of diagnosis by machine learning to users of the models would accelerate to develop the models as well as support farmers control health of dairy cows efficiently.

5 Conclusion

In this study, a method for explaining the outputs from machine learning models using ILP was suggested and evaluated when applied to the task of subclinical mastitis detection for dairy cows. For an earlier detection of the onset of subclinical mastitis, a model for subclinical mastitis detection trained using risk values based on LDH was proposed. Interpretable rules were then generated using ILP to interpret the trained models. The rules obtained indicate that the trained classifiers detect mastitis cases depending on a certain variation of the EC values and that the EC and LDH fluctuate in different ways. The interpretable rules help in understanding the outputs of machine learning models and encourage a practical introduction of models as tools for decision making.

References

[1] H. Motohashi, H. Ohwada, C. Kubota, “Early detection method for subclinical mastitis in auto milking systems using machine learning,” in 2020 IEEE 19th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), 76–83, IEEE, 2020, doi:10.1109/iccicc50026.2020.9450258.

[2] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, “Lightgbm: a highly efficient gradient boosting decision tree,” in Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17, 3149–3157, Curran Associates Inc., Red Hook, NY, USA, 2017, doi:10.5555/3294996.3295074.

[3] M. T. Ribeiro, S. Singh, C. Guestrin, “” Why should I trust you?” Explaining the predictions of any classifier,” in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 1135–1144, 2016, doi:10.1145/2939672.2939778.

[4] S. M. Lundberg, S.-I. Lee, “A unified approach to interpreting model predictions,” in Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17, 4768–4777, Curran Associates Inc., Red Hook, NY, USA, 2017, doi:10.5555/3295222.3295230.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, "Grad-cam: visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE international conference on computer vision, 618–626, 2017, doi:10.1109/iccv.2017.74.

[6] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, "Smoothgrad: removing noise by adding noise," arXiv preprint arXiv:1706.03825, 2017, doi:10.48550/arXiv.1706.03825.

[7] N. P. Martono, K. Abe, T. Yamaguchi, H. Ohwada, "An analysis of motion transition in subtle errors using inductive logic programming: a case study in approaches to mild cognitive impairment," International Journal of Software Science and Computational Intelligence (IJSSCI), **10**(1), 27–37, 2018, doi:10.4018/ijssci.2018010103.

[8] S. Sasaki, R. Hatano, H. Ohwada, H. Nishiyama, "Estimating productivity of dairy cows by inductive logic programming," in Proceedings of The 29th International Conference on Inductive Logic Programming, 2019, doi:10.1007/978-3-030-49210-6.

[9] D. B. Jensen, H. Hogeveen, A. De Vries, "Bayesian integration of sensor information and a multivariate dynamic linear model for prediction of dairy cow mastitis," Journal of Dairy Science, **99**(9), 7344–7361, 2016, doi:10.3168/jds.2015-10060.

[10] W. Steeneveld, L. C. van der Gaag, W. Ouweltjes, H. Mollenhorst, H. Hogeveen, "Discriminating between true-positive and false-positive clinical mastitis alerts from automatic milking systems," Journal of Dairy Science, **93**(6), 2559–2568, 2010, doi:10.3168/jds.2009-3020.

[11] J. Rabold, M. Siebers, U. Schmid, "Explaining black-box classifiers with ILP–empowering LIME with Aleph to approximate non-linear decisions with relational rules," in International Conference on Inductive Logic Programming, 105–117, Springer, 2018, doi:10.1007/978-3-319-99960-9_7.

[12] J. Rabold, G. Schwalbe, U. Schmid, "Expressive explanations of dnns by combining concept analysis with ilp," in KI 2020: Advances in Artificial Intelligence, 148–162, Springer International Publishing, Cham, 2020, doi:10.1007/978-3-030-58285-2_11.

[13] H. Nishiyama, H. Ohwada, "Parallel inductive logic programming system for superlinear speedup," in International Conference on Inductive Logic Programming, 112–123, Springer, 2017, doi:10.1007/978-3-319-78090-0_8.

[14] S. Pyörälä, "Indicators of inflammation in the diagnosis of mastitis," Veterinary research, **34**(5), 565–578, 2003, doi:10.1051/vetres:2003026.

[15] A. J. Schepers, T. J. G. M. Lam, Y. H. Schukken, J. B. M. Wilmink, W. J. A. Hanekamp, "Estimation of variance components for somatic cell counts to determine thresholds for uninfected quarters," Journal of Dairy Science, **80**(8), 1833–1840, 1997, doi:10.3168/jds.s0022-0302(97)76118-6.

[16] C. M. Duarte, P. P. Freitas, R. Bexiga, "Technological advances in bovine mastitis diagnosis: an overview," Journal of Veterinary Diagnostic Investigation, **27**(6), 665–672, 2015, doi:10.1177/1040638715603087.

[17] Y. C. Lai, T. Fujikawa, T. Maemura, T. Ando, G. Kitahara, Y. Endo, O. Yamato, M. Koiwa, C. Kubota, N. Miura, "Inflammation-related microRNA expression level in the bovine milk is affected by mastitis," PLoS One, **12**(5), e0177182, 2017, doi:10.1371/journal.pone.0177182.

[18] M. Åkerstedt, L. Forsbäck, T. Larsen, K. Svennersten-Sjaunja, "Natural variation in biomarkers indicating mastitis in healthy cows," Journal of Dairy Research, **78**(1), 88–96, 2011, doi:10.1017/S0022029910000786.

[19] E. Norberg, H. Hogeveen, I. R. Korsgaard, N. C. Friggens, K. H. M. N. Sloth, P. Løvendahl, "Electrical conductivity of milk: ability to predict mastitis status," Journal of Dairy Science, **87**(4), 1099–1107, 2004, doi:10.3168/jds.s0022-0302(04)73256-7.

[20] I. Tomek, "An experiment with the edited nearest-neighbor rule," IEEE Transactions on Systems, Man, and Cybernetics, **SMC-6**(6), 448–452, 1976, doi:10.1109/TSMC.1976.4309523.

[21] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," IEEE Transactions on Systems, Man, and Cybernetics, **SMC-2**(3), 408–421, 1972, doi:10.1109/TSMC.1972.4309137.

[22] M. Khatun, P. C. Thomson, K. L. Kerrisk, N. A. Lyons, C. E. F. Clark, J. Molfino, S. C. García, "Development of a new clinical mastitis detection method for automatic milking systems," Journal of Dairy Science, **101**(10), 9385–9395, 2018, doi:10.3168/jds.2017-14310.

[23] D. Cavero, K.-H. Tölle, C. Henze, C. Buxadé, J. Krieter, "Mastitis detection in dairy cows by application of neural networks," Livestock Science, **114**(2-3), 280–286, 2008, doi:10.1016/j.livsci.2007.05.012.

[24] F. Mizoguchi, H. Ohwada, "Constrained relative least general generalization for inducing constraint logic programs," New Generation Computing, **13**(3), 335–368, 1995, doi:10.1007/bf03037230.

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 2, 149-155 (2022)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

COVIDFREE App: The User-Enabling Contact Prevention Application: A Review

Edgard Musafiri Mimo [1,*], Troy McDaniel[1], Jeremie Biringanine Ruvunangiza[2]

[1] *Arizona State University, Systems Engineering, The Polytechnic School, Mesa, 85212, USA*

[3] *University of Mons, Computer Science, The Polytechnic School, Mons, 7000, Belgium*

ARTICLE INFO

Article history:
Received: 11 February, 2022
Accepted: 16 March, 2022
Online: 19 April, 2022

Keywords:
COVIDFREE App
Location-based services
Social distancing
Covid-19
Location tracking

ABSTRACT

The use of Covid-19 contact tracing applications has become almost irrelevant now that several flavors of Covid-19 vaccine have been developed and are constantly being distributed to people during the pandemic to help alleviate the need for lockdowns. Also, the availability of at-home testing kits and testing sites means that people do not need to contact trace as much since individuals can get tested and follow the health guidelines in ensuring their health and the safety of others around them. Nevertheless, governments around the world are still faced with the Covid-19 pandemic challenge because the virus is not yet controlled due to the different variants and the rapid contamination rate that outpaced the logistic supply chain processes in the distribution of the vaccines and the time it takes in convincing individuals to take the vaccine swiftly to reach herd immunity. Therefore, the current pressing need is that of addressing the infection rate by finding ways and solutions to minimize or slow down contamination among people especially with the increased number of variants. This paper is an extension of the "COVIDFREE App: The User-Enabling Contact Prevention Application" work originally presented in 2020 IEEE International Symposium on Technology and Society (ISTAS) conference that provided a smartphone application architecture with the goal of proactively enabling users to avoid encountering infected Covid-19 patients. This paper elucidates and discusses additional concerns not thoroughly addressed previously regarding the Covid-19 variants, vaccines, booster, and infection rates, and demonstrates the feasibility of the proposed architecture with a web application prototype. This paper also discusses the benefits of funding and developing contact tracing and prevention applications, such as the COVIDFREE App, to provide the needed ingredient in reducing the infection rate and provide citizens the needed preparedness and relief in actively fighting the virus.

1. Introduction

The world has made so many financial, social, political, and technological efforts and investments in fighting and controlling the Covid-19 virus spread as a remedy to end the virus fueled pandemic the world was forced in. All the efforts were made as more information about the virus became available starting with how one can avoid catching the virus to what one must do once infected to recover and not spread the virus to others and so on and so forth. The declared Covid-19 International Public Health Emergency Issue that began in 2020 has drastically changed the world, and there is no going back to normal anymore rather there is a contemplation of adopting a new normal. Covid-19 is a highly contagious respiratory virus with several unique features that have been identified with the ever-increasing number of infection cases to date.

With more detail about Covid-19 available today on how some infected individuals do not develop symptoms right away, while others stay asymptomatic throughout the duration of their treatment, and others are able to catch the different mutation and variation of the virus, it is important to evaluate all the alternative remediation solutions that can be instigated to manage the pandemic health crisis more efficiently. Several governments around the world have responded by exploring and seeking to harness technology in the fight against this fatal illness. Covid-19 has influenced the lives of almost everyone on the earth by forcing governments around the world to implement lockdowns and sanctions, and in certain situations, enforce measures that involve

* Corresponding Author: Edgard Musafiri Mimo, emusafir@asu.edu

work-from-home regulations, implement strong physical distancing safeguards, and set up emergency health responses that necessitate thorough rearrangement to perform mass testing, patient managing, and care giving [1].

These efforts are aimed at containing and controlling the virus's transmission until definitive cures or vaccines are produced and widely administered. Considering the effort and investment in producing the vaccines and the vaccine boosters, and getting people vaccinated and boosted, there is still a lot of work to be done in combating the virus and ensuring people's safety [2]. As a result, it is evident that the Covid-19 testing and vaccine solutions are maybe the remedy for the virus, but not the remedy of its infection rate. They simply cannot keep up with the pace of transmissibility of the virus as the supply chain logistic processes are also impacted by the pandemic. Hence, there is a need to address the contamination issues among the people by ensuring the prevention of contacts or encounters among people proactively because potential Covid-19 variants may continue to develop due to the constant virus reproduction among humans [3].

This is the intention of the COVIDFREE app because governments all over the world are inclined to adopt and employ mobile contact tracing and awareness applications to automatically manage, trace, and investigate recent interactions of both the newly tested Covid-19 infected individuals and those that are recovering from it regardless of vaccine status [4]. The potential use of such mobile and web applications has produced numerous discussions surrounding privacy, security, data management, contact projection algorithms, and cyber-attack vulnerabilities. The previously proposed smartphone application architecture that COVIDFREE employs aims to minimize the above-mentioned concerns by using only minimal information about the users and provide abstraction layers to ensure the security and the privacy of users are preserved in the process of giving the users more control and privacy of their data [1].

The COVIDFREE application aims to enhance users' situational mindfulness through communication of unsafe locations and proactively prompt them to avoid these locations while dynamically taking into the account the daily reported infection rate per location. Additionally, the COVIDFREE application provides the means of customizing by allowing users the control of considering their Overall Risk Density Safety Factor in the risk calculation based on health requirements and numerous customizable user-specific situations [1]. This paper discusses a prototype of a previously proposed proactive smartphone app with a centralized approach that aids uninfected people regardless of their vaccination status to overcome the stress and fear of getting contaminated by improving their situational awareness. The proposed proactive smartphone application enables users to informatively avoid congested places and tested infected individuals because the application notifies them whenever they are within 10 to 50 feet (depending on parameters) of an anonymous, confirmed infected person [1].

2. Insights and Concerns

2.1. General Anxieties

There are several unsolved questions about how technology can help provide a remedy in addressing the Covid-19 transmissibility issues even now that the vaccines and vaccine boosters are available, since they are not yet distributed everywhere due to their demands and the citizens' decision in taking them. The questions surrounding the quarantine period and the habits and behaviors of citizens in how they choose to live their lives, whether it is by masking up or just limiting their movements outside of their safe locations, cannot be overlooked. The simple fact that many Covid-19 patients are asymptomatic leaves the world with so little options to consider in providing the remediation needed to enable the minimization of the transmission of the virus from one patient to others [1], [5].

If some patients do not have symptoms right away and some are symptoms free while carrying the virus, then the Covid-19 infection rate would remain a burden to carry as more undercover carriers could prove difficult to identify and avoid [5]. As a result, Covid-19 testing for asymptomatic individual is problematic since infected asymptomatic individual may have already spread the virus to numerous others before being tested. Consequently, it is critical to guarantee that individuals feel secure going to and returning from testing centers by understanding how infectious their surrounding is. For instance, regardless of symptoms, it is beneficial for users to know how close they may be to someone who is contaminated and has had their infection confirmed by a previously taken Covid-19 test.

2.2. Tracking Complexity

The complexity of Covid-19 spread is vast, so there is a need of a proactive response strategy that locates confirmed infected people outside their safe location and provides means of avoiding them anonymously without any form of profiling through direct notification of users' surrounding exposure range [1], [6]. This proactive strategy is urgently needed to ensure people feel safer and more educated about their environments when it comes to Covid-19 health risks. We believe that proactively notifying people to avoid encountering a person that has tested positive for Covid-19 will help minimize the infection rate and allow the vaccines and boosters' effects to be effectively noticed as herd immunity is being achieved.

Reactive techniques of contact tracing or a passive approach of waiting on the vaccines and boosters only, would not provide the rapid remedy that people are waiting for as long as the contamination rate stays high and unaddressed in a proactive manner that of relying on informative individual discretional social distancing and isolation. The only possibilities left imminent involve the tradeoff between what privileges one can live with and what necessities one cannot live without as new Covid-19 demands need to be satisfied and new habits developed to cope with the world's present changes of the new normal. Hence, it is necessary to provide tools that give users insights and proactive prompt notifications to avoid the symptoms free carriers and minimize the potential risk of catching the virus. This is a great way to make use of the collected testing daily data to ensure the location of the individuals that test positive are activated until they test negative depending on their vaccination status and their quarantine periods. It will ensure other users would promptly avoid them should interactions' occasions arise.

Therefore, the method of preventing anyone from getting the virus from the afflicted person regardless of their vaccine status

that the COVIDFREE application provides, ensures people are informed and at peace with their movements while minimizing their concerns regarding the virus and their health and safety. With a couple of years in the pandemic, there are still some concerns regarding Covid-19 transmission and propagation that are yet to be completely understood regarding the variety of ways the virus can spread [6]. Thus, it is critical to access the user's physical surroundings, social connections, and health to determine the likelihood of being contaminated. The virus's complicated transmission characteristic generates a lot of concerns among the people. As a result, recommendations to maintain social distance and wear personal protective equipment (PPE) help to limit the virus's transmission. The case that someone can get the virus on their way to the testing site and then obtain a negative test report to become an undercover carrier in the testing process. Being uninformed about one's immediate surrounding adds to the concerns and does not help answer the question of when the virus is spread especially with the case of vaccinated individuals' potential to be contaminated again.

Another issue that makes the concerns worse is the amount of time the virus takes to disperse in the air, which varies based on area and spaces' ventilation systems. Thus, without the full visibility of the virus' transmitters, it's hard to identify, track, and locate unsafe sites accurately without depending on the tracking of individuals who have tested positive. This is a way that can help provide insights to address the question of where someone can potentially get Covid-19 by being in the vicinity of an active carrier (person to person) rather than a passive one (via the environment). It is convoluted to make the right informed decision regarding the virus's transmissibility during travel if critical information and notifications are not available to people in real time to support informed travel experiences.

3. Simplified Prototype Architecture

The overall density safety factor proposed earlier must be adjusted accordingly when considering the vaccine safety factor to determine how to account for the efficacy of the vaccine and their booster to provide a realistic vaccination safety factor [1]. The proposal may be using the normalized reported average efficacy value per age group from a credible vaccine approval institution like the Food and Drug Administration (FDA) agency. Nevertheless, the previously proposed normalized overall density safety factor along with all the considered factors as discussed in this paper remain a great indicator and tool to dynamically adjust per users' needs the notification distance range to facilitate users in avoiding encountering a Covid-19 carrier. The overall density safety factor can also be coupled with other factors like reported infection rates to provide an enhanced user's situational representation.

The previously proposed architecture required the necessity of the health center experts in providing the positive test data to the database and registering the Covid-19 carrier for tracking. To test the prototype, the requirement of the health center experts was circumvented, and the database was loaded with fictitious Covid-19 carriers with actual locations against which the actual devices were tried to ensure they are proactively notified to avoid coming in contact. The current prototype implementation provides a more privacy enhanced solution by circumventing the necessity of the health center interference of user's medical status as it pertains to Covid-19 and their personal risks. Thus, it increases users' privacy by eliminating the direct linkage of users' medical records and their current Covid-19 test results and vaccine status as well as facilitate the usability of the application's rapid prototype. Figure 1 below shows an adapted architecture of the previously proposed centralized architecture that is used for the prototype.

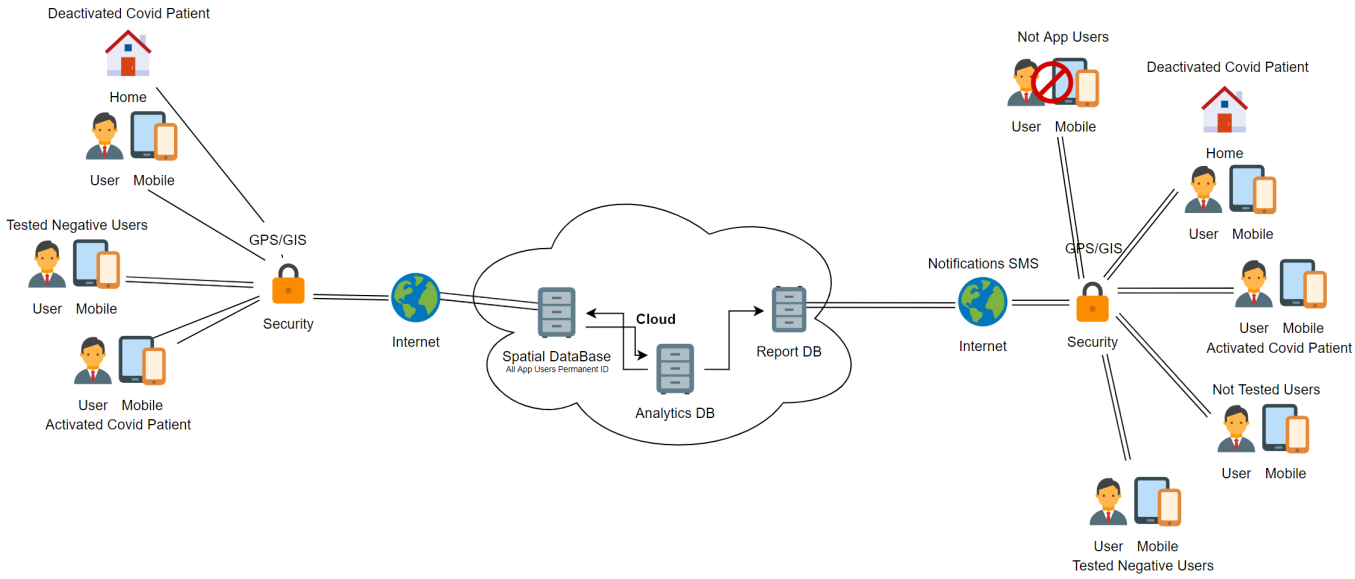


Figure 1: A diagram of the adapted architecture of the COVIDFREE centralized architecture prototype created using flat-color-icons.xml from the app.diagrams.net website.

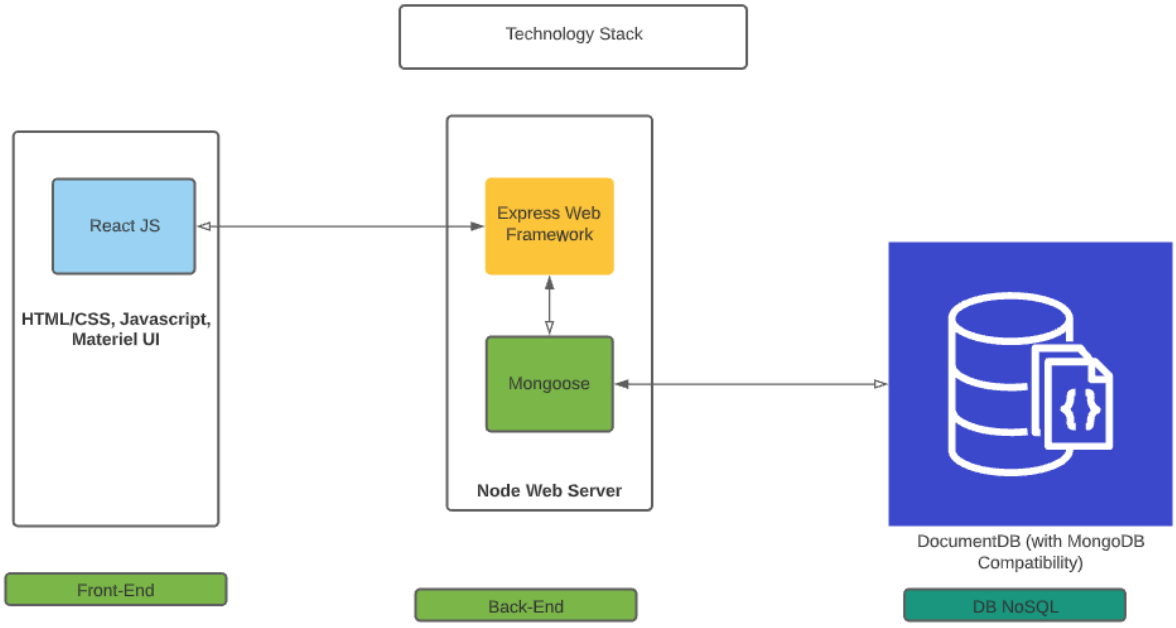


Figure 2: An illustration of the technology stack enabling the COVIDFREE application prototype created using flat-color-icons.xml from the app.diagrams.net website.

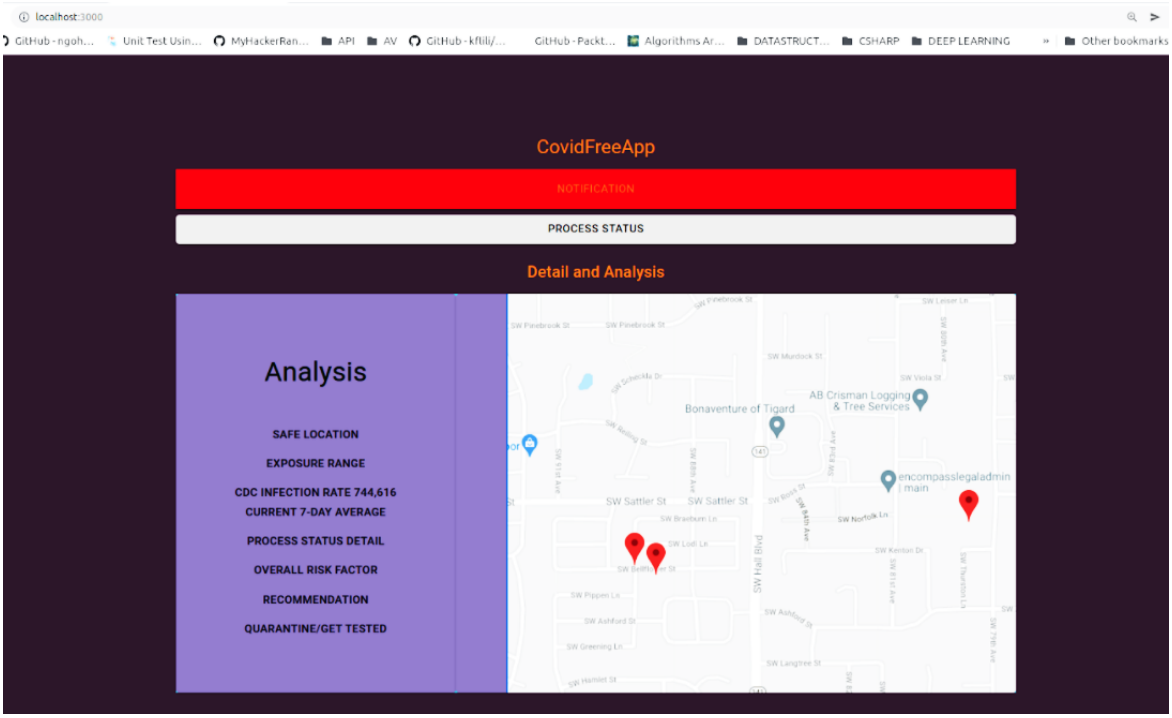


Figure 3: A screenshot showing the prototype of the COVIDFREE application's user interface

4. Used Technologies

To ensure the feasibility of the application in providing a prototype demo that other developers can use, many web application frameworks were considered for this project like Django, Ruby on Rails, MEAN and MERN frameworks. We decided to go with the MERN stack. MERN stands for MongoDB, Express, React, and Node, after the four key technologies that make up the stack. The technology stack enabling the prototype functionalities is shown in Figure 2 above.

The reason for choosing the MERN stack is that it provides an opportunity for rapid prototypes and proof of concepts. JavaScript is the primary programming language that is used in this project, and it offers the benefit of using one language for both the prototype's front-end and back-end. We used MongoDB to save the collected data. MongoDB is a document database model that maps to how developers think and code, and it provides a powerful, unified query API. MongoDB powers faster and more flexible applications. Node and Express are used for the back-end

implementation. Node is an asynchronous event-driven JavaScript runtime. It is designed to build scalable network applications.

5. App User Interface

Considering the front-end or the user interface UI, we used React. React is a free and open-source front-end JavaScript library for building user interfaces based on UI components.

We collect the data through the UI. Most data are collected dynamically without users' interaction. One of the most important data points that is collected is the user's location, which is associated with the user's registered phone number. We use the user's location to determine the safety of his position relative to nearby potential Covid-19 carriers while considering the infection rate in the user-specific location to determine the proper notification distance range.

The application's algorithm 1 shown below computes the best-case scenario for the end-user to avoid getting contaminated in the first place.

5.1. Algorithm

The application's algorithm as implemented in the code is designed with the virus prevention approach in mind. The source code is accessible on GitHub at link[1]. The prototype UI is shown Figure 3.

Algorithm 1: User's Covid-19 Prevention Situational Response

```
Result: Action Recommendation with Warnings
Get User Safe Location;
Get User Current Position;
Get Users' Geolocation in Same Zip Code;
while User Not in Safe Location do
    Get Closest Covid-19 Carrier Position;
    Calculate Social Distance;
    if Social Distance greater than 100 Feet
    then
    | GREEN Notification;
    else if Social Distance greater than 50 Feet
    then
    | ORANGE Notification;
    else if Social Distance less than 50 Feet
    then
        | RED Notification;
    else
        | Break;
    end
end
```

6. App Demo and Use Cases

The prototype explores four different scenarios as use cases to ensure proactive notification to users to enable enough time for users to take appropriate actions based on their specific scenarios.

6.1. User in Safe Location

The first scenario involves the user being in their safe location or within 25 feet of the safe location. When this is the case, the user is deactivated and is no longer tracked. The user is considered safe and providing no apparent threats to anyone regardless of their Covid-19 status since the safe location is also considered the quarantine location where the user is in isolation from the outside world. The user interface of this scenario shows the information and the notification that the users get in real time by having access to the app.

The notification button is triggered to a grey color notification, and only the user's home or safe location is shown on the map as the user is not in motion. The user can access all the analysis detail on what the process status means regarding all the notifications in detail by using the appropriate provided buttons on the application's analysis screen. The COVIDFREE application in this case recommends the user to avoid unnecessary trips to ensure they remain safe. The screenshot in Figure 4 below shows a representation of what the COVIDFREE application looks like with test demo data for this use case.

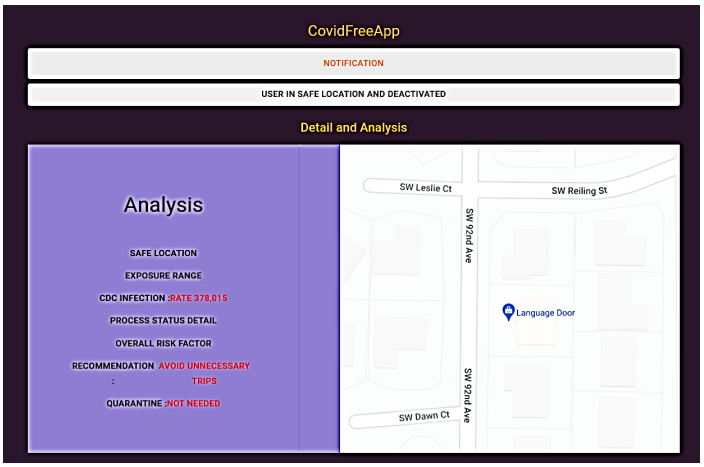


Figure 4: A screenshot showing the COVIDFREE application UI for a user in their safe location.

6.2. User out of Safe Location

The second scenario involves the user leaving their safe location and entering a state of motion moving from one location to another. When this is the case, the user is activated on the map and their position is tracked relative to his safe location and other potential Covid-19 threats agents or carriers. The user's test condition is considered when they are in motion relative to the potential Covid-19 carriers as the user runs a risk of interacting with a potential carrier. If the user is also a Covid-19 carrier then anonymously their location is also being flagged to be avoided by other users of the application providing a safe environment for all the application's users. The COVIDFREE application's user interface of this scenario shows the information and the notification that the users get in real time by having access to the app. The notification button is triggered to a green color notification if the user has not been in a vicinity of a Covid-19 carrier.

In this view, the user's current location and his safe location can be seen on the map interface of the application indicating the user is in motion and out of their safe location, and thus the user is

[1] https://github.com/Unitersity2021/Covid-free-app

active. The user can access all the analysis detail on what the process status means regarding all the notifications in detail by using the appropriate provided buttons on the application's analysis screen. The COVIDFREE application in this case recommends and reminds the user to social distance to ensure they remain safe throughout their trips. The screenshot in Figure 5 below shows a representation of what the COVIDFREE application looks like with test demo data for this use case.

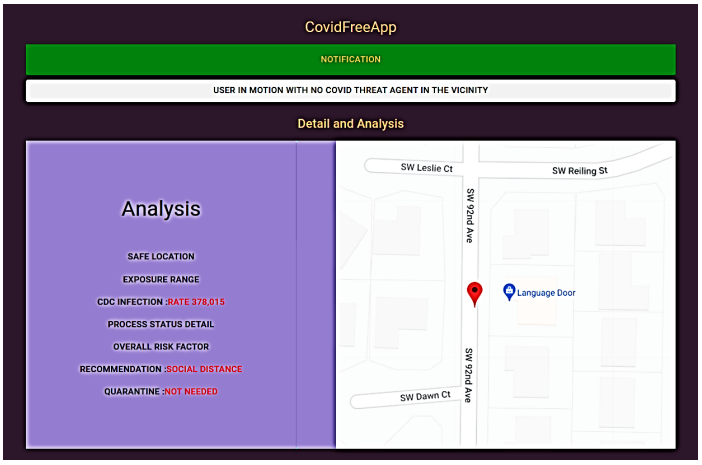


Figure 5: A screenshot showing the COVIDFREE application UI for a user outside their safe location and in motion, approaching no potential Covid-19 threat agents within their 100 ft radius.

6.3. User in motion with Covid-19 threat agents far away

The third scenario involves the user in a state of motion moving from one location to another and coming in the vicinity of an active Covid-19 carrier who is at a relatively far away distance of more than 50 feet from the user. When this is the case, the user remains activated on the map and their position is tracked relative to their safe location and the position of the potential Covid-19 threat agents or carriers. The user's Covid-19 test condition status is considered when they are in motion relative to the potential Covid-19 carriers as the user runs a risk of interacting with a potential Covid-19 carrier.

If the user is also a Covid-19 carrier then anonymously their location is also being flagged to be avoided by other users of the application providing a safe environment for all the application's users. The COVIDFREE application's user interface of this scenario shows the information and the notification that the users get in real time by having access to the app. The notification button is triggered to a yellow color notification since the user is in the vicinity of a Covid-19 carrier that is within their 100 feet radius but more than 50 feet away from them.

In this view, the user's current location and the Covid-19 carrier locations are shown on the map for the demonstration's sake and his safe location can also be seen on the map interface of the application when zoomed out indicating both that the user is in motion and out of their safe location as well as in the vicinity of Covid-19 threat carriers. The user can access all the analysis detail on what the process status means regarding all the notifications in detail by using the appropriate provided buttons on the application's analysis screen. The COVIDFREE application in this case recommends and reminds the user to consider quarantining when they return to their safe location. The screenshot in Figure 6 below shows a representation of what the COVIDFREE application looks like with test demo data for this use case.

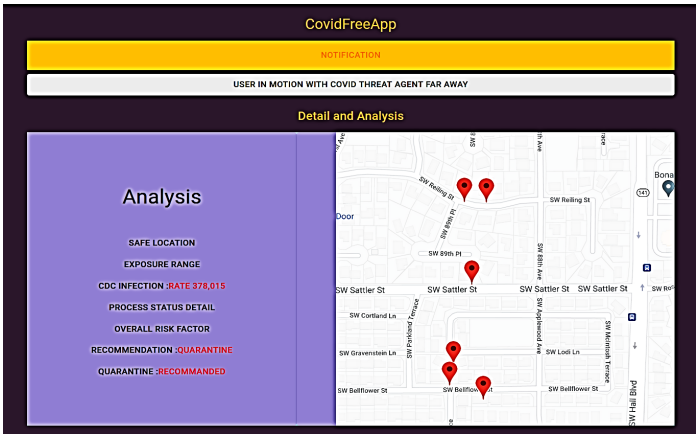


Figure 6: A screenshot showing the COVIDFREE application UI for a user outside their safe location and in motion, approaching potential Covid-19 threat agents that are relatively far away (more than 50 ft apart).

6.4. User in motion with Covid-19 threat agents nearby

The fourth scenario involves the user in a state of motion moving from one location to another coming in the vicinity of an active Covid-19 carrier who is relatively close by the user. When this is the case, the user remains activated on the map and their position is tracked relative to their safe location and potential Covid-19 threat agents or carriers. The user's test condition is considered when they are in motion relative to the potential Covid-19 carriers as the user runs a risk of interacting with a potential carrier. If the user is also a Covid-19 carrier then anonymously their location is also being flagged to be avoided by other users of the application providing a safe environment for all the application's users. The COVIDFREE application's user interface of this scenario shows the information and the notification that the users get in real time by having access to the app.

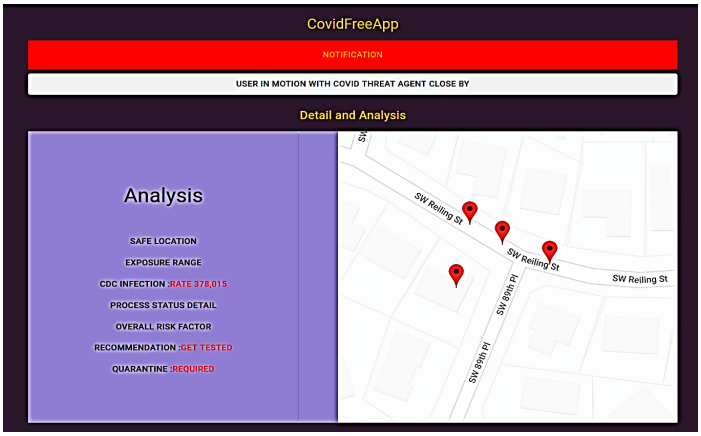


Figure 7: A screenshot showing the COVIDFREE application UI for a user outside their safe location and in motion, approaching potential Covid-19 threat agents that are relatively nearby (Less than 50 ft apart).

The notification button is triggered to a red color notification since the user is in the vicinity of a Covid-19 carrier that is relatively nearby. In this view, the user's current location and the Covid-19 carrier's location is shown on the map for demonstration's sake and their safe location can also be seen on

the map user interface of the application when zoomed out indicating both that the user is in motion and out of their safe location as well as in the vicinity of a Covid-19 threat carrier. The user can access all the analysis detail on what the process status means regarding all the notifications in detail by using the appropriate provided buttons on the application's analysis screen. The COVIDFREE application in this case recommends and reminds the user to consider getting tested and quarantining when they return to their safe location. The screenshot in Figure 7 below shows a representation of what the COVIDFREE application looks like with test demo data for this use case.

7. Conclusion

The current prototype presented in this paper demonstrates that the proof of concept for proactive notifications is feasible and can be further enhanced to provide efficient ways to process the users' data and provide swift notification to users while collecting the minimum amount of data from the users and providing multiple layers of abstraction to ensure security and privacy of the individuals and data quickly and optimally. To counteract the propagation of Covid-19 and increase citizens' safety and peace of mind, this article demonstrates the feasibility of COVIDFREE APP by creating a working prototype as a complementary and proactive technological solution that can minimize the likelihood of citizens contracting the Covid-19 virus.

The prototype application employs the previously proposed centralized architecture design to assist users in making educated decisions about how to comfortably and safely navigate from one location to another as well as when they can safely leave areas of isolation (such as their homes) and their immediate social groups. The prototype achieved the goal of improving users' situational alertness of high-risk sites around them. With better situational mindfulness, it is expected that users will likely feel more convinced and secure about their conduct, flexibility, and travel plans.

We hope this work stimulates parallel efforts to guarantee citizens leverage the available technologies to advance the citizen's safety and security, and eventually, to save citizens' lives. There are some opportunities for further developments. Our model gives a unique framework that is easy to use and configure for different machine learning models. For instance, one can implement a federated learning [7] model to optimize an independent user situational model and further enhance the user safety. As the application offers a real-time user notification to prevent contracting the virus and ensures safety, it also ensures the integrity of the information, and the trustworthiness of the data can be accessed using the zero-knowledge proof technique [8] even though the zero-knowledge proof technique is complex and in its early research stage. Nevertheless, it would provide a great way to gain public trust due to its security and privacy awareness implications.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors thank Arizona State University and the National Science Foundation for their funding support under Grant No. 1828010.

References

[1] E.M. Mimo, T. McDaniel, "COVIDFREE App: The User-Enabling Contact Prevention Application," in 2020 IEEE International Symposium on Technology and Society (ISTAS), IEEE: 452–456, 2020, doi:10.1109/ISTAS50296.2020.9462186.

[2] S. Shiehzadegan, N. Alaghemand, M. Fox, V. Venketaraman, "Analysis of the Delta Variant B.1.617.2 COVID-19," Clinics and Practice, **11**(4), 778–784, 2021, doi:10.3390/clinpract11040093.

[3] J.A. Plante, B.M. Mitchell, K.S. Plante, K. Debbink, S.C. Weaver, V.D. Menachery, "The variant gambit: COVID-19's next move," Cell Host & Microbe, 29(4), 508–515, 2021, doi:10.1016/j.chom.2021.02.020.

[4] N. Ahmed, R.A. Michelin, W. Xue, S. Ruj, R. Malaney, S.S. Kanhere, A. Seneviratne, W. Hu, H. Janicke, S.K. Jha, "A Survey of COVID-19 Contact Tracing Apps," IEEE Access, **8**, 134577–134601, 2020, doi:10.1109/ACCESS.2020.3010226.

[5] K. Michael, R. Abbas, R.A. Calvo, G. Roussos, E. Scornavacca, S.F. Wamba, "Manufacturing Consent: The Modern Pandemic of Technosolutionism," IEEE Transactions on Technology and Society, **1**(2), 68–72, 2020, doi:10.1109/TTS.2020.2994381.

[6] Y.-C. Wu, C.-S. Chen, Y.-J. Chan, "The outbreak of COVID-19: An overview," Journal of the Chinese Medical Association, **83**(3), 217–220, 2020, doi:10.1097/JCMA.0000000000000270.

[7] H.B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," 2016.

[8] S. Grzonkowski, W. Zaremba, M. Zaremba, B. McDaniel, "Extending web applications with a lightweight zero knowledge proof authentication," in Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology - CSTST '08, ACM Press, New York, New York, USA: 65, 2008, doi:10.1145/1456223.1456241.

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 3, 152-160 (2022)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Deep Learning Affective Computing to Elicit Sentiment Towards Information Security Policies

Tiny du Toit*, Hennie Kruger, Lynette Drevin, Nicolaas Maree

School of Computer Science and Information Systems, North-West University, Potchefstroom, 2531, South Africa

ARTICLE INFO

Article history:
Received: 31 January, 2022
Accepted: 06 June, 2022
Online: 27 June, 2022

Keywords:
Affective computing
Deep learning
Information security policies
Non-compliance
Sentiment analysis

ABSTRACT

Information security behaviour is an integral part of modern business and has become a central theme in many research studies. One of the essential tools available that can be used to influence information security behaviour is information security policies (ISPs). These types of policies, which is mandatory in most organisations, are formalised rules and regulations which guide the safeguarding of information assets. Despite a significant number of ISP and related studies, a growing number of studies report ISP non-compliance as one of the main factors contributing to undesirable information security behaviour. It is noteworthy that these studies generally do not focus on the opinion of users or employees about the contents of the ISPs that they have to adhere to. The traditional approach to obtain user or employee opinions is to conduct a survey and ask for their opinion. However, surveys present unique challenges in fake answers and response bias, often rendering results unreliable and useless. This paper proposes a deep learning affective computing approach to perform sentiment analysis based on facial expressions. The aim is to address the problem of response bias that may occur during an opinion survey and provide decision-makers with a tool and methodology to evaluate the quality of their ISPs. The proposed affective computing methodology produced positive results in an experimental case study. The deep learning model accurately classified positive, negative, and neutral opinions based on the sentiment conveyed through facial expressions.

1. Introduction

The importance of information security behaviour and the challenges associated with using information security policies (ISPs) as a management tool to ensure that employees and users comply with security requirements is a widely studied discipline. This paper addresses specific concerns and techniques that may assist in evaluating ISPs and is an extension of the work initially presented at the 2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC) [1]. This paper is also partially based on a master's degree study done in Computer Science [2].

Information security behaviour forms part of the general information security discipline and refers to the protection of information and information technology assets [3]. The human behaviour element of information security has become an integral part of modern enterprises, and considerable amounts of effort are often assigned to ensure that information security awareness, ISPs and other relevant human aspects are sufficiently addressed [4]. Technical solutions for undesirable human information security behaviour play an essential role [5] but are generally inappropriate on their own [6]. Additional measures to address the behaviour problem effectively are necessary. One approach often employed to influence security behaviour is ISPs [7], [8]. The popularity of ISPs as a control measure has inspired many studies with new research that is regularly added to the information security discipline [9]-[11].

Despite a large number of ISP and related studies, there is still a significant number of problems such as the inefficient use or non-compliance to ISPs that are regularly reported in the literature. Behavioural problems are evidenced by phenomena such as the privacy paradox [12] and the knowing-doing gap [13]. Users with a high level of information security awareness are easily persuaded to reveal personal or confidential information. Literature resources also indicate that one of the major contributing factors influencing the effective use of an ISP is the general lack of compliance [14], [15]. The work of [16] also presents a systematic overview of studies related to ISP compliance. Moreover, the lack of ISP compliance has also led to studies investigating the use of

*Corresponding Author: Tiny du Toit, North-West University, South Africa
Tel: +27828472512 E-mail: Tiny.DuToit@nwu.ac.za

https://dx.doi.org/10.25046/aj070317

psychological models to explain information security behaviour [17], [18].

It is clear from the above that many research projects are continuously conducted to evaluate and explain different aspects of ISP compliance. However, despite this large number of studies, little attention is given to the opinion of employees or users about the ISPs that they have to adhere to. For an ISP to be successful, employees should buy into the contents of the ISP and should have a positive attitude towards the contents – if not, non-compliance is likely to remain a reality. Two traditional methods to obtain the opinion of people or workers are to ask them or physically observe their behaviour. However, in addition to logistical difficulties (specifically to monitor employees), both techniques are subjected to biased results. During observation, users may comply with an ISP out of fear or merely because they know it is expected. Direct questioning through interviews or surveys also presents similar problems such as response bias, where answers may be faked [19]. In an attempt to address the bias problem, sentiment analysis, also known as opinion mining [20], is often employed. This technique enables decision-makers to determine whether someone has a positive, negative or neutral opinion or attitude about something through an analysis of personal sentiment information. Text-based sentiment analysis is a popular approach to determine someone's sentiment [21]. However, an ISP may still be subjected to response bias when a user simply writes down what is expected. To address this problem, affective computing may be used to perform sentiment analysis. Affective computing is a computational approach that aims to diagnose and measure emotional expression [22] and then use these measurements to evaluate human behaviour [23]. The technique can determine a user's opinion without asking any questions, thereby removing the risk of social desirability.

In this paper, the aim is to employ affective computing and sentiment analysis to address response bias problems and contribute to evaluating the quality of ISPs. The results would assist management in positively addressing challenges within ISPs and timely assessing and changing the contents of an ISP. The remainder of the paper is structured as follows. In Section 2, a brief overview of ISPs will be given, while background information on sentiment analysis and affective computing will be presented in Section 3. In Section 4, deep learning, which forms the basis of the experimentation, will be addressed. The experimental design of an illustrative case study will be discussed in Section 5, with the results and a reflection presented in Section 6. The paper will be concluded in Section 7 with some final remarks.

2. Information Security Policies

There are several definitions in the literature for an ISP. The authors of [24] provide a basic description by referring to an ISP as a set of rules and regulations that inform users of their responsibilities to safeguard information technology assets. A more formal definition at an organisational level is given by [25] as "a set of formalised procedures, guidelines, roles and responsibilities to which employees are required to adhere to safeguard and use properly the information and technology resources of their organizations". The importance of an ISP is also confirmed in internationally accepted information security standards such as the ISO/IEC 27002 standard which defines the objective of an ISP as "to provide management direction and support for information security in accordance with business requirements and relevant laws and regulations" (Source: www.iso.org/standards.html). These formal information security standards also prescribe ISPs as mandatory for information security management [26], and auditors are regularly advised to review the understanding and compliance of ISPs to ensure that users maintain acceptable levels of information security behaviour [27].

There is a general consensus that an ISP plays a critical role in any organisation. The researchers of [28] argue that effective information security management in organisations is largely dependent on the adherence to ISPs, while [4] state that the long-term success of any organisation in the current global and digitally driven economy is determined by the creation, deployment and enforcement of ISPs. However, there still seems to be an ongoing problem in ISP compliance. Large numbers of studies are found in the literature that try to explain and even predict the non-compliance of ISPs. Examples of such studies include the work of [24], who propose a model to raise the level of ISP compliance amongst end-users; [28], to predict ISP compliance, proposed a theoretical model that links security-related stress, discrete emotions, coping response and ISP compliance; and [29] who performed a study where aspects of the theory of planned behaviour and ISP compliance were investigated. Other examples of studies that employ psychological models to explain non-compliance can be found in [30], [31]. In addition to the existing non-compliance problem, it is also clear from the literature that employees and users are affected by the quality of an ISP. The scholars of [32] argue that the general quality of an ISP will affect employee satisfaction and ultimately plays a significant role in ISP compliance. This poses another question on how to determine employee or user satisfaction with an ISP. As alluded to in the introduction, the answer may be to simply ask employees for their opinion on the ISP. This, however, is not an easy task as different problems such as social desirability may render results invalid.

Social desirability is defined as the tendency to answer questions acceptable rather than truthful [33]. It is a significant problem in situations where opinions are solicited, and numerous studies exist on various aspects of applications and ways to address any adverse effects [34]-[36]. Social desirability is also applicable in information security, such as information security behaviour [19] and information security awareness evaluations [37]. The work by [38] is of particular interest as this research study has proved that response bias exists in current scale measurements used in compliance research. As a result, the findings of several studies in policy compliance may be questionable. To overcome these problems, this paper aims to introduce sentiment analysis and affective computing to exclude possible response bias when evaluating the quality of an ISP. A brief introduction to sentiment analysis and affective computing is presented in the next section.

3. Sentiment Analysis and Affective Computing

Opinions, like emotions, play an important role in human decision-making; thus, emotion recognition and sentiment analysis are critical for determining user or consumer preferences and opinions. Furthermore, sentiment analysis can enhance organizational functions such as sales and marketing by allowing

researchers to better understand consumers' preferences and behaviours [39]. Affective computing is a relatively recent method for computationally identifying and measuring emotions to adapt decisions to support people's emotional states. Therefore, in this paper, affective computing is suggested to analyse the opinions of employees or users towards ISPs. This will allow information security administrators to develop high-quality ISPs with high user satisfaction while excluding problems like social desirability from the opinion survey process.

3.1. Sentiment Analysis

Sentiment analysis is a method for analyzing people's feelings or opinions towards an entity [40]. Text-based sentiment analysis has an extensive body of knowledge, and studies in this field are performed regularly [40], [41]. These studies, however, remain difficult because they require a deep understanding of language, both in terms of semantics and syntax [42]. Therefore, it has become a more common practice to perform sentiment analysis using videos rather than text. The advancement and availability of communication technology (i.e. consumers who tend to record their opinions on products using a webcam and then upload the videos to social media platforms) are two reasons for this trend, according to [39]. Videos also provide multimodal data, such as vocal and visual modalities, contributing to more accurate emotion and sentiment models. The fundamental task of video sentiment analysis is to detect, model, and exploit the sentiment conveyed by facial gestures, as shown in numerous instances in the literature [42], [43]. Extracting emotions for sentiment analysis is a well-known task in affective computing, which will be addressed in more detail in the next section.

Figure 1: Emotions as represented by facial expressions.

3.2. Affective Computing

Affective computing is described by [44] as techniques for detecting, recognising, and predicting human emotions such as anger, fear, disgust, surprise, pleasure, and sadness. It is a branch of artificial intelligence dealing with creating or adapting computational systems to offer decision support depending on an individual's emotional state. Emotions may be identified by observing facial expressions, followed by a feature extraction process, which is then used to classify emotions. Figure 1 (Source: https://www.linkedin.com/pulse/scientific-tactics-boost-non-verbal-communication-body-rokham-fard/) is an example of the six fundamental universally distinctive emotions [45] as represented by facial expressions.

The data used in this paper's experimental case study is similar to the facial expressions presented in Figure 1 and consists of videos of people reading various text passages to prompt a particular sentiment. However, computational requirements dictate that the affective data be converted and represented quantitatively. This quantification process was performed using the Affectiva Software Development Kit (SDK) [46]. The Affectiva system is a reliable affective computing tool trained on more than 7.5 million faces. The Affectiva system processes information in four stages to classify emotional states in videos: detecting faces and 34 facial landmarks, feature extraction from face texture, classification of facial actions, and modelling emotion expression [46]. Figure 2 is an example of Affectiva's 34 identified landmarks, used to calculate 43 numeric metrics to classify emotions. Among the 43 metrics produced are seven emotions (the six identified by [45] in Figure 1 plus the emotion contempt), 21 facial expressions (e.g. brow raise, eye widen, jaw drop, etc.), 13 emojis (e.g. wink, smiley, etc.), and two additional values to represent valence and engagement.

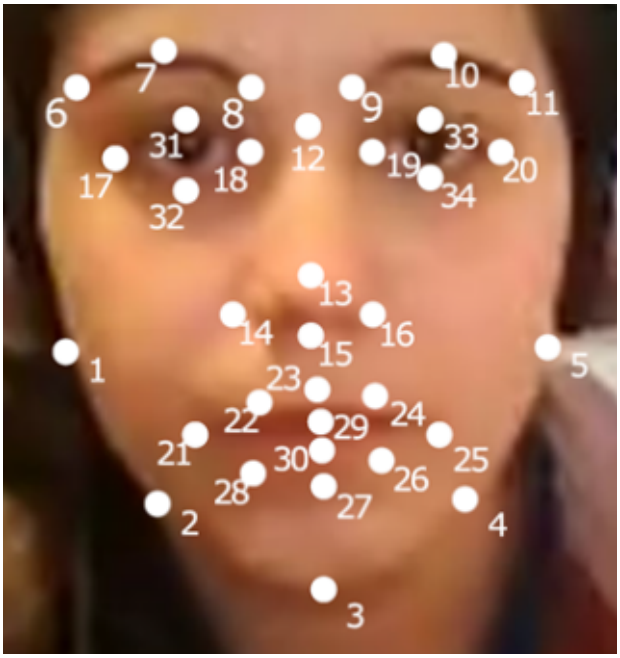


Figure 2: Facial landmarks identified by Affectiva.

Facial expression and emotion recognition research is widespread, and there are numerous relevant research projects in the literature [42], [47] and [48]. For example, two artificial intelligence researchers in Japan reported a practical application where facial expressions and emotion recognition were used to predict future policy changes. The Governor of the Bank of Japan's facial expressions at post-meeting news conferences were analysed in this study. Predicting an impending negative policy shift was possible based on observed signs of emotions such as anger and disgust, which were correlated with negative interest rates. The same researchers performed a follow-up and similar study (Source:

www.japantimes.co.jp), this time analysing the facial expressions of the European Central Bank's Chief. As in the first study, observing signs of sadness in videos recorded at previous press conferences enabled the prediction of negative changes in the bank's monetary policy.

These two examples clearly demonstrate the purpose of this current paper, which is to utilize identified emotions to evaluate if users or employees have a positive, negative, or neutral opinion of an organization's ISP. In this study, the identified emotions are obtained from video recordings of students reading known text passages which elicit specific emotions and their corresponding opinions from the subjects. Subsequently, a deep learning model is built to associate the elicited emotions obtained from facial expressions with the three opinion classes. First, background information on deep learning is presented in the following section. Then, the technique is applied in the illustrative case study of Section 5.

4. Deep learning

Machine learning is a subfield of artificial intelligence that focuses on constructing computer programs that can automatically adapt based on experience [49]. It has a broad field of applications, including, but not limited to, computer vision, speech recognition, natural language processing, and robotics. Until recently, research within machine learning generally employed shallow artificial neural networks, consisting of at most two hidden layers and one input layer [50]. These shallow models proved to be useful in solving basic and well-constrained problems. However, difficulties emerged when they were applied to problems with greater complexity levels, such as processing human voice, language, and real images and sceneries. The processing of raw natural data using shallow artificial neural networks was rather restricted [51]. Extensive domain expertise and careful engineering were necessary to create a machine learning system capable of extracting and transforming raw input data into an internal representation that the classifier could readily utilise to recognise and classify patterns in the input.

In 2006, deep learning originated from research in machine learning and artificial neural networks [50]. It was inspired by the deep architectures of human information processing mechanisms employed to extract complex structures and generate internal representations based on rich sensory inputs. Because deep learning models may convert a representation at one (lower) level into a higher abstracted representation, they can learn complex functions [51]. Starting with the raw input data, this transformation ensures that only the essential characteristics of the classification problem are highlighted while irrelevant aspects are ignored.

According to [52], artificial neural networks are structures of nodes or neurons (densely interconnected processing elements) that can perform many parallel computations. The architecture of a neural network is characterised by the pattern of connections between the neurons, the training or learning algorithm (the method for calculating the weights on the connections) and the activation function [53]. Deep learning is machine learning that uses neural networks with many layers of nonlinear nodes to solve problems. For feature extraction, supervised or unsupervised learning approaches are used at each of the successively higher levels of abstracted layers [42], [50]. In addition, in deep learning models, gradient-based optimisation algorithms such as the backpropagation algorithm modify the network's parameters depending on the output error rate [49]. The latter technique is discussed in more detail next.

4.1. Neural network training

The most fundamental deep learning neural network is a multilayer perceptron (MLP) neural network based on [53], [54]. An MLP comprises an input and an output layer and several hidden layers in between. It takes an input x and maps it to a category y by transferring the input values sequentially from one layer of nodes to the next and is represented as follows:

$$y = f(x, \theta), \tag{1}$$

where θ denotes the parameters, i.e. connection weights and biases, that the MLP uses to learn. It is important to notice that an MLP does not have any connections that transfer higher-level output values to lower-level nodes. Each layer of nodes has parameters that support the MLP in its learning process.

The term *learning* refers to the process of modifying the connection weights inside the MLP to minimise the difference between the desired and produced outputs [54]. The backpropagation algorithm [42], [55] is a frequently used method for training an MLP. The algorithm is given a collection of examples

$$\{\boldsymbol{p}_1, \boldsymbol{t}_1\}, \{\boldsymbol{p}_2, \boldsymbol{t}_2\}, \dots, \{\boldsymbol{p}_Q, \boldsymbol{t}_Q\}, \tag{2}$$

each of which comprises an input vector ($\boldsymbol{p}_Q$) that is mapped to a target output vector ($\boldsymbol{t}_Q$). The MLP adjusts its parameters in response to the calculated mean square error as it processes each of these inputs. This process can be summarised as follows:

1. Propagate the inputs forward through the MLP.
2. Calculate and propagate sensitivities backwards through the MLP.
3. Adjust the MLP's parameters accordingly.

For the first step, the outputs of a layer which is then used as input for the subsequent layer, is expressed as

$$\boldsymbol{a}^{m+1} = \boldsymbol{f}^{m+1}(\boldsymbol{W}^{m+1}\,\boldsymbol{a}^{m} + \boldsymbol{b}^{m+1}), \tag{3}$$

$$\text{for } m = 0, 1, \dots, M-1,$$

where $\boldsymbol{f}$ denotes the activation function, and $\boldsymbol{W}^n$ and $\boldsymbol{b}^n$ denote the weight vector and bias of layer n, respectively. M represents the number of layers in the MLP, and its starting point is denoted by

$$\boldsymbol{a}^0 = \boldsymbol{p}. \tag{4}$$

In (4) $\boldsymbol{p}$ denotes the original input vector, and the MLP's final layer's output represents the MLP's output, i.e.

$$\boldsymbol{a} = \boldsymbol{a}^M. \tag{5}$$

In the second step, the following equations are used to calculate the sensitivities:

$$\boldsymbol{s}^M = -2\dot{\boldsymbol{F}}^M(\boldsymbol{n}^M)(\boldsymbol{t} - \boldsymbol{a}), \tag{6}$$

where $\boldsymbol{n}$ denotes the net input, $\boldsymbol{t}$ represents the target or expected outputs, and

$$\dot{\boldsymbol{F}}^m(\boldsymbol{n}^m) = \begin{bmatrix} \dot{f}^m(n_1^m) & 0 & \cdots & 0 \\ 0 & \dot{f}^m(n_2^m) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \dot{f}^m(n_{S^m}^m) \end{bmatrix}, \quad (7)$$

where

$$\boldsymbol{s}^m = \dot{\boldsymbol{F}}^m(\boldsymbol{n}^m)(\boldsymbol{W}^{m+1})^T \boldsymbol{s}^{m+1}. \quad (8)$$

Finally, the MLP's biases and weights may be adjusted. This is accomplished via the use of the mean square error, which is calculated as follows:

$$\boldsymbol{W}^m(k+1) = \boldsymbol{W}^m(k) - \alpha \boldsymbol{s}^m(\boldsymbol{a}^{m-1})^T \text{ and} \quad (9)$$

$$\boldsymbol{b}^m(k+1) = \boldsymbol{b}^m(k) - \alpha \boldsymbol{s}^m, \quad (10)$$

at iteration k, with a learning rate represented by α.

More technical aspects of neural networks and deep learning are excluded due to the paper's scope. The work of [53] and [54] provide further details for interested readers. Constructing the best neural network model manually can be laborious. A neural architecture search methodology can alleviate this problem by finding architectures that perform well for the given data. This methodology is discussed in the following section.

4.2. Neural architecture search

The automation of machine learning model selection, hyperparameter optimization, and model search is called automated machine learning (AutoML) [56]. Neural architecture search (NAS), a subfield of AutoML that automates neural network architecture engineering, has resulted in models that outperform manually designed models [57]. The search space, search strategy, and performance estimation strategy are the three dimensions of a NAS method. Figure 3 depicts a simplified version of such a method.

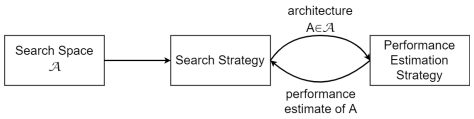


Figure 3: A high-level illustration of neural architecture search [57].

The search space ($\mathcal{A}$) defines all architectures that may be considered. Its size may be reduced by using previous knowledge about comparable task architectures, but this adds an undesired human bias. The maximum number of hidden layers (potentially unbounded), the operation of each layer, and the hyperparameters associated with the process define the search spaces of MLP neural networks and other chain-like neural networks. The choice of the search space determines the complexity of the architecture optimization problem, which is not continuous and has multiple dimensions.

A search strategy is used to explore the search space and identify an architecture $A \in \mathcal{A}$, which is then evaluated by the performance estimation strategy. Premature convergence to a region where suboptimal architectures exist should be avoided to find architectures that perform well. To find a suitable architecture inside the search space, approaches including random search, Bayesian optimisation, evolutionary methods, reinforcement learning, and gradient-based methods may be utilised. A reinforcement learning, evolutionary, and random search approach were compared in research by [58]. They discovered that the latter method outperformed the first two approaches. Furthermore, compared to the other two techniques, the evolutionary method created models with better accuracy throughout the early stages of the process. To develop and choose a suitable architecture in the experiment, a modified version of a regularised evolution approach given by [58] was implemented for the search strategy utilised in this work. This method is summarised in Algorithm 1.

Algorithm 1: Regularised evolution search strategy

```
Result: Highest accuracy model in history
population ← empty queue;
history ← empty list;
while | population | < P do
    model.arch ← RANDOM_ARCHITECTURE();
    model.accuracy ← TRAIN_AND_EVAL(model.arch);
    add model to right of population;
    add model to history;
end
while | history | < C do
    sample ← empty list;
    while | sample | < S do
        candidate ← distinct random element from population;
        add candidate to sample;
    end
    parent ← highest accuracy model in sample;
    child.arch ← MUTATE(parent.arch);
    child.accuracy ← TRAIN_AND_EVAL(child.arch);
    add child to right of population;
    add child to history;
    remove dead from the left of population;
    discard dead;
end
return highest accuracy model in history
```

Throughout the experiment, the method stores a population of previously trained models. At the start of the experiment P models with random architectures, based on the search space outlined above, are introduced to the population. The population is then mutated and added to the history list using C cycles. During each cycle, S candidates are selected at random from the population. After that, the candidate with the best accuracy is selected, mutated, and trained, resulting in a child model. A mutation performs a simple and randomised change in the chosen architecture. To achieve this, randomising one or more of the architecture's hyperparameters is done. The population and history are then updated to include the child model. Finally, the population is adjusted to exclude the oldest model. The performance estimation strategy is kept simple by maximising the model's validation loss. The generated models are configured to finish training when the model's accuracy begins to converge to guarantee that the NAS method makes optimal use of computing resources. In the next section, specific performance metrics used to evaluate the best neural network model found is addressed.

4.3. Performance metrics

According to [59], evaluating the performance of a machine learning model using just one aggregated measurement is insufficient. The researchers of [60], [61], [62], [63] and [64] all utilise or advise using different performance metrics. The following are some of the performance measures:

- Accuracy;
- Precision;
- Recall or sensitivity; and
- *F*-measure, also sometimes referred to as the F_1-measure.

Each sample in a testing process is always labelled with a real and a predicted label [61]. The real label identifies the real class to which the testing sample belongs. The predicted label is the predictor's output. As shown in Table 1, a multiclass confusion matrix can visually represent these label counts.

Table 1: Multiclass confusion matrix [65].

		Predicted		
		Class$_1$ - Class$_{k-1}$	**Class$_k$**	**Class$_{k+1}$ - Class$_n$**
Real	**Class$_{k+1}$ - Class$_n$**	tn_1	fp_1	tn_2
	Class$_k$	fn_1	tp	fn_2
	Class$_1$ - Class$_{k-1}$	tn_3	fp_2	tn_4

All of the above performance measures are based on the values represented by the multiclass confusion matrix. Each of the measures is discussed briefly below, along with a definition. The most common metric is accuracy, which determines how well the model can correctly classify positive and negative samples. To calculate accuracy, the number of correctly classified samples, positive and negative, are divided by the total number of samples.

As a result, it can be formalised as follows:

$$\text{Average accuracy} = \left(\sum_{i=1}^{n} \frac{tp_i + tn_i}{tp_i + tn_i + fp_i + fn_i}\right) / n. \quad (11)$$

The error rate is a measurement of how frequently errors occurred during the prediction phase. It is given as

$$\text{Average error rate} = \left(\sum_{i=1}^{n} \frac{fp_i + fn_i}{tp_i + tn_i + fp_i + fn_i}\right) / n. \quad (12)$$

The precision measure can be used to calculate the proportion of correctly classified true positives versus the total number of predicted positives. As a result, its definition is as follows:

$$\text{Precision}_M = \left(\sum_{i=1}^{n} \frac{tp_i}{tp_i + fp_i}\right) / n. \quad (13)$$

The recall measure calculates the proportion of samples labelled as positive compared to all truly positive samples. Consequently, this metric denotes the model's completeness. It can be defined as follows:

$$\text{Recall}_M = \left(\sum_{i=1}^{n} \frac{tp_i}{tp_i + fn_i}\right) / n. \quad (14)$$

Finally, the *F*-measure, also known as the harmonic mean of precision and recall, is a metric for determining how accurately a model performed on a test. The metric is defined as

$$F_M = \frac{(\beta^2+1) \times Precision_M \times Recall_M}{\beta^2 \times Precision_M + Recall_M}, \text{where } 0 \leq \beta \leq +\infty. \quad (15)$$

The β value is used to balance the importance of precision and recall. *F* becomes the harmonic mean of precision and recall if β is equal to 1 because both measures have the same weight. When β is greater than 1, *F* becomes more recall-oriented. In contrast, *F* becomes more precision-oriented when β is less than 1.

The following section will describe the experimental design to illustrate how deep learning affective computing and sentiment analysis may assist in solving response bias issues in the context of ISPs.

5. Experimental Design

A deep learning neural network approach is proposed to illustrate the concept of affective computing and sentiment analysis. This experimental approach is divided into two components: dataset acquisition and the building and testing of a deep learning neural network architecture.

5.1. Data acquisition

Instead of using publicly accessible videos, it was decided that a small video dataset would be generated as an initial experiment. A group of nine postgraduate Computer Science students agreed to participate in the study and help create facial expression videos. The nine participants were instructed to read three text passages while being recorded. The three text passages were selected to prompt a particular sentiment from the participants, and they were classified as positive, neutral, or negative. A collection of jokes was used to elicit a positive sentiment, and an ordinary neutral news article was used to evoke a neutral sentiment. Finally, a news article about consequences for unlawfully copying online material (which most students frequently do) was used to elicit a negative sentiment. The participants were informed that they would be recorded. Still, the objective of the exercise was not revealed until after the recording to ensure that they were not influenced to respond in a particular manner. The participants were offered the option of withdrawing from the experiment after they learned the purpose of the recordings. Despite this, they all decided to continue to be involved in the research.

The Affectiva SDK [46] was then used to extract 42 features from the 27 videos that were annotated based on the desired sentiment of the text passages. The emotion contempt was omitted from the dataset since it did not correlate with the sentiment. The pre-processing yielded 132 261 data records extracted from the videos. Positive sentiment was represented by 41 934 records,

neutral sentiment by 54 873 records, and negative sentiment by 35 454 records. The complete set of records was randomized and divided into three datasets: training (70%), validation (20%), and test (10%), all of which were utilized to build the deep neural network model.

5.2. Deep learning neural network architecture

To identify and select a suitable deep learning neural network architecture, the Google Colab cloud service was utilized. Then, model search, model selection, and hyperparameter optimisation were performed using the NAS methodology [57] described in Section 4.2. This method yielded a deep learning feed-forward neural network architecture with 42 input nodes (the 42 extracted facial expressions) and three output nodes (positive, neutral, and negative). The final layer used a softmax activation function to determine the sentiment of each input sample. In addition, five hidden layers were constructed, each using the ReLU activation function. Figure 4 shows a graphical representation of the deep learning model that was selected.

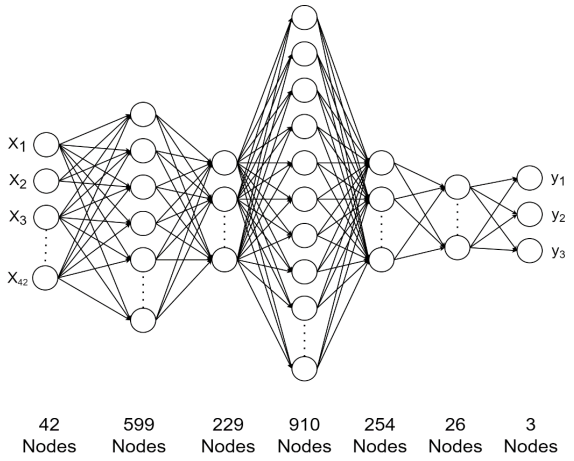


Figure 4: Deep learning neural network architecture.

The model was trained on the Google Colab cloud service for 2 hours and 16 minutes with a batch size of 9376 and 751 epochs. The accuracy was fairly high, as discussed further in the following section.

6. Results and Discussion

The selected model was evaluated on the test dataset to predict the out-of-sample class of each data record after the training and validation process. These predictions had an average accuracy of 96.23 percent, according to the results. The high levels of accuracy achieved with the selected architecture are detailed in a confusion matrix (Table 2) of the test dataset (13226 records). In addition, other calculated metrics, such as precision (average of 94.43 percent), recall (average of 94.19 percent), and *F*-measure (average of 94.31 percent), support the above-average results and the model's ability to perform sentiment classification. The high precision value shows that the model is very effective. In addition, the high recall value indicates that a high fraction of the total number of relevant instances was correctly classified.

The results will be discussed next regarding the selected deep learning model and the implications for ISP compliance.

Table 2: Confusion matrix for the test dataset.

		Predicted		
		Positive	**Neutral**	**Negative**
Real sentiment	**Positive**	3971	65	65
	Neutral	65	5237	213
	Negative	27	306	3277

6.1. Reflection on the deep learning model

The high accuracy result indicates that affective computing and sentiment analysis based on video analysis and an appropriate deep learning neural network architecture is feasible, supporting previous literature studies in this field. However, despite the high accuracy and excellent performance metrics achieved, the results of the particular illustrative experiment and the selected deep learning neural network model reported in this study should be interpreted with caution.

A variety of factors may impact the results, which will be considered in a follow-up study. The exceptionally high accuracy might be attributed to the limited number of participants utilized to create the videos. This means that a dataset with minimal variation was produced, which may aid the learning process in achieving high accuracy results. The minimal variation in the data may be contributed to the fact that all participants had the same study background. It is also uncertain if reading text is the most effective method of prompting a sentiment; maybe viewing a video would provide a more reliable dataset. Further experiments with splitting the dataset into training, validation, and test datasets may reduce overfitting.

Nonetheless, the objective was to show how a dataset including facial expressions might be generated and then used to perform sentiment analysis using a deep learning neural network. The experiment conducted in this paper achieved above-average results, demonstrating the feasibility of the suggested techniques.

6.2. Reflection on information security compliance

As explained previously, non-compliance with ISPs may be attributed partly to employees or users who negatively react to a policy because they disagree with its contents. Employee opinions may be obtained via surveys or text-based sentiment analysis; however, both methods might be biased since opinions can be expressed in a fake manner to meet expectations. When prompting employees for their opinions on the contents of an ISP, affective computing, which is based on emotional expression, offers a different approach that may be utilized to reduce the response bias problem. The dataset generated in this study, together with the selected deep learning neural network model, may be used to address social desirability problems in a similar way as predicting the sentiment of a bank governor based on facial expressions (see Section 3.2). It is no longer necessary to ask individuals their opinions; instead, one may deduce an opinion from their facial expressions. This may be especially significant when it comes to ISP compliance. Management will now understand whether or not employees are satisfied with the context of an ISP in general. It

may also assist in a more specific way by identifying particular areas of concern, leading to new or extra information security training opportunities.

A dataset acquired in the context of ISPs, i.e. employees reading an ISP, would be ideal for training a deep learning neural network model. This is unrealistic, however, since gathering a big enough sample of individuals who read an ISP would be difficult if not impossible. Furthermore, to create a dataset that can be utilized in a supervised learning environment, readers will be asked to indicate whether they found the ISP positive or negative, which puts one back to the response bias problem. The approach used in this paper is similar to that used in practice, i.e., in the example of bank governors, the training set was not constructed using a large number of bank governors but rather a large dataset of everyday videos from which facial expressions could be extracted. This implies that a model trained on regular individuals in videos may detect sentiment based on facial expressions in any other video.

This paper provided an example of the proposed concept. The following steps would be to collect a more extensive and more diverse dataset and test the model on employees that read an ISP.

7. Conclusion

This paper argues that the opinion of users and employees is essential in the creation and maintenance of ISPs. Employees should have a positive attitude toward an ISP and buy into the contents of the ISP to avoid non-compliance. However, obtaining user input on an ISP often poses a social desirability problem. Users are more likely to answer questions in an acceptable rather than truthful way. This study suggested sentiment analysis and affective computing to exclude possible fake responses while evaluating the contents of an ISP to minimize this problem. A deep learning neural network model was constructed to classify sentiment as positive, neutral, or negative in a real-world scenario. The model was trained using a video dataset of individuals reading various text passages to elicit multiple facial expressions. The suggested method proved to be an acceptable choice after achieving high accuracy. The experiment's findings may significantly affect how ISPs are evaluated since it would no longer be required to ask consumers for their opinions, which risks social desirability. Applying the suggested affective computing and sentiment analysis improves the policy evaluation process by making it simpler to gather opinions without the risk of fake answers. Management may identify areas of concern that may be addressed by either changing or correcting the policy's contents or giving extra training to particular (negative) users, or training on specific topics.

The research presented in this paper is an exploratory study, and many opportunities for future investigation have been identified. For example, experiments involving larger populations (participants being recorded), various methods of evoking emotions (i.e. viewing a video instead of reading text), and the use of different neural network architectures are all possibilities.

References

[1] H. Kruger, T. du Toit, L. Drevin, N. Maree, “Acquiring sentiment towards information security policies through affective computing,” in 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC), 1-6, 2020, doi:10.1109/IMITEC50163.2020.9334134.

[2] N. Maree, Affective computing and deep learning to perform sentiment analysis, M. Sc. Thesis, North-West University, South Africa, 2020.

[3] R.E. Crossler, A.C. Johnston, P.B. Lowry, Q. Hu, M. Warkentin, R. Baskerville, "Future directions for behavioural information security research," Computers & Security, **32**, 90-101, 2013, doi:10.1016/j.cose.2012.09.010.

[4] W.A. Cram, J. D'Arcy, J.G. Proudfoot, "Seeing the forest and the trees: a meta-analysis of the antecedents to information security policy compliance," MIS Quarterly, **43**(2), 525-554, 2019, doi:10.25300/MISQ/2019/15117.

[5] V.T. Patil, P.R. Patil, V.O. Patil, S.V. Patil, "Performance and information security evolution with firewalls," Journal of Scientific Computing, **8**(4), 1-6, 2019, doi:16.10089.ISC.2019.V8I5.285311.2630.

[6] M. Butavicius, K. Parsons, M. Lillie, A. McCormac, M. Pattinson, D. Calic, "When believing in technology leads to poor cyber security: Development of a trust in technical controls scale," Computers & Security, **98**, 102020, 2020, doi:10.1016/j.cose.2020.102020.

[7] G.D. Moody, M. Siponen, S. Pahnila, "Toward a unified model of information security policy compliance," MIS Quarterly, **42**(1), 285-311, 2018, doi:10.25300/MISQ/2018/13853.

[8] J. C. Sipior, D.R. Lombardi, "The impact of employee organisational commitment on compliance with information security policy," in Proceedings of the 2019 Southern Association for Information Systems Conference (SAIS), 2019.

[9] M. Kang, A. Hovav, "Benchmarking methodology for information security policy (BMISP): Artifact development and evaluation," Information Systems Frontiers, **22**, 221-242, 2020, doi:10.1007/s10796-018-9855-6.

[10] M. Karjalainen, M.T. Siponen, S. Sarker, "Toward a stage theory of the development of employees' information security behaviour," Computers & Security, **93**, 101782, 2020, doi:10.1016/j.cose.2020.101782.

[11] A. Vance, M.T. Siponen, D.W. Straub, "Effects of sanctions, moral beliefs, and neutralization on information security policy violations across cultures," Information & Management, **57**(4), 103212, 2020, doi:10.1016/j.im.2019.103212.

[12] S. Kokolakis, "Privacy attitudes and privacy behavior: a review of current research on the privacy paradox phenomenon," Computers & Security, **64**, 122-134, 2017, doi:10.1016/j.cose.2015.07.002.

[13] J.A. Cox, "Information systems user security: a structured model of the knowing-doing gap," Computers in Human Behavior, **28**(5), 1849-1858, 2012, doi:10.1016/j.chb.2012.05.003.

[14] K.L. Gwebu, J. Wang, M.Y. Hu, "Information security policy noncompliance: An integrative social influence model," Information Systems Journal, **30**(2), 220-269, 2020, doi:10.1111/isj.12257.

[15] J.H. Nord, A. Koohang, K. Floyd, "Impact of habits on information security policy compliance," Issues in Information Systems, **21**(3), 217-226, 2020, doi:/10.48009/3_iis_2020_217-226.

[16] R.A. Alias, "Information security policy compliance: Systematic literature review," Procedia Computer Science, **161**(2019), 1216-1224, 2019, doi:10.1016/j.procs.2019.11.235.

[17] P. Ifinedo, "Understanding information systems security policy compliance: An integration of the theory of planned behavior and the protection motivation theory," Computers & Security, **31**(1), 83-95, 2012, doi:10.1016/j.cose.2011.10.007.

[18] T.B. Lembcke, K. Masuch, S. Trang, S. Hengstler, P. Plics, M. Pamuk, "Fostering information security compliance: Comparing the predictive power of social learning theory and deterrence theory," in Proceedings of the 2019 American Conference on Information Systems (AMCIS), Information Security and privacy (SIGSEC), 2019.

[19] D.P. Snyman, H.A. Kruger, W.D. Kearney, "The lemming effect in information security," in Proceedings of the 2017 International Symposium on Human Aspects of Information Security & Assurance (HAISA), 91-103, 2017.

[20] S. Redhu, S. Srivastava, B. Bansal, G. Gupta, "Sentiment analysis using text mining: a review," International Journal on Data Science and Technology, **4**(2), 49-53, 2018, doi:10.11648/j.ijdst.20180402.12.

[21] G.S. Murthy, S.R. Allu, "Text based sentiment analysis using LSTM," International Journal of Engineering Research & Technology, **9**(5), 299-303, 2020, doi:10.17577/IJERTV9IS050290.

[22] E. Yadegaridehkordi, N.F.B.M. Noor, M.N.B. Ayub, H.B. Affal, N.B. Hussin, "Affective computing in education: a systematic review and future research," Computers & Education, **142**, 2019, doi:10.1016/j.compedu.2019.103649.

[23] S. Richardson, "Affective computing in the modern workplace," Business Information review, **37**(2), 78-85, 2020, doi:10.1177/0266382120930866.

[24] M.J. Alotaibi, S. Furnell, N. Clarke, "A framework for reporting and dealing

with end-user security policy compliance," Information & Computer Security, **27**(1), 2-25, 2019, doi:10.1108/ics-12-2017-0097.
[25] P.B. Lowry, G.D. Moody, "Proposing the control-reactance compliance model (CRCM) to explain opposing motivations to comply with organisational information security policies," Information Systems Journal, **25**(5), 433-463, 2015, doi:10.1111/isj.12043.
[26] H. Paananen, M. Lapke, M. Siponen, "State of the art in information security policy development," Computers & Security, **88**, 2020, doi:10.1016/j.cose.2019.101608.
[27] T. Stafford, G. Deitz, Y. Li, "The role of internal audit and user training in information security policy compliance," Managerial Auditing Journal, **33**(4), 410-424, 2018, doi:10.1108/MAJ-07-2017-1596.
[28] J. D'Arcy, P. The, "Predicting employee information security policy compliance on a daily basis: the interplay of security-related stress, emotions and neutralization," Information & Management, **56**(7), 2019, doi:10.1016/j.im.2019.02.006.
[29] T. Sommestad, H. Karlzen, J. Hallberg, "The theory of planned behaviour and information security policy compliance," Journal of Computer Information Systems, **59**(4), 344-353, 2019, doi:10.1080/08874417.2017.1368421.
[30] M. Rajab, A. Eydgahi, "Evaluating the explanatory power of theoretical frameworks on intention to comply with information security policies in higher education," Computers & Security, **80**, 211-223, 2019, doi:10.1016/j.cose.2018.09.016.
[31] S. Trang, B. Brendel, "A meta-analysis of deterrence theory in information security policy compliance research," Information Systems Frontiers, **21**(6), 1265-1284, 2019, doi:10.1007/s10796-019-09956-4.
[32] A. Alzahrani, C. Johnson, S. Altamimi, "Information security compliance: investigating the role of intrinsic motivation towards policy compliancy in the organisation," in Proceedings of the 2018 International Conference on Information Management (ICIM), 125-132, 2018, doi:10.1109/INFOMAN.2018.8392822.
[33] R.J. Fisher, "Social desirability bias and the validity of indirect questioning," Journal of Consumer Research, **20**(2), 303-315, 1993, doi:10.1086/209351.
[34] N. Bergen, R. Labonte, "Everything is perfect and we have no problems: Detecting and limiting social desirability bias in qualitative research," Qualitative Health Research, **30**(5), 783-792, 2020, doi:10.1177/1049732319889354.
[35] D. Burchett, Y.S. Ben-Porath, "Methodological considerations for developing and evaluating response bias indicators," Psychological Assessment, **31**(12), 1497–1511, 2019, doi:10.1037/pas0000680.
[36] D. Kwak, P. Holtkamp, S.S. Kim, "Measuring and controlling social desirability bias: Applications in information systems research," Journal of the Association for Information Systems, **20**(4), 2019, doi:10.17705/1jais.00537.
[37] A. McCormac, D. Calic, M. Butavicius, K. Parsons, T. Zwaans, M. Pattinson, "A reliable measure of information security awareness and the identification of bias in responses," Australasian Journal of Information Systems, **21**, 1-12, 2017, doi:10.3127/ajis.v21i0.1697.
[38] S. Kurowski, "Response biases in policy compliance research," Information & Computer Security, 2019, doi:10.1108/ICS-02-2019-0025.
[39] S. Poria, N. Majumder, E. Cambria, A. Gelbukh, A. Hussain, "Multimodal sentiment analysis: addressing key issues and setting up the baselines," IEEE Intelligent Systems, **33**(6), 17-25, 2018, doi:10.1109/MIS.2018.2882362.
[40] J.K. Rout, K.-K.R. Choo, A.K. Dash, S. Bakshi, S.K. Jena, K.L. Williams, "A model for sentiment and emotion analysis of unstructured social media text," Electronic Commerce Research, **18**(1), 181-199, 2018, doi:10.1007/s10660-017-9257-8.
[41] D.P. Alamanda, A. Ramdhani, I. Kania, W. Susilawati, E.S. Hadi, "Sentiment analysis using text mining of Indonesia tourism reviews via social media," International Journal of Humanities, Arts and Social Sciences, **5**(2), 72-82, 2019, doi:10.20469/ijhss.5.10004-2.
[42] N. Maree, T. du Toit, L. Drevin, H. Kruger, "Affective computing and deep learning to perform sentiment analysis," in Proceedings of the 2019 Southern Africa Telecommunication Networks and Applications Conference (SATNAC), 94-99, 2019.
[43] S. Poria, E. Cambria, N. Howard, G.-B. Huang, A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," Neurocomputing, **174**, 50-59, 2016, doi:10.1016/j.neucom.2015.01.095.
[44] B. Kratzwald, S. Ilic, M. Kraus, S. Feuerriegel, H. Prendinger, "Deep learning for affective computing: text-based emotion recognition in decision support," Decision Support Systems, **115**, 24-35, 2018, doi:10.1016/j.dss.2018.09.002.
[45] P. Ekman, Basic emotions. Handbook of cognition and emotion, **98**(45-60), 16, 1999.
[46] D. McDuff, M. Mahmoud, M. Mavadati, J. Amr, J. Turcot, R. Kaliouby, "AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit," in Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems, 3723-3726, 2016, doi:10.1145/2851581.2890247.
[47] S.E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, et al., "Emonets: multimodal deep learning approaches for emotion recognition in video," Journal on Multimodal User Interfaces, **10**(2), 99-111, 2016, doi:10.1007/s12193-015-0195-2.
[48] O.M. Nezami, M. Dras, P. Anderson, L. Hamey, "Face-cap: image captioning using facial expression analysis," Joint European Conference on Machine Learning and Knowledge Discovery in Databases: Springer, 226-240, 2018, doi:10.1007/978-3-030-10925-7_14.
[49] M.I. Jordan, T.M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, **349**(6245), 255-260, 2015, doi:10.1126/science.aaa841.
[50] L. Deng, D. Yu, "Deep learning: methods and applications. Foundations and trends in signal processing," **7**(3–4), 197-387, 2014, doi:10.1561/2000000039.
[51] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," Nature, **521**(7553), 436, 2015, doi:10.1038/nature14539.
[52] I.A. Basheer, M. Hajmeer, "Artificial neural networks: fundamentals, computing, design, and application," Journal of Microbiological Methods, **43**(1), 3-31, 2000, doi:10.1016/S0167-7012(00)00201-3.
[53] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016.
[54] H. Ramchoun, M.A.J. Idrissi, Y. Ghanou, M. Ettaouil, "Multilayer Perceptron: Architecture optimization and training," IJIMAI, **4**(1), 26-30, 2016, doi:10.9781/ijimai.2016.415.
[55] M.T. Hagan, H.B. Demuth, M.H. Beale, O. De Jesus, Neural Network Design, Martin Hagan, 2014.
[56] I. Guyon, K. Bennett, G. Cawley, H.J. Escalante, S. Escalera, T.K. Ho, N. Macia, B. Ray, M. Saeed, A. Statnikov, "Design of the 2015 ChaLearn AutoML challenge," in Proceedings of 2015 International Joint Conference on Neural Networks (IJCNN), IEEE, 1-8, 2015, doi:10.1109/IJCNN.2015.7280767.
[57] T. Elsken, J.H. Metzen, F. Hutter, "Neural architecture search: A survey," Journal of Machine Learning Research, **20**(55), 1-21, 2019, doi:10.5555/3322706.3361996.
[58] E. Real, A. Aggarwal, Y. Huang, Q.V. Le, "Regularized evolution for image classifier architecture search," in Proceedings of the 2019 AAAI Conference on Artificial Intelligence, **33**(1), 4780-4789, 2019, doi:10.1609/aaai.v33i01.33014780.
[59] P. Flach, "Performance evaluation in machine learning: The good, the bad, the ugly and the way forward," in Proceedings of 2019 AAAI Conference on Artificial Intelligence, 2019, doi:10.1609/aaai.v33i01.33019808.
[60] A. Tripathy, A. Agrawal, S.K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," Expert Systems with Applications, **57**, 117-126, 2016, doi:10.1016/j.eswa.2016.03.028.
[61] Y. Jiao, P. Du, "Performance measures in evaluating machine learning based bioinformatics predictors for classifications," Quantitative Biology, **4**(4), 320-330, 2016, doi:10.1007/s40484-016-0081-2.
[62] E. Gokgoz, A. Subasi, "Comparison of decision tree algorithms for EMG signal classification using DWT," Biomedical Signal Processing and Control, **18**, 138-144, 2015, doi:10.1016/j.bspc.2014.12.005.
[63] D.M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," Journal of Machine Learning Technologies **2**(1), 37-63, 2011, doi:10.48550/arXiv.2010.16061.
[64] M. Sokolova, G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing & Management, **45**(4), 427-437, 2009, doi:10.1016/j.ipm.2009.03.002.
[65] F. Krüger, Activity, context, and plan recognition with computational causal behaviour models, Ph.D Thesis, Universität Rostock, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 2, 156-170 (2022)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Leakage-abuse Attacks Against Forward Private Searchable Symmetric Encryption

Khosro Salmani[*,1], Ken Barker[2]

[1]*Assistant Professor, Department of Mathematics and Computing, Mount Royal University, Calgary, AB T3E 6K6, Canada*

[2]*Professor, Department of Computer Science, University of Calgary, Calgary, AB T2N 1N4, Canada*

ARTICLE INFO

Article history:
Received: 12 January, 2022
Accepted: 07 April, 2022
Online: 19 April, 2022

Keywords:
Dynamic SSE
Cloud security
Access pattern leakage
Search pattern leakage
Data privacy
Leakage-abuse attacks

ABSTRACT

Dynamic Searchable Symmetric Encryption (DSSE) methods address the problem of securely outsourcing\updating private data into a semi-trusted cloud server. Furthermore, Forward Privacy (FP) notion was introduced to limit data leakage and thwart the related attacks on DSSE approaches. FP schemes ensure previous search queries cannot be linked to future updates and newly added files. Since FP schemes use ephemeral search tokens and one-time use index entries, many scholars conclude that privacy attacks on traditional SSE schemes do not apply to SSE approaches that support forward privacy. However, to obtain efficiency, all FP approaches accept a certain level of data leakage, including access pattern leakage. Here, we introduce two new attacks on forward-private schemes. We demonstrate that it is still plausible to accurately unveil the search pattern by reversing the access pattern. Afterward, the attackers can exploit this information to uncover the search queries and consequently the documents. We also show that the traditional privacy attacks on SSE schemes are still applicable to schemes that support forward privacy. We then construct a new DSSE approach that supports parallelism and obfuscates the search and access pattern to thwart the introduced attacks. Our scheme is cost-efficient and provides secure search and update. Our performance analysis and security proof demonstrate our approach's practicality, efficiency, and security.

1 Introduction

Cloud service providers offer various services that attract users and encourage them to outsource their personal data to reduce maintenance costs and increase user satisfaction, convenience, and flexibility. Nevertheless, these services come at the cost of losing complete control over the outsourced data, which raises security and privacy concerns. Although encrypting data before uploading it into the cloud addresses privacy concerns, it suffers from low efficiency. Keyword search is an essential requirement in these systems which is not supported by traditional encryption schemes. A naive solution is to download and decrypt all the encrypted documents to search for a keyword. Obviously, this solution suffers from excessive communication overhead and is inefficient.

Hence, Searchable Symmetric Encryption (SSE) schemes [1]–[4] were introduced to tackle this challenge. In SSE schemes, a cloud can perform search queries on user's outsourced data while the queries, results, and data are encrypted. In the other words, SSE schemes address both privacy challenges and the searchability requirement. However, the early approaches only work for static data which means that no additions, updates, and deletions are feasible in a low-cost and efficient manner after the setup phase (securing and storing data into the cloud). Later, researchers proposed Dynamic SSE (DSSE) approaches [5]–[8]. These schemes enable users to update\modify the outsourced corpus arbitrarily in addition to performing search queries.

Moreover, SSE schemes support Multi-keyword [4, 9] search or Boolean [7, 10, 11] . Boolean schemes search for a single keyword and returns all of the documents that contain the queried keyword. Alternately, Multi-keyword search supports multiple keywords search which prevents unacceptably coarse results and improves result accuracy. Moreover, some of the SSE schemes support *ranked* search [4, 9] which means the cloud returns most relevant files by ranking them based on their relevance to the query.

However, DSSE approaches leak sensitive information such as search and access pattern. These methods employ deterministic queries \search tokens, which allow a server to determine if multiple queries consist of the same keyword (*search pattern*). furthermore, the matching document identifiers (*access pattern*) will be leaked af-

[*]Corresponding Author: Khosro Salmani, B113F- 4825 Mt Royal Gate SW, Calgary, AB T3E 6K6, (+1) 403-440-6492 & ksalmani@mtroyal.ca

ter each query. Many scholars have shown [12]–[14] that revealing such important meta-data can be used to obtain critical information and even to expose underlying plaintext data. Note that, in theory it is possible to design SSE scheme with no information leakage using several cryptographic primitives such as oblivious RAM, homomorphic encryption, and secure two-party computation[15]–[17]. However, these methodologies suffer from excessive computation power, low efficiency, and large storage costs.

Moreover, in Dynamic SSE approaches, it is still feasible to link previous search and update requests to a file that is recently added. For example, if we add a new document to the corpus, the server can determine whether the new document contains the keyword that we searched for in the past. Moreover, the cloud can execute the queries on deleted documents. To tackle these challenges, researchers introduced *Backward* and *Forward privacy* [5]. Backward privacy guarantees the privacy of deleted documents while forward privacy guarantees the privacy of newly added documents. No *efficient* approach currently provides full backward privacy.

Forward-private (FP) approaches employ one-time use search tokens to preserve the newly added documents' privacy [7, 6]. This means the search token for a keyword changes after being used in a query. Moreover, server index entries are ephemeral, which means the user generates new encrypted index entries after each time of access. Hence, the server cannot track the queries. As a result, Scholars believe that privacy attacks on traditional SSE approaches do not apply to forward-private schemes. In particular, in[14] the author believes these features "*highlights the importance of forward privacy*", and in [18] the author believes with forward privacy these attacks can be thwarted and prevented. Furthermore, applying these attacks will become a cumbersome task, considering that forward-private schemes are primarily *dynamic* and provide *update* functionalities (including *add* and *delete*). Therefore, monitoring and linking the queries turns significantly harder, if multiple update requests occur between two search queries.

Nevertheless, this paper extends work initially presented in the Eleventh ACM Conference on Data and Application Security and Privacy [10] and shows that it is still possible to reveal the documents and queries accurately. All DSSE approaches accept a certain level of leakage to achieve an acceptable level of performance\efficiency [2, 7, 13, 19]. Hence, the primary objective is to increase the performance as high as possible while decreasing the leakage. In particular, *access pattern* leakage is among the acceptable leakages [5]–[7] and, thus, one of the open challenges that has not been addressed among the forward-private and traditional approaches.

In this paper with introducing two attacks we show that it is possible to retrieve the search pattern with high accuracy that can be exploited by previous attacks to unveil the search tokens and consequently the documents in FP approaches. Our introduced attacks **reverse-analyze** (see Section 4) the access pattern to recover the search pattern. The first attack is applicable on the forward-private DSSE approaches that only provide "add" functionality such as [7]. We modified the the first attacked (based attack) and introduced the advanced attack which can invade the forward-private DSSE approaches that provides both "add" and "del" functionalities such as [18].

In this paper, a new forward-private DSSE approach is introduced to tackle this problem. In contrary with other scheme, our approach hides and obfuscates the access and search pattern and employs non-deterministic search tokens. All these features thwart and prevent the introduced attacks in this paper and also previous privacy and security attacks. Particularly, our contributions are:

1. Defining two concrete attacks and demonstrating its potential threats and privacy\ security risks.
2. Tackling the access pattern leakage challenge in DSSE approaches with forward privacy with a novel method.
3. Constructing an forward private DSSE scheme that support search and update (add and delete) operation. Furthermore, our *efficient* approach supports *parallelism* which is an important efficiency factor [7].
4. Providing a security proof against adaptive adversaries which verify the privacy and security of our method.
5. Demonstrating the efficiency of our approach in real-world by implementing it using real-world datasets.

The rest of this article is organized as follows. We present related work and the state-of-the-art work in Section 2. We then state the preliminaries in Section 3. In Section 4, we introduce two new attacks on current forward-private DSSE schemes, and in Section 5 we construct a new scheme that prevents the introduced attacks in Section 4. Experiments and evaluation are detailed in Section 6, and the security proof is provided in Section 7. Finally, we conclude our paper in Section 8.

2 Related Work

During the last decade, various privacy constructions and security definitions have been proposed for searchable encryption. Efficiency has always been a key requirement and a primary challenge in this research area. For example, Oblivious RAM [15] achieves full privacy and security without leaking any information to the server, but it is impractical for real-life applications because of its excessive computation costs. Hence, several approaches were designed [1, 2, 4, 9, 19] that selectively leak information (*e.g.,* search and access pattern). This means, these schemes accept a low level of information leakage to gain a higher level of efficiency.

Searchable Symmetric Encryption (SSE) and Public-key Encryption with Keyword Search (PEKS) are the two main divisions of searchable encryption schemes. In [20], the author introduced the notion of the public key encryption with keyword search, which followed by several methods [21]–[23] to improve the system cost and efficiency of PEKS approaches. In particular, these methods use one key for encryption and another key for decryption. Hence, only data users who possess the private-key can search the encrypted outsourced data. In this paper we focus on the SSE schemes and our introduced construction is build on symmetric security primitives.

The first SSE scheme was introduced by [1]. They employed a two-layered encryption to encrypt each keyword. However, they suggest a sequential search which impacts the search time (makes it linear to the document size). Later, in [2] the author used a secure

index structure called Bloom filters to address this issue. In [19], the author proposed a scheme which preserves the security of the outsourced data against an adaptive adversary. However, this comes at the cost of higher communication overhead and requires more memory space on the server side.

Nevertheless, traditional SSE approaches provide exact keyword search and cannot tolerate any imperfections or format inconsistency. In [24], the author addressed this issue and proposed a method in which resultant documents are selected based on the keyword similarity and closest possible matching documents. In [25], the author tackled the same challenge and proposes a scheme that decreases system cost and provides more efficiency.

Moreover, SSE schemes support Boolean [7, 10, 11] or Multi-keyword [4, 9] search. Boolean schemes search for a single keyword and returns all of the files that contain the respective keyword. On the other hand, multi-keyword ranked search solutions [4, 8, 26, 27] enhance the result accuracy by supporting multiple keywords search. In [4], the author introduced the notion of "coordinate matching" which is a similarity measure that matches as many keywords as possible. They also constructed a multi-keyword search approach using coordinate matching. However, previous methods only support single data owner. In [26], the author designed a new scheme with a trusted proxy that supports multiple data owners. In [8], [27] the author considered a system model with semi-honest cloud server and proposed verifiable SSE approaches that can detect a malicious server. Moreover, in [9] the author propose a multi-keyword ranked search scheme that solve the problem of search pattern, and co-occurrence information leakage. They introduce a novel chaining encryption notation which prevent the aforementioned information leakages.

Dynamic Searchable Symmetric Encryption (DSSE) methods were introduced to support *add*, *delete*, and *update* operations in an efficient manner. In particular, the author in [28] introduced a DSSE approach that preserve users' privacy and security against adaptive chosen keyword attacks. In [29], the author proposed a new DSSE method called "Blind Storage" that hinders leaking sensitive information such as the size and number of stored documents. DSSE approaches employ interactive protocols which results in leaking more information about the outsourced data in compare with traditional SSE approaches. In [5], the author introduced the notion of forward-privacy and designed a forward private DSSE construction to address this issue. However, their proposed method suffers from low efficiency. In [6], the author improved the system efficiency by using trapdoor permutations and designed a more efficient forward private DSSE scheme. However, these approaches use sequential scan to execute a query which makes palatalization impossible. In [7], the author addressed this issue with designing a new forward private DSSE method that provides parallelism by design.

3 Problem Formulation

3.1 Preliminaries

For a finite set X, we employ $x \leftarrow X$ to represent that x is sampled uniformly from the set X. λ is the security parameter, and $\|$ shows concatenation. Function $\mathsf{neg}(k) : \mathbb{N} \rightarrow [0, 1]$ is negligible if for all positive polynomial p, there exists a constant c such that: $\forall\ k > c,\ neg(k) < 1/p(k)$.

Definition 1 *(Symmetric-key Encryption). A symmetric encryption scheme is a set of three probabilistic polynomial time (PPT) algorithms* $\mathsf{SE} = (\mathsf{Gen}, \mathsf{Enc}, \mathsf{Dec})$ *such that* Gen *takes an unary security parameter λ and generates a secret key k;* Enc *takes a key k and n-bit message m and returns a ciphertext c;* Dec *takes in a key k and a ciphertext c, and returns m if k was the key under which c was generated. The* SE *is required to be secure against chosen plaintext attack (CPA). We refer to [19] for formal definitions.*

Definition 2 *(Pseudorandom function). Let $F : \{0,1\}^{\lambda} \times \{0,1\}^{l} \rightarrow \{0,1\}^{l'}$ be a deterministic function which maps l-bit strings to l'-bit strings. We define $F_s(x) = F(s, x)$ as a pseudorandom function* (PRF) *if: $\forall$ PPT distinguishers $\mathcal{D}$: $|Pr[\mathcal{D}^{F_s(.)}(1^{\lambda}) = 1] - Pr[\mathcal{D}^{f(.)(1^{\lambda})} = 1]| \leq neg(\lambda)$, where $f(.)$ is a truly random function, and λ is the security parameter.*

In Definition 3, the notation $(c_{out}, s_{out}) \leftarrow \mathsf{protocol}(c_{in}, s_{in})$ denotes an interaction between client and server where c_{in} and s_{in} are the client and server input, and the c_{out} and s_{out} are the output of client and server after performing a protocol.

Definition 3 *(DSSE Scheme). Let $D = \{D_1, \ldots, D_n\}$ be a corpus of n documents, a Dynamic Searchable Symmetric Encryption consists of five PPT algorithms:*

- $(sk, \perp) \leftarrow \mathsf{GenKey}(1^{\lambda}, 1^{\lambda})$*: In this algorithm, the data owner (client) generates a secret key sk using the security parameter λ.*
- $(\mathcal{I}_c, \mathcal{I}_s) \leftarrow \mathsf{BuildIndex}((sk, D), \perp)$*: In this algorithm the client's secret key sk and document collection D are used to produce a client-index $\mathcal{I}_c$, and server outputs index $\mathcal{I}_s$.*
- $(\perp, C) \leftarrow \mathsf{Encryption}((sk, D), \perp)$*: The client inputs secret key sk, and document collection D, and outputs the encrypted corpus $C = \{C_1, \ldots, C_n\}$.*
- $((\mathcal{I}'_c, D_w), \mathcal{I}'_s) \leftarrow \mathsf{Search}((sk, \mathcal{I}_c, w), (\mathcal{I}_s, C))$*: In this algorithm the client inputs the secret key sk, index $\mathcal{I}_c$, and query w; and it outputs the updated index $\mathcal{I}_c$, and resultant documents D_w. The server also, inputs the index I_s, and the encrypted document collection C and outputs the updated index $\mathcal{I}'_s$.*
- $(\mathcal{I}'_c, (\mathcal{I}'_s, C')) \leftarrow \mathsf{Update}((sk, \mathcal{I}_c, \mathsf{op}, , in), (\mathcal{I}_s, C))$*: In this algorithm the client inputs the secret key sk, index $\mathcal{I}_c$, and an operation* op *= add or* op *= del, and an input " in", which is parsed as a set of keywords $\boldsymbol{w}_{in}$ and a document identifier id_{in}. It outputs the updated index $\mathcal{I}'_c$. The server inputs the index $\mathcal{I}_s$, and the encrypted document collection C; it outputs the updated index $\mathcal{I}'_s$ and updated encrypted document collection C'.*

  *We call the first three protocols (*GenKey, BuildIndex, Encryption*) the* ***Setup phase****.*

3.2 Our System Architecture

Our system architecture, as illustrated in Figure **??**, consists of two parties: a *cloud server* and a *client (data owner - user)*. The client is the actual owner of the data and intends to outsource its personal corpus into a cloud server for several reasons including maintenance costs. The client first creates an inverted index for each keyword.

Each entry in this index maps the respective keyword, w_i, to the documents IDs that contain w_i. The client then encrypts the documents and index entries and outsources them into the cloud server. Once the cloud receives a search request, it performs the query using the provided index and outputs the resultant files. Note that the documents and index entries are all encrypted, so the cloud server will not know the content of search tokens or the documents. Nevertheless, in see Section 3.5 we explain that like other related work [5]–[8], some meta-data may leak over time and after executing a number of queries.

3.3 Threat Model

In our approach, the server follows the prescribed protocol, however, it is keen to gather meta-data and information about the client. This type of cloude server is called *honest-but-curious* and is employed in many related work such as [5]–[7]. In addition, we suppose the server knows the encrypted index, documents, queries, and the employed encryption scheme, but it does not know the secret key.

3.4 A short overview of our approach

The client initiate the protocol by extracting keywords, $\Delta = \{w_1, w_2, \ldots, w_m\}$, and creating the inverted plain-index. Each entry (id_i, L) in the index is a pair of an id_i and a list of L. Each keyword, w_i in the corpus corresponds to an id_i in the index, and L consists of all the files that contain w_i. To achieve our primary objectives which are hiding and obfuscating the access and search pattern, we inject random files IDs (noise) among the nodes in each list. The client is the only party who can distinguish the noise nodes. To monitor the lists, the user must keep a small index, $\mathcal{I}_c$ (see Section 5.2) on her side. Then, the index entries and files will be encrypted and transfered to the cloud server. Upon receiving the data, the server stores the encrypted index entries, $\mathcal{I}_s$, and the encrypted corpus C and stands by for the first search or update request. Every query in our approach, $\mathbf{q}$, consists of a limited number of sub-queries $\mathbf{q} = \{q_1, \ldots, q_k\}$. The fake\noise sub-queries are added to hide and obfuscate the search and access pattern. Once a query is received, $\mathbf{q} = \{q_1, \ldots, q_k\}$, the server performs the sub-queries one-by-one, or parallelly if we employ the parallel algorithm, and returns the results. The user can retrieve the real results and discard the noise. Lastly, using new keys and IDs, new encrypted index entries will be created and sent to the cloud server. Note that the user ($\mathcal{I}_c$) and cloud ($\mathcal{I}_s$) indexes will be updated respectively.

3.5 Security Definitions

To gain efficiency, most of the SSE schemes leak some meta-data such as number of keywords, file size, and file IDs [4, 5, 19]. In addition, more meta-data may leak after performing each query. Thus, we start this sections by defining the leakage functions that show the leaked meta-data to the cloud server after executing each step of the protocol.

Definition 4 *(Search pattern). Let $Q = (\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_t)$ be the query list over t queries. The search pattern over a query list Q is a tuple, $\mathsf{SP} = (\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_t)$, where $\hat{w}_i, 1 < i < t$ is the encrypted keyword (or its hash) in the i-th query.*

Definition 5 *(Access pattern). The access pattern over a query list Q is a set, $\mathsf{AP} = (R(\boldsymbol{q}_1), R(\boldsymbol{q}_2), \ldots, R(\boldsymbol{q}_t))$ over t queries, where $R(\boldsymbol{q}_i), 1 < i < t$ is the i-th query's resultant document identifiers (result set).*

To demonstrate the leakage to the server, we employ leakage function $\mathcal{L}^{\mathsf{op}}$ which indicates the information revealed to the adversary after executing operation op. We first define the $\mathcal{L}^{\mathsf{Setup}}$ and $\mathcal{L}^{\mathsf{Search}}$, and then demonstrate the information leakage of Update through Definition 6.

- $\mathcal{L}^{\mathsf{Setup}}(D) = \{N, n, (id(D_i), |D_i|, C_i)_{1 \le i \le n}\}$, where N is the number of server index entries, $\mathcal{I}_s$; n is the number of documents, $id(D_i)$ is the document D_i's identifier, $|D_i|$ is the size of document D_i, and C_i is the encrypted corpus.
- $\mathcal{L}^{\mathsf{Search}}(\mathbf{q}_i) = \{R(\mathbf{q}_i), |R(\mathbf{q}_i)|, C_{\mathbf{q}_i}\}$ where $\mathbf{q}_i$ is a client's query, and $R(\mathbf{q}_i)$ is the resultant document identifiers, and $C_{\mathbf{q}_i}$ is the encrypted resultant documents.

Definition 6 *(Forward privacy). A SSE scheme is forward private if the update leakage function $\mathcal{L}^{\mathsf{Update}}$ is limited to: $\mathcal{L}^{\mathsf{Update}}(in, D_i, \mathsf{op}) = \{id_{in}(D_i), |\boldsymbol{w}_{in}|, |D_i|, \mathsf{op}\}$.*

To define the security of the DSSE scheme we employ the standard simulation model which requires a real-ideal simulation [5, 7]:

Definition 7 *(DSSE Security). Let $\mathsf{DSSE} = (\mathsf{GenKey}, \mathsf{BuildIndex}, \mathsf{Encryption}, \mathsf{Search}, \mathsf{Update})$ be a DSSE scheme. Let $\mathcal{A}$ be an adversary (server), and the $\mathcal{L}^{\mathsf{Setup}}$, $\mathcal{L}^{\mathsf{Search}}$, and $\mathcal{L}^{\mathsf{update}}$ be the leakage functions. The following describes the real and ideal world:*

- **Ideal**$^{\mathsf{DSSE}}_{\mathcal{F},\mathcal{S},\mathcal{Z}}(\lambda)$*: An environment $\mathcal{Z}$ sends the client a set of documents D to be outsourced. The client forwards them to the ideal functionality $\mathcal{F}$. A simulator $\mathcal{S}$ is given $\mathcal{L}^{\mathsf{Setup}}$. Later, the environment $\mathcal{Z}$ asks the client to run an* Update *or a* Search *protocol by providing the required information. The search request is accompanied with a keyword w. For an update request, $\mathcal{Z}$ picks an operation from $\{add, del\}$. Add requests are accompanied with a new document and del requests contain a document identifier. The client prepares and sends the respective request to the ideal functionality $\mathcal{F}$. Using $\mathcal{L}^{\mathsf{Update}}$ and $\mathcal{L}^{\mathsf{Search}}$, $\mathcal{F}$ notifies $\mathcal{S}$ of leakages. $\mathcal{S}$ sends $\mathcal{F}$ either* abort *or* $\mathsf{continue}$*. The ideal functionality $\mathcal{F}$ sends the client either* abort *or "success" for* Update*, or set of matching document identifiers for* Search*. Finally, the environment $\mathcal{Z}$ outputs a bit as the output of the experiment.*

- **Real**$^{\mathsf{DSSE}}_{\Pi_F,\mathcal{A},\mathcal{Z}(\lambda)}$*: An environment $\mathcal{Z}$ sends the client a set of documents D to be outsourced. Then, the client executes the* $\mathsf{GenKey}(1^\lambda)$ *to generate the key sk and starts the* $\mathsf{BuildIndex}$ *and* $\mathsf{Encryption}$ *protocols with the real world adversary $\mathcal{A}$. Later, the environment $\mathcal{Z}$ provides the required information and asks the client to run a* Search *or an* Update *request. The search request contains a keyword w to search for. $\mathcal{Z}$ picks an operation from $\{add, del\}$ for an update request. Add requests are accompanied with a new document and del requests contain a document identifier. The client then executes*

the real-world protocols with the server on the inputs that are selected by $\mathcal{Z}$. The client outputs either abort *or "success" for* Update*, or a set of matching document ids for* Search*. $\mathcal{Z}$ observes the output. Finally, outputs a bit b as the output of the experiment.*

We say that a DSSE *scheme (Π_F) emulates the ideal functionality $\mathcal{F}$ in a semi-honest model, if for all PPT real world adversaries $\mathcal{A}$, there exists a PPT simulator $\mathcal{S}$ such that for all polynomial-time environments $\mathcal{Z}$, there exists a negligible function* negl(λ) *on the security parameter λ such that [5]:*

$$|Pr[\mathbf{Real}^{\mathsf{DSSE}}_{\Pi_F,\mathcal{A},\mathcal{Z}(\lambda)} = 1] - Pr[\mathbf{Ideal}^{\mathsf{DSSE}}_{\mathcal{F},\mathcal{A},\mathcal{Z}}(\lambda) = 1]| \leq \mathsf{negl}(\lambda).$$

4 Attack methods

The first step of the attacker to launch an attack is to put the file IDs in a random order. For instance, $(id(D_5), id(D_7), id(D_4))$ is a valid order for a corpus with three documents $\{D_4, D_5, D_7\}$. Based on the chosen arbitrary order, the server creates a bit-string after executing each query. Each bit will be set to one if the corresponding file exists in the result set and to zero otherwise. For instance, suppose after executing a query, $\mathbf{q}_i$, the results set is the $R(\mathbf{q}_i) = \{D_4\}$. Thus, 001 is the corresponding bit-string that is generated based on result set of the current query. Moreover, to keep track of the frequency of each bit-string, the cloud server creates a Search Pattern Map (SPM) which is a hash map data structure. The attacker's main challenge is to track the queries and since the search tokens are one-time use, storing them is pointless. However, the bit-strings that are created in our attacks can be employed as search token identifiers. Hence, the attacker stores them in the SPM along with the number of times that each token is searched.

In other words, the search tokens are ephemeral and change after each use, but the result set for each keyword remain the same and it becomes a major vulnerability for forward private schemes because of the access pattern leakage. For instance, $(11, \{001, 7\})$ can be possible element in SPM that demonstrates the keyword with 001 bit-string has queried seven times. The "11" number is the hash map key that starts from zero and increments by one after adding a new element. The complexity (number of elements) of the SPM is $O(m)$ where m is the number of keywords.

Like other related work [30], in our attacks it is assumed that the bit-strings are unique. To challenge this assumption, we extracted and studies 1927 keywords from the 50,000 files in Enron email dataset [31]. The results shows a scarce 0.2% conflict probability. In other words, there were only 2 conflicts among the investigated files. As a result, the search pattern can be recovered with 99.8% accuracy using our attack. In addition, remark that the keywords that had conflicts were among the very low frequency keywords, thereby, perhaps the cloud server is not interested in. Furthermore, the conflict probability significantly decreases as the number of files in the corpus increase. This is because the state space of bit-string's , all possible bit strings set, expands and becomes larger. Nevertheless, we later address this attacker's challenge and describe how keywords with unique bit-string can be distinguished. We emphasize that all assumption in related work [13, 14, 30, 32] and our work are consistent and we add no new assumption in this attack.

The basic attack is explained in Algorithm 1. Briefly, once each query is executed, the cloud server looks in the SPM to find a match for the resultant bit-string (r_bitString). If there is a match, the server increments the respective frequency by one, otherwise, the new bit-string will be added to SPM with frequency of one (line 24).

Algorithm 1 Basic Attack

input: SPM, r_bitString
output: updated SPM

```
1:  found = false
2:  for each e ∈ SPM & until !found do
3:      e_tmp = e.bitString
4:      r_tmp = r_bitString
5:      flag = true
6:      while flag & e_tmp > 0 do
7:          e_rem = e_tmp  mod 2
8:          r_rem = r_tmp  mod 2
9:          if e_rem ≠ r_rem then
10:             flag = false
11:         end if
12:         e_tmp / = 2
13:         r_tmp / = 2
14:     end while
15:     if flag then
16:         found = true
17:         match = e
18:     end if
19: end for
20: if found then
21:     match.bitString = r_bitString
22:     match.frequency++
23: else                                   ▷ new keyword found
24:     Add (r_bitString, 1) to SPM
25: end if
```

Nevertheless, the length of the bit-string can be affected by "add" operation. In other words, adding a new file increases the length of the bit-string. To tackle this issue, the new file ID will be added to the left of the arbitrary order by the cloud server. Recall the previous example and suppose the server has received a new request to add D_6 to the dataset. The updated arbitrary order will be $(id(D_6), id(D_5), id(D_7), id(D_4))$. Hereafter, SPM bit-strings are a bit shorter than queries' bit-strings. However, this does not stop the server\attacker from recovering the search pattern, because, the attack algorithm compare the SPM and query bit-strings bit-wisely starting from left bit to the right. The algorithm halts (line 6) when it achieves the last bit of the respective SPM bit-string. For instance, suppose $\{D_4\}$ and 001 are the result set and respective bit-string of query, $\mathbf{q}_i$. If the user issues the same query $\mathbf{q}_i$ again, after adding D_6 and of course with a new search token, the resultant bit-string would be either 0001 or 1001. Remark that only one can happen at a time, because either the new file, in this case D_6, contains respective keyword in $\mathbf{q}_i$ or not. Hence, if the server detects a bit-string in SPM that matches the first three bits (from right-side) of the resultant

bit-string, it can be confident that these two token IDs refer to same keyword. Remark that, the respective bit-string will be updated to $n+1$ from n-bit string in line 21 of the algorithm where n is the number of files. This means, in our example the bit-string will be updated to a four-bit from a 3-bit string.

The traditional attacks are not effective on schemes that support forward privacy. The main reason is that forward-private approaches hide and obfuscate the search pattern to a certain extent. However, by applying our attack on schemes with forward privacy, we reveal the search pattern. Once the attacker possess the search pattern, forward-private approaches will become susceptible against previous attacks. The output of our attack can be exploited by frequency-based attacks such as [12]–[14], [32]. Moreover, after applying a small modification, our attack can be used by occurrence-based attacks such as [30]. To support the occurrence-based attacks the attacker creates a $n \times m$ matrix $\mathcal{M}$ instead of SPM. In this matrix each column represents a bit-string\keyword, and each row corresponds to a document. We set the value of an entry to zero if the respective keyword does not exist in the corresponding document, and to one otherwise. Once a query is executed, the cloud server updates the value of the respective entry, If it finds the same bit string, or it adds the bit-string as a new keyword otherwise. If a new file is added, the cloud server append a new row to the matrix $\mathcal{M}$.

To address the problem of distinguishing the keywords with the same result set, the cloud server can inject a limited numbers of documents into the corpus (keywords with the same result set). For instance, suppose $\{k_1, k_2\}$ and $\{k_3, k_4, k_5\}$ are two groups of keywords that have the same result set. The attacker can distinguish k_1 from k_2 by injecting a file that contains either of the keywords. The same method can be used to make the other group keywords distinguishable. To maximize the efficiency, we should minimize the number of injected files. Hence, the attacker creates new files that contains only one keyword from each group. For instance, the attacker creates a file that contains k_1 and k_3 from the first and second group. It also generates a another file which only contains k_4. With injecting only two files these keywords will become distinguishable. Generally, suppose we have l groups of keywords that possess the same result set, $\{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_l\}$, the minimum number of required injected documents is $\max_{j=1}^{l}\{|\mathcal{P}_j|\} - 1$, in which $|\mathcal{P}_j|$ shows the cardinality of j-th group.

This techniques is also employed in many related work such as [13, 32, 14] in which the cloud server sends the documents of its choice to the user. The client then encrypts and transfers them back to the cloud [14]. For example, consider a company that uses an automatic email process. Remark that in both occurrence-based and occurrence-based attacks the attacker commonly benefits from public information and auxiliary knowledge that rectifies the indistinguishable keywords challenge in advance. For instance, the attack in [12] benefits from public web facility services such as Google Trends®.

Nevertheless, the basic attack is not applicable on DSSE approaches that provide "*del*" functionality. We modified the basic attack, see Algorithm 2, that can successfully attack DSSE schemes that provide both *add* and *del* functionality. In our *advanced attack*, a new bit-string, *d_bitString*, will be generated by the cloud server to monitor the deleted documents. Each bit in the *d_bitString* represents a document in the corpus (considering the same arbitrary order). Each bit will be set to zero if the respective file still exists in the corpus, and to one if it is deleted. After executing a delete request, the cloud updates this bit-string accordingly. Once a query is executed, the resultant bit-string will be bit-wisely compared to the data in SPM, but this time, we ignore the bits that their corresponding files are deleted. We demonstrate the changes with red lines in Algorithm 2.

Algorithm 2 Advanced Attack

input: SPM, r_bitString, d_bitString
output: updated SPM

```
1:  found = false
2:  for each e ∈ SPM & until !found do
3:      e_tmp = e.bitString
4:      r_tmp = r_bitString
5:      d_tmp = d_bitString
6:      flag = true
7:      while flag & e_tmp > 0 do
8:          d_rem = d_tmp mod 2
9:          if !d_rem then
10:             e_rem = e_tmp mod 2
11:             r_rem = r_tmp mod 2
12:             if e_rem ≠ r_rem then
13:                 flag = false
14:             end if
15:             e_tmp / = 2
16:             r_tmp / = 2
17:         end if
18:         d_tmp / = 2
19:     end while
20:     if flag then
21:         found = true
22:         match = e
23:     end if
24: end for
25: if found then
26:     match.bitString = r_bitString
27:     match.frequency++
28: else                                   ▷ new keyword found
29:     Add (r_bitString, 1) to SPM
30: end if
```

4.1 An example of our attack with "del" operation

Assume there are four documents, $\{D_1, D_2, D_3, D_4\}$, and four keywords, $\Delta = \{w_1, w_2, w_3, w_4\}$, in the user's corpus. Moreover, $(id(D_4), id(D_3), id(D_2), id(D_1))$ is the arbitrary order that the attacker\cloud uses to create the bit-strings. Suppose D_1 contains $\{w_1, w_2, w_3\}$, D_2 includes $\{w_2, w_3, w_4\}$, D_3 has $\{w_1, w_3\}$, and D_4 contains $\{w_2, w_4\}$. Furthermore, we assume the setup phase is successfully executed, and the encrypted index and documents are outsourced.

Case 1: Searching for a keyword for the first time. The user searches for all documents that contain w_1, so he generates an ephemeral search token ($\mathbf{q}_1$) and send it to the server. Upon

receiving the search token, the server finds the related index entries and the respective encrypted documents ($R(\mathbf{q}_1) = \{D_1, D_3\}$). Since there is no bit-string in the SPM that matches the current bit-string, the server adds the respective bit-string $(0101, 1)$ to the SPM (1 is the frequency). The user then generates and sends an encrypted query ($\mathbf{q}_2$) to the server to search for w_3. After returning the results ($R(\mathbf{q}_2) = \{D_1, D_2, D_3\}$), the server adds $(0111, 1)$ to the SPM.

Case 2: Searching again for a keyword that exists in the SPM. Now imagine the user searches again for w_1 with a new ephemeral search token ($\mathbf{q}_3$). Once the query executed, the server searches for the resultant bit-string (0101) in the SPM. Since the bit-string already exists in the SPM, the server only updates its frequency $(0101, 2)$.

Case 3: Adding a new document. The user then adds a new document (D_5) that contains (w_1, w_2, w_4). The server updates the arbitrary order to $\{id(D_5), id(D_4), id(D_3), id(D_2), id(D_1)\}$ respectively. Now the user issues a new query ($\mathbf{q}_4$) to search for w_4 for the first time. The resultant bit-string will be 11010. Hence, $(11010, 1)$ will be added to the SPM. Note that at this point the bit-strings in the SPM may have different length (4 and 5). The user may also searches again for a keyword after adding a new document. For example, suppose the user searches again for w_3. The new resultant bit-string is 00111. The server looks for a bit-string in the SPM that is either equal to our current bit-string or is equal to the first four bits (from right-side) of the resultant bit-string. In this case the sever will find the $(0111, 1)$ entry and will update it to $(00111, 2)$ respectively.

Case 4: Deleting a document. To monitor the deleted documents, the attacker creates a bit-string, $d\_bitS\,tring$, in which each bit represent a document and its value demonstrates whether the respective file is deleted (=1) or not. This vector will be updated after each delete request. Once a query is executed, the resultant bit-string will be bit-wisely compared to the data in SPM, but this time, we ignore the bits that their corresponding files are deleted.

Assume the user asks server to delete D_3. The server deletes the document and updates the $d\_bitS\,tring$ to 00100. If the user searches for w_1 again, the resultant bit-string will be 10$*$01 ("$*$" means its value is not important and can be 0 or 1). Since D_3 is deleted (based on the $d\_bitS\,tring$), the server ignores the value of that position when it is searching for the current bit-string in the SPM.

5 Construction

To prevent the attacks that we introduced in the previous section, we build and construct a new dynamic SSE scheme that supports forward privacy. Our construction hides and obfuscates the access pattern to thwart the above privacy attacks. In our approach, the client first creates an inverted index for each keyword in the corpus. In an inverted index, every entry maps a keyword to the corresponding document IDs that contains the respective keyword. We add fake\noise document IDs among each keyword result-set to hide and obfuscate the access pattern. The fake IDs can only be distinguished by the client. In addition, in our approach each query $\mathbf{q}_i$ consists of a limited number of sub-queries $\mathbf{q}_i = \{q_{i1}, q_{i2}, \ldots, q_{ik}\}$. All sub-queries except one are noise which can only be recognized by the user and sub-query searches for a keyword. We propose two methodologies to inject the noise file IDs:

1. **Random injection.** We first determine a threshold , τ_d, that represents the lower and upper bound of noise injections in a specific result-set. Once τ_d is set, we arbitrarily inject file IDs into the result set of the every keyword.

2. **Aforethought injection.** Our main objective is to flatten the the number of times that each document accessed. With this strategy the number of times that each document is accessed will be in the same range as others. This makes it much harder for the attacker to gain information about the data, search tokens, and the files by using the access pattern meta-data. To achieve this objective, the client generates an Access Pattern Vector (APV) that stores the number of times that each file is accessed. For instance, $< 1, 0, 4, 2 >$ demonstrates D_1, D_3, and D_4 are accessed one, four, and two times since launching the system. This information is valuable in our approach and will help to choose the fake sub-queries wisely. For each query, the user monitor and analyze the APV vector and selects the keywords that are accessed less in compare with others. Considering our earlier example ($< 1, 0, 4, 2 >$), the client can inject keywords that return D_1 and D_2 to straighten the access pattern vector.

Creating the noise sub-queries are a significant challenge in the aforethought injection. To query the less requested files in the corpus, the user must be able to identify the keywords that return a specific file ID in their resultant-set. The number of files in the corpus is myriad and increasing over time, thereby keeping this information on the users' devices is not a realistic approach. In addition, to hide the clients' foot tracks (activities), we intend to put each file ID into more than one keyword's result-set. This makes the above solution more infeasible.

To tackle this problem, we first label all of the files with a number in a random order. Note that even storing these labels on the users' machines are not practical. Hence, we then generate an arbitrary number called δ. Afterward, we begin with δ and label each document successively. remark the files are shuffled before being labeled with successive numbers, so, the labels will not leak additional meta-data about the files. To identify the keyword w that a specific file ID, D_i, is injected in, we employ the $G_w(.)$ PRF. $G_w(.)$ accepts a key k and a document label and returns a keyword number, m, where $m \in [1..m]$. For instance, in $G_n(k, 7654321) = 123$ the user is searching for a keyword number (123) that holds a document that is labeled as 7654321. In other words, after calculating $G_n(k, 7654321) = 123$, the user learns that the file with 7654321 label is injected in the w_{123}' result set.

To inject each file ID a number of times, we define a new parameter called step, s_δ. Step (s_δ) shows the number of result-sets that each file ID is injected into. As a result, we increase the label number by s_δ instead of labeling them consecutively. By exploiting this technique we expedite the flattening process of the APV vector. Therefore, instead of possessing one label, each file will have s_δ labels that is reserved for the respective file. For instance, suppose $s_\delta = 3$ and we want to search for keyword w_i that returns the document that is labeled as 7654321. Hence, the client executes $G_n(k, 7654321) = 123$, $G_n(k, 7654322) = 77$, and $G_n(k, 7654323) = 2105$. This means, keywords w_{123}, w_{77}, and w_{2105} are having file 7654321 in their result-sets. Consider that, due to the

definition of the pseudo random function the keyword numbers are not necessarily successive even if the the input labels are.

5.1 Our linked-list structure

In this section we explain a customized linked-list data structure that we employ in our scheme. Our linked-list consists of 4 tuples: $(id, type, data, next)$. In particular, *id* refers to the ID of a node, *type* shows whether a node is real *R* fake\noise (F), *data* is a file identifier, and *next* refers to the ID of the next node in the respective linked-list. The red elements will be secured and encrypted using the user's key, while an ephemeral key will be used to encrypt the orange elements. Hence, *type* will be encrypted using the user's key and, we will be using an ephemeral key to encrypt *data* and *next*. Note that elements in black (*i.e.*, *id*) is plaintext data. To prevent leaking any additional meta-data, we use a random even number for real and a random odd number for noise nodes. In addition we set the *next* to null if a node does not have a successive node. Consider that, the ephemeral key will be provided for the server if the respective keyword is being queried, but we never share the user's secret key with server. Hence the server will never know which nodes are fake and which ones are real.

- AddNode($\mathbb{L}, k_1, k_2, id, type, data$): This function employs the input data to append a node to the beginning of the current linked-list, $\mathbb{L}$.
- RestoreList(k, id_h): This function looks for the node with the provided ID; it then retrieves *type* and decrypts *next* and *data* and looks for the next node in the linked-list. The process stops when the algorithm reaches the last node in the linked-list (*i.e.*, $next = null$).

In our approach, all of the secure inverted index is constructed using the aforementioned linked-list data structure. To add more security, the client injects fake\noise nodes in each and every linked-list to hide and obfuscate the relevance between a keyword, w_i, and its corresponding linked-list, L. Note that the noise will not be injected if it already exists in the linked-list. for instance, assume that one of the inverted index entries is $(w_7, \{D_3, D_6, D_5\})$. The objective is to inject D_6 and D_9 in random positions in the respective linked-list. however, we only inject D_9 because D_6 is already in the linked-list. Now suppose after injecting the noise node(s) the index entries will be $(w_7, \{D_3, D_6, D_9, D_5\})$. Hence, the user generates four nodes $(id_1, R, id(D_3), id_2)$, $(id_2, R, id(D_6), id_3)$,$(id_3, F, id(D_9), id_4)$ $(id_4, R, id(D_5), null)$. Consider that, since the nodes are encrypted and will be sent in a random order to the server, the attacker is not able to link them.

5.2 Our scheme

In this section we demonstrate and describe how each algorithm in Definition 3 operates:

- GenKey: let SE = (Gen, Enc, Dec) be a CPA-secure symmetric encryption scheme. Let $G_n(.)$, $G_{id}(.)$ and $G_w(.)$ be three PRFs and GenPK(1^λ) be a key generator function. The following describes $(sk, \perp) \leftarrow$ GenKey($1^\lambda, 1^\lambda$):

  1: k_{SE} = SE.GEN(1^λ)
  2: $k_G \leftarrow$ GenPK(1^λ)
  3: *return* $sk = (k_{SE}, k_G)$

  We employ a secret key, *sk*, to fulfill the encryption objectives. *sk* is a tuple of two, k_{SE} and k_G. The former will be used to encrypt the documents and latter for $G_w(.)$, $G_n(.)$ and $G_{id}(.)$ functions.

Algorithm 3 $(\mathcal{I}_c, \mathcal{I}_s) \leftarrow$ BuildIndex($(sk, D, s_\delta), \perp$)

Client
1: Δ = ExtractKeywords(D)
2: δ = Rand()
3: $lbl_{next} = \delta$
4: $D' = \{\}$
5: **while** $D \neq empty$ **do**
6: D_{cur} = randomly choose one doc and assign lbl_{next} to it
7: $D' \cup D_{cur}$
8: $id_{next} += s_\delta$
9: **end while**
10: PI = BuildPlainIndex(Δ, D', s_δ)
11: Create $\mathcal{I}_c$ and APV and initialize all elements to zero
12: $\mathbb{L}$ = GenLinkedList(sk, PI)
13: Send $\mathbb{L}$ to server
Server
14: Generate $\mathcal{I}_s$ using $\mathbb{L}$

- BuildIndex. In this algorithm (see Algorithm 3), after extracting the keywords, $\Delta = \{w_1, \ldots, w_m\}$, we assign an label\id to each file according to the value of the δ (the starting point), and s_δ (step). Afterward, the client index $\mathcal{I}_c$, and its corresponding linked-lists will be created. To enable the client to generate the search queries, the client must store a $m \times 2$ look-up table (index), $\mathcal{I}_c$. In particular, this table stores length of the list, len_i and the number of nodes, cnt_i, that is generated for each keyword. Moreover, the APV (access pattern vector) will be created and initialized to zero. Once the index entries are generated, they will be sent to the cloud server. Note that a random number will be assigned to δ which is generated by Rand().

  We explain generating the plain-text inverted index *PI* in Algorithm 4. Using the aforethought injection method, we first generates the noise nodes (line 5-9), and then we append the real nodes to their corresponding list (line 11-15). Note that as we mentioned in Section 5.1, every entry in *PI* consists of two tuples which are a keyword and the receptive list (w_i, L). recall that beside the file label, each node in the list keeps a type (F\R). For instance, D_2 is fake, and D_4 and D_8 are real nodes in $(w_{23}, \{\{D_8, R\}, \{D_2, F\}, \{D_4, R\}\})$. In addition, to decide the lists that each file ID must be injected in, we employ $G_w(k_G, doc_{id})$ function. Consider that this is process happen in the setup phase and only for once. Next, a linked-list will be generated for each generated list in the previous step. The node IDs are one-time use and will be generated by the G_{id} function. In particular, the G_{id} function uses a counter, *ctn*, which shows the number of nodes that are created for the respective keyword, w_i. The client store this number in the client index ($\mathcal{I}_c[i][0]$).

Algorithm 4 BuildPlainIndex

```
1:  procedure BuildPlainIndex(Δ, D′, s_δ)
2:      PI = {}
3:      for all w_i in Δ creates an empty L_i
4:      for all D_j in D′ do
5:          for k = 0 to s_δ-1 do
6:              i = G_w(k_G, (lbl(D_j) + k))
7:              if D_j ∉ L_i then
8:                  L_i ∪ {D_j, F}
9:              end if
10:         end for
11:         for all w_i in Δ do
12:             if w_i ∈ D_j then
13:                 L_i ∪ {lbl(D_j), R}
14:             end if
15:         end for
16:     end for
17:     Shuffle and Add all (w_i, L_i) to PI
18:     return PI
19: end procedure
```

Algorithm 5 GenLinkedList

```
1:  procedure GenLinkedList(sk, I_c, PI)
2:      L = {}
3:      for all e ∈ PI do
4:          w_i = e.w_i
5:          L = e.L_i
6:          len = 0
7:          cnt = I_c[i][0]
8:          for all lbl_doc & type ∈ L do
9:              id_n = G_id(k_G, w_i||cnt)
10:             if len == 0 then
11:                 k_h = G_n(k_G, id_n)
12:                 L_s = {}
13:             end if
14:             AddNode(L_s, k_SE, k_G, id_n, type)
15:             len + +
16:             cnt + +
17:         end for
18:         L ∪ L_s
19:         I_c[i][0] = cnt
20:         I_c[i][1] = len
21:     end for
22:     return L
23: end procedure
```

We employ an ephemeral key to encrypt the private data in each node. This key will be generated using the receptive function, $k_h = G_n(k_G, id_n)$, which is shown in line 10-13 of Algorithm 5. All of the data in a linked-list will be encrypted using the same ephemeral key except the type data. We employ the secret key, k_{SE}, to encrypt the type field, because the client should be the only party who can decrypt this data. For instance, assume the client intends to create a secure linked-list for w_{21}'s list, $\{D_4, D_7\}$. Assuming $cnt = 0$, we generate the node IDs, $id_{10} = G_{id}(k_G, w_{21}||0)$, $id_{11} = G_{id}(k_G, w_{21}||1)$. Afterward, we create an ephemeral key $k_h = G_n(k_G, id_{10})$ to encrypt the nodes.

Remark that for the whole linked-list, we only generate one ephemeral key (see line 10). Lastly, using AddNode, we create and encrypt the required nodes and add them to the corresponding linked-list. We demonstrate how we create a linked-list from an inverted plain-index in Algorithm 5. Once all of the nodes and required linked-list are created, the user sends them to the cloud server to be stored on the server index, $\mathcal{I}_s$. Remark that the index entries are encrypted and sent in an arbitrary order and the server cannot link them.

- Encryption. Using the secret key k_{SE}, the client encrypts the entire corpus (all of the files), transfers them to the cloud server including the file IDs.

Algorithm 6 $((\mathcal{I}'_c), (\mathcal{I}'_s, C')) \leftarrow$ Update$((sk, \mathcal{I}_c, add, \text{in}), (\mathcal{I}_s, C))$

```
    Client
1:  Δ_D = ExtractKeywords(D_{n+1})
2:  L = {}
3:  for all w_i ∈ Δ_D do
4:      cnt = I_c[i][0]
5:      len = I_c[i][1]
6:      id_h = G_id(k_G, w_i||cnt)
7:      id_n = G_id(k_G, w_i||cnt − len)
8:      k_h = G_n(k_G, id_n)
9:      L ∪ (id_h, k_h)
10:     I_c[i][0] + +
11:     I_c[i][1] + +
12: end for
13: C_{n+1} = Enc(k_SE, D_{n+1})
14: Send L, C_{n+1} to server
    Server
15: C ∪ C_{n+1}
16: Update I_s using L
```

- Search. Obfuscating and hiding the access and search pattern is our primary goal. To achieve this objective, we append a bounded number of sub-queries, $\mathbf{q}_i = \{q_{i1}, q_{i2}, \ldots, q_{ik}\}$, to each query $\mathbf{q}_i$. Every sub-query consists of two tuples, (k_h, id_h). k_h is an ephemeral key that decrypts a linked-list starting with id_h (linked-list's header). Employing $\mathcal{I}_c$ and k_G enable the user to re-generate k_h and id_h which are required for generating the query. In addition, the noise sub-queries will be added to boost the security and privacy of the outsourced data and obfuscate the access and search pattern.

  To flatten the APV and amplify the privacy of the outsourced files, we use a biased random generator function called GenRandom(). This function chooses high accessed documents with lower probability, so the less accessed files has more opportunity to be selected which impact directly the pace of flattening the APV. Note that we assign s_δ number of labels to each file. Hence, along with APV, GenRandom() inputs s_δ to randomly choose one of the available labels for the file of interest. ν holds the number of fake\noise queries that can be determined randomly. Before

sending the query to the cloud server, we insert the sub-queries in arbitrary positions in query $\mathbf{q}_i$.

The query generation process for a keyword w is demonstrated in Algorithm 7. Remark that the node ID (of the linked-list header) and its respective ephemeral key are regenerated in line 15. Since the users may possess various devices with different level of computation power and resources, a number of parameters including ν and s_δ can be set by the client.

Once a query, $\mathbf{q}_i$, is received, the cloud server runs each sub-query $q_{ij} = (id_h, k_h)$ by finding id_h (header of the requested linked-list) in $\mathcal{I}_s$. Employing the k_h, the server then decrypts all of the nodes (except the type field) and discards all of the used index entries. All of the extracted document IDs and their respective type fields will be added to the result set. Note that the server can store the used index entries, however, it is pointless because they are already leaked and enclose no new meta-data. Lastly, the server returns a result set consists of $(R(q_{i1}), \ldots, R(q_{ik}), bag)$ where $R(q_{ij})$, $1 < j \leq k$, is the q_{ij}'s resultant file IDs; and the bag is $\mathbf{q}_i$'s resultant encrypted files.

Once the result set of $\mathbf{q}_i$ is received, the client who is aware of the location of the noise and real sub-queries, decrypts and separates the real results form the bag. Next, the GenLinkedList algorithm will be called to generate new entries for queried keywords in $\mathbf{q}_i$.

The forward privacy of our approach is guaranteed by using non-deterministic search tokens and adding random noise sub-queries. The client then updates its index, $\mathcal{I}_c$, and creates new node IDs and ephemeral keys for each linked-list. Note that, since the value of the cnt is updated, brand new keys and node IDs will be generated. To track the access frequency of each file, the client then updates the APV. lastly, the new index entries will be sent to the cloud server to be stored on the server index, $\mathcal{I}_s$.

Albeit the cloud server is aware of the relation between the last query and new entries, it cannot determine the noise keywords from the real search keyword. In addition, the node IDs and their respective keys are one-time use and vary after each search. Furthermore, there exist fake\noise nodes among the actual nodes in every linked-list. As a result, it is impossible for the server to realize the actual search and access pattern. All of these specifications in our approach guarantee the forward privacy requirement and preserving the access and search pattern. The search process is described in detail in Algorithm 7.

- Update. The update algorithm consists of two functions, *del* and *add*, as follows:
  - *add*. In *add* algorithm, we first extract the keywords from the new file. The algorithm then generates a node for each keyword and adds them to the respective linked-list. Next, we encrypt the the new file using the user's secret key and transfer it to the cloud server. On the other side, the server updates the $\mathcal{I}_s$, once the the Update request is received. The *add* function is described in detail in Algorithm 6.
  - *del*. The user creates a Update request and sets the operation to *del* and includes the file ID in the request to delete a file, D_k. Upon receiving the *del* inquiry, the cloud server deletes the respective file form its storage. Nevertheless, the corresponding index entries cannot be removed because the server does not possess the keys.

Algorithm 7 $((\mathcal{I}'_c, D_w), (\mathcal{I}'_s)) \leftarrow$ Search$((sk, \mathcal{I}_c, w, \nu, APV), (\mathcal{I}_s, C))$

Client
1: $counter = 0$;
2: $\Delta_q = \{w\}$
3: **while** $counter < \nu$ **do**
4: lbl = GenRandom(APV, s_δ)
5: $i = G_w(k_G, lbl)$
6: **if** $w_i \notin \Delta_q$ **then**
7: $\Delta_q \cup w_i$
8: $counter++$
9: **end if**
10: **end while**
11: $\mathbf{q} = \{\}$
12: **for all** w_i **in** Δ_q **do**
13: $cnt = \mathcal{I}_c[i][0]$
14: $len = \mathcal{I}_c[i][1]$
15: $id_n = G_{id}(k_G, w_i || cnt - len)$
16: $id_h = G_{id}(k_G, w_i || cnt)$
17: $k_h = G_n(k_G, id_n)$
18: $\mathbf{q} \cup (id_h, k_h)$
19: **end for**
20: Shuffle$(\mathbf{q})$
21: Send $\mathbf{q}$ to server
Server
22: $bag = \{\}$
23: **for all** q_i **in q do**
24: Find respective $node$ with id_h
25: **while** $node \neq null$ **do**
26: Decrypt the $node$ using k_h in q_i
27: Add lbl and $type$ to $R(q_i)$
28: Find $next\ node$
29: **end while**
30: Add files corresponds to $R(q_i)$ to bag
31: **end for**
32: Send $(R(q_1), \ldots, R(q_k), bag)$ to client
Client
33: Decrypt results R
34: $PI = \{\}$
35: update APV based on the results
36: **for all** w_i **in** Δ_q **do**
37: $L =$ All doc-ids contain w_i in R
38: $PI \cup (w_i, L)$
39: **end for**
40: $\mathbb{L}$ = GenLinkedList$(sk, \mathcal{I}_c, PI)$
41: Send $\mathbb{L}$ to server
42: Delete noise results
43: Consume real results
Server
44: Update $\mathcal{I}_s$ using $\mathbb{L}$

However, the index entries will be removed over time and after receiving a number of queries. The server simply removes the nodes that are pointing to a deleted file. To incorporate this

feature in Algorithm 7, the server first investigate the availability of the a file extracted from a node (line 28). The server removes them from the result set, if they do not exist.

6 Experimental results and complexity analysis

To assess the efficiency of our approach, we study and compare the complexity of the state-of-the-art methods [7], [6], [5] with our approach. Lastly, we finish this section by demonstrating the experimental results that are obtained using real world datasets.

6.1 Analyzing the complexity of our proposed algorithm

Required storage space for client and server. We first show that the amount of data that the server and especially the client should store is reasonable and manageable. Recall that user must store a dictionary, $\mathcal{I}_c$, on her side which holds the number of nodes (a counter) and the length of each linked-list for each keyword. Hence, the client index look likes a table with two column and m rows where m is the number of keywords. Hence, $\mathcal{I}_c$ is an $O(m \times 2) \approx O(m)$ dictionary. Assume the user's dataset consists of 1M keywords. Moreover, suppose each integer requires 4 bytes on the memory and each keyword has an average 10 bytes. In this scenario, the user needs to store a 18 MB dictionary on her side ($1\text{M} \times (10 + 4 + 4) \approx 18\text{MB}$). Considering resource-constrained devices such as cellphones which have limited memory space and constrained computations, 18 MB is rational, cost-efficient, and manageable. As an alternative, by using the method in [7] also proposed, it is feasible to outsource the user index. In comparison to other work, in [7] the author needs $O(m + n)$, in [5] the author requires $O(\sqrt{N})$, and in [6] the author occupies $O(m)$, where n is the number of documents and N is the number of $(keyword, doc\ id)$ tuples. Regarding the size of the server index, our method needs $O(N + k)$, in which k is the number of fake\noise nodes. All state-of-the-art methods that we mentioned above require a space with size of $O(N)$ to store the server index.

Supporting parallelism by design. Beside our approach, this requirement is also fulfilled in [7] among the Dynamic SSE schemes that support forward privacy. Since in our approach the node IDs are generated by a pseudo-random function and the server index entries are independent, it is possible to distribute the sub-queries among the processors to expedite the update and search process, and achieve parallelism. The complexity of our search method is $O(d + k_d)/p$ and our update (add) system-cost is $O(r/p)$, where p is the number of cores\CPUs, d holds the number of a files containing a keyword, k_d shows the number of fake nodes in a keyword list, and r holds the number of keywords in a file. The best-case scenario happens when the number of sub-queries are equal to the number of available cores\CPUs. As a result, all sub-queries will be executed concurrently. The search cost in [7] is $O(d + n_d)/p$ and the add\update cost is $O(r/p)$, where n_d shows the number of times that a keyword has been affected by file deletions since last search. Table 1 shows our complexity analysis.

Table 1: Complexity Analysis of Related Work and Our Approach

Approach	$\mathcal{I}_c$	$\mathcal{I}_s$	Parallelism	$Search$	$Update$
Stefanov[5]	$O(\sqrt{N})$	$O(N)$	–	$O(d)$	$O(r)$
Bost[6]	$O(m)$	$O(N)$	–	$O(d)$	$O(r)$
Etemad[7]	$O(m + n)$	$O(N)$	✓	$O(d + n_d)/p$	$O(r/p)$
Ours	$O(m)$	$O(N + k)$	✓	$O(d + k_d)/p$	$O(r/p)$

6.2 Experimental results

We implemented a prototype and conducted a through and comprehensive evaluation to study our approach using real-life datasets. We employed Java (JDK 1.8) as the programming language and Crypto packages for the encryption process. The server and client connect and communicate through a TCP connection. Moreover, Windows machines were used for both server and client. Each machine came with 8GBs RAM and a Corei7 CPU at 3.6 GHz. To assess our scheme, we used the real-world Enron email dataset [31]. We ran each experiment ten times and the output is the average of all trials. The variance of the 10 trials were very low to be notable. We implemented the search\query algorithm twice, once using a parallel algorithm (four cores) and another time in a sequential manner. We call the former *multi-threaded* and the latter *single-threaded*. The results shows an admissible and reasonable overhead on the system that even a user with a resource-constrained device can benefit from our approach.

Table 2: # of server index entries

#Docs	#server entries
10000	829799
20000	1571676
30000	2568438
40000	3548027
50000	4404160
60000	5341524
70000	6194452

Setup time. We first started by studding the setup time per various number of files. As we discussed in Section 5.2, the setup phase includes several steps including the encryption process, creating the plain index, and the encrypted linked-lists. Our results indicate that a dataset with 20K files requires less than minute (≈ 59 *sec*), while the same experiment, setup process, for a corpus with 50K needs less than seven minutes to be finished (see Figure 2). Remark that, this process only happens at the beginning of our approach, so it is a one-time process. In addition, we investigated the number of server index entries. Our study shows that around 4.4×10^6 entries were generated for 50K files, and 1.57×10^6 for 20K documents (see Table 2).

Query generation process time. To search for a keyword, the user needs to create a query. Each query consists of numerable search tokens\sub-queries. Every search token includes the header ID of a linked-list and its receptive key. To study the impact of number of fake\noise sub-queries on the query generation process, we queried the same keyword several times but with various number of fake keywords. The results demonstrate that the system requires less

than 1.5 **milli**seconds to generate a query which contains 50 fake keywords. Remark that we used 50000 files for this experiment.

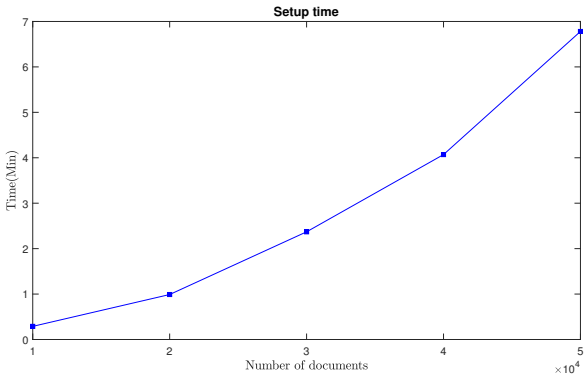


Figure 2: Client setup time

Search time. Once the server receives a search request, it will look for all files that match the search token. In this experiment we aimed to measure the amount of the time that each query requires. However, the size of the result-set (number of resultant files for a specific query) is a crucial factor in this experiment. Hence, we created ten keywords and injected them into our corpus with frequencies from 100 to 1000. These new injected keywords will enable us to investigate the effects of the number of resultant documents on the search time. Our results from the single threaded algorithm show that it takes around two seconds from the server to execute a query with 1000 resultant files. Moreover, multi-threaded algorithm requires considerably less time(around 45 percent) to answer the same query. Note that, the noise was set to three in both experiment settings.

Update (Add\Delete) requests. The user may request for an update on the corpus that can be a delete or an add request. Creating a delete query is a very low-cost operation, and takes less than 1ms in our approach. To remove a file, the user should sets the operation mode to *del* and embeds the file ID in the query. To add a new file, we first remove the stop words and extracts the main keywords. We then creates the index entries, plain index, encrypted linked-lists, and encrypt the file. Lastly, we transfer it to the cloud server. Once the cloud server receives the add query, it adds the encrypted file to the corpus and store the index entries. Note that because all of the index entries are encrypted with a new ephemeral key, the cloud server cannot determine the relation between current files in the corpus and the new file. To run our experiment, a file with 155 words and 68 keywords (excluding the stop words) from Enron dataset was selected. The operation lasts around 1.8 ms.

Obfuscating the Access Pattern. The most important goal in our approach is preserving the search and access pattern. To achieve this goal, we injected noise nodes among each linked-list's nodes, and also added noise sub queries to the main query. With this strategy, the access pattern vector (APV) become flattened and obfuscated. To measure how flattened\uniform the APV become before and after applying our approach, we used the Shannon entropy. Due to not having access to the real-life search requests, we randomly selected the queries from the keyword-set. We set the noise to three and issued 1000 queries. To calculate the entropy improvement we employed $((e_{our} - e_{org})/e_{org}) \times 100$, where e_{org} and e_{our} are the Shannon entropy of the calculated APV before and after applying our approach. Figure **??** demonstrates our results in detail. To illustrate, applying our approach on a corpus with 30K files flattens the APV more than two times. This means our approach has made the access pattern more secure and private more than twice.

For example, exploiting our approach on a corpus of 30K documents flattens the access pattern vector more than two times. That is, the access pattern is two times more private than before applying our scheme.

7 Security proof

We defined the DSSE Security in Definition 7 and designed our dynamic SSE scheme in Section 5.2. Here, we prove that our scheme is secure using the standard simulator model.

Theorem 1 *Let* $\mathsf{SE} = (\mathsf{Gen}, \mathsf{Enc}, \mathsf{Dec})$ *be a CPA- secure symmetric encryption scheme, and* G_n, G_{id}, *and* G_w *be three pseudo-random functions, our DSSE scheme in Section 5.2 is secure under Definition 7.*

Proof 1 *We demonstrate how the ideal world is indistinguishable from the real world by any probabilistic polynomial time (PPT) distinguisher to prove that our dynamic and forward private SSE scheme is secure. We illustrate and explain a PPT simulator* $\mathcal{S}$ *that imitate the user actions using the provided leakage functions that are defined and provided in Section 3.5. In other words, we explain how a simulator,* $\mathcal{S}$*, can adaptively mimic the user behavior including generating the encrypted indexes, queries, and documents:*

Setup. *In the first step, the simulator* $\mathcal{S}$ *generates the encrypted document set , C, simulates N index entries, and creates a secret key,* k_{SE}*. To generate the simulated data,* $\mathcal{S}$ *employs leakage function* $\mathcal{L}^{\mathsf{Setup}}(D) = \{N, n, (id(D_i), |D_i|)_{1 \le i \le n}\}$*. Note that all data including the index entries,* $\mathcal{I}_s$*, and generated files, C, are encrypted with the secret key that was generated earlier. This means, the simulator* $\mathcal{S}$ *does not require to have access to the contents of the files, and as a result, it encrypts strings of size* $|D_i|$ *containing all zeros to create the encrypted files. Note that no probabilistic polynomial time distinguisher (attacker) can detect and discern this behavior due to the CPA security of the applied symmetric encryption scheme. Moreover, the simulator requires to generate and keep two dictionaries, keyDict and* Δ_s*, to answer the* Update *and* Search *queries.* Δ_s *simply keeps track of the simulated keywords. For each linked-list, KeyDict dictionary stores a key, keyword, and the first node identifier of the respective linked-list. The simulator then generates the keywords and arbitrary values for linked-lists which are selected from a keyword distribution based on the range of the encryption scheme. To facilitate generating the search and update tokens, the simulator,* $\mathcal{S}$*, updates the* Δ_s *and keyDict dictionaries adaptively. We explained the setup phase in Algorithm 8.*

Add. *To add a new file, the simulator* $\mathcal{S}$ *uses the update leakage function,* $\mathcal{L}^{\mathsf{Update}}(in, D_i, \mathsf{op}) = \{id_{in}(D_i), |\boldsymbol{w}_{in}|, |D_i|, \mathsf{op}\}$*, and employs the same keyword distribution and* Δ_s*. First, the simulator randomly selects* $|\boldsymbol{w}_{in}|$ *keywords to be assigned to the new document. It then*

generates an encrypted file C_i and the respective linked-lists. Lastly, $\mathcal{S}$ updates the dictionaries respectively for future references. This process, add simulation, is shown in Algorithm 9. Remark that to follow our scheme's architecture, beside using a new key, every keyword is appended as a new linked-list to the cloud index. As a result, it is impossible for the server to link the newly added file to the previous search tokens even if the simulator generate a query that has the new file among the results.

Algorithm 8 Simulator's setup phase

Simulator
$k_{SE} \leftarrow \mathsf{SE.Gen}(1^\lambda)$
Simulate C as $\{C_i \leftarrow \mathsf{SE.Enc}(k_{SE}, \{0\}^{|D_i|})_{1 \le i \le n}\}$
Create *keyDict* dictionary.
Create keyword dictionary Δ_s
$\mathbb{L} = \{\}$
$node\_cnt = 0$
$word\_cnt = 0$
while $node\_cnt \neq N$ **do**
  $word\_flag = 1$
  $\Delta_s \cup w_{word\_cnt}$
  while $word\_flag$ & $node\_cnt \neq N$ **do** ▷ Randomly generates a linked-list
    $list\_flag = 1$
    $L = \{\}$
    $k_G \leftarrow \{0,1\}^\lambda$ ▷ k_G is used to encrypt the current linked-list
    $id_{node} \leftarrow \{0,1\}^{l'}$
    Add $(w_{word\_cnt}, id_{node}, k_G)$ into *keyDict*
    while $list\_flag$ & $node\_cnt \neq N$ **do** ▷ Adds new nodes to L until flag becomes false
      $id_{doc} \leftarrow \{id(D_i) | id(D_i) \notin L, 1 \le i \le n)\}$
      $\mathsf{AddNode}(k_G, L, id_{node}, id_{doc})$
      $node\_cnt++$
      $list\_flag \leftarrow \{0,1\}$
      if $list\_flag$ **then**
        $id_{node} \leftarrow \{0,1\}^{l'}$
      end if
    end while
    $\mathbb{L} \cup L$
    $word\_flag \leftarrow \{0,1\}$
  end while
  $word\_cnt++$
end while
Send $\mathbb{L}$ to server
Server
Generate $\mathcal{I}_s$ using $\mathbb{L}$

Search. *$\mathcal{L}^{\mathsf{Search}}$ is the leakage function that the simulator $\mathcal{S}$ uses to imitate the search function. This information provides enough data for the simulator to randomly selects a required number of keywords from Δ_s. In the next step, the sub-queries will be created using keyDict dictionary. This means, the simulator should look in the keyDict to find the key and node IDs for each keyword that is being searched. Once the simulator receives the results, it generates new index entries for the queried keywords and updated the respective dictionary. We explained every step in detail in Algorithm 10. Since the queries\search tokens are non-deterministic and ephemeral, it is not feasible to unfold the search pattern using the search tokens. Moreover, recall that each sub-query can be real or fake\noise (known only to the user\simulator) which makes more difficult for the attacker to ascertain the search pattern.*

Algorithm 9 Add simulation

Simulator
Simulate new file as $\{C_i \leftarrow \mathsf{SE.Enc}(k_{SE}, \{0\}^{|D_i|})\}$
$\mathbb{L} = \{\}$
$\Delta_{tmp} = \{\}$
for $i = 1$ to $i < |\mathbf{w}_{in}|$ **do**
  $w \leftarrow \{w | w \in \Delta_s, w \notin \Delta_{tmp}\}$
  $\Delta_{tmp} \cup w$
  $L = \{\}$
  $k_G \leftarrow \{0,1\}^\lambda$
  $id_{node} \leftarrow \{0,1\}^{l'}$
  Add (w, id_{node}, k_G) into *keyDict*
  $\mathsf{AddNode}(k_G, L, id_{node}, id_{doc})$
  $\mathbb{L} \cup L$
end for
Send $\mathbb{L}$ to server
Server
Update $\mathcal{I}_s$ using $\mathbb{L}$

Algorithm 10 Search simulation

Simulator
Generate a random value k which shows the number of keywords in the current search
$\Delta_{tmp} = \{\}$
$\mathbf{q} = \{\}$
for $i = 1$ to $i \le k$ **do**
  $w \leftarrow \{w | w \in \Delta_s, w \notin \Delta_{tmp}\}$
  $\Delta_{tmp} \cup w$
  Find w entries in *keyDict* and add them to $\mathbf{q}$
end for
$\mathsf{Shuffle}(\mathbf{q})$
Send $\mathbf{q}$ to server
Server
Perform $\mathbf{q}$ and return the result $R = (R(q_1), \ldots, R(q_k), bag)$
Simulator
Generate new $\mathcal{I}_s$ entries based resultant *bag* from the server
Update *keyDict* respectively
Send new entries to the server
Server
Update $\mathcal{I}_s$ according to the new entries

Hence, we programmed a simulator that mimics our approach's operations with the defined leakage functions in consideration. Remark that all simulated operations in Algorithm 8, 9, 10 are executing in polynomial time where a polynomial number of queries exists. Thus, the cloud\attacker is unable to discern the output generated by a real user from a simulator's output unless with a $neg(\lambda)$ amount or

it shatters the employed pseudo random functions or the encryption scheme.

8 Conclusions

In this paper, first, we demonstrated that DSSE schemes with forward privacy are vulnerable to leakage-abuse attacks. Moreover, we introduced two new attacks to demonstrate the vulnerability of the forward-private approaches. All SSE schemes, including approaches with forward privacy, allow a defined level of information leakage (*e.g.,* access\search pattern) to acquire more efficiency. In our introduced attacks, we showed by reverse analyzing the access pattern, it is feasible to recover the search pattern accurately. The recovered data can be used by traditional attacks to reveal the queries, search tokens, and as a result the documents in approaches with forward privacy. Our research demonstrates that the former attacks on traditional SSE schemes are adequate to methods that follows forward privacy principals.

We then addressed this problem by constructing a new Dynamic SSE approach that support update, search, and parallelization. Our method also obfuscates the search and access pattern. In our approach, we first create an inverted-index that maps each keyword to the documents IDs containing the respective keyword. We inject fake documents' IDs in the result-set of each keyword to hide the access pattern. Only the user can discern the fake IDS from real ones. Furthermore, each search request consists of a number of sub queries where all except one are noise which is only known to the user.

Last, using a standard simulation model, we provided the security proof of our approach. Moreover, we conducted a through performance analysis on the implemented prototype that demonstrates the efficiency and low system-cost of our proposed method. As a future work, we plan to upgrade our scheme to support semi-honest cloud servers.

References

[1] D. X. Song, D. Wagner, A. Perrig, "Practical techniques for searches on encrypted data," in Security and Privacy, 2000. S&P 2000. Proceedings. 2000 IEEE Symposium on, 44–55, IEEE, 2000.

[2] E.-J. Goh, et al., "Secure indexes." IACR Cryptology ePrint Archive, **2003**, 216, 2003.

[3] Y.-C. Chang, M. Mitzenmacher, "Privacy preserving keyword searches on remote encrypted data," in International Conference on Applied Cryptography and Network Security, 442–455, Springer, 2005.

[4] N. Cao, C. Wang, M. Li, K. Ren, W. Lou, "Privacy-preserving multi-keyword ranked search over encrypted cloud data," IEEE Transactions on parallel and distributed systems, **25**(1), 222–233, 2014, doi:10.1109/TPDS.2013.45.

[5] E. Stefanov, C. Papamanthou, E. Shi, "Practical Dynamic Searchable Encryption with Small Leakage." in NDSS, volume 71, 72–75, 2014.

[6] R. Bost, "Forward Secure Searchable Encryption," in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 1143–1154, ACM, 2016, doi:10.1145/2976749.2978303.

[7] M. Etemad, A. Küpçü, C. Papamanthou, D. Evans, "Efficient dynamic searchable encryption with forward privacy," Proceedings on Privacy Enhancing Technologies, **2018**(1), 5–20, 2018, doi:10.48550/ARXIV.1710.00208.

[8] X. Liu, G. Yang, Y. Mu, R. Deng, "Multi-user verifiable searchable symmetric encryption for cloud storage," IEEE Transactions on Dependable and Secure Computing, 2018, doi:10.1109/TDSC.2018.2876831.

[9] K. Salmani, K. Barker, "Leakless privacy-preserving multi-keyword ranked search over encrypted cloud data," Journal of Surveillance, Security and Safety, 2020, doi:10.20517/jsss.2020.16.

[10] K. Salmani, K. Barker, "Don't Fool Yourself with Forward Privacy, Your Queries STILL Belong to Us!" in Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy, CODASPY '21, 131–142, Association for Computing Machinery, New York, NY, USA, 2021, doi:10.1145/3422337.3447838.

[11] K. Salmani, K. Barker, "Dynamic Searchable Symmetric Encryption with Full Forward Privacy," in 2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP), 985–995, 2020, doi:10.1109/ICSIP49896.2020.9339338.

[12] C. Liu, L. Zhu, M. Wang, Y.-A. Tan, "Search pattern leakage in searchable encryption: Attacks and new construction," Information Sciences, **265**, 176–188, 2014.

[13] D. Cash, P. Grubbs, J. Perry, T. Ristenpart, "Leakage-abuse attacks against searchable encryption," in Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, 668–679, ACM, 2015, doi:10.1145/2810103.2813700.

[14] Y. Zhang, J. Katz, C. Papamanthou, "All Your Queries Are Belong to Us: The Power of File-Injection Attacks on Searchable Encryption," in 25th USENIX Security Symposium (USENIX Security 16), 707–720, USENIX Association, Austin, TX, 2016.

[15] O. Goldreich, R. Ostrovsky, "Software protection and simulation on oblivious RAMs," Journal of the ACM (JACM), **43**(3), 431–473, 1996.

[16] M. Naveed, "The Fallacy of Composition of Oblivious RAM and Searchable Encryption." IACR Cryptology ePrint Archive, **2015**, 668, 2015.

[17] R. Canetti, U. Feige, O. Goldreich, M. Naor, "Adaptively Secure Multi-party Computation," in Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing, STOC '96, 639–648, ACM, New York, NY, USA, 1996, doi:10.1145/237814.238015.

[18] X. Song, C. Dong, D. Yuan, Q. Xu, M. Zhao, "Forward private searchable symmetric encryption with optimized I/O efficiency," IEEE Transactions on Dependable and Secure Computing, 2018, doi:10.1109/TDSC.2018.2822294.

[19] R. Curtmola, J. Garay, S. Kamara, R. Ostrovsky, "Searchable symmetric encryption: improved definitions and efficient constructions," Journal of Computer Security, **19**(5), 895–934, 2011.

[20] D. Boneh, G. Di Crescenzo, R. Ostrovsky, G. Persiano, "Public key encryption with keyword search," in International conference on the theory and applications of cryptographic techniques, 506–522, Springer, 2004.

[21] M. Bellare, A. Boldyreva, A. O'Neill, "Deterministic and efficiently searchable encryption," in Annual International Cryptology Conference, 535–552, Springer, 2007.

[22] N. Attrapadung, B. Libert, "Functional encryption for inner product: Achieving constant-size ciphertexts with adaptive security or support for negation," in International Workshop on Public Key Cryptography, 384–402, Springer, 2010.

[23] A. Boldyreva, N. Chenette, Y. Lee, A. O'neill, "Order-preserving symmetric encryption," in Annual International Conference on the Theory and Applications of Cryptographic Techniques, 224–241, Springer, 2009.

[24] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, W. Lou, "Fuzzy keyword search over encrypted data in cloud computing," in INFOCOM, 2010 Proceedings IEEE, 1–5, IEEE, 2010.

[25] M. Kuzu, M. S. Islam, M. Kantarcioglu, "Efficient similarity search over encrypted data," in Data Engineering (ICDE), 2012 IEEE 28th International Conference on, 1156–1167, IEEE, 2012, doi:10.1109/ICDE.2012.23.

[26] Z. Guo, H. Zhang, C. Sun, Q. Wen, W. Li, "Secure multi-keyword ranked search over encrypted cloud data for multiple data owners," Journal of Systems and Software, **137**, 380–395, 2018, doi:https://doi.org/10.1016/j.jss.2017.12.008.

[27] S. K. Kermanshahi, J. K. Liu, R. Steinfeld, S. Nepal, "Generic Multi-keyword Ranked Search on Encrypted Cloud Data," in European Symposium on Research in Computer Security, 322–343, Springer, 2019.

[28] S. Kamara, C. Papamanthou, T. Roeder, "Dynamic Searchable Symmetric Encryption," in Proceedings of the 2012 ACM Conference on Computer and Communications Security, CCS '12, 965–976, ACM, New York, NY, USA, 2012, doi:10.1145/2382196.2382298.

[29] M. Naveed, M. Prabhakaran, C. A. Gunter, "Dynamic Searchable Encryption via Blind Storage," in 2014 IEEE Symposium on Security and Privacy, 639–654, 2014, doi:10.1109/SP.2014.47.

[30] J. Ning, J. Xu, K. Liang, F. Zhang, E.-C. Chang, "Passive attacks against searchable encryption," IEEE Transactions on Information Forensics and Security, **14**(3), 789–802, 2018, doi:10.1109/TIFS.2018.2866321.

[31] "Gutenberg Publication," https://www.cs.cmu.edu/~enron/, accessed: 2019-11-08.

[32] M. S. Islam, M. Kuzu, M. Kantarcioglu, "Access Pattern disclosure on Searchable Encryption: Ramification, Attack and Mitigation." in Ndss, volume 20, 12, Citeseer, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 2, 179-190 (2022)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Cloud-Based Hierarchical Consortium Blockchain Networks for Timely Publication and Efficient Retrieval of Electronic Health Records

Alvin Thamrin, Haiping Xu*, Rui Ming

Computer and Information Science Department, University of Massachusetts Dartmouth, Dartmouth, MA 02747, USA

ARTICLE INFO

Article history:
Received: 28 February, 2022
Accepted: 17 April, 2022
Online: 22 April, 2022

Keywords:
Hierarchical blockchain
Cloud Computing
Timely publication
Electronic health records
Consensus mechanism

ABSTRACT

Blockchain technology is seeing a trend of popularity and adoption in many different application areas. One such area is healthcare, as there is a need to develop a system that can reliably store and share electronic health records (EHRs) among hospital-based health facilities. In this paper, we present a cloud-based hierarchical consortium blockchain framework for storing and sharing EHRs in a scalable, secure, and reliable manner. The framework enables data sharing between local hospital blockchain networks (HBNs) through high-level blockchain networks, namely, city blockchain networks (CBNs) and a state blockchain network (SBN). To support the timely publication of EHRs in HBNs, we adopt a temporary and permanent block scheme in hospital blockchains. In addition, we develop role-based access control (RBAC) policies for data authorization and procedures for concurrent search and retrieval of EHRs across cities and states. The experimental results show that our proposed approach is feasible and supports timely publication and efficient retrieval of EHRs in cloud-based hierarchical blockchain networks.

1. Introduction

Blockchain technology was originally proposed in 2008 as a decentralized and distributed digital ledger mechanism for the peer-to-peer electronic cash system called Bitcoin [1]. A blockchain stores data in blocks that are cryptographically chained together in the form of a linked list. Thus, blocks can be used to store and record transactions in a tamper-proof and immutable manner. Unlike public blockchains, a consortium blockchain is defined as a permissioned blockchain, and access to it is usually restricted to a specific number of "permissioned" nodes [2]. In recent years, the popularity of consortium blockchain has increased due to its potential use in many different application areas, including healthcare [3], [4]. A consortium blockchain-based system can be implemented within the healthcare domain to enable and support the storage and sharing of healthcare data among health institutions or hospitals. In our earlier work, we introduced a cloud-based blockchain solution for storing and sharing electronic health records (EHRs) while enabling data accessibility, redundancy, and security on a local scale [5]. This solution allows storing big data, such as EHRs with multimedia files, in a cloud-based blockchain, while storing their metadata in a lite blockchain for efficient information retrieval. However, due to the big data involved, the solution can only be effective when implemented on a small/local scope. This is because the growth potential of the blockchain increases dramatically with the large number of hospitals participating in the network. This can lead to a very unsustainable expansion of the blockchain in terms of size, which constitutes a major scalability issue.

In this paper, we present a cloud-based hierarchical consortium blockchain framework to address the above scalability issue. The framework consists of three layers of blockchain networks, namely hospital blockchain networks (HBNs), city blockchain networks (CBNs), and a state blockchain network (SBN). An HBN is designated as a blockchain network at the first layer and is shared by hospitals that are geographically close to each other in a local area or a city. To simplify matters, in this paper we use the term *city* to refer to a city, a local area or any form of governmental jurisdiction below the state level. A CBN is designated as a blockchain network at the second layer. Unlike an HBN, a CBN is shared by all cities located within a state as participants. Each city in a CBN is also connected to an HBN as the network regulator, which allows agents from different HBNs within the same state to communicate with each other for data sharing purposes. Finally, An SBN is designated as a blockchain network at the third layer. The SBN is shared by all states located within a country. Each state in the SBN is also connected to a CBN and acts as the network regulator of the CBN. The SBN is designed to allow agents from

*Corresponding Author: Haiping Xu, University of Massachusetts Dartmouth, Dartmouth, MA 02747, Email: hxu@umassd.edu

different CBNs across the country to communicate with each other for data sharing purposes, similar to the sharing relationship between a CBN and the HBNs connected under it. The implementation of these network layers enables all hospital peers across the country to communicate and share data with each other in a scalable, secure, and reliable manner.

Another challenging issue we face concerns the publishing of EHRs to the blockchain in a timely and space-efficient manner. Whenever data are made available, they can either be published to the blockchain immediately in the form of block records stored in a block, or they can be accumulated until the block contains a sufficient number of block records to be published. The first method excels in terms of timeliness but is inefficient in terms of space/memory usage because this method generates many blocks containing a single or very few records. On the other hand, the second method is more spatially efficient compared to the first one because fewer and denser blocks are generated; however, it has a significant drawback in terms of timeliness that may affect the effectiveness of a blockchain-based system for storing and sharing EHRs. In this paper, we present an approach that facilitates the timely and space-efficient publication of new block records using a temporary and permanent block scheme. As demonstrated in previous work [6], a new block record can be published immediately in a temporary block after being approved using a temporary block consensus mechanism. Once a predefined number of temporary blocks have been published, they can be merged into a permanent block and published to the blockchain.

This work significantly extends our previous proposed framework for healthcare data storage using hierarchical cloud-based blockchains. In our previous work [7], we focused on the structural design of the storage system and did not fully consider the scalability of HBNs with many peers and the timely publication of EHRs in HBNs. To address these issues, we now limit the number of hospital super-peer agents in an HBN to reduce the redundancy of big data storage. Furthermore, by using a temporary and permanent block scheme, EHRs can be efficiently published in HBNs. Finally, in previous work [7], new access control policies need to be established and approved after the search results are returned. In this work, we require that access control policies be established prior to the doctor's visits. Thus, the search and retrieval steps of EHRs can be combined to achieve an efficient information search and retrieval process.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 presents a cloud-based hierarchical consortium blockchain framework. Section 4 introduces the block structures and the processes of generating and publishing new blocks in different blockchains. Section 5 describes the search and retrieval process of EHRs in details. Section 6 presents the case studies and their analysis results. Section 7 concludes the paper and mentions future work.

2. Related Work

There are various studies and explorations on blockchain technology to develop a decentralized storage and sharing system for the healthcare sector [3]-[5]. Blockchain technology has been shown to be a viable technology as it allows sharing of medical data among approved healthcare providers while maintaining patient privacy [8]. Further research has also addressed the challenge of storing big data such as images and videos in the blockchain; however, these studies have typically utilized off-chain approaches to store big data, rather than on-chain solutions. In [9], the authors proposed a storage model based on blockchain and InterPlanetary File System (IPFS) to store transactions efficiently in blockchain. In their design, the actual patient reports are stored in distributed off-chain storage using IPFS, while the blockchain stores only hash values of the reports, thereby reducing the overall block size in the blockchain. In [10], the authors developed a decentralized and permissible blockchain-based application for storing and accessing satellite task-scheduling schemes using the Hyperledger Fabric framework. They used IPFS for off-chain storage to reduce the asset size of the transaction and increase the transaction throughput of the network. In [11], the authors introduced a video surveillance storage and sharing system using blockchain technology. In their approach, videos received from the camera are encrypted and stored off-chain through distributed IPFS system with their metadata stored in the blockchain. Some researchers also proposed a secure data sharing solution for sensitive financial data using blockchain and proxy re-encryption technology [12]. Access control rules, hash values, and storage addresses of financial data are stored in the blockchain, while the actual financial data are stored off-chain in distributed databases. In [13], the authors proposed a blockchain framework using an attribute-based cryptosystem for the development of a secure EHR storage and sharing system. In their approach, large-scale medical data are stored in the cloud and the blockchain stores only the metadata of EHRs. Unlike these approaches, our cloud-based blockchain solution enables all healthcare data, including multimedia files, to be stored in the blockchains. Thus, our on-chain data storage approach provides the benefits of a complete blockchain storage solution in terms of data immutability, integrity and availability.

There are other studies focusing on the design of new blockchain architecture, which are summarized below. In [14], the authors proposed Fortified-Chain, a decentralized EHR and blockchain-based distributed data storage system (DDSS). They designed a global DDSS network that facilitates communications between local DDSS networks consisting of hospitals and third-party health services that store patient medical data. In [15], the authors developed a simplified version of a scalable blockchain architecture for sharing EHRs among patients, healthcare professionals and health institutions. In their approach, each health facility implements a local blockchain network connected to a global blockchain system to allow interaction between different health institutions. Some researchers also studied and introduced a blockchainless approach based on directed acyclic graphs (DAGs) for trusted public construction bidding to ensure fairness in the bidding process [16]. The DAG-based approach differs from the traditional blockchain because in the chainless approach, the DAG links the transaction containing its parents, documents, and a list of transaction signatures to other transactions through a less complex verification process. In a more recent effort, researchers designed a Compacted DAG-based blockchain protocol (CoDAG), used in the field of Industrial Internet of Things (IIoT) [17]. They developed protocols and algorithms to secure the network and confirm transactions within a specified time. The aforementioned blockchain-based approaches either use off-chain storage, e.g.,

[14] and [15], or do not address scalability issues, e.g., [16] and [17]. Unlike the above approaches, our novel cloud-based hierarchical blockchain architecture not only supports on-chain storage of big data, but also allows interaction and data sharing between peers located in different cities and states. Since EHRs from different cities are stored in different HBNs, our cloud-based hierarchical blockchain approach provides a scalable solution for storing sensitive information and big data in a nationwide network of connected consortium blockchains.

Previous research efforts on implementing access control mechanisms in blockchain networks have focused on preventing unauthorized access to confidential data stored in the blockchains. In [4], the authors proposed MeDShare, a blockchain-based system that enables peer-to-peer medical data sharing in a trustless environment. They used smart contracts and access control mechanisms to monitor and track the behavior of storing data in the blockchain. If any form of data permission violation is detected, the system revokes the access rights of the offending user. In [18], the authors proposed a Blockchain-as-a-Service based solution for Health Information Exchange (BaaS-HIE) activities to deal with security issues including patient privacy, integrity of medical records, and fine-grained access control. Their approach involves the use of a private blockchain based on the Ethereum protocol and smart contracts as access control management for medical records. In a similar way, other researchers designed an access control mechanism on managing user access to ensure efficient and secure sharing of EHRs on mobile devices by leveraging smart contracts on the Ethereum blockchain [19]. In [20], the authors proposed the use of blockchain and edge nodes to facilitate attribute-based access control and storage of EHR data. They used smart contracts to enforce access control of EHR data stored in off-chain edge nodes. In their subsequent work, encryption for data stored at the edge nodes was further developed [21]. The multi-authority attribute-based encryption (ABE) scheme and attribute-based multi-signature (ABMS) scheme were used to encrypt the EHR data stored at the edge nodes and verify users' signatures, respectively. In contrast to the above work, our approach involves the implementation of different scopes of role-based access control (RBAC) policies that restrict user access to various healthcare facilities in different cities and states. We define three layers of the networks, each implementing its own RBAC policies – local hospital-wide policies, city-wide policies, and statewide policies, respectively. As a result, our approach provides a more comprehensive and reliable mechanism than other methods because it is designed to work in a much larger environment.

3. A Framework for Hierarchical Blockchains

The framework for cloud-based hierarchical consortium blockchain networks consists of three layers. As shown in Figure 1, these layers are the *Hospital Layer*, the *City Layer*, and the *State Layer*, which contain multiple HBNs, multiple CBNs, and an SBN, respectively. An HBN in the hospital layer covers multiple hospitals from the same city or local area, represented by hospital super-peer agents β_{HOSs} or hospital regular-peer agents β_{HREPs}. A CBN in the city layer covers multiple cities from the same state. A city super-peer agent β_{CIT} acts as a representative of a city and a network regulator for the HBN belonging to the city. Finally, an SBN in the state layer covers all states of a country. A state super-peer agent β_{STA} acts as a representative of a state and a network regulator for the CBN belonging to the state.

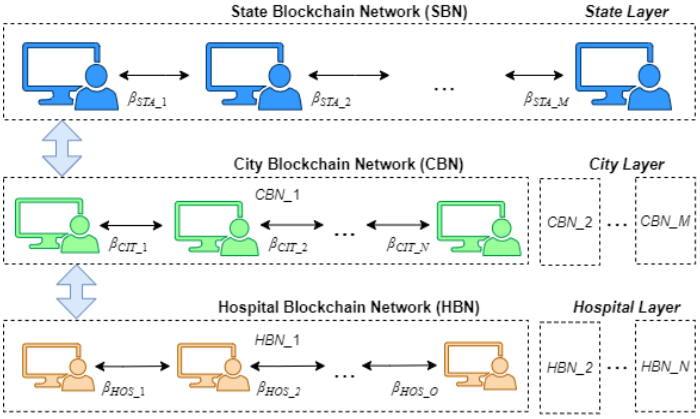


Figure 1: The Architecture of Cloud-Based Hierarchical Consortium Blockchains

To enable big data storage in blockchain networks, a cloud-based blockchain scheme is implemented in the hospital layer, i.e., HBNs. Unlike previous work [5], which requires all participating hospitals in an HBN to implement cloud-based blockchains, in this study, the HBN consists of three types of agents, namely the hospital super-peer agents β_{HOSs}, representing designated hospitals, hospital regular-peer agents β_{HREPs}, representing general hospitals, and regular-peer agents β_{REPs}, representing end users including doctors, nurses, and patients. We define general hospitals as those that do not have the required infrastructure to implement cloud-based blockchain storage or choose not to do so. Figure 2 shows an example of an HBN where hospital *A* and *B* are designated hospitals that offer private cloud services, while hospital *C* is a general hospital that do not provide such services.

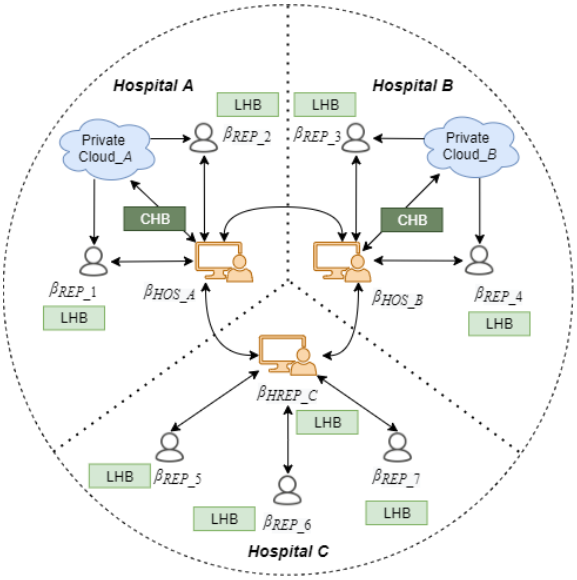


Figure 2: The Relationships Between Participants in an HBN

As shown in Figure 2, a cloud-based hospital blockchain (CHB) is implemented in the designated hospitals with their private clouds. A CHB is managed by a hospital super-peer agent β_{HOS} and stores all data, including EHRs in multimedia file format. To avoid excessive redundancy of big data in an HBN, we limit the number of hospital super peer agents in an HBN to no more

than 6-10. A lite hospital blockchain (LHB) is implemented on the server of a general hospital, managed by a hospital regular-peer agent β_{HREP}, or on the local machine of an end user, managed by a regular-peer agent β_{REP}. An LHB stores all data in its corresponding CHB, except for big data such as multimedia files, for which only their metadata are stored in the LHB. Access to confidential data, i.e., a patient's EHRs, stored in a CHB is managed by a hospital super-peer agent β_{HOS}, while access to confidential data stored in an LHB is managed by either a hospital super-peer agent β_{HOS} or a hospital regular-peer agent β_{HREP}.

Figure 3 shows the general blockchain structure and the similarity between CHB and LHB. Let the length of a LHB and its corresponding LHB be h. A cloud-based block CB_i and a lite block LB_i, where $1 \leq i \leq h$, contain the same information except for the multimedia files. This scheme allows an end user or a general hospital to use the metadata stored in its LHB to submit a request to a relevant hospital super-peer agent β_{HOS} through its regular-peer agent β_{REP} or hospital regular-peer agent β_{HREP} and retrieve the corresponding multimedia files stored in the CHB.

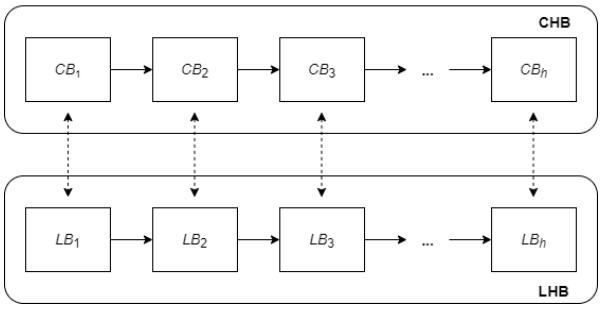


Figure 3: The General Blockchain Structure of a CHB and an LHB

To support the timely publication of EHRs in an HBN, we introduce a temporary and permanent block scheme based on earlier work [6]. Due to the need to publish EHRs, including their associated multimedia files in a timely manner, temporary blocks are only included in the hospital layer of our cloud-based hierarchical blockchain networks; however, they are not required in the city and state layers, as access control policies and access records do not need to be published immediately. Figure 4 shows an example of an CHB with temporary and permanent blocks.

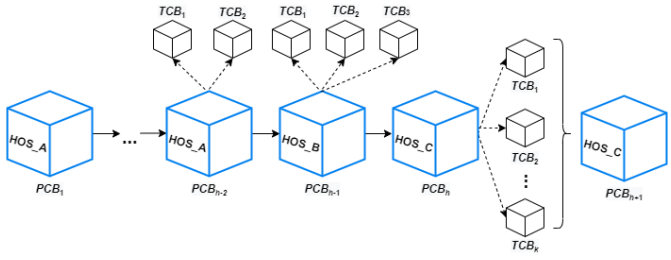


Figure 4: A CHB with Temporary and Permanent Blocks

As shown in Figure 4, a block PCB_i, where i is the height of the block in the blockchain, is a cloud-based permanent block. To support efficient information retrieval, a *PCB* contains only EHRs and other related records from the same hospital. On the other hand, a block TCB_j, where j denotes the order in which the temporary blocks are attached to a *PCB* according to their publishing time, is a cloud-based temporary block. To support the timely publication of block records, a *TCB* stores only one record from a hospital and must be attached to the latest *PCB* published for the same hospital. As shown in the figure, the latest *PCB* for a hospital can be attached by multiple *TCB*s that are numbered in the order in which they appear. Note that the LHB corresponding to a CHB shares the same structure but have different notations, i.e., permanent lite block *PLB* and temporary lite block *TLB*.

Since a *TCB* contains only one block record, whenever a block record is generated and submitted to a super-peer agent, the agent can immediately publish the block record to the blockchain as a temporary block through the temporary block generation process. Meanwhile, a permanent block stores multiple block records as in a typical blockchain. Once enough *TCB*s are generated and published to the blockchain by a hospital super-peer agent, the agent can consolidate them into a new *PCB* through the permanent block generation process. For example, as in Figure 4, when the number or total size of *TCB*s attached to PCB_h reaches a threshold, the agent β_{HOS_C} merges the list of *TCB*s and forms a new cloud-based permanent block PCB_{h+1} for publishing. When PCB_{h+1} is published, all *TCB*s attached to PCB_h are removed from the blockchain. Note that other *TCB*s that are attached to *PCB*s other than PCB_h will remain in the blockchain until they are merged into new *PCB*s.

4. Publication of New Blocks in the Blockchain Networks

An HBN, a CBN or the SBN maintains its own blockchain, namely hospital blockchain, city blockchain or state blockchain, respectively. Blockchain networks of the same type, such as two HBNs, are independent, but they can communicate through a higher-level blockchain network, e.g., a CBN if the two HBNs belong to the same state, or the SBN if the two HBNs are in different states. Hospital, city, and state blockchains can store different types of block records for different purposes. In this section, we describe the types of block records used in different blockchains and the procedures for generating and publishing new blocks in different types of blockchains.

4.1. Hospital Block and its Block Record Types

There are four different types of block records that can be used in a CHB or an LHB, namely HR_{UPR}, HR_{ACP}, HR_{MER}, and HR_{AR}. To simplify matters, we define a hospital blockchain as a general term that can refer to a CHB or an LHB. We now describe the four types of block records as follows.

- HR_{UPR} is a record that stores the account information and user profile of an end user, represented by regular-peer agent β_{REP}. An HR_{UPR} is defined as a 6-tuple (I, N, R, U, S, T), where I is the identification of β_{REP} in the HBN; N is the full name of β_{REP}; R, U and S are β_{REP}'s private key, public key and secret symmetric key, respectively; and T is the timestamp when the HR_{UPR} is created. Whenever a new user joins the HBN or an existing user's profile is updated, a new HR_{UPR} is created.
- HR_{ACP} is a record that stores access control policies and is used to conduct permission checks on requests to access EHRs stored at hospitals within the same city. An HR_{ACP} is defined as a triple (P, H, T), where P is a set of policies; H is a set of hospital where the policies are executed; and T is the timestamp when the policies are created.
- HR_{AR} is a record that stores information on access requests to a patient's EHRs stored at hospitals within the same city where

the patient resides. HR_{AR} is created as a log of access requests for accountability purposes. An HR_{AR} is defined as 4-tuple (N, D, O, T), where N is the request number; D is the detail of the request; O is the outcome of the request; and T is the timestamp when the request is created.

- HR_{MER} is a record that stores medical information, including patient reports and metadata for any related multimedia files generated after a doctor's visit. An HR_{MER} is defined as 5-tuple (I, H, X, M, T), where I are the identifications of all peers involved in the doctor's visit, including the patient, the nurse and the doctor; H is the name of the hospital where the patient visited; X includes a summary of the visit and any text-based medical data; M is the metadata of any multimedia files generated after the doctor's visit; and T is the timestamp when the HR_{MER} record is created.

Since both permanent and temporary blocks in a CHB may contain multimedia files, the blocks *PCB* and *TCB* consist of two major components: the block component and the multimedia file component. Figure 5 shows the block structure of a new temporary cloud-based block TCB_j with three sections in the block component and one section in the multimedia file component. These are header, hospital block records, verification information, and multimedia files in an EHR.

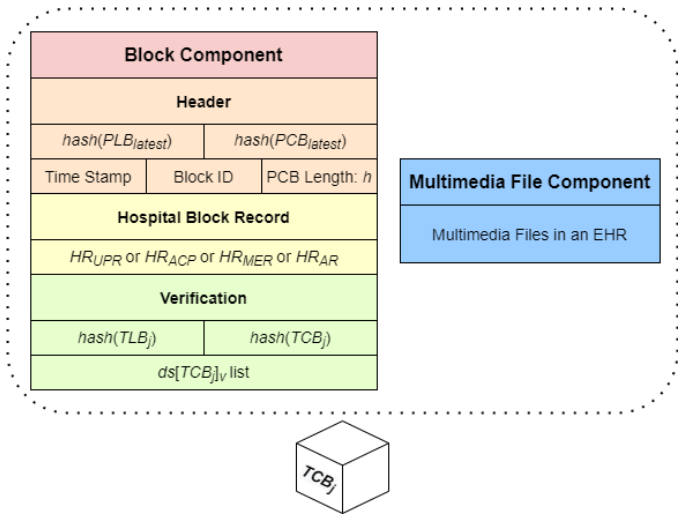


Figure 5: The Structure of a New Temporary Cloud-Based Block TCB_j

As shown in Figure 5, the header section contains the hash values of the latest *PLB* and *PCB*, published by the same hospital super-peer agent who generates TCB_j, the timestamp when TCB_j was created, the block ID, and the length h of the current blockchain. The hospital block records section contains only a single block record of HR_{UPR}, HR_{ACP}, HR_{MER}, or HR_{AR} as TCB_j is a temporary block. Consequently, the multimedia file section can only store multimedia files from one doctor's visit, if any, while their metadata is recorded and stored in the relevant HR_{MER} in the hospital block record section. Lastly, the verification section contains the hash values of the current block, including the hash value of the header and hospital block records, denoted as $hash(TLB_j)$, and the hash value of the header, hospital block records and the multimedia files, denoted as $hash(TCB_j)$. The verification section also contains a list of digital signatures $ds[TCB_j]_v$, where each peer v is an agent β_{HOS} who approves TCB_j during the temporary block consensus process. Note that the structure of the temporary lite block TLB_j is similar to that of TCB_j but does not include the multimedia file component.

Figure 6 shows the block structure of a new permanent cloud-based block PCB_{h+1}. The block structure of *PCB* is similar to that of *TCB*, but a *PCB* can accommodate multiple block records and EHRs in its hospital block records section and multimedia file component, respectively.

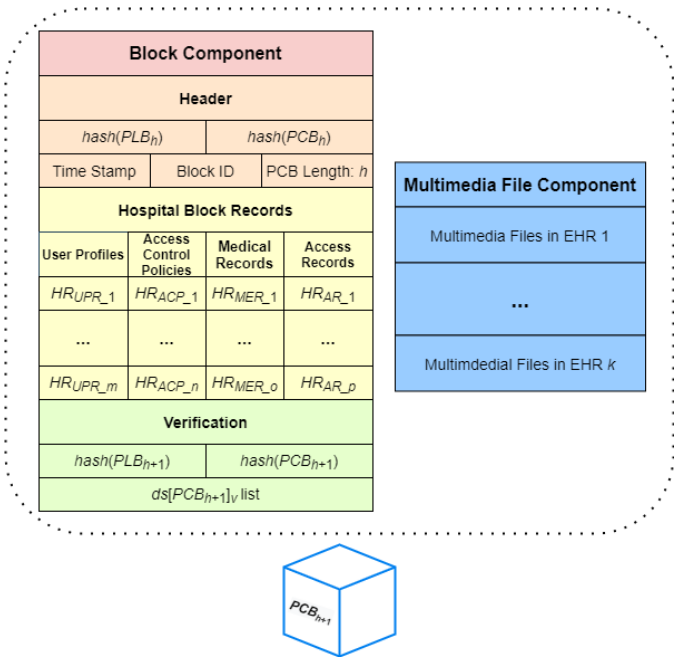


Figure 6: The Structure of a New Permanent Cloud-Based Block PCB_{h+1}

In a *PCB*, the verification section contains the hash values of the permanent lite block PLB_{h+1} and the permanent cloud-based block PCB_{h+1}. It also contains a list of digital signatures $ds[PCB_{h+1}]_v$, where each peer v is an agent β_{HOS}, who approves PCB_{h+1} during the permanent block consensus process. Note that while a new *TCB* is attached to the latest *PCB*, published by the hospital super-peer agent who generates the *TCB*, a new *PCB* must be attached to the last *PCB* of the cloud-based blockchain, i.e., PCB_h, where h is the height of the current blockchain. As with temporary blocks, the structure of a permanent lite block PLB_{h+1} is similar to that of PCB_{h+1} except for the inclusion of the multimedia file component in PCB_{h+1}.

4.2. Hospital Temporary and Permanent Block Generation

Let hospital super-peer agent $\beta_{HOS\text{-}\Psi}$ be the one who creates a new temporary cloud-based block TCB_j. Algorithm 1 shows how the new block TCB_j is generated by agent $\beta_{HOS\text{-}\Psi}$. According to the algorithm, agent $\beta_{HOS\text{-}\Psi}$ first creates an empty temporary cloud-based block TCB_j. All attributes in TCB_j's header section are then created and added. These include the hash values of the latest permanent blocks, previously published by $\beta_{HOS\text{-}\Psi}$, i.e., $hash(PCB_{latest})$ and $hash(PLB_{latest})$, the timestamp when TCB_j is created, TCB_j's block ID, and the blockchain length h. After that, $\beta_{HOS\text{-}\Psi}$ processes the hospital block record φ given in the input list accordingly. If φ is an HR_{UPR} and $\varphi.S$ is *null*, it indicates that φ records the account information and user profile of a new end user. In this case, a secret symmetric key is automatically generated and added it to $\varphi.S$. Then φ is encrypted using the public key of $\beta_{HOS\text{-}}$

$_\psi$ and added to the hospital block record section of TCB_j. If φ is a medical block record HR_{MER} and a list ρ of multimedia files is included, $\beta_{HOS\text{-}\psi}$ encrypts the files in ρ using the associated patient's secret symmetric key retrieved from the patient's latest HR_{UPR}. The encrypted files are then added to the multimedia file component of TCB_j. The metadata of the encrypted files are also recorded and added to $\varphi.M$ of the medical block record. Finally, φ itself is encrypted, except for $\varphi.M$, before it is added to the hospital block record section of TCB_j. Note that if φ is an HR_{ACP} or HR_{AR}, it is added directly to the hospital block record section of TCB_j in plaintext. Once the header section and hospital block record section are established, $\beta_{HOS\text{-}\psi}$ calculates the hash values $hash(TLB_j)$ and $hash(TCB_j)$, as well as the digital signature $ds[TCB_j]_\psi$ using $hash(TCB_j)$. All these elements are then added to the verification section of TCB_j. Note that while not shown in Algorithm 1, a new temporary lite block TLB_j can be created by simply removing the multimedia file component from TCB_j.

Algorithm 1: Generating a New Temporary Block TCB_j

Input: A hospital block record φ containing record HR_{UPR}, HR_{ACP}, HR_{AR}, or HR_{MER}, and an optional list ρ of multimedia files.
Output: A new temporary cloud-based block TCB_j.

1. Create an empty temporary cloud-based block TCB_j
2. Verify and add $hash(PCB_{latest})$, $hash(PLB_{latest})$, time stamp, block ID and current blockchain length h to the header section of TCB_j
3. **if** φ is an HR_{UPR} and $\varphi.S$ is *null* // indicates a new end user
4. Generate a secret symmetric key, add it to $\varphi.S$, and encrypt φ
5. **else if** φ is an HR_{MER} and ρ is not empty
6. Encrypt the multimedia files in ρ
7. Add the encrypted files to the multimedia file section of TCB_j
8. Add the metadata of ρ to $\varphi.M$ and encrypt φ, except for $\varphi.M$
9. Add φ to the hospital block record section of TCB_j
10. Calculate the hash values $hash(TCB_j)$ and $hash(TLB_j)$
11. Add the hash values to the verification section of TCB_j
12. Create digital signature $ds[TCB_j]_\psi$ using $hash(TCB_j)$
13. Add $ds[TCB_j]_\psi$ to the $ds[TCB_j]_v$ list in the verification section
14. **return** TCB_j

Once enough *TCB*s are generated and published to the blockchain by $\beta_{HOS\text{-}\psi}$, the *TCB*s can be consolidated into a new permanent block *PCB* through a permanent block generation process. Algorithm 2 shows how a new permanent cloud-based block PCB_{h+1} is generated by agent $\beta_{HOS\text{-}\psi}$. According to the algorithm, agent $\beta_{HOS\text{-}\psi}$ first creates an empty permanent cloud-based block PCB_{h+1}. All attributes in PCB_{h+1}'s header section, including $hash(PCB_h)$ and $hash(PLB_h)$, the timestamp, the block ID, and the blockchain length h, are then created and added. For each temporary block τ in the temporary block list Ξ, $\beta_{HOS\text{-}\psi}$ verifies it using information stored in τ's header and the verification section and transfers all relevant information from the hospital block record section of τ to the hospital block records section of PCB_{h+1} as a new block record. If τ contains a block record HR_{MER} and a list of encrypted multimedia files ρ, $\beta_{HOS\text{-}\psi}$ moves files in ρ to the multimedia file component of PCB_{h+1} and adds the metadata of ρ to the relevant block record $HR_{MER}.M$. This ensures that all previously stored information in the temporary blocks from the list Ξ is transferred to PCB_{h+1}. Finally, $\beta_{HOS\text{-}\psi}$ calculates the hash values $hash(PCB_{h+1})$ and $hash(PLB_{h+1})$, as well as the digital signature $ds[PCB_{h+1}]_\psi$ using $hash(PCB_{h+1})$. All these elements are then added to the verification section of PCB_{h+1}.

Similar to the generation of TLB_j, a new permanent lite block PLB_{h+1} can be created by simply removing the multimedia file component from PCB_{h+1}.

Algorithm 2: Generating a New Permanent Block PCB_{h+1}

Input: A list of blocks Ξ containing k temporary cloud-based blocks.
Output: A new permanent cloud-based block PCB_{h+1}.

1. Create an empty permanent cloud-based block PCB_{h+1}
2. Verify and add $hash(PCB_h)$, $hash(PLB_h)$, time stamp, block ID, and current blockchain length h to the header section of PCB_{h+1}
3. **for** each temporary cloud-based block τ in Ξ
4. Verify τ and add all relevant parts from the hospital block record section in τ to the hospital block records section in PCB_{h+1}
5. **if** τ contains HR_{MER} and a list of encrypted multimedia files ρ
6. Add files in ρ to PCB_{h+1}'s multimedia file component
7. Add the metadata of ρ to the corresponding $HR_{MER}.M$
8. Calculate hash values $hash(PCB_{h+1})$ and $hash(PLB_{h+1})$
9. Add the hash values to the verification section of PCB_{h+1}
10. Create digital signature $ds[PCB_{h+1}]_\psi$ using $hash(PCB_{h+1})$
11. Add $ds[PCB_{h+1}]_\psi$ to the $ds[PCB_{h+1}]_v$ list in the verification section
12. **return** PCB_{h+1}

4.3. City and State Block and their Block Record Types

Unlike the hospital blockchain, there is only one variant of the city and state blockchains due to the absence of regular peers and big data. Thus, implementing cloud-based versions of city and state blockchains is not necessary. For city blockchain, there are two types of block records that can be stored in the blockchain. These are city-wide record for access control policies CR_{ACP} and city-wide access record CR_{AR}. CR_{ACP} is a record that stores the access control policies implemented in a CBN and enforced by the relevant city super-peer agent β_{CIT}. CR_{ACP} has the same structure as HR_{ACP}, except that the $CR_{ACP}.H$ contains additional information such as the names of cities and hospitals where the policies are enforced. CR_{ACP} is created to check any requests regarding access to patient EHRs stored in HBNs across cities within the same state. On the other hand, CR_{AR} is a record that stores information on access requests to patient EHRs in hospitals across cities within the same state. The structure of a CR_{AR} is also similar to that of an HR_{AR}.

For state blockchain, a state block shares the same structure as that of a city block and stores statewide records for access control policies SR_{ACP} and statewide access record SR_{AR}. SR_{ACP} is a record that stores the access control policies implemented in the SBN and enforced by the relevant state super-peer agent β_{STA}. SR_{ACP} is established to check for any access requests to patient EHRs stored in HBNs across states; while SR_{AR} is a record that stores information on access requests to patient EHRs in hospitals across states. Figure 7 shows the structure of a new city or state block B_{h+1} in a city or state blockchain. From the figure, we can see that block B_{h+1} consists of only one component as city and state blockchains do not store EHRs. There are three sections present in block B_{h+1}, namely header, state or city block records, and the verification section. The header section contains the previous city or state block's hash value $hash(B_h)$, the timestamp when B_{h+1} is created, the block ID of B_{h+1}, and the current blockchain length h. The city or state records section contains a list of block record CR_{ACP} and/or CR_{AR}, or SR_{ACP} and/or SR_{AR}, respectively. The verification section contains the hash values of B_{h+1} and a list of

digital signatures $ds[B_{h+1}]_v$, where each peer v is a city or state super-peer agent who approves B_{h+1} during the consensus process.

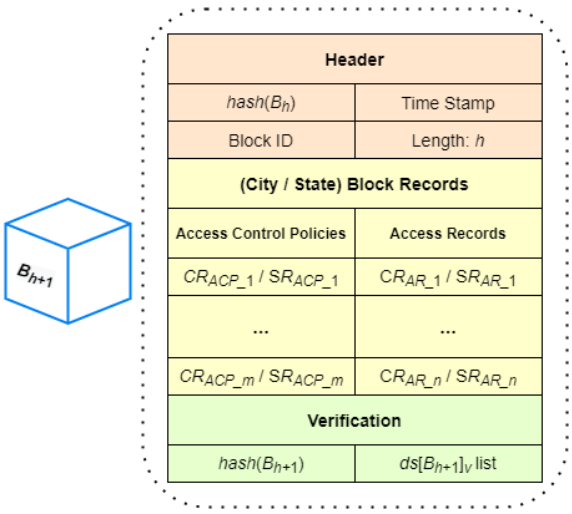


Figure 7: The Structure of a New City or State Block B_{h+1}

4.4. City and State Block Generation

The process of generating a new city or state block is similar to that of generating a lite hospital block, although it is simpler due to the absence of big data and end users. Let the city super-peer agent $\beta_{CIT\text{-}\Psi}$ be the one who creates the new city block B_{h+1}. Algorithm 3 shows the procedure for generating B_{h+1}, which is then approved and added to the city blockchain through a city block consensus process.

Algorithm 3: Generating a New City Block B_{h+1}

Input: A list of city block records Φ containing CR_{ACP} and/or CR_{AR}
Output: A new city block B_{h+1}

1. Create an empty city block B_{h+1}
2. Verify and add $hash(B_h)$, time stamp, block ID, and current blockchain length h to the header section of B_{h+1}
3. **for** each record φ in the list Φ of city block records
4. **if** φ is an CR_{ACP}
5. Add φ to the city block records section of B_{h+1}
6. **else**
7. Encrypt φ and add it to the city block records section
8. Calculate $hash(B_{h+1})$ and add it to the verification section
9. Create digital signature $ds[B_h]_\Psi$ using $hash(B_{h+1})$
10. Add $ds[B_h]_\Psi$ to the $ds[B_h]_v$ list in the verification section
11. **return** B_{h+1}

According to the algorithm, agent $\beta_{CIT\text{-}\Psi}$ first creates an empty city block B_{h+1}. All attributes in B_{h+1}'s header section are then created and added. These include the previous block's hash value $hash(B_h)$, the timestamp when B_{h+1} is created, the block ID, and the blockchain length h. After that, $\beta_{CIT\text{-}\Psi}$ processes all records in the city block record list Φ accordingly before they are added to the city block records section of B_{h+1}. If a city block record φ is a CR_{ACP}, it is simply added to B_{h+1}'s city block records section without being encrypted. Afterwards, $\beta_{CIT\text{-}\Psi}$ calculates $hash(B_{h+1})$ and $ds[B_{h+1}]_\Psi$ before adding them to the verification section of B_{h+1}. Note that the algorithm for generating a new state block is similar to Algorithm 3 due to the shared structure of the city and state blocks.

4.5. Temporary and Permanent Block Consensus Process

In our approach, we implemented a simple majority vote consensus mechanism for publishing new hospital, city, and state blocks. The consensus processes implemented in HBN, CBN and SBN function similarly. Let λ be the total number of super-peer agents from a blockchain network who participate in a consensus process. The block announcer, the super-peer agent who is responsible for initiating the consensus process, must broadcast the new block to other super-peer agents in the network and gather at least $\lambda/2$ approvals from super-peer agents within the same blockchain network.

Figure 8 shows a general illustration of the consensus process for approving a new temporary block in an HBN. From the figure, we can see that the temporary block consensus process consists of 7 steps. The first step is the announcement of a newly created temporary block TCB_j by the block announcer β_{HOS_A} to the super-peer agents of other hospitals in the network. To simplify matters, we show only one such agent in the figure, i.e., β_{HOS_B}. Note that hospital regular-peer agent β_{HREP_C} does not participate in the consensus process as it does not have direct access to the CHB. Once the announcement is broadcast and received, β_{HOS_B} retrieves TCB_j from the block announcer in step 2. After that, in step 3, β_{HOS_B} verifies the validity of TCB_j by checking the integrity of TCB_j and the digital signature of the block announcer in the block. If TCB_j is considered valid, β_{HOS_B} creates its digital signature and sends it back to the block announcer as an approval vote in step 4. The block announcer waits for a certain amount of time in step 5 until either a timeout is reached, or a majority of approval votes are collected. All valid digital signatures are added to the digital signature list of $ds[TCB_j]_v$. If a majority vote is received by the block announcer β_{HOS_A}, block TCB_j is considered complete and can be added to the CHB. In this case, agent β_{HOS_A} notifies β_{HOS_B} that block TCB_j has successfully passed the consensus process in step 6. Finally, in step 7, each hospital super-peer agent with a completed TCB_j can generate a lite temporary block TLB_j and broadcast it to its respective regular-peer and hospital regular-peer agents for inclusion of TLB_j in their LHBs.

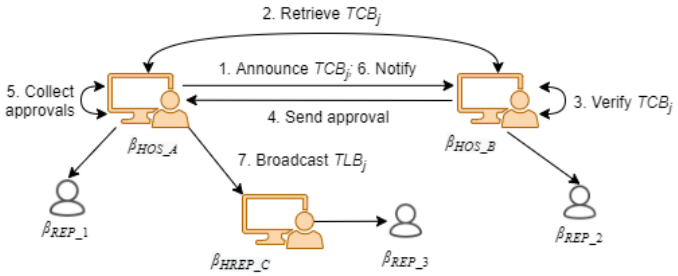


Figure 8: The Consensus Process for Approving a New Temporary Block

As more and more temporary blocks are added to the blockchain by a hospital super-peer agent β_{HOS} through the temporary block consensus process, agent β_{HOS} can decide to merge all its own published temporary blocks into one permanent block by initiating the permanent block consensus process. Note that when a permanent block consensus process is initiated, no other permanent block or temporary block consensus processes are allowed to occur at the same time and vice versa. The consensus process for approving a new permanent block in an HBN is similar to that for approving a new temporary block in an

HBN, depicted in Figure 8, but it requires the deletion of all temporary blocks that have been merged into a new permanent block in its last step. Finally, the consensus process for approving a new city or state block in a CBN or the SBN, respectively is also similar. More details can be found in a recent work [7].

5. The Search and Retrieval Processes for EHRs

There are numerous peers involved in the blockchain networks, playing different roles such as doctors, nurses and patients. Before retrieving EHRs from the blockchains, it is critical to assign appropriate permissions to each role to access the EHRs stored in the blockchains and protect them from unauthorized access [22]. In this section, we first describe the RBAC policies used in our approach, and then present our integrated search and retrieval process for EHRs.

5.1. Role-Based Access Control Policy

In our previous work, we implemented RBAC policies in an HBN as mandatory rules that specify which data in a blockchain can be accessed by participants based on their credentials [5]. With the introduction of hierarchical blockchain framework, RBAC policies are required to function effectively across all three layers of the blockchain networks. In other words, RBAC policies must be defined to grant access to a patient's EHRs across hospitals, cities, or states. These RBAC policies are stored in HR_{ACP}, CR_{ACP} and SR_{ACP} of a hospital blockchain, a city blockchain and a state blockchain, respectively. A patient is required to decide whether to allow or deny the sharing of their medical data with other hospitals within the city, state, or country prior to a doctor's visit. Any relevant access control policies are then created and added to the appropriate hospital, city, or state blockchains. A regular peer agent β_{REP} that represents an end user (e.g., a doctor), must seek permission from a hospital, city, or state super-peer agent for access to a patient's EHRs stored within hospitals either locally or across the country. An example policy H1 is shown below, which is stored as an HR_{ACP} in a hospital blockchain and enforced by the hospital super-peer agents within the corresponding HBN.

policy H1 {
  summary: Doctor D#111 from Hospital_1 is allowed access to Patient P#112's EHRs in Hospital_2.
  hospitals: Hospital_1; Hospital_2
  role: doctor (Doctor D#111), patient (Patient P#112)
  condition: doctor ∈ Hospital_1 && patient ∈ Hospital_2
  owners: $\beta_{HOS\text{-}HOSPITAL\_1}$; $\beta_{HOS\text{-}HOSPITAL\_2}$
  expiration: 01/01/2031
}

Access control policy H1 specifies that doctor D#111 is allowed to access patient P#112's EHRs in Hospital_2. Note that since a hospital access control policy HR_{ACP} specifies access rights within an HBN, we can safely assume that Hospital_1 and Hospital_2 are located in the same city. When doctor D#111 makes a request to access patient P#112's EHRs, both hospital super-peer agents $\beta_{HOS\text{-}Hospital\_1}$ and $\beta_{HOS\text{-}Hospital\_2}$ attempt to verify the request by checking policy H1 stored in their blockchains. If approved, doctor D#111 is granted access to patient P#112's EHRs maintained by Hospital_2. A city access control policy stored as a CR_{ACP} is similar to a hospital access control policy stored as an HR_{ACP}, but it must specify the cities where the hospitals are located because the hospitals belong to different cities within the same state; otherwise, if the hospitals belong to the same city, the access control policy shall be recorded as an HR_{ACP}. An example policy C1 is shown below, which can be stored in a city blockchain as a CR_{ACP} and enforced by city super-peer agents in CBNs for the HBNs under their jurisdiction.

policy C1 {
  summary: Doctor D#111 from Hospital_1 (City_1) is allowed access to Patient P#112's EHRs in Hospital_3 (City_3).
  hospitals: City_1.Hospital_1; City_3.Hospital_3
  role: doctor (Doctor D#111), patient (Patient P#112)
  condition: doctor ∈ City_1.Hospital_1 && patient ∈ City_3.Hospital_3
  owners: $\beta_{CIT\text{-}City\_1}$; $\beta_{CIT\text{-}City\_3}$
  expiration: 02/02/2032
}

Access control policy C1 specifies that doctor D#111 from Hospital_1 in City_1 is allowed to access patient P#112's EHRs at Hospital_3 in City_3. Different from policy H1, when doctor D#111 makes a request to access patient P#112's EHRs located in a different city, both hospital super-peer agents $\beta_{HOS\text{-}Hospital\_1}$ and $\beta_{HOS\text{-}Hospital\_3}$ forward the request to their city super-peer agents $\beta_{CIT\text{-}City\_1}$ and $\beta_{CIT\text{-}City\_3}$ to check against policy C1 stored in their city blockchains. If approved, doctor D#111 is granted access to patient P#112's EHRs at Hospital_3 in City_3.

A state access control policy stored as an SR_{ACP} is similar to a city access control policy stored as a CR_{ACP}, but it must specify both the cities and the states where the hospitals are located. an example policy S1 is shown below, which can be stored as an SR_{ACP} in a state blockchain and enforced by state super-peer agents in the SBN for CBNs and HBNs under their jurisdiction.

policy S1 {
  summary: Doctor D#111 from Hospital_1 (City_1, State_1) is allowed access to Patient P #112's EHRs in Hospital_4 (City_4, State_4).
  hospitals: State_1.City_1.Hospital_1; State_4.City_4.Hospital_4
  role: doctor (Doctor D#111), patient (Patient P#112)
  condition: doctor ∈ State_1.City_1.Hospital_1 && patient ∈ State_4.City_4.Hospital_4
  owners: $\beta_{STA\text{-}State\_1}$; $\beta_{STA\text{-}State\_4}$
  expiration: 03/03/2033
}

Access control policy S1 specifies that doctor D#111 from Hospital_1 (City_1, State_1) is allowed to access patient P#112's EHRs stored in Hospital_4 (City_4, State_4). In a similar nature to policy C1, when doctor D#111 makes a request to access patient P#112's EHRs located in a different state, both hospital super-peer agents $\beta_{HOS\text{-}Hospital\_1}$ and $\beta_{HOS\text{-}Hospital\_4}$ forward the request to their state super-peer agents $\beta_{STA\text{-}State\_1}$ and $\beta_{STA\text{-}State\_4}$, through their city super-peer agents, $\beta_{CIT\text{-}City\_1}$ and $\beta_{CIT\text{-}City\_4}$. The request is then checked against policy S1 stored in their state blockchains. If approved, doctor D#111 is granted access to patient P#112's EHRs from Hospital_4 (City_4, State_4).

Note that to support efficient access authorization and avoid duplication of an access control policy across multiple access control policy records, access control policies are no longer encrypted as in our previous work [7]. For more examples of access control policies at hospital, city and state levels, refer to earlier work [5], [7].

5.2. Integrated Search and Retrieval of EHRs

Once the required access control policies have been created and stored in the relevant hospital, city and state blockchains, the

associated data can now be opened and shared with other hospitals across the country. This data sharing is supported by an integrated EHRs search and retrieval process that enables those with the proper authorization to retrieve all EHRs of a patient from any hospitals, regardless of which HBNs they participate in. This process involves all three layers of our hierarchical blockchain framework, as search requests are forwarded and executed concurrently across all super-peer agents in the hierarchical network structure. The concurrent search and retrieval process is defined by three procedures, which are searching and retrieving EHRs across hospitals within the same city, searching and retrieving EHRs across cities within the same state, and searching and retrieving EHRs across states within a country. We now describe each of the three procedures as follows.

The procedure of searching and retrieving EHRs across hospitals within the same city is presented in Algorithm 4. The algorithm is initiated by a hospital super-peer agent β_{HOS} on behalf of its end user (e.g., a doctor) to search and retrieve patient p's EHRs from other hospitals within the same city (i.e., within the same HBN). Agent β_{HOS} sends this request to its city super-peer agent β_{CIT} to start the process.

Algorithm 4: Searching and Retrieving a Patient's EHRs from All Hospitals within the Same City by a City Super-Peer Agent β_{CIT}

Input: A retrieval request for hospitals containing patient p's EHRs
Output: A list of links to patient p's EHRs

1. Let ρ_{h_list} be the list of hospital super-peers under β_{CIT}'s jurisdiction
2. Let η_{ehr_hlist} be an empty list of links to EHRs; *nResponse* = 0
3. **for** each γ_h in ρ_{h_list}
4. forward the retrieval request to γ_h asynchronously, which invokes a search process at hospital h based on the established policies
5. **while** (*not* timeout) or *nResponse* != | ρ_{h_list} |
6. **if** γ_h returns a link to p's EHRs
7. add the link to the list η_{ehr_hlist}; *nResponse*++
8. **else** *nResponse*++; **continue** // γ_h returns no link to EHRs
9. **return** the list η_{ehr_hlist}

According to the algorithm, agent β_{CIT} sends concurrent requests in its HBN to all hospital super-peer agents under its jurisdiction and waits for responses or until the timeout. Each hospital super-peer agent γ_h who receives this request will perform a permission checking based on the established access control policies stored as HR_{ACP} in its hospital blockchain. If valid, γ_h creates a link that allows access to patient p's EHRs and sends it back to β_{CIT}. If β_{CIT} receives a response from γ_h with this link, the link is added to list η_{erh_hlist}; otherwise, β_{CIT} continues to wait. When all hospital super-peer agents have responded or timed out, the list η_{erh_hlist} is returned and sent back to β_{HOS}. Upon receiving η_{erh_hlist}, β_{HOS} can then use the links to access and retrieve patient p's EHRs on behalf of the end user.

The procedure of searching and retrieving EHRs across cities within the same state, is presented in Algorithm 5. Similar to the procedure of searching and retrieving EHRs across hospitals within the same city, Algorithm 5 is initiated by a hospital super-peer agent β_{HOS} on behalf of its end user (e.g., a doctor) to search and retrieve patient p's EHRs from hospitals in different cities within the same state (i.e., within different HBNs connected under the same CBN). Agent β_{HOS} sends this request to its city super-peer agent β_{CIT}, who forwards it to its state super-peer agent β_{STA} to start the process. According to the algorithm, agent β_{STA} sends concurrent requests to all city super-peer agents under its jurisdiction in its CBN and waits for responses from them or until it times out. Upon receiving the search request, each city super-peer agent γ_c performs a permission check based on the access control policies stored as CR_{ACP} in its city blockchain. If valid, γ_c executes Algorithm 4 to forward the search and retrieval requests to the hospital super-peer agents under its jurisdiction. If γ_c returns a list η_{ehr_hlist} containing links to p's EHRs, η_{ehr_hlist} is appended to the list η_{ehr_clist}; otherwise, β_{STA} continues to wait. When all city super-peer agents have responded or it times out, the list η_{ehr_clist} is returned and sent back to β_{CIT}, who further sends it back to β_{HOS}. Upon receiving η_{erh_clist}, β_{HOS} can then use the links to access and retrieve patient p's EHRs on behalf of the end user.

Algorithm 5: Searching and Retrieving a Patient's EHRs from All Hospitals within the Same State by a State Super-Peer Agent β_{STA}

Input: A retrieval request for hospitals containing patient p's EHRs
Output: A list of links that can be used to access patient p's EHRs

1. Let ρ_{c_list} be the list of city super peers under β_{STA}'s jurisdiction
2. Let η_{ehr_clist} be an empty list of links to EHRs; *nResponse* = 0
3. **for** each γ_c in ρ_{c_list}
4. forward the retrieval request to γ_c asynchronously, which invokes *Algorithm* 4 to search in city c based on the established policies
5. **while** (*not* timeout) or *nResponse* != | ρ_{c_list} |
6. **if** γ_c returns η_{ehr_hlist} that contains links to p's EHRs
7. append η_{ehr_hlist} to η_{ehr_clist}; *nResponse*++
8. **else** *nResponse*++; **continue** // γ_c returns an empty list
9. **return** the list η_{ehr_clist}

Finally, the procedure of searching and retrieving EHRs across states within a country is similar to Algorithm 5, where the retrieval request is sent from a hospital super-peer agent β_{HOS} to its state super-peer agent β_{STA}. Agent β_{STA} then initiates the concurrent searches by broadcasting the request to all other state super-peer agents. Each state super-peer agent γ_s, representing state S, performs a permission checking based on the access control policies stored as SR_{ACP} in the state blockchain. If valid, γ_s executes Algorithm 5 to search and retrieve EHRs of patient p from all cities within state S. The return result η_{erh_clist} is appended to η_{erh_slist} if it is not empty. When all state super-peer agents have either responded or timed out, the list η_{erh_slist} is returned and sent back to β_{HOS} via β_{CIT}. β_{HOS} can then use the links to access and retrieve patient p's EHRs on behalf of the end user.

6. Case Study

To demonstrate the feasibility and efficiency of our proposed approach, we conducted experiments and evaluated the performance of our hierarchical approach based on the settings and results of each simulation. In our experimental environment, we utilized multiple servers and computers connected under the same network. The specifications of the servers include Intel® Core™ i7-4790k CPU @ 3.60GHz (4 CPU Cores); 16 GB RAM, Windows 10 OS (64-bit, x64-based processor); and 256 SSD Hard Drive. Our experimental environment also had a recorded Internet speed of 600 Mbps.

6.1. Numbers of Published Blocks During a Week

In the first case study, we test our temporary and permanent block approach by conducting simulations to evaluate the need to

use temporary blocks. We simulate and analyze the number of permanent blocks that can typically be created each day of a week based on predefined threshold values. These threshold values are the maximum total size of 2GB for all accumulated temporary blocks and a maximum of 100 new blocks added during a single day. If neither of the thresholds is reached, a permanent block is always created at the end of the day. The number of temporary blocks added during a day is determined by the number of patient visits. We assume that a patient visit always results in the generation of an EHR, which is saved as an HR_{MER} and immediately published to the blockchain as a temporary block. To simplify our experiments, we focus only on HR_{MER} rather than other record types, i.e., HR_{AR}, HR_{UPR}, and HR_{ACP}, because in real-world scenarios, HR_{MER} is the main contributor of block content in hospital blockchains. For the content or nature of the EHRs stored in each HR_{MER}, we use the following experimental settings. Each HR_{MER} includes text-based reports in the size range of [5, 10] KB, while there is also less than a 10% probability of including multimedia files in the size range of [10, 500] MB. Thus, an HR_{MER} must contain text-based report along with possible multimedia files of different sizes. The range of patient visits are based on hospital sizes, where we simulate three different sizes of hospitals. The first type of hospitals has daily patient visits of [10, 100] and is categorized as a small hospital. The second type of hospitals has daily patient visits of [50, 500] and is categorized as a medium hospital. The third type of hospitals has daily patient visits of [100, 1000] and is categorized as a large hospital. Table 1 shows the number of patient visits for each day of a week at each simulated hospital.

Table 1: Numbers of Patient Visits per Day

Hospital Size	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Large	900	700	500	550	700	750	850
Medium	450	300	150	200	300	350	400
Small	100	70	50	55	60	75	90

We now conduct experiments to generate permanent blocks based on the number of patient visits per day. Figure 9 shows the average number of permanent blocks that can be formed at each hospital based on the experimental settings and the numbers of daily patient visits listed in Table 1. As we can see from the figure, even for a large hospital, the number of permanent blocks published per day is limited. The time interval between each addition of permanent blocks can be several hours or even longer, depending on the number of permanent blocks added that day. This indicates a critical need to use temporary blocks to publish data to the blockchain in a timely manner for immediate access without delay. Based on the results in Figure 9, we conclude that medium and small hospitals would benefit the most from our temporary and permanent block approach, as they generate the fewest average numbers of permanent blocks. Figure 10 shows the relationship between the number of permanent blocks formed vs. the number of temporary blocks added during a day at a medium-sized hospital. Since each patient visit results in a new temporary block being created, the number of newly added temporary blocks is equal to the number of new patient visits. According to the simulation results, in medium-sized hospitals, when the number of patient visits increases, the number of new permanent blocks also increases. When the maximum number of patient visits is reached, i.e., 500, the number of new permanent blocks formed daily is between 4 and 7, which is considered to be very acceptable in terms of spatial efficiency.

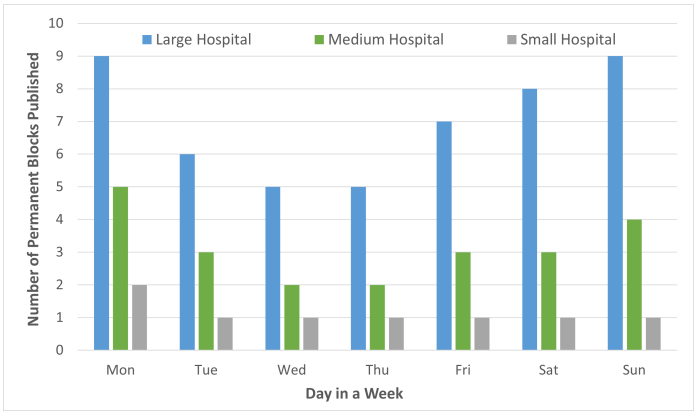


Figure 9: Average Number of Permanent Blocks Published During a Week

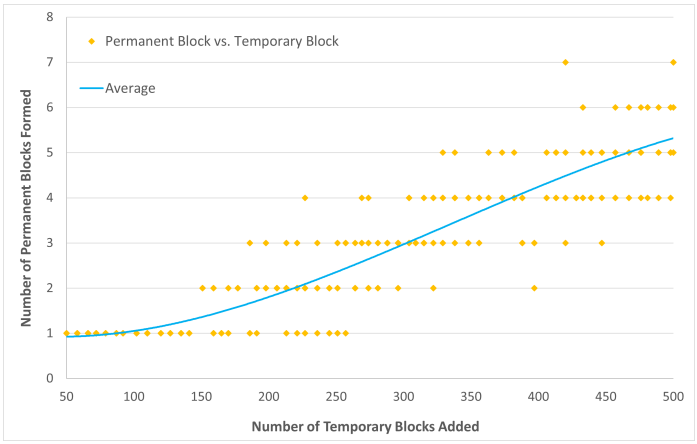


Figure 10: Number of Permanent Blocks Formed vs. Number of Temporary Blocks Added During a Day in a Simulated Medium-Sized Hospital

6.2. Latency of Publishing Temporary and Permanent Blocks

In this experiment, we simulate the creation and addition of permanent and temporary blocks to the blockchain. We record and analyze the time to create, broadcast, and publish temporary and permanent blocks based on a varying number of hospital super-peer agents within an HBN. We use the same settings established in the first case study for this experiment. This means that in a typical day, a new permanent block will have either 100 records stored or 2 GB in size, according to the given thresholds. Each temporary block includes only HR_{MER} containing text-based reports and potentially (10%) multimedia files of [10, 500] MB in size. In addition, we add random delays in the range of [100, 3000] milliseconds to simulate network congestion during the consensus process. Figure 11 shows the experimental results and the efficiency of our approach using temporary blocks. From the figure, we can see that at most, it takes about less than half a minute to create a temporary block and add it to the blockchain. Due to the large variation in the potential size and frequency of temporary blocks containing multimedia files in our experimental settings, the range between each case can vary considerably. In contrast, for permanent blocks, the range is more consistent for each case due to previously determined size and number thresholds. In general, the overall time for a permanent block to

be created and added to the blockchain is less than one minute. While there is no big data transferred during the consensus process, additional validation is required by the other super-peer agents to verify the permanent block broadcast by the block announcer and the associated temporary blocks previously stored in their local copies of the blockchain. This significantly increases the overall time required for the consensus process, which takes longer time when compared to the publication process involving temporary blocks only. Nevertheless, it takes no more than 20 seconds to create and add a temporary block to the blockchain, which allows many consensus processes for new temporary blocks to be performed during a single day without encountering significant delays. Therefore, we can conclude that our approach supports efficient creation and addition of temporary and permanent blocks to the blockchain.

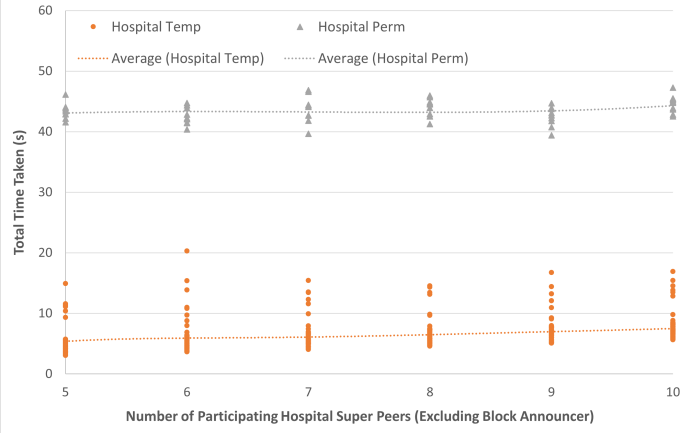


Figure 11: Total Time Taken to Create, Broadcast, and Publish Hospital Temporary and Permanent Blocks

6.3. Search and Retrieval Time in Hierarchical Blockchains

In this experiment, we simulate the integrated concurrent search and retrieval processes for patient EHRs. We record and analyze the total time taken to search and retrieve a number of patient's EHRs in the hierarchical blockchain networks. Our experimental environment consists of three layers of fully simulated hierarchical networks with a varying number of HBNs, CBNs, and an SBN. We also have a range of [6, 10] hospital super peers within an HBN, [100, 500] city super peers within a CBN, and 50 state super peers within the SBN. The contents of our hospital, city and state blockchains contain all the necessary or relevant access control policy records to enable our searching process. To simplify the overall downloading process, the hospital blockchain contains only EHRs with multimedia files. Each EHR has a range of [10, 500] MB in size, similar to our previous case studies. The EHRs are stored in different hospital blockchains within different HBNs to simulate a patient who visits multiple hospitals in different cities. We also introduce a random delay with a range of [100, 3000] milliseconds to simulate network congestions. In addition, we assume that all hospitals have the required infrastructure to allow multiple concurrent file downloads to mitigate throttling when any number of peers download multiple files simultaneously. Each hospital agent also maintains a separate local index file for efficient responses to any EHR-related inquiries. Figure 12 shows the total time taken to search and retrieve different numbers of EHRs. Based on the figure, we can see that the search time remains relatively constant regardless of the number of EHRs to be searched. Several factors, such as the use of separate index files to track patient EHRs for fast response and the small size of the metadata involved during the concurrent search process, contribute significantly to this stability. However, when the total numbers of EHRs to be retrieved increases, the time to retrieve those EHRs also increases. This result is consistent with the experimental results we reported in our previous work [7]. However, the current approach is more efficient compared to the previous method as the waiting time for creating new access policy records is not needed any more after the search process. This leads to an overall time improvement, which allows multiple EHRs to be searched and retrieved in one integrated process.

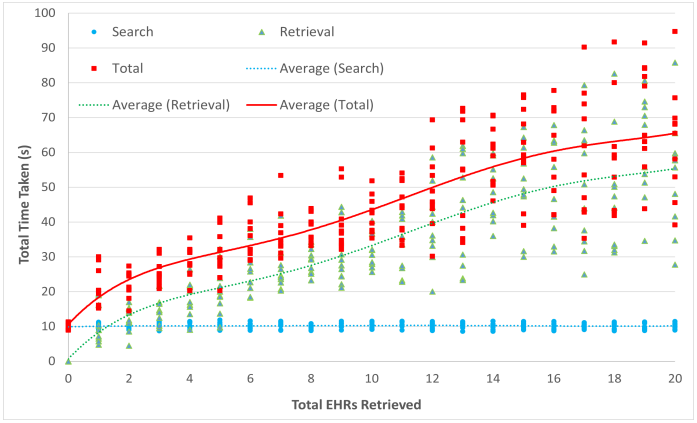


Figure 12: Time Taken to Search and Retrieve Varying Number of EHRs

7. Conclusions and Future Work

In this paper, we build on the concepts and methods of previous work [7] and further explore them by introducing several design changes. These changes include limiting the number of hospital super-peer agents in an HBN, adopting a temporary and permanent block scheme, and integrating the search and retrieval processes for EHRs. The number of hospital super-peer agents in an HBN is limited to minimize the redundancy of big data stored in the blockchain. This also ensures better scalability, as we limit the growth potential of hospital blockchain to a more manageable level. The use of the temporary and permanent block scheme in our hierarchical blockchain approach ensures timely publications of EHRs in an HBN. Any urgent data can be published in a temporary block immediately, while once a certain threshold has been reached, a permanent block consisting of a number of temporary blocks can be formed. This scheme allows for timely and space-saving publications of EHRs to the cloud-based hospital blockchains. Finally, the search and retrieval processes for EHRs have been integrated to be more efficient. As the experimental results show, our new hierarchical blockchain approach is efficient and effective, allowing for a timely and spatially efficient publication of EHRs to the hospital blockchain as well as a better overall performance in the integrated search and retrieval process of EHRs.

In future work, we plan to perform an in-depth comparison of our cloud-based on-chain blockchain approach with IPFS-based off-chain approaches [10], [11] and mechanisms for reliable and secure distributed cloud data storage [23]. We will focus on the

redundancy and efficiency aspects of such comparisons and evaluate the performance of our cloud-based hierarchical blockchain mechanism. In addition, we plan to further improve and develop our approach to defend against real-world attacks, such as DDOS attacks and insider threats [24], [25], [26]. An emphasis will be on evaluating the performance of our consensus process and cryptographic procedures against potential attacks and improving them if necessary.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] S. Nakamoto, "Bitcoin: a peer-to-peer electronic cash system," October 2008. Retrieved on January 15, 2021 from https://bitcoin.org/bitcoin.pdf.

[2] O. Dib, K.-L. Brousmiche, A. Durand, E. Thea, E. B. Hamida, "Consortium blockchains: overview, applications and challenges," International Journal on Advances in Telecommunications, **11**(1&2), 51-64, 2018.

[3] M. T. de Oliveira, L. H. A. Reis, R. C. Carrano, F. L. Seixas, D. C. M. Saade, C. V. Albuquerque, N. C. Fernandes, S. D. Olabarriaga, D. S. V. Medeiros, D. M. F. Mattos, "Towards a blockchain-based secure electronic medical record for healthcare applications," in Proceedings of the 2019 IEEE International Conference on Communications (ICC), 1-6, Shanghai, China, May 2019, doi: 10.1109/ICC.2019.8761307.

[4] Q. Xia, E. B. Sifah, K. O. Asamoah, J. Gao, X. Du, M. Guizani, "MeDShare: trust-less medical data sharing among cloud service providers via blockchain," *IEEE Access*, **5**, 14757-14767, July 2017, doi: 10.1109/ACCESS.2017.2730843.

[5] A. Thamrin, H. Xu, "Cloud-based blockchains for secure and reliable big data storage service in healthcare systems," in Proceedings of the 15th IEEE International Conference on Service-Oriented System Engineering (IEEE SOSE 2021), 81-89, Oxford Brookes University, UK, August 2021, doi: 10.1109/SOSE52839.2021.00015.

[6] R. Ming, H. Xu, "Timely publication of transaction records in a private blockchain," 2020 IEEE 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C), 116-123, Macau, China, December 2020, doi: 10.1109/QRS-C51114.2020.00030.

[7] A. Thamrin, H. Xu, "Hierarchical cloud-based consortium blockchains for healthcare data storage," in 2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C), 644-651, Hainan Island, China, December 2021, doi: 10.1109/QRS-C55045.2021.00098.

[8] S. Alexaki, G. Alexandris, V. Katos, N. E. Petroulakis, "Blockchain-based electronic patient records for regulated circular healthcare jurisdictions," in Proceedings of the 23rd IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 1-6, 2018, doi: 10.1109/CAMAD.2018.8514954.

[9] R. Kumar, N. Marchang, R. Tripathi, "Distributed off-chain storage of patient diagnostic reports in healthcare system using IPFS and blockchain," in Proceedings of the 2020 International Conference on COMmunication Systems & NETworkS (COMSNETS), 1-5, Bengaluru, India, 2020, doi: 10.1109/COMSNETS48256.2020.9027313.

[10] D. Li, W. E. Wong, M. Zhao, Q. Hou, "Secure storage and access for task-scheduling schemes on consortium blockchain and interplanetary file system," 2020 IEEE 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C), 153-159, 2020, doi: 10.1109/QRS-C51114.2020.00035.

[11] Y. Jeong, D. Hwang, K. Kim, "Blockchain-based management of video surveillance systems," in Proceedings of the 2019 International Conference on Information Networking (ICOIN), 465-468, Kuala Lumpur, Malaysia, 2019, doi: 10.1109/ICOIN.2019.8718126.

[12] Z. Su, H. Wang, H. Wang, X. Shi, "A financial data security sharing solution based on blockchain technology and proxy re-encryption technology," in Proceeedings of the IEEE 3rd International Conference of Safe Production and Informatization (IICSPI), 462-465, Chongqing City, China, 2020, doi: 10.1109/IICSPI51290.2020.9332363.

[13] H. Wang, Y. Song, "Secure cloud-based EHR system using attribute-based cryptosystem and blockchain," Journal of Medical Systems, **42**(152), 1-9, July 2018, doi: 10.1007/s10916-018-0994-6.

[14] B. S. Egala, A. K. Pradhan, V. R. Badarla, S. P. Mohanty, "Fortified-chain: a blockchain based framework for security and privacy assured Internet of medical things with effective access control," IEEE Internet of Things Journal, **8**(14), 11717-11731, July 2021, doi: 10.1109/JIOT. 2021.3058946.

[15] A. Fernandes, V. Rocha, A. F. d. Conceicao, F. Horita, "Scalable architecture for sharing EHR using the Hyperledger blockchain," in Proceedings of the IEEE International Conference on Software Architecture Companion (ICSA-C), 130-138, Salvador, Brazil, March 2020, doi: 10.1109/ICSA-C50368.2020.00032.

[16] N. Nicol, H. Xu, "A blockchainless approach for trusted public construction bidding," Computer and Information Science Technical Report, Computer and Information Science Department, University of Massachusetts Dartmouth, December 2018.

[17] L. Cui, S. Yang, Z. Chen, Y. Pan, M. Xu, K. Xu, "An efficient and compacted DAG-based blockchain protocol for industrial Internet of things," IEEE Transactions on Industrial Informatics, **16**(6), 4134-4145, 2020, doi: 10.1109/TII.2019.2931157.

[18] A. Buzachis, A. Celesti, M. Fazio, M. Villari, "On the design of a blockchain-as-a-service-based health information exchange (BaaS-HIE) system for patient monitoring," in Proceedings of the IEEE Symposium on Computers and Communications (ISCC), 1-6, Barcelona, Spain, July 2019, doi: 10.1109/ISCC47284.2019.8969718.

[19] D. C. Nguyen, P. N. Pathirana, M. Ding, A. Seneviratne, "Blockchain for secure EHRs sharing of mobile cloud based e-health systems," IEEE Access, **7**, 66792-66806, May 2019, doi: 10.1109/ACCESS. 2019.2917555.

[20] H. Guo, W. Li, M. Nejad, C. Shen, "Access control for electronic health records with hybrid blockchain-edge architecture," in Proceedings of the 2019 IEEE International Conference on Blockchain (Blockchain), 44-51, July 2019, doi: 10.1109/Blockchain.2019.00015.

[21] H. Guo, W. Li, E. Meamari, C. Shen, M. Nejad, "Attribute-based multi-signature and encryption for EHR management: a blockchain-based solution," in Proceedings of the 2020 IEEE International Conference on Blockchain and Cryptocurrency (ICBC), 1-5, Toronto, ON, Canada, May 2020, doi: 10.1109/ICBC48266.2020.9169395.

[22] M. Meingast, T. Roosta, S. Sastry, "Security and privacy issues with health care information technology," in Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society, 5453-5458, New York, NY, USA, September 2006, doi: 10.1109/IEMBS.2006.260060.

[23] H. Xu, D. Bhalerao, "Reliable and secure distributed cloud data storage using Reed-Solomon codes," International Journal of Software Engineering and Knowledge Engineering (IJSEKE), **25**(9&10), 1611-1632, 2015, doi: 10.1142/S0218194015400355.

[24] S. Northcutt, J. Novak, Network intrusion detection, 3rd Edition, Sams Publishing, August 2002.

[25] H. Xu, A. Reddyreddy, D. F. Fitch, "Defending against XML-based attacks using state-based XML firewall," Journal of Computers (JCP), **6**(11), 2395-2407, November 2011, doi: 10.4304/jcp.6.11.2395-2407.

[26] J. Mirkovic, P. Reiher, "A taxonomy of DDoS attack and DDoS defense mechanisms," **34**(2), 39-53, April 2004, doi: 10.1145/997150.997156.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 2, 191-197 (2022)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Towards a Framework for Organizational Transformation through Strategic Design Implementation

Lynne Whelan[1,*], Louise Kiernan[2], Kellie Morrissey[2], Niall Deloughry[2]

[1]*South East Technological University Ireland, Carlow, R93V960, Ireland*

[2]*Univeristy of Limerick, School of Design, V94T9PX, Ireland*

ARTICLE INFO

Article history:
Received: 28 January, 2022
Accepted: 07 April, 2022
Online: 22 April, 2022

Keywords:
Transformational
Design
Strategy

ABSTRACT

The aim of the research is to contribute to the emergent field of strategic design as an approach to creating transformational impacts in organizations. This is driven by international strategies to promote sustainable business environments which are innovative and adaptive to change. The literature identifies a gap in the knowledge in relation to how knowledge flows within an organization from knowledge producing activities and back. This flow between management and operational activity is identified as the ongoing innovation capability and ultimately the area of transformational impact within the organization. However, the structures which support this flow and how it is measured are ambiguous and ill defined. Strategic design is a holistic approach to developing business strategies which may provide methods which support the flow of knowledge and provide the transformational impacts. A 'research through design' approach was taken to collect data from small to medium enterprises engaging in a series of strategic design workshops. Analysis was carried out through visual mapping of how strategic design is applied in an organization at management and operational level. This identifies the nuanced differences in application and design tools used to develop management strategy versus operational strategy. It also identifies the points of knowledge creation and transfer which builds business intelligence to inform both business and operational strategies. This is presented as a contextual framework which provides the basis for understanding the complexity of the flow between the two. This contributes to informing organizations of where the links between these strategies may be built, resulting in a more dynamic organization which is nimble, innovative, and adaptive to change.

1. Introduction

This research is an extension of work originally presented in 2021 IEEE Technology & Engineering Management Conference. The title of the previous paper was 'Measuring the Success Factors of Strategic Design Implementation'[1]. This paper presents a literature review of discourse on the development and application of strategic design across the business landscape with a focus on transformational impacts. The gap in the knowledge is identified in relation to how the flow of organizational knowledge which is both tacit and explicit is supported and leveraged to create strategic growth and innovation. A 'research through design' approach is used for data collection and analysis through a series of design workshops. The findings are presented in a contextual framework of where and how strategy is used in an organization and where and how a design approach supports the transformational flow of knowledge. The research began with seeking to understand methods of determining success factors of strategic design implementation. A design approach is a holistic view of a situation or environment. It looks at the usability, feasibility, and viability factors to solve problems [2]. This can be a people centered approach, a technological approach, and a business approach. Therefore, to consider best practices for measuring success factors in a way that reflects this, different disciplines were considered in the literature review such as social science, data science and business. The literature review noted that across these disciplines they all referred to a flow of knowledge between management strategy and operational strategy. The gap in the knowledge is in relation to how the flow of knowledge operates within an organization. Knowledge management is described as the process of acquisition, storage, distribution and the use of knowledge [3]. The research design was to collect data from seven small to

*Corresponding Author: Lynne Whelan, Lynne.Whelan@itcarlow.ie

medium enterprises (SMEs) in research through design approach as they engage in a series of strategic design workshops within a Technology Gateway in an Irish University. The data is analysed through visual mapping to provide insights into the nuanced differences in approach to business strategy and operational strategy and the relevance of knowledge exchange between the two. The visual mapping highlights the tools used across a four-stage process of strategic design application. The findings highlight the process as on ongoing cycle of innovation and identifies the key points of knowledge exchange. It is from this that the author has developed a framework of organizational strategy depicting the strategic design process, tools used and knowledge acquisition and transfer points. The research aims to inform both the design and business communities nationally and internationally, in the actionable methods, tools, and processes for accessing and implementing a strategic design approach for transformational impacts.

2. Literature Review

There are clear international drivers to promote innovation on economic, social, cultural and educational sectors. Innovation is recognized as a means to create sustainable organizations which can not only problem solve to adapt to change but also become the drivers of change. The US Chamber of Commerce Global Innovation Policy Centre (GIPC) champions innovation to "create jobs, save lives, advance global economic and cultural prosperity, and generate breakthrough solutions to global challenges" [4]. This highlights the significant and varied applications for innovation. The Organization for Economic Co-operation and Development (OECD) work to improve innovation across 38-member countries to stimulate economic progress and world trade. The core mission of OECD is not static but needs to respond to an evolving and challenging world and notes that this is a time when the expanding integration and influence of new technologies are disrupting both advanced and emerging economies [5]. It is highlighted that adaptability to change, and innovation are at the core of global economic strategies to develop competitive and sustainable organizations. This paper is set within the context of Irish Higher Education (HEI) engagement with industry which has long been identified as a model to foster and promote research and development, innovation and faster knowledge exchange between researchers and industry. The publication of 'Winning by Design' marks the clear linking of design and Innovation by the Irish State [6]. Design was recognized as a process of innovation and design thinking as a strategic tool for innovative business development. It is highlighted that it will require more than transactional interventions and that the focus should be on longer term developmental programmes to achieve transformational impacts.

"Mobilizing universities needs to be addressed in a holistic way and not just by focusing on transactional interventions such as consultancy services for local companies. It is tempting to focus on transactional mechanisms as they have clear outputs such as number of firms assisted. However, they are less likely to have the longer-term outcomes and impacts that can be achieved with 'transformational' and more developmental programmes" [7]

In seeking to understand transformational engagements within a design context, strategic design processes and approaches were examined through literature. Roberto Verganti presents a theory of design driven innovation as radically innovating what things mean by bringing the designer into the corporate space [8]. This approach however is referenced in the context of new product development within the business model. Giulio Calabretta who writes specifically about strategic design describes it as "the professional field in which designers use their principles, tools and methods to influence strategic decision making within an organization" [9]. However, Calabretta also leans towards the strategic input as an underpinning of the product brief "designers are no longer mere executors of design briefs-they are involved in the crafting of the briefs and guide the strategic decisions that underlie them" . This research paper argues that the impacts may be further reaching and more transformational than product strategy alone. It presents strategic design as a holistic approach which when applied to industry can result in not only new products but new services, new markets, or simply new ways of doing things, impacting an organization across business and operational strategy. Strategic design is the application of design thinking to develop strategies, in a business setting this is primarily strategies for growth and innovation. An approach to better understanding the application of design on both product development and business model strategy is to consider the outcomes and the measures of success factors. In other words what are the outcomes of a strategic designer's engagement when applied to new product development or when applied to the broader business model strategy. A new product is an explicit and tangible outcome of a design engagement but how do we measure the success factors of the business model strategy. This area is more difficult to communicate and measure in terms of impacts as it may be of a tacit nature, less tangible and more subjective. It would appear that a more qualitative method of assessing the outcomes may be of value.

2.1. Social science, data science and business

Design approaches encompass both tacit and explicit knowledge to form holistic understanding. Whilst explicit knowledge is objective and easy to quantify, tacit knowledge is subjective and difficult to quantify and therefore often overlooked. Literature was therefore considered from disciplines which measure values in different ways, such as social science, data science and business. In social science, Etienne Wenger was one of the first to use the term 'communities of practice' – groups of people who together accumulate and share their collective learning [10]. In relating to organizations, Wenger considers ways to measure the value of the organizations community and proposes that good measurement has to follow the course of the story. He refers to the process of analysing the stories as "systematic anecdotal evidence" and that by considering communities of practice, organizations have the chance to see the value in qualitative not quantitative terms. Wenger concludes "in fact , the value of knowledge is a flow from knowledge producing activity to performance and back". Wenger specifically links social practice and behaviours with a transformative business model. He refers to the generation of social practice and changing the designs of our organizations so that they are more in line with our behaviours. Geoff Walsham also looks at information flow within an organization but considers it from a context of information systems such as data processing. "in order for all types of

organizations to succeed, they need to be able to process data and use information effectively" [11]. Walsham refers to this information as informing everyday operations such as planning, controlling, organizing and decision making. This process of information gathering and application, whilst dealing with data science, is also reflective of Wenger's communities of practice and the information flow between operations and business strategy. As far back as the 1970s, Mintzberg was one of the first to highlight the 'flaw' in organizational structure which is how work, information and decision processes actually flow through it [12]. in other words, the interconnectedness and flow between business strategy and operations. Wenger follows this understanding and links the communities of practice to managing the knowledge in an organization, between employees and external stakeholders as peer exchange. He is clear however that to nurture knowledge resources a CEO must understand the broad strategy but also needs to be in contact with the practitioners who manage that knowledge. This infers a more involved and human influence on building the strategy. In exploring these elements, we can consider praxeology, the theory of human action and purposeful behaviour, and axiology, the philosophical study of value. Dr Marina Pankina, presents a paper specifically looking at axiology and praxeology of design thinking. Pankina states that "design combines an objective and subjective approach. This is what distances it from the classical science that aims for objectivity and elimination of everything subjective, as a key to the validity of knowledge" [13]. This is particularly relevant in a design context as design is focused on understanding cultural, ethical, and economic values, along with human drivers, motivations, wants and needs. It may be that these human influences are the tacit links to successful transformative strategies. Additionally, this would indicate that to measure the success factors of design engagements it is also appropriate to combine objective and subjective results. Reflecting on this literature we can see there are, according to Walsham and Wenger, two key tiers of approach to organizational strategy; a managements' strategic level of broad initiative and an operational level of knowledge creation and that the flow between the two is essential (Figure 1). Measuring the success factors can be approached subjectively based on systematic anecdotal evidence or objectively based on data collection, application, and management. However, Wenger clearly links social practice and behaviours with transformative business models. Pankina presents design as a methodology which combines both an objective and subjective approach and that designers work to understand the human factors of behaviour and value to bring meaning.

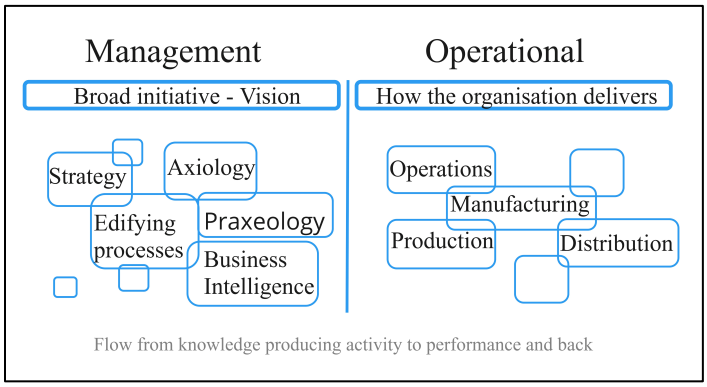


Figure 1. Two tiers of approach for organizational strategy (authors own 2021)

3. Research Question

This combined objective and subjective approach that design offers, applied holistically across an organization at management and operational levels, may result in transformational impacts. The research question is; How can the strategic design approach support the flow of knowledge (both tacit and explicit) across management and operational strategies resulting in transformational impacts?

4. Methodology

The research is approached in a 'research through design' methodology. This is a research approach that uses design practice to inform research [14]. This research was carried out in the Design + Technology Gateway which specializes in design research, design driven innovation and strategic design. The research design was to participate in the facilitation of the design strategy workshops with a series of small to medium enterprises (SMEs). Each SME owner/manager engaged in four workshops facilitated by two design strategists as laid out in Table 1.

Table 1. Research design

Location	Design Studio & remote video workshops
Participants	Design strategists x2 SME manager x1
Time	2 hours x 4 weeks
Duration	6-8 weeks per SME (x7 SMEs)
Methods	Visual drawing/text on white wall, post-it notes, colored pens. Use of design toolkit
Total Organization	7
Sectoral range	Corporate service, Brewing and distilling, IOT, Health care, Retail, Events, Technology sector

The workshops were carried out in the design studio with large whiteboard walls which were used to capture the data as the participants engaged in the process. Colored markers, post it notes, print outs and digital images were used as a mixed media approach to capture data (Figure 4). Strategic design, as delivered by the Design+ team consists of a four-stage process (Figure 3). This process is derived from the underlying theoretical double diamond design process [2]. This is an approach which is internationally recognised by both industry and academic realms and is based on a series of divergent and convergent thinking (Figure 2).

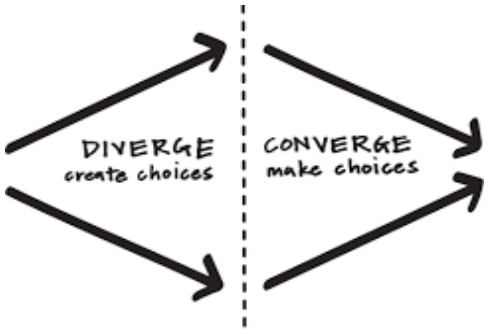


Figure 2. Divergent and convergent thinking [IDEO 2019]

The theoretical approach of convergent and divergent thinking has been translated by the design strategy team into the business landscape as an actionable four stage process for the development of innovation strategies.

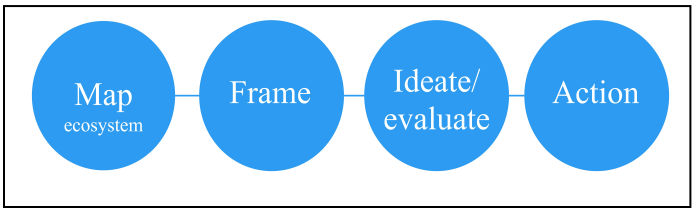


Figure 3. Design+ Technology Gateway, four stage strategic design process (authors own 2020)

The process is supported by a design strategy toolkit. The toolkit combines business tools, design tools and tools newly developed by the team which responded to an identified business need. Examples of these are user journey maps, role reviews, Ansoff matrix, leadership wheel, strategic roadmap. The tools are generally A3 infographic prompts to ensure the thinking is being challenged at each step of the way and that all things are being considered (Figure 5). The tools differ from templates in that they are flexibile and can be used in different ways unlike a template which is predefined .

Figure 4. Mixed media approach to data capture (authors own 2021)

Figure 5. Strategic design toolkits (authors own 2021)

4.1. Data collection

The research began with three SMEs and was specifically looking at ways to measure the success factors of a strategic design engagement. The objective was to analyze the outcomes of the engagement which were captured in a strategic roadmap, to better understand the resulting values (Figure 5).

Table 2. SMEs 1-3 Background

SME1 was as a legacy business on the point of handover to the next generation which incorporated three retail outlets. A new strategy was required based on a new vision by the next generation, considering challenges within today's markets but also with a consideration for future proofing and creating a sustainable business model fit for purpose. As part of the future vision and in developing options for a new way of doing things, a specific scanning technology was identified around which there were many uncertainties which required evaluation.
SME2 was a start-up business with a specific offering in the technology sector that was not clearly linked to a specific target market. There was therefore no clear understanding of individual user needs. Their request on engaging with the team was to frame key market sectors and develop a strategy for responding to their needs.
SME3 was an established events-based company who was seeking a new growth strategy. The company required support with ideation and rationale building around diversification. They had also identified an opportunity to leverage the existing company data to improve business intelligence.

The research captured data from each stage of the process through visual mapping on whiteboard walls and digital whiteboards. The background context to the engagement requirements was recorded in Table 2. The key area of data capture was in the culminating strategic roadmaps for each of the SMEs. The strategic roadmaps present the innovation or the new strategy through steps to implementation, resources required, participation scope, timeframes, key decision points and milestones. This provided the research with an insight into a variety of outcomes across the SMEs such as collaborate with new partner, run a pilot study, engage with data analytics, or apply for funding support. In addition, there were noted new approaches to strategy such as the introduction of new strategy meetings within the organization, who would be involved and what tools could be used to support those meetings. The research also recorded any identified new ways of doing things such as introduction of digital whiteboards and visual mapping techniques as an approach to dealing with complex data. Also considered was the introduction of new platforms for creating shared context and knowledge exchange utilizing tools from the toolkit. Material outputs were also recorded such as the digital whiteboards, infographics, the toolkit, the roadmap, and a presentation of the process. In building on the first three SMEs phase of research an additional four SMEs were engaged in a series of strategic design workshops. This brings a total of seven SMEs and a cumulative time of fifty-six hours of workshops from which data was collected. The objective of this additional research was to begin assessing the nuanced differences in approach to applying strategic design to a broad business strategy and to an operational level strategy. The data collected focused on the tools used when applying strategic design to both the broad business model and on an operational level. The tools represent the area of focus pertinent to the objective and were based on the Design+ strategic design toolkit which includes almost 40 tools.

4.2. Analysis

Analysis of the first three SMEs data collected was carried out through visual mapping. Visual mapping enables grouping of commonalities and differences and themes to emerge as visual ques. The workshops act as a process of co design therefore the roadmaps are developed and validated with the business owners. The actions identified in the roadmaps were listed and grouped into tacit and explicit responses and then further categorized into physical material outputs, tangible outcomes, and potential tacit impacts as presented in Figure 6. It was also possible to consider the transactional elements and the transformational elements of the resulting outcomes. The transactional elements are those which are task related and focused on planning and execution, as opposed to the transformational elements which influence change of the existing culture of the organization.

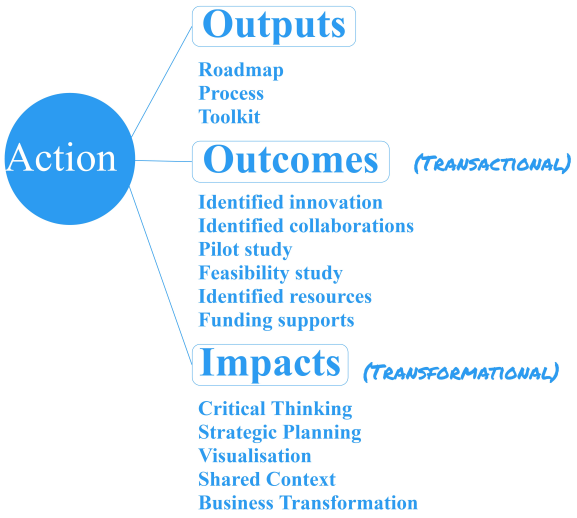


Figure 6. Analysis of roadmap conclusions as outputs, outcomes, and impacts (authors own 2021)

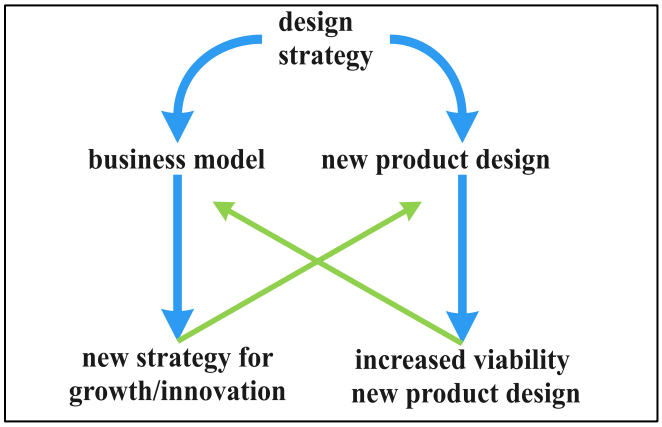


Figure 7. Garden gate model of flow between business and product strategies (authors own 2021)

The analysis of the additional four SMEs which was the extended research, was also carried out through visual mapping of all of the tools used throughout the process with the aim of detecting the nuanced differences in approach to business level and operational strategy development (Figure 9). Each of the tools highlighted in the state of play and action stages represent the questioning and data gathering activity. As noted in the literature review, strategic design is commonly associated with new product development. Considering the models from this context, and analyzing both approaches, it became apparent that the business strategy may result in a new product concept which then must undertake the operational strategy route to develop the product. Similarly, if a new product concept is developed, the new business model must go through the business strategy development, represented in the 'Garden Gate' model (Figure 7).

This is a potentially ongoing cycle of adaptability and innovation which reflects the flow between business and operations. By considering this in the broader context of business and operations, a model was created, the triple transverse model, (Figure 8) which provides a contextual representation of organizational strategy. In building on this, the activities that happen when applying the strategic design process across business and operational strategies were mapped. The author developed this with the addition of the knowledge creation and transfer points that occur between business and operations to produce the triple transverse framework (Figure 9). This is a key framework for the research and for our understanding of transformational impacts. It provides a contextual referece of how strategic design is applied and where the knowledge is acquired, distributed, and leveraged for ongoing adaptability to change and innovation within the organization.

Figure 8. Triple transverse model of organizational strategy (authors own 2021)

5. Findings

The aim of the research is to contribute to the emergent field of strategic design as an approach to creating transformational impacts in organizations. The findings show that firstly, there are nuanced differences in the application of strategic design to develop business strategy or to develop an operational strategy. Capturing this through visual mapping and as the triple transverse framework will inform designers and managers of the approach and the specific tools to ensure all opportunities are presented when innovating. Secondly, innovation by its nature has unknown outcomes, having an indicative guide to outputs, potential outcomes and typical impacts will assist in communicating the type of results to expect from a strategic design engagement. Thirdly, there are specific points in the process which, if the appropriate structures and supports are in place, will produce an ongoing cycle of innovation and adaptability, a transformational impact for an organization.

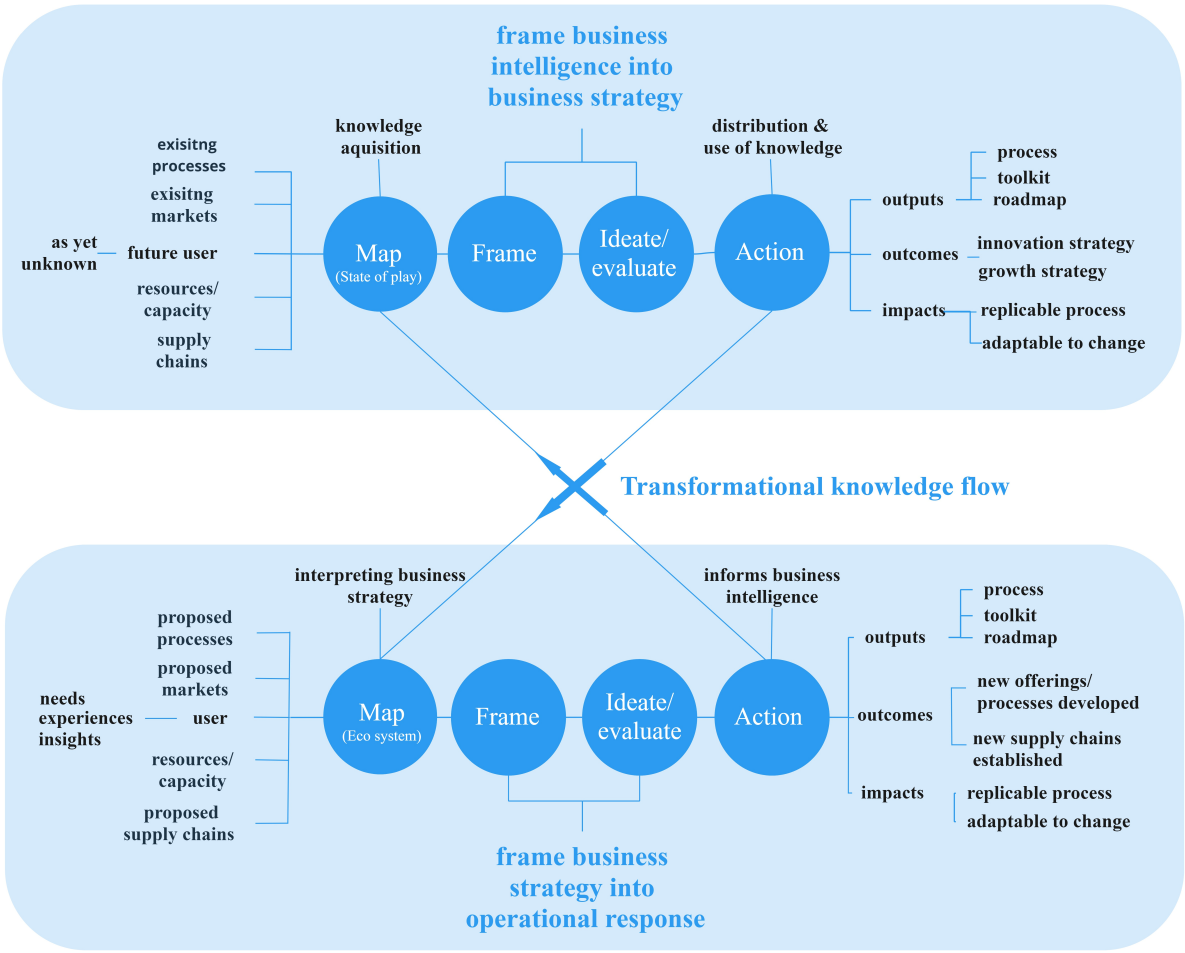


Figure 9. Triple transverse framework of strategic design applied to organizational strategy (authors own 2022)

The categorization of the results into outputs, outcomes and impacts was based on the physical material results classed as outputs, the tangible new direction as outcomes, and the tacit elements as impacts. Throughout the process SME1 framed the human factors and needs to assist in identifying the appropriate scanning technology. They were then introduced to the technical and financial resources within the research network to pursue a feasibility study which included a pilot of the technology. They also formulated a business strategy based on a framed vision for innovation and adaptability. The company introduced strategic meetings between management teams and identified the specific tools to use which could provide a *neutral* voice within family business management. SME2 identified a new application of a technology which was the most suitable in aligning with owner/managers skill sets. As part of the evaluation around viability, a growth plan was formulated and the steps to implementation added to the strategic roadmap. The company vision was framed and the strategy to achieve it was captured. A collaboration was initiated within in the newly identified market sector bringing together the technology and finance sector for a unique new offering. SME3 mapped out the user and business needs for improved interaction and engagement. A CRM system could then be identified appropriate to those needs. Improved business intelligence and decision making ability was framed into usable dashboards for the new system. The company introduced strategy sessions to continue working towards their operational goals with the use of the toolkit. All three SMEs who participated shared similar material outputs form the engagement such as; a four-stage process, toolkit, infographics, and roadmap. The outcomes varied pertaining to the nature of the innovation, but included results such as; an identified innovation, pilot study, feasibility study, identified collaboration, or new process introduced. The impacts are more difficult to communicate as they are of a tacit nature. However, based on the data collection and analysis we can gain insights to key elements such as: a developed understanding of human drivers and aspirations, the introduction of critical thinking, the use of visualization techniques, and the development of shared context.

In building on this, the triple transverse framework highlights that to measure the success factors of the transformational elements of strategic design engagements, we need to measure the structures within the organization that engage in strategy. The three approaches to strategy are identified as business strategy, operational strategy, and transformational strategy. The transformational strategy depends on the flow of knowledge exchange between management and operations. The research aimed to map the strategic design approach for business strategy development and operational strategy development. The objective was to identify the nuances of implementation from each other

and from traditional business modelling. The findings from mapping the two processes across four SMEs, highlight that within a design strategy approach to operations, the eco system is mapped around the end user needs. This is to produce insights that can be translated into a meaningful product/service response. In a design strategy approach to business strategy development, the future user is unknown. If the process is restricted to existing users, it will limit the opportunities for transformational big system changes which may result in new ways of doing things, new market arenas or new product arenas to develop into. Therefore, the business eco system is mapped to identify areas of opportunity. The mapping includes the interplay between the business model, the existing user groups, the broad stakeholder needs, environments, processes, and experiences. This is a key insight to the nuanced difference in approaches. However, a cycle of interplay between business and operations is apparent. The strategic design approach to developing a business strategy reaches an 'action' stage, and a roadmap is produced. The roadmap provides the steps of implementation which forms the basis to develop the operational strategy. Equally when the operational strategy is developed and reaches the 'action' stage, the information and experiences of implementation are captured in the 'state of play' tools which informs the ongoing business strategy development. This is the ongoing cycle of knowledge within the organisation. The tools used by strategic designers support this process in a holistic way ensuring the capture of both tacit and explicit knowledge and the use of visual prompts and methods to easily exchange information and support the flow of knowledge. Critical thinking, for example, is supported using multiple dialectic tools which look at different perspectives of the same problem. Developing a shared context is supported with tools such as the roadmap tool which provides visual representation of complex data. The roadmap will mark the steps to implementation, participation scope, timelines, milestones, key decision points, resources required etc. This provides a shared context for all stakeholders when undertaking a new direction. It also enables quick changes to be made as everyone can see where the change is required and why and who it may impact. This cycle of innovation between business and operational strategy provides an adaptability and sustainability of a business model which is innovative, competitive, and responsive to change. This is the ongoing cycle of knowledge within the organization.

6. Conclusion

International strategies strive to promote sustainable business practices which are innovative and adaptive to change. The literature review highlights that understanding and leveraging the flow of knowledge within an organisation has a transformational impact. The gap in the knowledge is in understanding the structures that support this transformational dynamic. The research captured and analysed the specific nuances in the application of strategic design as an approach to developing innovation strategies, applied at both management and operational level. The findings have been developed into a triple transverse framework which demonstrates the application of the process and tools at both management and operational level. The framework also highlights where knowledge is acquired and disseminated forming an ongoing cycle of innovation and adaptability to change. This framework can be utilised in organisations globally, with design facilitation, across all sectors such as business, policy, education, and social applications. The framework contributes to the emergent strategic design sector providing a contextual reference for the application of strategic design for transformational impacts across an organization. The framework is currently being developed into a strategic design for innovation accredited training programme with an accompanying toolkit to supports the development of the transformational strategy of organizations.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] L. Whelan, "Measuring the Success Factors of Strategic Design Implementation" IEEE Technology and Engineering Management Conference Europe 2021 https://doi.org/10.1109/TEMSCON-EUR52034.2021.9488632.

[2] T. Brown, "Change by Design" Journal of Product Innovation Managemnt 2019 https://doi.org/10.1111/j.1540-5885.2011.00806.x

[3] R. Grant, Knowledge Management Theories. In: Augier M., Teece D.J. (eds) The Palgrave Encyclopedia of Strategic Management. Palgrave Macmillan, London.2018 https://doi.org/10.1057/978-1-137-00772-8_492

[4] GIPC Global Innovation Policy Centre , Chamber of Commerce. Washington DC 2021 URL: https://www.theglobalipcenter.com/

[4] D. Runde, "The OECD Faces a Decision Point in 2021" Centre for Strategic & International Studies. Washington DC 2020 URL: https://www.csis.org/analysis/oecd-faces-decision-point-2021

[6] EGFSN. Winning by Design. Dublin: Expert Group on Future Skills Needs. National Skills Council, Dublin 2017 URL: https://enterprise.gov.ie/en/Publications/Winning-by-Design.html

[7] J. Goddard, Connecting Universities to Regional Growth. Brussels: European Commission. 2011 doi:10.1093/cje/bes005 http://ec.europa.eu/regional_policy/sources/docgener/presenta/universities2011/universities2011_en.pdf

[8] R. Verganti, Design Driven Innovation. The Oxford Handbook of Innovation Management 2014 DOI: 10.1093/oxfordhb/9780199694945.013.006

[9] G. Calabretta, Strategic Design. Amsterdam: BIS Publishers 2018 ISBN 978 90 6369 445 6

[10] E. Wenger, Cultivating Communities of Practice. Harvard Business Review 2002. ISBN 1-57851-330-8

[11 G. Walsham, G., Interpreting Information Systems in Organization. Wiley New York. 1994 https://doi.org/10.1177/017084069401500614

[12] M. Henry. The Rise and Fall of Strategic Planning. New York 1994 : Toronto :Free Press ; Maxwell Macmillan Canada

[13] M. Pankina, "Axiology and Praxeology of Design Thinking. Knowledge E. Ural Federal University Russia 2020 DOI: 10.18502/kss.v4i11.7559

[14] C. Frayling, "Research in Art and Design. London: Royal College of Art 1993. ISBN 1-874175-55-1

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 5, 213-221 (2021)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Electrification of a Bus Line in Savona Considering Depot and Opportunity Charging

Michela Longo[*,1], Carola Leone[1], Luise Lorenz[1], Andrea Strada[1], Wahiba Yaici[2]

[1]*Politecnico di Milano, Energy Department, Milan, 20056, Italy*

[2]*CanmetENERGY Research Centre - Natural Resources Canada, Ottawa, Ontario, Canada*

ARTICLE INFO

Article history:
Received: 06 August, 2021
Accepted: 24 September, 2021
Online: 30 September, 2021

Keywords:
Electrification
Publica Transport
E-bus

ABSTRACT

A transition towards electrification of the public transport sector is ongoing in many cities around the world, as a response to global warming and pollution. However, the question is whether the current state of technology is already sufficient to replace the conventional buses with electric ones and if the existing charging facilities are appropriate to ensure the smooth operation of the buses. Therefore, this work aims to verify the technical feasibility of the electrification of an existing urban line. The purpose is achieved by evaluating a case study on a public transport bus line in the city of Savona, Italy. The average energy consumption of an electric bus operating in the considered line path is estimated in order to investigate the possible locations and sizes of the charging systems to install. The results show that the correct service operation of the electric buses can be achieved by installing one opportunity charger of at least 300 kW in one of the terminals or by installing three 43 kW charging ports in the depot.

1. Introduction

With the Paris Agreement in 2015, the members of the UNFCCC (United Nations Framework Convention on Climate Change) agreed to undertake ambitious efforts to keep the global average temperature rise well below 2°C above pre-industrial levels and to strive for an increase of less than 1.5°C within this century [1]. To reach this ambitious long-term goal, the nations must take action and reduce their overall carbon dioxide (CO_2) emissions drastically. Regarding the European Union (EU) total emissions, the transportation sector makes a share of 21%, whereas road buses in combination with other heavy-duty vehicles contribute to 6% of the total emitted CO_2 [2]. Following the electrification of the public transport sector could have a significant role in the transition towards more sustainable mobility. Evaluating the trend in the last years is quite clear how electric buses continue to witness a dynamic development with more than 460 000 vehicles on the world's road in 2018, almost 100 000 more than in 2017 [3]. The electrification of public transport is not just an important step towards a free-CO_2 transport sector, but it also entails positive effects on the living conditions of the inhabitants of the city. Indeed, it enhances air quality as well as noise pollution. [4]. Therefore, the transformation of the public transport sector is not just an important step towards sustainability but also an important sign to the future generation as a livable city of the future.

This electric transition however is bringing, opportunities as well as challenges. As a matter of fact, this technology, in addition to having very high capital costs, also required the installation of a proper design charging infrastructure [5]. In this global context, this paper proposes the electrification of a currently operated bus line in Savona, a small city in the northern part of Italy. More in detail, the different possible charging infrastructures that can be used to correctly electrified the line are evaluated. To achieve this scope, first of all, the energy consumed by an electric bus running on the considered line must be carried out. In the literature, the estimation of an electric bus energy consumption along a route is typically obtained considering standard driving cycles, for instance in [6] the energy consumption of city transit electric buses is computed using four international driving cycles. However, these general driving cycles do not take into account all the statistical uncertainties of the studied bus route. In other works, driving cycles are developed from real-world data, as reported in [7] presents a freeway driving cycle developed based on the traffic information in California. In the present work, instead, the energy consumption is found by simulating the driving cycle of an electric bus on two different sections of the considered line. The test sections are chosen to represent the entire characteristics of the road structure in the overall line. Therefore, this method allows us to develop a very precise driving cycle for the studied route, which considers also factors such as road slope and signs.

This paper is organized as follows. In section 2 the methodology used to estimate the consumption of electric buses along the considered lines is proposed. The methodology is applied

[*]Corresponding Author: Michela Longo, via La Masa 34, 20156, Milan (Italy), michela.longo@polimi.it

to the case study in section 3 and the energy consumption of the line is estimated based on the previous considerations. In section 4 the possible charging systems solutions are discussed. Finally, in section 5 the conclusions are presented.

2. Methodology to assess the energy consumption

The overall energy consumption is determined by different factors. Not only the technical characteristics of the bus itself define the consumption of the bus, but also external factors and the characteristic of the route have a huge influence [8]. The following section describes the technical properties of the chosen bus, as well as the assumptions and equations used to calculate the overall energy consumption.

2.1. Test conditions

To get a proper assessment of the e-bus consumption, it is necessary to define the operating conditions assume for this study, such as vehicle occupancy level, service constraints, and climatic conditions.

1. Occupancy rate: the e-bus is assumed to be ballasted at 2/3 of its payload which means it is transporting about 60 passengers (n_{pass}). Therefore, the overall mass of the loaded vehicle (m) is obtained in (1), by adding to the mass of the vehicle (m_v), the mass due to the presence of passengers, assuming an average weight of 70 kg for each passenger (m_{pass}).

$$m = m_v + n_{pass} \cdot m_{pass} \quad (1)$$

2. The external temperature is assumed to be equal to 5°C (average minimum temperature in Savona). This choice has been made to consider the worst-case scenario in terms of auxiliary power consumption, as a matter of fact, it has been widely proved in the literature that the highest secondary is achieved by the heating system.
3. The actual existing service provided by TPL (Trasporto Pubblico Locale) for Line 6 cannot be degraded by the electrification of the line.

2.2. Energy consumption computation model

In the following the model and the equations necessary to evaluate the speed, power, and energy consumption profiles of the e-bus over the chosen route are described. The theoretic model used is based on the motion equation reported in (2) [9].

$$F - R = m_e \cdot a \quad (2)$$

where F is the resultant of active forces, we will refer to this quantity as tractive effort, R is the resultant of passive forces also called resistances, m_e is the equivalent mass and a is the vehicle acceleration.

Through (3) the equivalent mass of the e-bus is calculated. Therefore, the equivalent mass is the gross mass of the vehicle m augmented by a factor β which must weigh up the inertial momentum of all the rotating masses inside the vehicle such as wheels, shafts, and axles [10].

$$m_e = m \cdot (1 + \beta) \quad (3)$$

As regards the total resistance to the bus motion, in the present analysis three terms have been taken into account: the rolling resistance, the aerodynamic resistance, and the grade resistance shown respectively in (4), (5), and (6).

The rolling resistance is due to the continuous deformation of the tire/wheel during its rotation and is computed through (4).

$$R_1 = \left[0.005 + \left(\frac{1}{p}\right)\left(0.01 + 0.0095\left(\frac{v_1}{100}\right)^2\right)\right] mg \quad (4)$$

where p is the pressure of the wheels in bars, v_1 is the speed in m/s, m is the vehicle mass in kg and, g is the acceleration of gravity in m/s^2.

The aerodynamic resistance represented in (5), also called wind or air resistance, depends on the square of the vehicle speed (v), the density of the medium in which the vehicle is moving which in this case is air (ρ_{air}), on the frontal area of the vehicle (A) and its shape, considered through a drag coefficient (C_x). In the calculations, for reasons of simplicity, the air density is assumed constant and the wind speed equal to zero.

$$R_2 = \frac{1}{2}\,\rho_{air}\,A\,C_x\,v_2^2 \quad (5)$$

Finally, the grade resistance appears when the vehicle is traveling over a surface with a positive or negative slope α and it is computed by (6).

$$R_i = mg\,sin\,\alpha \quad (6)$$

Regarding the analysis of the power consumption of the bus, the formula varies depending on the sign of the tractive effort. If a positive tractive effort is applied by the motors to the wheels, it means that the vehicle is accelerating, on the contrary, a negative tractive effort corresponds to a braking condition; finally, a nil tractive effort implies that the motor is turned off (or decoupled from the wheels) and the vehicle naturally decelerated because of the external resistances. Equations (7) and (8) describe the interrelations that apply in each case in order to compute the electric power P_e provided by the motor.

$$P_e = F_T \frac{v}{\eta_T} + P_{aux} \qquad for\ F_T > 0 \quad (7)$$

$$P_e = F_T\,v\,\eta_B\,\eta_{RBS} + P_{aux} \qquad for\ F_T < 0 \quad (8)$$

Where F_T is the value of the tractive effort, v is the speed, η_T represents the efficiency in traction, P_{aux} is the value of the power absorbed by the auxiliaries, η_B is the efficiency in braking and η_{RBS} represents the amount of recoverable energy. Since during braking the value of the tractive effort is negative, consequently the value of the power is also negative and this means that there is a recovery of energy in that instant of time.

The value of the current provided by the supply system (in this case the battery) is calculated in (3), where V is the supply battery voltage.

$$I = \frac{P_e}{V} \quad (9)$$

Lastly, the cumulative energy required by the bus is obtained by taking the integral of the electric power over time. However, since in this study, the procedure is based on discrete-time values with a time interval Δt of 1 s, for the energy computation a backward integration is used and reported in (10).

$$E_n = P_{e-n}\frac{\Delta t}{3600} + E_{n-1} \quad (10)$$

where n is the discretization index and P_{e-n} indicates the electric power exchanged at the n time instant.

2.3. Characteristics of electric bus

The chosen electric bus model, for the analysis, is the eCitaro developed by Mercedes-Benz. This model is a 12 m long bus with a mass of 15800 kg. It is completely electric and equipped with an onboard lithium-ion battery composed of 15 modules, each one made of 12 battery cells. The nominal maximum power this bus model can provide is 250 kW, while the supply battery voltage is 400 V. The traction battery must provide the energy also to supply the consumption of the auxiliary service, i.e., rear and front lights, climatization, and radio [11]. The maximum amount of recoverable energy through the regenerative braking system is set equal to 65 %, while the efficiency in traction and braking are considered both equal to 80% in order not to have unbalanced distributions of power between the acceleration mode and the deceleration one. The mass of the vehicle, with the assumed number of passengers declared previously, is 20 t, thus leading to a value for the equivalent mass equal to 23.2 t. For safety and comfort reasons, the maximum acceleration of the bus is limited to 0.8 m/s^2 while for braking deceleration the constraint is less restrictive, and its value can reach values up to 1.5 m/s^2. Given the values of the equivalent mass and maximum acceleration, it is possible to compute the maximum tractive effort at the starting phase by (11), which is equal to 18.56 kN.

$$F_T^{max} = m_e \cdot a_{max} \quad (11)$$

Lastly, the obtained maximum tractive effort must be compared with the maximum adherence force, which is a limit of the tractive effort which must not be overcome otherwise slippage occurs. However, since the speed of the bus is rather small because it is limited to 50 km/h by the Italian laws about speed limits in the urban environment, and because the value of adhesion coefficient is high (equal to 0.6) due to the contact between the tires and the street, the adhesion limit value is never reached and so the bus will never work in slip conditions.

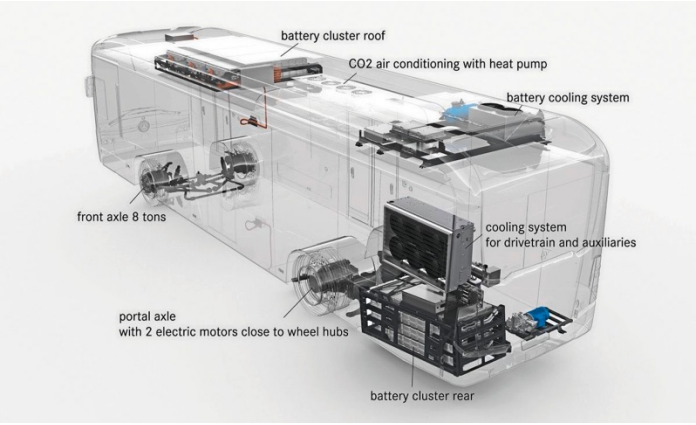


Figure 1: e-Citaro scheme [11].

The eCitaro manufactured by Mercedes-Benz has been considered in this study, since during the last two years, this model has taken over the market of electric buses and it is common in different European countries like Germany, Luxembourg, Sweden, and Norway [12]. Moreover, it has the most similar characteristics to the conventional bus currently employed by TPL in the selected line. Indeed, the buses currently working in Savona for the public transport services are the Citaro C2 models, produced by Mercedes-Benz too. They are gasoline buses having an internal combustion engine able to provide a nominal power of 220 kW.

The main advantages of choosing an electric bus fleet instead of a gasoline one are the reductions in pollution and fuel costs. Furthermore, the electric motors are less loud than the internal combustion engine and if well-displayed, the recharging process for an electric bus is more efficient and safer than the gasoline one in terms of reliability. The feature of regenerative braking must not be forgotten as it represents a great advantage compared to the conventional buses which are not able to recover energy braking. All these advantages combined confirm that if a modernization of the bus lines is needed, one better option will be to replace the gasoline buses with electric ones.

All the parameters value necessary for the simulation are reported in Table I. Using an Excel Spreadsheet, the overall electric energy consumed by the vehicle along the chosen sections was estimated.

Table 1: Simulation values

l	Bus length	12 m
m_v	Mass vehicle	15 800 kg
n_{pass}^{max}	Maximum number of passengers	85
m_{pass}	Average mass of passengers	70 kg
a_{br}^{max}	Maximum comfortable deceleration	-1.5 m/s^2
a_{max}	Maximum comfortable acceleration	0.8 m/s^2
P_e^{max}	Maximum motor power	230 kW
V	Battery voltage	400 V
η_{RBS}	Percentage of recoverable energy	65%
β	Equivalent mass factor	0.16
η_B	Braking efficiency	80%
η_T	Traction efficiency	80%
F_T^{max}	Maximum tractive effort	18.56 kN
P_{aux}	Power absorbed by auxiliaries	22 kW
C_x	Shape coefficient	0.6
m	Overall mass of the vehicle (considering 60 passengers)	20 000 kg
A	Frontal area	8.7 m^2
p	Tyre pressure	9 bar

3. Energy consumption estimation

The considered Line 6, depicted in Figure 2, connects the bus stations in Porto Vado to “Via Alessandria” in Savona for an overall trip 7.9 km long. While the first part of the line is passing through the city center, the other and longer part of the line heads to the west following the seafront of Savona. For the energy consumption estimation, two representative and very different sections in terms of ground characteristics have been analyzed. We will refer to the first one (between the yellow pointers) as Section A and to the second one (between the red pointers) as Section B.

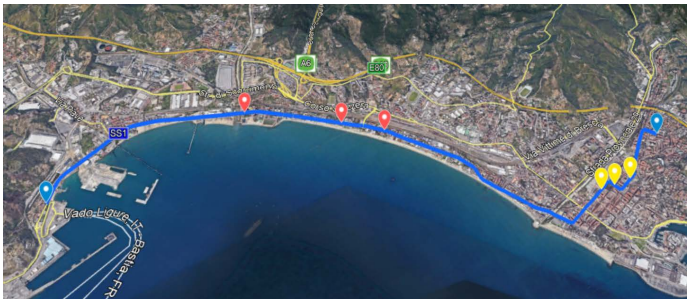


Figure 2: Bus Line 6: From Porto Vado to Via Alessandria in Savona (Italy).

3.1. Analysis of Section A

The section analyzed in the following paragraph is an extract from the downhill part of Line 6 TPL, starting from Via Paolo Boselli and reaching its destination of "Via XX Settembre", with an intermediate stop in "Piazza del Popolo" after 300 m. In the 700 m long section, some crossings and turnings are present. Therefore, to make the simulation more realistic and consequently much more consistent and reliable, some braking sections are considered in correspondence of pedestrians crossing and traffic lights. For instance, as can be seen from the speed profile reported in Figure 4, 3 deceleration and 2 complete stops are simulated along the section. The first deceleration is performed after 140 m from the beginning, and it is due to the fact that the bus must decrease its speed to perform a right turn. Then it accelerates until it reaches a speed of about 40 km/h but at that moment another deceleration phase starts after 270 m because the bus must stop at the intermediate station of "Piazza del Popolo". The stopping time at this station has been set equal to 20 sec. After few seconds from the restart, the bus must stop again at a red traffic light. This second stopping time is set equal to 5 seconds. Finally, in the last part of Section A, the driver can reach the maximum speed of 50 km/h and keep it for a few seconds before arriving at the final stop of this section.

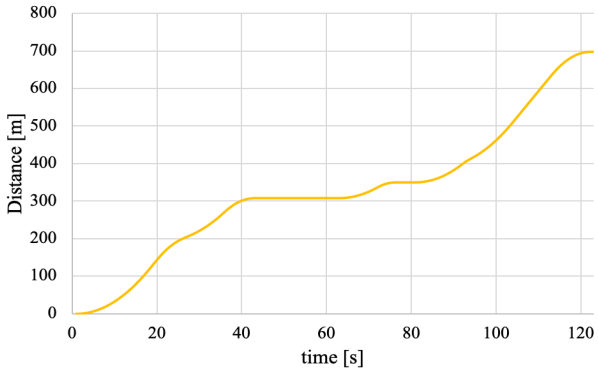


Figure 3: Space Profile of Section A

In Figure 3 the space profile of the section is shown, and we can see that the total time spent to complete the path is equal to 123 seconds.

In Figure 5 the evolution of the electric power as a function of time, from which the electric energy consumption is derived. As expected, in the first part electric power increases up to its maximum value since at this moment the motor must provide the highest tractive effort to let the vehicle reach the speed limit of 50 km/h. Instead, negative values of electric power indicate that at that moment the regenerative braking phase is ongoing.

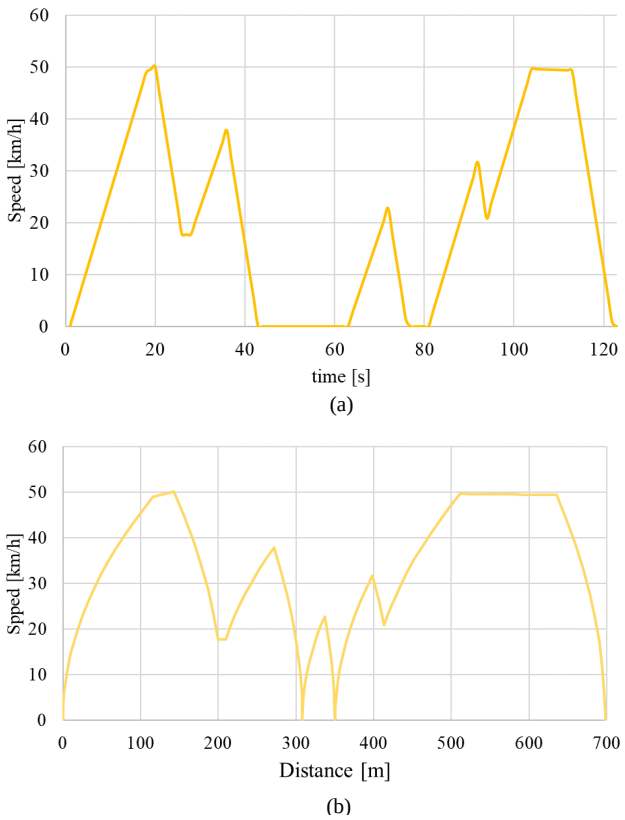


Figure 4: Speed Profile of Section A a) speed vs time and b) speed vs distance.

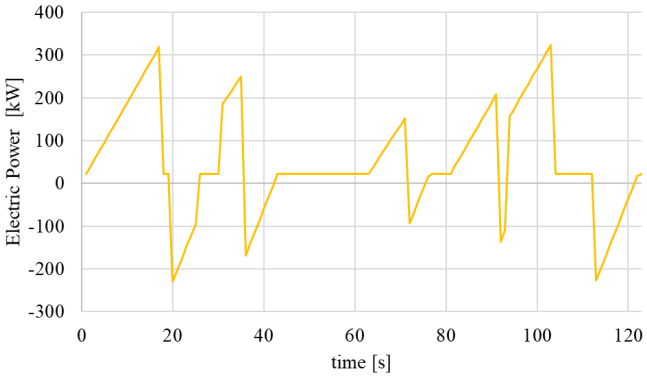


Figure 5: Electric Power profile of Section A

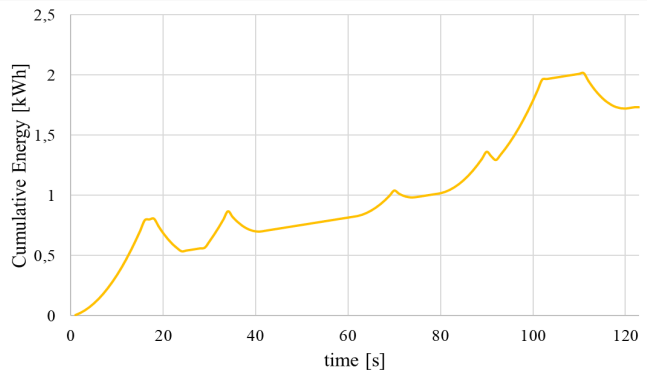


Figure 6: Cumulative energy consumption in Section A.

Finally, the total energy required by the bus is reported in Figure 6 and it can be concluded that to cover the 700 m related to

Section A, 1.73 kWh are necessary, which correspond to consumption of about 2.47 kWh per km, in line with that find in [13].

3.2. Analysis of Section B

The second considered section of Line 6 TPL goes from the bus stop Villa Zanelli to Via Nizza (Camping) until the final destination Via Nizza 126. This path differs from section A primarily because of its location along the seaside, on Strada Statale 1 Via Aurelia (SS 1), instead of the city center. Consequently, Section B is flatter and characterized by a straight path on which, however, the speed limit is 50 km/h, since it is always an urban road [14]. The whole Section B has a total length of 1,450 m and the middle bus station Via Nizza (Camping) is located after 270 m starting from the first bus station Villa Zanelli.

The whole Section B lays on a priority road with no traffic lights intersections but with two rotaries and several crosswalks. Within the model, a deceleration down to 20 km/h is assumed before entering the rotaries that are located after 200 m and 968 m starting from the first bus stop 'Villa Zanelli'. In the whole path of Section B, there are ten crosswalks on which the bus has to decelerate in case of a pedestrian wants to cross the street. However, since the path does not have any curves, the driver of the vehicle has a wide forward view on the road, therefore, in the computation, it is not assumed to stop on every crosswalk but only to slow down to 10 km/h or less. Finally, before the final bus stop, the bus must pass a fork in the road. However, as the bus has the right of way the intersection is not taken further into account.

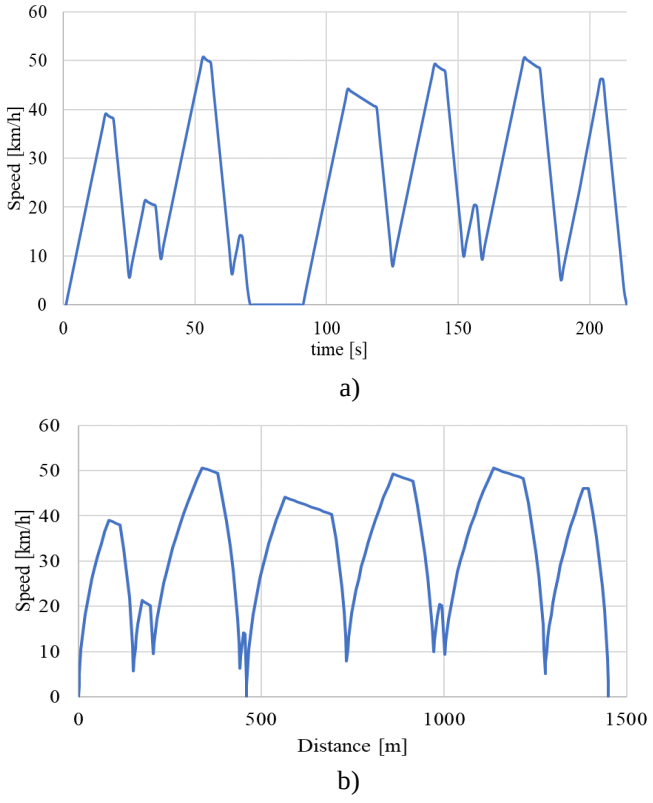


Figure 7: Speed profile in Section B a) speed vs time and b) speed vs distance.

An approximated gradient profile is included in the calculation, even though the slope is very low. The complete profile is generated by Google Earth and then is approximated with a positive value of 0.23% for the first 1300 m and a negative value (which represent a downhill) of 0.67% for the resting 300 m. Finally, the bus is assumed to stop at 'Via Nizza (Camping)' for 20 seconds.

From the results reported from Figure 7 to Figure 9, it is possible to conclude that the bus needs 3 min and 34 seconds to cover this section. In particular, the bus reaches Via Nizza (Camping) after 1 min and 11, and after a 20 second stop, it takes another 2 minutes to run the last part. The speed profile in Figure 7 is characterized by many peaks, which means many changes in velocity caused by the rotaries and crossings the bus is passing by. The maximum speed of 50 km/h is reached three times.

Overall it is visible, that due to the number of crossings and rotaries, the bus does not have the driving distance in between the traffic interactions to accelerate much further than 50 km/h. Regarding the fact that the road is located close to the beach and right next to camping opportunities, it can be assumed that it is a rather touristic area that many walkers pass. Additionally, the fact that there is a cycle path right next to the road clarifies that driving faster than 50 km/h would bring dangers with it.

As Figure 8 shows, the electric power taken by the supply system (the battery) is strongly varying over the path. The fluctuation can be traced back to the situation, whether the bus is in traction (tractive effort > 0) or braking (tractive effort < 0) – as it can be observed in Figure 9. If the bus is in traction mode, the supply system is providing not only the needed power for the wheels but also the power absorbed by the auxiliaries. The bus is always consuming the power needed for its climate system and other extra facilities, even though the tractive effort is zero or the bus is not moving (speed = 0 km/h). The electric power assumes negative values when the bus is braking because in those moments a percentage of energy is sent back from the wheels to the battery thanks to the regenerative braking technology. However, due to technical limitations, there is no regenerative braking when electric vehicles are running at low speed; in this study, this limit is set at 10 km/h.

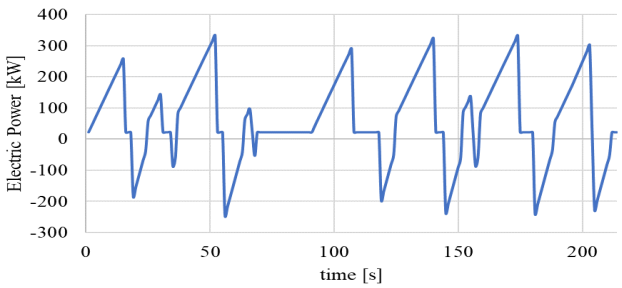


Figure 8 Electric power along Section B.

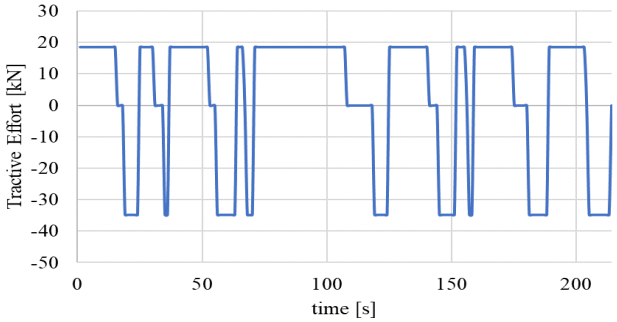


Figure 9: Tractive effort along Section B.

As can be seen from Figure 10, the overall energy consumption along Section B is 3.8 kWh which corresponds to 2.62 kWh per km. Compared to the average energy consumption of 2 kWh/km found in [15], it is a quite higher result which can be justified by the presence of many influencing factors. First of all, as already mentioned, the bus is consuming power, even when it is stopped at the bus station because it has to supply the auxiliaries (22 kW). For instance, the 20 seconds stopping time at the station Via Nizza (Camping) results in power consumption of 0.12 kWh. Additionally, the high number of crossings and rotaries present along the path forces the bus to continuously decelerate and accelerate. The high power required to accelerate leads to high energy consumption. A smoother driving style that reduces the deceleration and acceleration rates could lead to energy savings.

In order to extrapolate the energy consumption of the entire line, a weighted average of the two energy consumptions found for Section A and Section B is computed. Since the line is prevalently located along the seaside rather than in the center of the city, for the computation of the overall consumption Section B weights more (75%) than section A (25%). Therefore, the total energy consumed along the 7.9 km trip between Porto Vado and Via Alessandria is estimated to be about 20.3 kWh. The results coming from this section are summarized in Table.

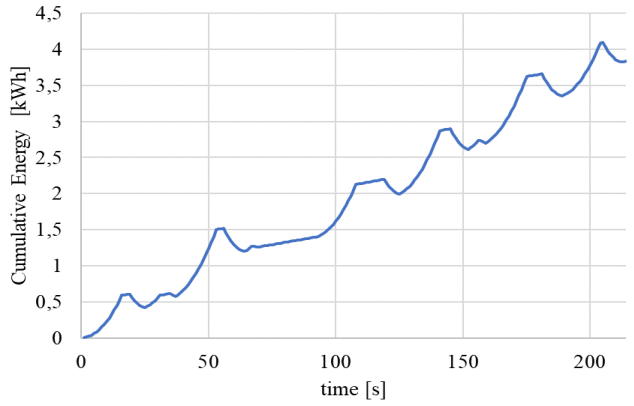


Figure 10: Cumulative energy consumption at section B.

Table 2: Average energy consumption

Section	*Length*	*Request [kWh/km]*	*% overall path*	*Weighted Average [kWh/km]*
A	700	2.47	25%	2.57
B	1 450	2.62	75%	

4. Charging Systems Sizing

An electric bus can be charged with two main charging technologies: depot and opportunity charging. Depot charging technology fills the fleet of electric buses at the depot through plug-in connectors [16]. The charging power for each connector ranges from 40 up to 150 kW. The rated power of this charging system mainly depends on whether AC or DC charging system is installed. If an AC technology is chosen, the onboard charger of the vehicle must be used to operate the AC-DC conversion, and the off-board structure will only include power and communication cables, metering, and protection devices. AC charging systems can provide a maximum power of 43 kW (86 kW if two Type2 plugs are inserted in the vehicle). If higher charging rates are necessary or if the electric buses which compose the fleet are not equipped with the onboard charger, a DC charging system (mode 4) must be used. In this case, the conversion stage is performed offboard the vehicle, inside the charging system. On one side this will result in higher achievable power rates on the other hand this will lead to higher investment costs. In DC charging systems, the power limit is mainly imposed by the cable and the connector; as a matter of fact, the Combo 2 (CCS2) connector can provide a maximum of 200 A without the need for special and expensive liquid cooling systems. To not have a too high impact on the public utility grid, the output current is limited to 150 A, which corresponds to a nominal power of about 100 kW considering the average battery voltage of the actual electric buses on the market. It can occur that to complete and provide a proper service electric buses require additional occasional recharges that take advantage of the halt times at the terminus and/or at the stops. The charging system which provides this type of facility takes the name of opportunity charging. The opportunity charging uses overhead pantographs which can support charging powers up to 600 ÷ 750 kW [17].

In this paragraph, the two charging technologies for the chosen electric buses of Line 6 TPL will be sized. The most important data necessary to display the analysis are the total energy consumption on the entire line, its timetable, and the battery and charging characteristics of the vehicle.

Following the information found on the timetable of Line 6 TPL, it can be understood that to ensure the service 3 buses must be used at the same time. In the following, we will refer to these three buses as Bus 1, Bus 2, and Bus 3. All the buses start and end in the two stations of Via Alessandria and Porto Vado, however, some of them perform small deviations along the day. The total number of one-way trips between Porto Vado and Via Alessandria for these three buses are respectively equal to 32, 30, and 25.

The eCitaro buses are provided with lithium-ion batteries having different levels of battery capacity: 150, 200, and 250 kWh, in fact, they are made of a variable number of modules ranging from 6 up to 10. The standard charging system for eCitaro buses is the Combo2 plug-in system, but there is also the possibility to adopt the pantograph system which is sold as special equipment.

4.1. Depot charging

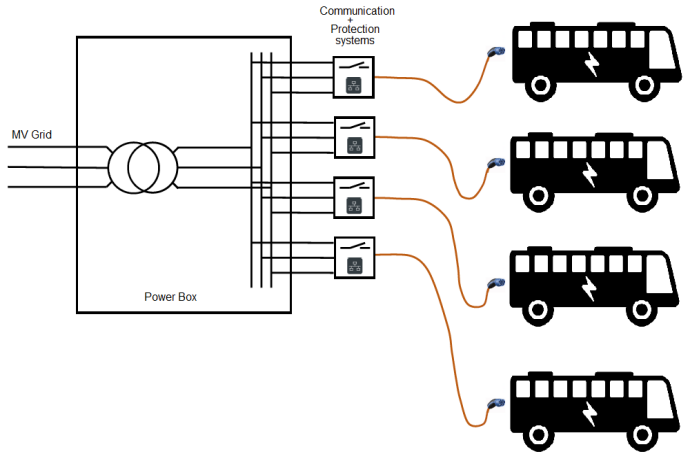
In this section, the charging system that must be installed in the depot in order to assures a correct service of Line 6 is analyzed. An example of the layout that the depot charging infrastructure can assume is shown in Figure 11 b). In Figure 11 a), instead, the energy distribution scheme is sketched.

TPL already rents a depot in a public area installed in Via Valletta San Cristoforo, Savona, approximately located in the midpoint of the trip. Therefore, this depot will be the location of the charging system. Since the depot is located in a public area, some technical issues are to be discussed properly [13]. First, it must be considered that other large consumers may also take energy from the same grid thus limiting the amount of energy that can be used to recharge the fleet. Then, the European Union stated restrictions on the maximum power to be installed in a charging depot and this threshold cannot be overcome for legal reasons. Consequently, in the depot, it will be installed an AC charging system with rated power for each port equal to 43 kW.

To not oversize the number of required charging ports, and hence to not waste money, the schedule of all the electric buses recharging times becomes a crucial feature of this study. In order to size the depot charging, the following assumptions are made:

1. All the electric buses of the fleet made have the same battery capacity which in the first case is set at 250 kWh, then to 200 kWh, and finally to 150 kWh.
2. A safety SoC threshold has been set at 20%, and this amount of energy is necessary for example to return to the depot in case of emergency conditions.
3. All buses leave the depot fully charged (SoC 100%).
4. The buses rely only on the depot charging for the operation.

(a)



(b)

Figure 11: Example of depot charging infrastructure a) layout and b) electric scheme.

CASE 1: The buses are equipped with a 250 kWh battery. Under the before-mentioned assumptions, a single bus is able to do nine one-way trips before reaching a battery percentage level below the threshold of 20%. As a bus can approximately drive four hours before having to drive back to the depot, another bus can use this time for charging its battery. Within that time, the battery level of the bus can reach values up to 80%, considering that the bus arrives at the depot having 20% of battery and the depot has a recharging system of 43 kW with a charging efficiency equal to 88%. So, after the second bus has covered its part of the runs, the bus that was waiting at the depot becomes the one in service while the other heads to the depot to recharge its batteries. By keeping this alternate on during the whole day, two electric buses for each currently employed diesel bus are needed, so the fleet will be composed of six electric buses in total. Three charging ports must be installed in the depot, one for every two electric buses.

CASE 2: The battery capacity is decreased at 200 kWh. In this case, a bus can cover only seven trips without reaching the level of 20% of battery and according to the timetable, this means a service time equal to three hours. By repeating the same calculations as in the previous case, the fleet will be composed of thirteen buses. If instead the daytime charging at the depot is allowed, the number of buses needed will decrease as described. In three hours, the battery of 200 kWh can recharge to a level of 80% by making the same assumption made before for a battery of 250 kWh. So, after the second block of three hours, the two buses overlap and this continuous cycle is repeated throughout all the daytime, leading to the conclusion that again two electric buses for each currently employed diesel bus are needed. Therefore, even using electric buses equipped with batteries having a capacity of 200 kWh, the electric fleet must be composed of six vehicles.

CASE 3: finally, the case of batteries with a capacity of 150 kWh is brought into the analysis. In this case, a single bus can cover five trips with a corresponding amount of time of two hours. Therefore, a bus can recharge up to 70% of its battery at the 43kW charging pole at the depot with an efficiency of 88%. Since the battery capacity is very low, the overlapping of two buses is not fitting for the case, as the iterative decrease of the battery percentage during the service for the two buses leads to an incompletion of the daytime service (two buses are able to cover only 65% of the daytime service). For this reason, three buses are needed, with two buses recharging simultaneously while the other one is providing the service. In this case, the buses can always run into service having a 100% battery capacity, thanks to the fact that having two buses waiting in the depot, the overlapping of one of the two with the exhausted one enables the second bus to further recharge its batteries for supplementary two hours. By extending this observation to the three currently employed buses, the result is an electric fleet composed of nine buses, which means 3 electric buses for each diesel one.

Table 2: Results depot charging

#Case	*Battery Capacity*	*# necessary e-buses*	*# installed 43 kW charging ports*
Case 1	250 kWh	6	3
Case 2	200 kWh	6	3
Case 3	150 kWh	9	3

In conclusion, the battery capacity has a significant impact on the number of buses needed to substitute a conventional bus if the electrification relies only on the depot charging technology.

4.2. Opportunity charging

As an alternative to the traditional depot charging system, opportunity charging is gaining in popularity. The difference between the two charging technologies can be simplified in a way that opportunity charging is taking place during the service at on-street charging stations [18]. Given the higher power level, this charging system can provide, it follows that the charging duration gets significantly shortened [17]. The main components of an opportunity charging system are sketched in Figure 12.

Regarding the timetable of the 3 routes of Line 6 TPL, the vehicles are having regular breaks between five and twenty minutes at the bus station Porto Vado. The parking area in that bus

station is spacious, and hence the buses can take their break without disturbing the traffic. Therefore, the installation of a pantograph for opportunity charging at this place is technically possible. To electrify the routes, without passengers noticing a change in the schedule of Line 6 TPL, the break at Porto Vado is consequently chosen for opportunity charging.

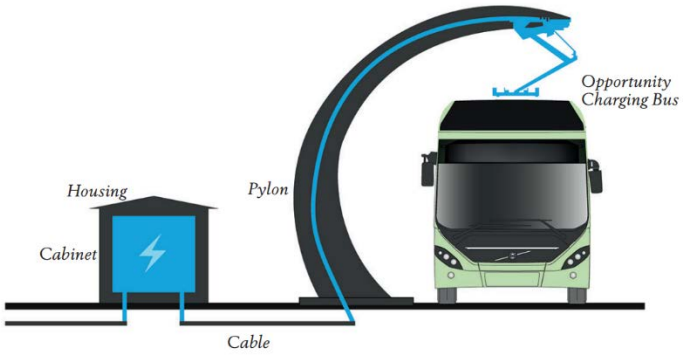


Figure 12: Example of opportunity charging infrastructure.

To determine the necessary battery capacity of the bus as well as the charging power of the opportunity charging station, the timetables of the three routes are evaluated. In particular, the overall timetable is characterized by regular stops at the station Porto Vado for 10 minutes. Between two consecutive stopping times in Porto Vado, therefore the vehicle must perform a round trip for an overall length of 15.8 km (2 times 7.9 km). Based on the energy consumption coming from the previous section, to cover this path the bus consumes 40.6 kWh (2 times 20.3 kWh).

The charging time is determined by the breaking times scheduled through the timetable of Line 6 TPL. This time is assumed to be 2 min shorter than the scheduled one in order to take into account the connection and the disconnection maneuvers to connect to the pantograph. Therefore, the bus charges every time it reaches Porto Vado for an average time of 8 minutes (10 min minus 2 min as a buffer). Therefore, in this time interval, the charging system must provide the bus enough energy to manage the next trip without exceeding the minimum battery SoC limit of 20%. The opportunity charging system can usually have different power rated: 150 kW, 300 kW, 450 kW up to 600 kW [17]. Considering the available charging time in Porto Vado t_{ch}, and the required energy to cover a round trip E_{cons}, in (12) the necessary charging power is computed. The result shows that a pantograph of at least 300kW is needed.

$$P_{ch} = \frac{E_{cons}}{t_{ch}} = \frac{40\ kWh}{8\ min} \approx 300\ kW \qquad (12)$$

Looking at the overall path of the buses operating on line 6 reported in [19], as before mentioned, the three conventional buses must make some small deviations from the usual path. However, all the three buses during their operation stop at Porto Vado station for some minutes, and this interval of time is used to charge the battery through the installed opportunity charger. Figure 13 illustrates the SoC of each of these three buses as a function of the covered distances. Each vertical upwards course represents a charging operation at Porto Vado. As the halt time is most of the time 10 minutes long, the trend of the battery SoC is characterized by a regular pattern due to the same amount of energy supplied in each charging event. While the charging-discharging cycle of the Bus 3 is repetitive, Bus 2 is showing more irregularities. Those can be traced back to the higher number of deviations that this bus has to perform, which will cause more irregularities in the stopping times at Porto Vado. For instance, Bus 2 is facing the absence of stops in Porto Vado in the evening runs, so the longer stop it has after 250 km is used to counterbalance the lack of charging events from 213 km up to 244 km. Bus 1, instead, is facing shorter breaks in the morning hours. To ensure sufficient SoC, Bus 1 must completely charge its battery at Porto Vado before starting its last trip in the evening. Finally, Bus 3, which is starting its journey from Porto Vado in the morning has to arrive at the bus station around 15 minutes earlier in order to completely fill its battery. The extra time needed for charging at the Pantograph at Porto Vado as well as driving there off-schedule causes extra costs for personnel that needs to be considered. For the calculations, it is, in addition, assumed that all the buses during night drive to their depot at Via Valletta San Cristoforo, so the extra energy necessary to cover this path to go to the depot has been considered.

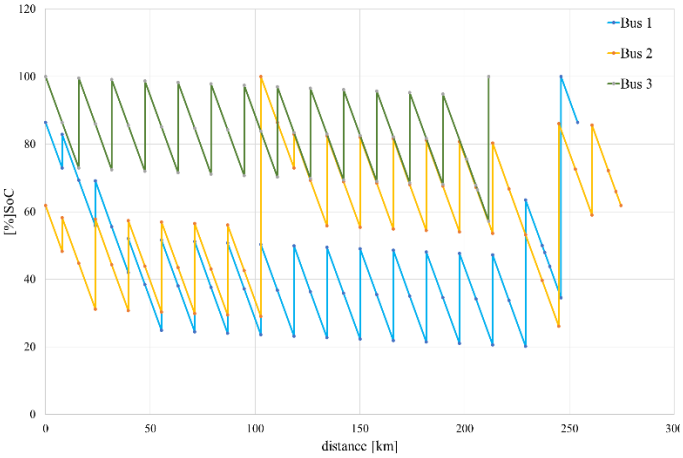


Figure 13: Battery Charge Level of the buses on the investigated routes.

The eCitaro is available in different versions and with different battery capacities. As Figure 13 shown, the basic model (150kWh) of the lithium-ion battery is already sufficient to meet the requirements of Line 6 TPL. As well it can be observed that according to the results the battery SoC is never lower than 20 %. So, the opportunity charging system fulfills the set safety requirements.

Concluding, it is technically achievable to fulfill the timetable using only one opportunity charging station with a capacity of 300kW at Porto Vado in combination with the eCitaro equipped with a standard battery of 150 kWh.

5. Conclusion

This work aims to analyze the electrification of a conventionally operated bus line. Precisely, the focus is on the bus Line 6 operating between the stations Via Alessandria and Porto Vado in Savona, Italy. Therefore, the overall energy consumption on this specific bus route is first computed and then used to evaluate possible depot and opportunity charging infrastructure.

The model results show that it is possible to substitute the conventional buses of Line 6 TPL connecting Porto Vado and Via Alessandria with electric buses and corresponding charging infrastructure. Based on the energy consumption of two characteristically different sections on the assigned Line 6 TPL path, the overall energy consumption of an eCitaro shuttling between Porto Vado and Via Alessandria is upscaled. The

calculations lead to the energy consumption of 20.3 kWh for the 7.9 km long one-way trip between Porto Vado and Via Alessandria. Corresponding the electric bus has an average energy consumption of 2.57 kWh/km.

In order to stick to the present timetable of the chosen routes, there are three possible solutions for electrification. According to the first approach, the conventional buses could be replaced by e-buses using depot charging technology. The depot charging system is composed of three charging ports of 43 kW each. From the results, it can be seen that this scenario would imply that each conventional bus needs to be replaced by at least two electric ones with a battery capacity of 250 kWh. As an alternative to reduce the number of needed vehicles, opportunity charging is introduced as the second possible electrification approach. Based on the halt times available at the terminals in the timetable of Line 6 and the energy consumption along the route, the results prove that it is possible to substitute each conventional bus with one e-bus exclusively relying on opportunity charging systems installed in Porto Vado terminal with a power rate of at least 300 kW. Merging both approaches leads to a third possible solution which could ensure sufficiently charged batteries, increase the safety against failures as well as the efficiency in operation. The analysis of this mixed scenario as well as the cost analysis of the different solutions will be part of further research.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Adoption of the Paris Agreement, Paris, 2015.

[2] Regulation (EU) 2019/631 of the European Parliament and of the Council, 2019.

[3] IEA, Global Ev Outlook 2019 – Analysis - IEA, Paris, 2019.

[4] The Impact of Electric Buses On Urban Life, 2019.

[5] S. Majumder, K. De, P. Kumar, B. Sengupta, P.K. Biswas, "Techno-commercial analysis of sustainable E-bus-based public transit systems: An Indian case study," Renewable and Sustainable Energy Reviews, 144, 111033, 2021, doi:10.1016/J.RSER.2021.111033.

[6] X. Zhao, Y. Ye, J. Ma, P. Shi, H. Chen, "Construction of electric vehicle driving cycle for studying electric vehicle energy consumption and equivalent emissions," Environmental Science and Pollution Research 2020 27:30, 27(30), 37395–37409, 2020, doi:10.1007/S11356-020-09094-4.

[7] H. He, J. Guo, N. Zhou, C. Sun, J. Peng, "Freeway Driving Cycle Construction Based on Real-Time Traffic Information and Global Optimal Energy Management for Plug-In Hybrid Electric Vehicles," Energies 2017, Vol. 10, Page 1796, 10(11), 1796, 2017, doi:10.3390/EN10111796.

[8] X. Ma, R. Miao, X. Wu, X. Liu, "Examining influential factors on the energy consumption of electric and diesel buses: A data-driven analysis of large-scale public transit network in Beijing," Energy, 216, 2021, doi:10.1016/J.ENERGY.2020.119196.

[9] M. Brenna, F. Foiadelli, D. Zaninelli, "Electrical railway transportation systems."

[10] P. Fajri, R. Ahmadi, M. Ferdowsi, "Equivalent vehicle rotational inertia used for electric vehicle test bench dynamic studies," IECON Proceedings (Industrial Electronics Conference), 4115–4120, 2012, doi:10.1109/IECON.2012.6389231.

[11] eCitaro – Mercedes-Benz Buses, Sep. 2021.

[12] Export success for the Mercedes Benz eCitaro with fully electric drive: orders from Luxembourg, Norway and Sweden - Daimler Global Media Site, Sep. 2021.

[13] C. Leone, M. Longo, F. Foiadelli, S. Bracco, G. Piazza, F. Delfino, "Opportunity fast-charging of e-buses: A preliminary study for the city of Savona," 2020 AEIT International Conference of Electrical and Electronic Technologies for Automotive, AEIT AUTOMOTIVE 2020, 2020, doi:10.23919/AEITAUTOMOTIVE50086.2020.9307434.

[14] Going abroad - European Commission, Sep. 2021.

[15] S. Ceccon, M. Longo, R. Mazzoncini, A. Panarese, "Analysis of the Implementation of Full Electric and Hydrogen Hybrid Buses in Two Lines of the City of Milan," in 2020 15th International Conference on Ecological Vehicles and Renewable Energies, EVER 2020, Institute of Electrical and Electronics Engineers Inc., 2020, doi:10.1109/EVER48776.2020.9243074.

[16] S.M. Arif, T.T. Lie, B.C. Seet, S.M. Ahsan, H.A. Khan, "Plug-In Electric Bus Depot Charging with PV and ESS and Their Impact on LV Feeder," Energies **13**(9), 2139, 2020, doi:10.3390/EN13092139.

[17] M.M. Hasan, M. Ranta, M. El Baghdadi, O. Hegazy, "Charging management strategy using ECO-charging for electric bus fleets in cities," 2020 IEEE Vehicle Power and Propulsion Conference, VPPC 2020 - Proceedings, 2020, doi:10.1109/VPPC49601.2020.9330970.

[18] O. Barraza, M. Estrada, "Battery Electric Bus Network: Efficient Design and Cost Comparison of Different Powertrains," Sustainability **13**(9), 4745, 2021, doi:10.3390/SU13094745.

[19] LINEA 6 - LINEA 6/ Timetable, Sep. 2021.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 5, 291-296 (2021)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Ensemble Learning of Deep URL Features based on Convolutional Neural Network for Phishing Attack Detection

Seok-Jun Bu[*,1], Hae-Jung Kim[2]

[1]*Deptartment of Computer Science, Yonsei University, Seoul 03722, Korea*

[2]*Department of Computer Science, Kyungil University, Daegu 38428, Korea*

ARTICLE INFO

Article history:
Received: 20 July, 2021
Accepted: 10 October, 2021
Online: 14 October, 2021

Keywords:
Phishing detection
Deep learning
Ensemble learning
Convolutional neural network
Recurrent neural network

ABSTRACT

The deep learning-based URL classification approach using massive observations has been verified especially in the field of phishing attack detection. Various improvements have been achieved through the modeling of character and word sequence of URL based on convolutional and recurrent neural networks, and it has been proven that an ensemble approach of each model has the best performance. However, existing ensemble methods have limitations in effectively fusing the nonlinear correlation between heterogeneous features extracted from characters and the sequence of sub-domains. In this paper, we propose a convolutional network-based ensemble learning approach to systematically fuse syntactic and semantic features for phishing URL detection. By learning the weights that integrating the heterogeneous features extracted from the URL, an ensemble rule that guarantees the best performance was obtained. A total of 45,000 benign URLs and 15,000 phishing URLs were collected and 10-fold cross-validation was conducted for quantitative validation. The obtained classification accuracy of 0.9804 indicates that the proposed method outperforms the existing machine learning algorithms and provides plausible solution for phishing URL detection. We demonstrated the superiority of the proposed method by receiver-operating characteristic (ROC) curve analysis and the case analysis and confirmed that the accuracy improved by 1.93% compared to the latest deep model.

1. Introduction

Network security based on information technology for protecting personal information and system resources from various types of threats may be defined through policies and methods. Various methods for network administrator have been developed to protect networks and cyber assets including detection mechanism against active attacks [1]. However, few studies have been conducted to analyze the characteristics of phishing attacks, which steal entire input information from users. Phishing attack in its broadest sense can be defined as a scalable act of deception whereby impersonation is used by an attacker to obtain the information from an individual [2]. Considering that the most common form of online phishing attack is malicious hyperlinks embedded in messages, the recent technological trend in which personal connections are reinforced due to the explosive growth of social media services is particularly vulnerable [3].

Existing security systems primarily conduct rule-based detection mechanism using phishing databases to identify malicious URLs [4]. However, phishing URLs based on web applications have zero-day exploit characteristics that frequently involve novel attack instances, as URLs can be generated very conveniently in such applications. For this reason, phishing URLs hardly detected by predefined databases or simple detection rules [2, 5, 6].

Meanwhile, previous study based on ensemble of the convolutional neural network (CNN) and recurrent neural network (RNN) for the modeling the character and word-level features found that classification of malicious URLs was improved [7,8].

In Figure 1, we visualize the phishing URLs into feature space generated by the t-SNE dimension reduction method. Blue and red dots represent normal URLs and phishing URL instances, respectively. The Euclidean distance was determined based on the similarity of character combinations constituting the URL, and a cluster of short and regular URLs was mainly formed at the

*Corresponding Author: Seok-Jun Bu, Yonsei University, sjbuhan@yonsei.ac.kr

bottom. On the other hand, in the center, instances where it is difficult to distinguish between normal and phishing URLs due to subdomains are intricately confused.

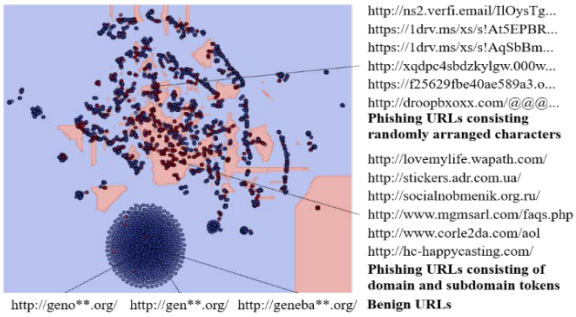


Figure 1: The feature space of phishing URLs and the necessity of ensemble learning

Phishing URL features can be distinguished into a syntactic feature consisting of a sequence of randomly arranged characters and a semantic feature consisting of a sequence of domain and subdomain words. In the existing deep learning-based ensemble approach, a simple rule-based ensemble that averages the output of syntactic-semantic convolution and recurrent networks at log-scale was applied, but it showed the limitation to effectively model the complex nonlinear correlation of features and resulted the degradation of accuracy and recall.

Taken together, we propose an ensemble learning network based on CNN that can systematically utilize the syntactics and semantics of URL features using CNN and RNN. The proposed ensemble network is a deep learning algorithm that can extract effective features for phishing URL classification using filter operations that can be trained using data. The joint learning of ensemble rule based on deep representations of URLs provides plausible solution for phishing detection. We collected a total of 45,000 benign URLs and 15,000 phishing URLs and the proposed method was validated through 10-fold cross-validation, and chi-squared test. The analytic results indicated the best performance among the machine learning-based phishing detector. To the best of our knowledge, this is the first attempt that convolutional neural network is incorporated to learn the ensemble rule for phishing detection. The main findings of this research can be summarized as follows:

- The convolutional neural network works well for learning the ensemble rule of fusing heterogeneous features of URL representations, resulting the best accuracy and recall for phishing detection.
- We categorized the features of URLs into character and word levels, and demonstrated the convolutional and recurrent neural networks to effectively model each feature.

The remainder of this paper is organized as follows. In Section 2, we review the previous URL modeling methods based on machine learning and clarify the contributions of this paper by discussing the differences between them. In Section 3, we illustrate how the heterogeneous URL features are extracted by the deep learning and fused with convolutional neural network. The performance of the model is evaluated in Section 4 through various experiments, including the 10-fold cross-validation and ROC curve analysis. Finally, section 5 concludes the paper with some discussion of future directions.

2. Related Works

Previous studies on phishing URL classification can be classified into the following categories as summarized in Table 1: those on phishing URL detection based on the blacklist, which were mainly performed before 2010; those on modeling of words extracted from the text based on traditional machine learning; and those on text feature extraction through the latest deep learning algorithms.

The author proposed a system that extracts lexical features from the text according to ex-pert-defined rules, constructs a blacklist on known phishing URLs, and detects new phishing URLs through a simple comparison algorithm [9]. However, this method has the limitation of detecting new phishing URLs in terms of generalization performance. To confirm the validity of the machine learning method in the field of phishing URL classification, the authors applied fundamental machine learning methods including naive Bayes classifiers to the word combination found in URLs and classified phishing URLs that were not included in training datasets [10]. The authors enhanced the performance of phishing URL classification systems based on machine learning by applying a support vector machine (SVM), which is widely known to perform more complex nonlinear mapping [11]. Verma significantly increased phishing URL classification accuracy through the implementation of a random forest algorithm that was designed to perform effective modeling of hierarchical elements of lexical features in the URLs [12].

The researchers extracted semantic features from phishing URLs using a word-to-vector model capable of embedding word vectors based on their statistical meaning using deep learning algorithms. Furthermore, they applied long short-term memory (LSTM) and gated re-current unit (GRU) deep learning algorithms specialized for time series modeling, including gate operations, to enhance the phishing URL classification performance of existing modeling methods [8,13]. It was proposed a convolution-recurrent network to effectively model semantics extracted from the word-to-vector model [14].

The majority of the current research in deep learning-based phishing detection focuses mainly on optimizing the operation of the neural network [16]. In particular, the comparative study in [17] proves the superiority of the ensemble approach based on CNN variations. This motivates our decision to consider the ensemble learning approach proposed in this paper. The proposed method deviates from existing work in that it implements and learns the ensemble rule with convolutional operation based on CNN to consider the heterogeneous URL features.

Table 1: Related works on phishing URL detection with respect of URL features and modeling method.

URL Features	Method	Author
Bag-of-words	Naive Bayes	Prakash [9]
Lexical Features	Matching Rules	Ma [10]
Bag-of-words	SVM	Le [11]
Lexical Features	Random Forest	Verma [12]
Word embeddings	LSTM	Bahnsen [8]
Word embeddings	GRU	Zhao [13]
Lexical Features	Generative adversarial network (GAN)	Anand [15]
Word embeddings	CNN-LSTM	Yang [14]

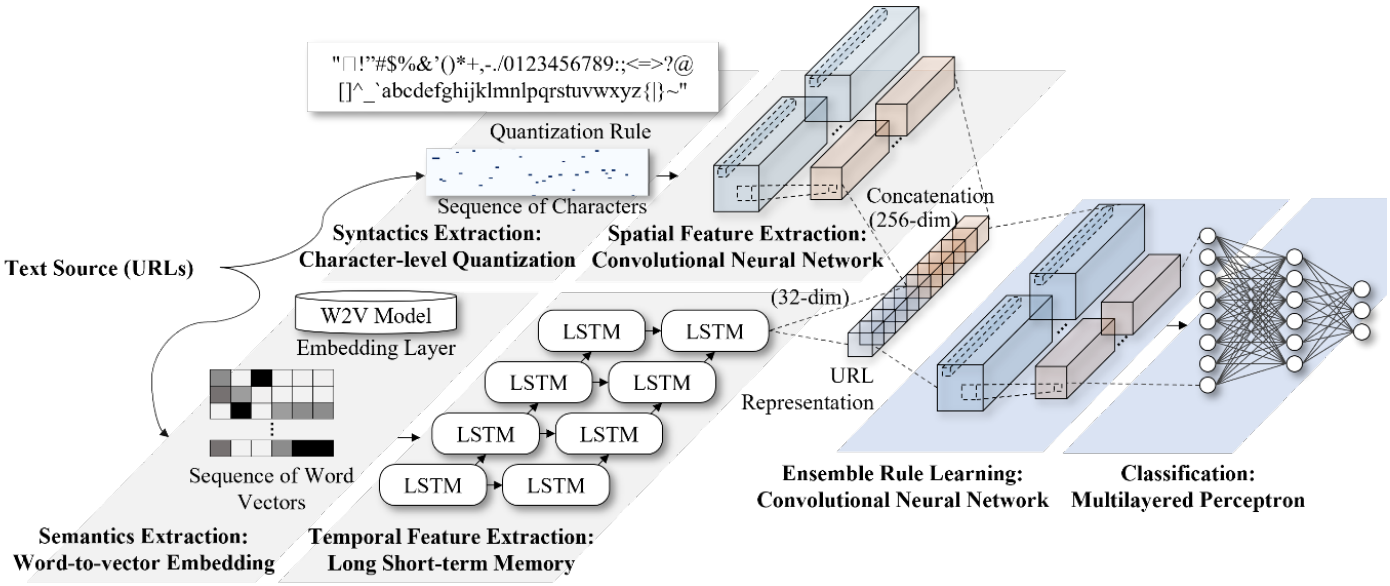


Figure 2: Proposed CNN-based ensemble learning method for phishing URL detection

3. Proposed Method

In this section, we describe the combination of the convolutional neural network and recurrent neural network to extract URL features and its ensemble learning method. Figure 2 visualize the diagrams of deep learning-based phishing classifier that extract the syntactic and semantics from URLs, as well as the proposed CNN-based ensemble learning network for the late-fusion of URL features.

3.1. Deep Learning-based Phishing URL Feature Extraction

Two types of deep learning algorithms and individual preprocessing steps were applied to conduct the modeling of syntactic and semantic features of phishing URLs. First, an integer was assigned to each character, and modeling of a low-level signal obtained through this process was performed by the CNN to model the syntactic features of random characters, including enumerated special characters, which are frequently observed in phishing URLs. Second, each word was embedded based on the word-to-vector model, and the modeling of a sequence of words obtained through this process was performed by the LSTM to model the semantic features of domains and sub-domains composing the internal URLs.

In detail, a preprocessing step for each character was performed to replace the characters with their unique Unicode values based on UTF-8 encoding, and an integer sequence of up to 100 characters was extracted in consideration of the average length of URL characters in the datasets collected. In total, 139 types of characters were used, and a vector in the dimension of n ×100 ×139 based on n of observations were inputted into the character-level CNN.

The convolution operation $\phi_c^l(\cdot)$ in Equation 1 applies a parameterized filter to the input vector and extracts syntactics from sequence of characters in URL. A filter size $m \times m$ is applied to the ith row and jth column nodes of the lth layer.

$$\phi_c^l(x_{ij}) = \sum_{a=0}^{m-1} \sum_{b=0}^{m-1} w_{ab} x_{(i+a)(j+b)} \quad (1)$$

The pooling operation $\phi_p^l(\cdot)$ in the lth pooling layer performs the extraction of representative value and defined as Equation 2, with the pooling distance τ in the region $k \times k$ among the input vectors, and outputs the maximum activation value from the region.

$$\phi_p^l(x_{ij}) = \max_{\tau \in R} x_{ij \times \tau} \quad (2)$$

The learning of the convolution operation is the process of optimizing the weight of the filter w that extracts the syntactics while preserving the spatial correlation between characters, and the pooling operation is based on extraction of emphasized features.

Meanwhile, representative features of URLs include semantics that can be derived from a sequence of words such as domains and sub-domains. Phishing URL classification accuracy can be enhanced through the parallel utilization of deep learning algorithms for additional modeling of a sequence of subdomains [18].

The modeling of semantics of phishing URLs was carried out through word embedding based on the word-to-vector model and LSTM deep learning algorithm application for time series modeling. Moreover, 20 words that appeared in sub-domains were additionally extracted since phishing URLs generally included various sub-domains. Each word was replaced as vectors in 32 dimensions using the word-to-vector model, and URLs formed as n ×20 ×32 sized vector according to n observations were input in the phishing word-level LSTM.

The LSTM network is a type of RNN in which three types of nonlinear gates are implemented. The LSTM $\phi_L^l(\cdot)$ performs the time-series modeling of sequence of domain and subdomains.

$$\phi_L^l(x_{ij}) = o_t \odot \tanh(c_t) \quad (3)$$

The input gate(i), forget gate (f), output gate (o), and LSTM cell state (c) were defined based on the input domain sequence of

x = (x(t),…, x(t-ω)) with word sequence length ω, as shown in Equation 4. b, σ and ⊙ refer to the bias added to each neural network, the sigmoid activation function of neural networks, and Hadamard multiplication, respectively.

$$\begin{aligned} i_t &= \sigma(W_{ix}x(t) + W_{im}h_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x(t) + W_{hf}h_{t-1} + b_f) \\ o_t &= \sigma(W_{xo}x(t) + W_{ho}h_{t-1} + b_o) \\ g_t &= tanh(W_{xh}x(t) + W_{hc}h_{t-1} + b_o) \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t \end{aligned} \quad (4)$$

3.2. Ensemble Learning based on the Convolutional Network

The proposed ensemble network utilizes the deep URL representations with a size of (n × 256) and (n × 32) with n observations from the intermediate layer of CNN and LSTM in Section 3.1. Contrary to the existing CNN-LSTM ensemble-based phishing URL detector, the model is optimized to weight the outputs from multi-level URL representations.

The character-level and word-level representations of phishing URL derived from the character-level CNN and word-level RNN were concatenated to form a vector of size (n × 288). The proposed fusion neural network was trained to minimize errors that might occur in the process of mapping the input vector to the benign or phishing URLs.

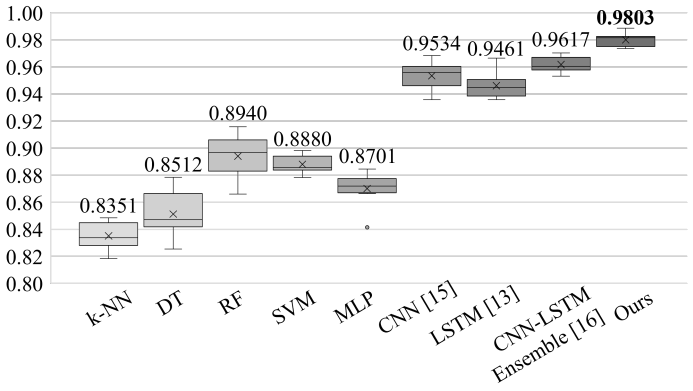


Figure 3: Results of 10-fold cross validation comparing the proposed ensemble learning with other existing methods (k-NN: k-nearest neighbor; DT: decision tree; RF: random forest; SVM: support vector machine; MLP: multilayered perceptron; CNN: convolutional neural network; LSTM: long short-term memory).

As the proposed method contains a convolution layer in order to systematically fuse level-based features, effective features are selected from input vectors from 288 dimensions and output the predictive label $\hat{y}$, as shown in Equation 5.

$$p(\hat{y}_i|x_i) = argmax \frac{exp(\phi^{l-1}(x_i)w^l + b^l)}{\sum exp(\phi^{l-1}(x_i)w^l + b^l)} \quad (5)$$

At this stage, the Softmax function, which is an activation function of the neural network, was applied to facilitate the encoding of the output vector at the probability of [0,1] range and to promote the differentiation process that operates during the optimization of the loss function. The entire mapping results obtained from inputs to outputs in the character-level and semantic-level neural networks, including the proposed ensemble network, is differentiable and can be learnt by the massive URL observations.

The entire weights of neural networks are tuned by applying a backpropagation algorithm based on gradient descent to the cross-entropy loss function L_{CE} shown in Equation 6.

$$L_{CE} = -\sum_i y_i log(\hat{y}_i) \quad (6)$$

The proposed ensemble network, which fuses features according to deep URL representations, performs the optimization of ensemble rules in consideration of joint learning of CNN and LSTM.

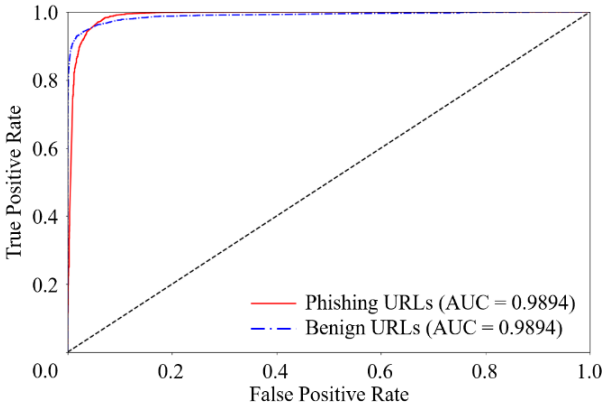


Figure 4: ROC curve and area under the curve (AUC) of classification result based on ensemble learning approach.

4. Experimental Results

4.1. Phishing URL Dataset

A total of 45,000 of URLs were collected by crawling method and 15,000 of phishing URLs were collected from Phishtank [19,20] where provides a blacklist of phishing URLs. Benign URLs were mainly collected from open directory project(ODP), which is the URL database to categorize URLs. The number of observations was intentionally adjusted because of the data imbalance issue, to reflect the conditions in which the number of phishing URLs was much lower than that of benign URLs. We noted the length of the URL is one of the critical phishing features, given that the average lengths of phishing and benign URLs are 75.74 and 35.83, respectively.

4.2. Phishing URL Classification Performance

Figure 3 shows the result of 10-fold cross validation on the proposed ensemble network and other machine learning-based models to verify the phishing URL classification performance of the proposed method. The average accuracy of the random forest algorithm, CNN, and RNN were achieved 0.8940, 0.9534, and 0.9461, respectively, with the CNN and RNN exhibiting significantly higher accuracy performance than the random forest algorithm.

The comparative result based on the ensemble of CNN [7] and RNN [8] achieved 0.9641. Based on the performance improvement, we confirmed the significance of proposed ensemble learning approach in consideration of character and word-level URL modeling. Regarding the proposed ensemble network designed to fuse Multi-level URL features, the best classification accuracy was achieved as 0.9803.

Table 2: Qualitative evaluation of complementarity based on the case analysis of CNN and LSTM (0: benign, 1: phishing).

Category		URL (accessed date: 19-10-2020)	CNN	LSTM
CNN Advantages	Phishing	https://1drv.ms/xs/s!AhtvzT3KrwqMZzLMKnTc8clHnRA?wdFormId=%7BA0F7982D%2D71A4%2D4DE0%2DB4C4%2DC16A0F044	0.9874	0.7385
	Benign	http://market.security***.net	0.0031	0.8441
LSTM Advantages	Phishing	http://bitcoin24-wallet.site	0.0722	0.9837
	Benign	http://www.knightfeatu***.com/kfweb/content/features/kffeatures/puzzlesandcrosswords/kf/sudoku/sudoku_classic/sudoku_classic.html	0.8384	0.0073
Misclassified	Benign	http://archives.seattletimes.nwsou***.com/cgi/bin/texis.cgi/web/vortex/display?slug=will&date=199903	0.8815	0.8764
	Phishing	http://tesla-present.site/ethereum/	0.0584	0.0354

Table 3: A confusion matrix for phishing URL classification based on the ensemble network

		Predicted (w/o ensemble learning)		
		Benign	Phishing	**Recall**
Actual	Benign	9035 (8902)	74 (207)	**0.9919 (0.9773)**
	Phishing	115 (266)	2776 (2625)	**0.9602 (0.9080)**
	Precision	**0.9874 (0.9710)**	**0.9740 (0.9269)**	Accuracy: 0.9843 (0.9606)

Since it is essential to minimize false negatives and improve recall in the field of phishing URL detection, the ROC curve and AUC are described in Figure 4. Table 3 summarizes the classification results based on the model that exhibited the best accuracy. Considering the false negatives and the recall of phishing instance of 0.9602, it is inferred that additional modeling should be carried out mainly focusing on the generalization strength of the model.

4.3. Discussion

Table 2 indicates the advantages of each model based on practical classification cases, to aid in the classification of the performance of deep learning models according to multi-level URL representations. The two upper rows show the robustness against random character enumeration of CNN. The character-level CNN classified URL as phishing instance with a probability of 0.9874, considering that a phishing URL feature is hidden in the sequence of random characters. On the other hand, in the second case, the word-based LSTM misclassified benign URL as phishing with a probability of 0.8441 because the benign words are included in the URL.

The word-level LSTM supplements the entire system by reflecting a sub-domain that was not used by the character-level CNN in the form of words. The LSTM was able to classify certain words such as ‘security’ and ‘bitcoin’ based on the massive observations that such words are frequently used in phishing URLs. The CNN, however, misclassified benign URLs as phishing based on the number of sub-domains.

5. Concluding Remarks

5.1. Conclusion

This study introduced a character-level CNN and word-level RNN for phishing URL representation and proposed an ensemble network that can effectively fuse the syntactics and semantics of phishing URLs extracted from each model. The proposed ensemble network, implemented as convolutional neural network, provide the plausible solution of parameterization and optimization of the ensemble rule. Specifically, it exhibited a classification accuracy of 0.9804, which is the highest compared to other machine learning methods including deep learning models.

5.2. Future Work

Further studies should be performed to guarantee the generalization strength of model, in consideration of zero-day attack characteristics of phishing attacks. The latest deep learning algorithms, such as one-shot learning, should be thoroughly examined. Meanwhile, the scope of this study is limited to the modeling of syntactics and semantics of URL and optimizing the ensemble rule. In this regard, a symbolic AI approach to fully exploit and utilize the domain knowledge is promising. In addition, the neural-symbolic integration approach that can calibrate the deep learning model should be additionally considered in the future studies.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work has supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1F1A1063085).

References

[1] H.-J. Kim, Image-based malware classification using convolutional neural network, Springer: 1352–1357, 2017, doi:10.1007/978-981-10-7605-3_215.
[2] S.-J. Bu, S.-B. Cho, “Deep Character-Level Anomaly Detection Based on a Convolutional Autoencoder for Zero-Day Phishing URL Detection,” Electronics, **10**(12), 1492, 2021, doi:10.3390/electronics10121492.
[3] V. Suganya, “A review on phishing attacks and various anti phishing techniques,” International Journal of Computer Applications, **139**(1), 20–23, 2016, doi:10.5120/ijca2016909084.
[4] K.L. Chiew, K.S.C. Yong, C.L. Tan, “A survey of phishing attacks: Their types, vectors and technical approaches,” Expert Systems with Applications, **106**, 1–20, 2018, doi:10.1016/j.eswa.2018.03.050.

[5] I. Qabajeh, F. Thabtah, F. Chiclana, "A recent review of conventional vs. automated cybersecurity anti-phishing techniques," Computer Science Review, **29**, 44–55, 2018, doi:10.1016/j.cosrev.2018.05.003.

[6] S.-J. Bu, S.-B. Cho, "Integrating Deep Learning with First-Order Logic Programmed Constraints for Zero-Day Phishing Attack Detection," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE: 2685–2689, 2021, doi: 10.1109/ICASSP39728.2021.9414850.

[7] H. Le, Q. Pham, D. Sahoo, S.C.H. Hoi, "URLNet: Learning a URL representation with deep learning for malicious URL detection," ArXiv Preprint ArXiv:1802.03162, 2018, doi:10.475/123_4.

[8] A.C. Bahnsen, E.C. Bohorquez, S. Villegas, J. Vargas, F.A. González, "Classifying phishing URLs using recurrent neural networks," in 2017 APWG symposium on electronic crime research (eCrime), IEEE: 1–8, 2017, doi:10.1109/ECRIME.2017.7945048.

[9] P. Prakash, M. Kumar, R.R. Kompella, M. Gupta, "Phishnet: predictive blacklisting to detect phishing attacks," in 2010 Proceedings IEEE INFOCOM, IEEE: 1–5, 2010, doi:10.1109/INFCOM.2010.5462216.

[10] J. Ma, L.K. Saul, S. Savage, G.M. Voelker, "Beyond blacklists: learning to detect malicious web sites from suspicious URLs," in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 1245–1254, 2009, doi:10.1145/1557019.1557153.

[11] A. Le, A. Markopoulou, M. Faloutsos, "Phishdef: Url names say it all," in 2011 Proceedings IEEE INFOCOM, IEEE: 191–195, 2011, doi:10.1109/INFCOM.2011.5934995.

[12] R. Verma, K. Dyer, "On the character of phishing URLs: Accurate and robust statistical learning classifiers," in Proceedings of the 5th ACM Conference on Data and Application Security and Privacy, 111–122, 2015, doi:10.1145/2699026.2699115.

[13] J. Zhao, N. Wang, Q. Ma, Z. Cheng, "Classifying malicious URLs using gated recurrent neural networks," in International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, Springer: 385–394, 2018, doi:10.1007/978-3-319-93554-6_36.

[14] W. Yang, W. Zuo, B. Cui, "Detecting malicious URLs via a keyword-based convolutional gated-recurrent-unit neural network," IEEE Access, **7**, 29891–29900, 2019, doi:10.1109/ACCESS.2019.2895751.

[15] A. Anand, K. Gorde, J.R.A. Moniz, N. Park, T. Chakraborty, B.-T. Chu, "Phishing URL detection with oversampling based on text generative adversarial networks," in 2018 IEEE International Conference on Big Data (Big Data), IEEE: 1168–1177, 2018, doi:10.1109/BigData.2018.8622547.

[16] F. Tajaddodianfar, J.W. Stokes, A. Gururajan, "Texception: A character/word-level deep learning model for phishing URL detection," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE: 2857–2861, 2020, doi:10.1109/ICASSP40776.2020.9053670.

[17] D. Vasan, M. Alazab, S. Wassan, B. Safaei, Q. Zheng, "Image-Based malware classification using ensemble of CNN architectures (IMCEC)," Computers & Security, **92**, 101748, 2020, doi:10.1016/j.cose.2020.101748.

[18] Q. Li, M. Cheng, J. Wang, B. Sun, "LSTM based phishing detection for big email data," IEEE Transactions on Big Data, 2020, doi:10.1109/TBDATA.2020.2978915.

[19] L.L.C. OpenDNS, "PhishTank: An anti-phishing site," Online: https://Www. Phishtank. Com, 2016 (accessed: 1 Oct. 2021).

[20] Q. Cui, G.-V. Jourdan, G. V Bochmann, R. Couturier, I.-V. Onut, "Tracking phishing attacks over time," in Proceedings of the 26th International Conference on World Wide Web, 667–676, 2017, doi:10.1145/3038912.3052654.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 5, 348-355 (2021)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Neural Network for 2D Range Scanner Navigation System

Giuseppe Spampinato*, Arcangelo Ranieri Bruna, Ivana Guarneri, Davide Giacalone

STMicroelectronics, System, Research and Application, Catania, 95100, Italy

ARTICLE INFO

Article history:
Received: 02 July, 2021
Accepted: 06 October, 2021
Online: 23 October, 2021

Keywords:
Navigation
Localization
Laser Scans
Dataset Generation
Neural Network

ABSTRACT

Navigation of a moving object (drone, vehicle, robot, and so on) and related localization in unknown scenes is nowadays a challenging subject to be addressed. Typically, different source devices, such as image sensor, Inertial Measurement Unit (IMU), Time of Flight (TOF), or a combination of them can be used to reach this goal. Recently, due to increasing accuracy and decreasing cost, the usage of 2D laser range scanners has growth in this subject. Inside a complete navigation scheme, using a 2D laser range scanner, the proposed paper considers alternative ways to estimate the core localization step with the usage of deep learning. We propose a simple but accurate neural network, using less than one hundred thousand overall parameters and reaching good precision performance in terms of Mean Absolute Error (MAE): one centimeter in translation and one degree in rotation. Moreover, the inference time of the neural network is quite fast, processing eight thousand scan pairs per second on Titan X (Pascal) GPU produced by Nvidia. For these reasons, the system is suitable for real-time processing and it is an interesting complement and/or integration for traditional localization methods.

1. Introduction

This paper is an extension of the work originally presented in ICARA [1]. Further investigations to increase performance of the approach are done changing deep learning parameters, moreover problem simplification and data augmentation have been tested.

The field of the proposed paper is navigation system and related localization, which is still considered a challenging task [2]. The core function of these kind of systems is for sure the localization itself. Main goal of this vital function is the correct estimation of the step-by-step position of the moving robot in unknown scenes. To perform this task different data sensors can be used. Moreover, to reach better estimation usually previous step data are stored inside an updating map.

About localization approaches proposed in literature, two main localization groups can be recognized: vision-based and laser-based.

Vision-based localization techniques just use images to achieve their goal. Usually, features inside previous and current image are calculated and matched to retrieve the global displacement. In literature, different vision-based techniques have been proposed: effective prioritized matching [3], ORB-SLAM [4], monocular semi-direct visual odometry (SVO) [5], camera pose voting [6], localization based on probabilistic feature map [7], etc. More sophisticated solutions, like multi-resolution image pyramid methods, have been proposed to reach more robust feature matching [8]. These approaches are usually robust, but they do not have useful distance information, i.e., it is difficult to map the estimated trajectory (in pixel) in the real world (in cm).

On the other hand, laser-based localization techniques just use laser scans to achieve their goal. A features matching approach, as in vision-based localization, is not simple to be implemented. This is mainly due to the poor information of laser scans compared to images. In fact, in the case of image details (e.g., corners), we can note only weak variations in range measurements and then a lack of distinctive features to be analyzed [9].

Usually, Bayesian filtering are largely used in laser-based techniques to consider the robot position as a problem of probability distribution estimation based on grid maps [10,11]. Apart Bayesian filtering, in literature other laser-based proposed techniques are: iterative closest point (ICP) [12] and related variants [13], which minimize the matching error between two point-clouds estimating the related transformation, perimeter based polar scan matching (PB PSM), Lidar odometry and mapping (LOAM) [15], which achieves real time processing by running in parallel two different algorithms, and so on. Even if these approaches usually achieve precise localization, since distance information is available, they can fail in scene changing conditions. In fact, when an object is moving in the scene, due to

*Corresponding Author: Giuseppe Spampinato, giuseppe.spampinato@st.com

the occlusions, we can have lack of information in the moving object area and then the estimated localization can be wrong.

Recently, in vision-based localization, to extract and match image features, deep learning approaches have been successfully used. These promising approaches allow to estimate camera position. A lot of deep learning approaches have been proposed in literature: PoseNet [16], which for pose regression task uses the convolutional neural networks (CNNs), Deepvo [17], which for the same task uses recurrent neural networks (RNNs), undeepVO [18], which estimates the monocular camera pose using a deep learning unsupervised method, and so on. Unfortunately, at moment the deep learning-based methods do not achieve the same pose estimation accuracy of classical vision-based localization approaches.

Inspired by vision-based localization approaches based on deep learning algorithms, few attempts of deep learning methods have also been suggested for laser-based localization: in [19] authors estimate odometry processing 3D laser scanner data with a series of CNNs, in [20] authors trained for giving steering commands a navigation model target-oriented, in [21] authors performed loop closure and matching of consecutive scans making use of a CNN network, in [22] authors improved the odometry estimation considering also temporal features using a RNN, able to model sequential long-term dependencies, and so on.

Regarding the deep learning laser-based localization, as indicated in the case of vision-based localization, unfortunately they still do not achieve the accuracy of the pose estimation compared to classical laser-based localization. For this reason, in literature some authors propose the integration of deep learning approaches with the classical ones: in [23] the authors make use of Inertial Measurement Unit (IMU) in combination with CNNs for 3D laser scanners for assisted odometry, in [24] the authors use the result of vision-based localization approaches based on CNN as starting seed for Monte Carlo localization algorithm, to speed-up algorithm convergence, also increasing robustness and precision, and so on.

It is easy to understand that the field of deep learning localization, in particular about laser-based approaches, has not yet been intensively explored and, at moment, it is still considered a challenging process. In fact, just a small number of papers discuss about this subject [22]. Our choise is to go further in this investigation, to obtain a simple deep learning laser-based localization, using only data taken by 2D laser scanners. Moreover, we tried to reduce as far as possible the number of parameters used by the proposed network, to deal with the low-cost resources constraint.

Our contribution to the research in the field of navigation system, using deep learning approaches, with only 2D laser scan input is firstly the exploration of state of art algorithms. Another contribution is to indicate a methodology to generate the ground truth for the neural network without using real sensors but simulating them with existing powerful navigation tools.

Novelties of the proposed system are in both neural network dataset generation and training. In particular, in data generation we indicate a methodology to choose properly the angle resolution trying to reduce the collisions per frame (to avoid loss of important data) and to maximize array density (to avoid working with sparse data). In this way, the neural network was more able to solve the regression problem.

In the training phase, the novelty is the demonstration with real tests that in regression problems the choice of input/output values scale is vital to let the neural network working. In fact, after several experiments, we obtained the correct scale measures for distances and angles.

At last, the great contribution was to find, after lots of experiments with various neural network hyper parameters, a really light network to solve the localization problem with good performances in terms of mean absolute error between estimated positions and ground truth.

The proposed research is composed by the following Sections: Section 2, where the proposed deep learning laser-based localization is described; Section 3, where the experimental results obtained are deeply described; Section 4, where final considerations are remarked.

2. Proposed Approach

A typical navigation system is described in Figure 1. A starting moving object position (x,y,α) is considered, where (x,y) are the horizontal and vertical position in the cartesian axis and α is the orientation angle. Usually, at the beginning the position is assumed to be (0,0,0). Every time laser scan data is available (θ, d), where θ is the angle and d is the distance from object in front of the laser beam, the localization step will calculate the new position (x',y',α'). The system will then decide next movement. Depending on how it is programmed the moving object (for example the robot), the system can decide to continue moving (in the case an obstacle is not found) or to stop motors (when an obstacle is found). The correct command are then send to the motor control (which interact with the IMU) to update the movement. Positions and movements are updated each time.

Inside the navigation system, the proposed deep learning approach is applied on the core localization step. From this point, this article will focus only on the localization step and all the research will be focused on this particular block of the navigation system.

To reach this objective we used a wheeled robot equipped with the laser scanner A2 RPLidar on the top. This rotating laser scanner has twelve meters as maximum range, view at 360 degrees, running up to fifteen Hz. Thanks to the robot, we acquired a custom dataset in various environments (apartment, laboratory and office).

The ground truth generation schema used is shown in Figure 2. At the beginning, the dataset acquisition is needed to record the input Lidar dataset. Each scan is composed by multiple couples (d, θ), where d is the distance from the object and θ is the related angle. Once the dataset was obtained, we needed to generate the ground truth position (x, y, α) for each sample taken. Since we do not have the real position of the robot for each scan contained in the acquired dataset, we needed a simulation environment to obtain a ground truth.

For this purpose, we make usage of the MATLAB Navigation Tool. The generated ground truth was tested using a simple Mat2Map program to display the path of obtained positions (x, y, α) for each sample taken and the map generated by Lidar scans. In this way, we also checked the robustness of Navigation Tool. Even if it is a very slow method, we tested it in different conditions and we conclude to be very precise, so it was used as reference. It is based on Google Cartographer [25], which builds multiple

submaps and try to align upcoming scans with previous nearby submaps, generating the constrains on a graph.

Once we generated the custom dataset and related ground truth, we perform our experiments using the TensorFlow framework with Keras wrapper in a Python environment. In this configuration, to obtain the best compromise between quality and complexity, we tried different data binarization and augmentation with various neural network configurations.

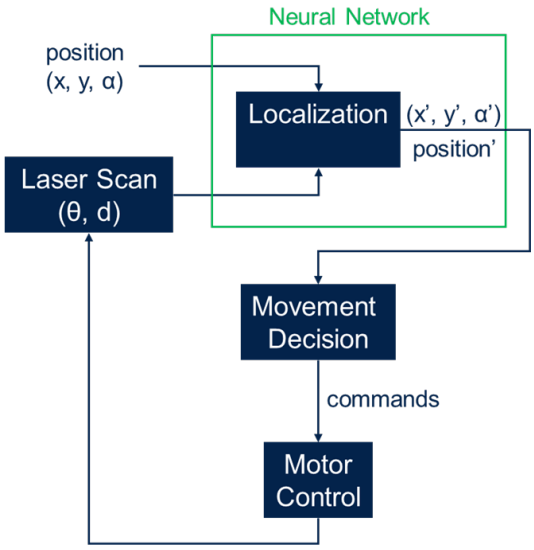


Figure 1: Navigation system.

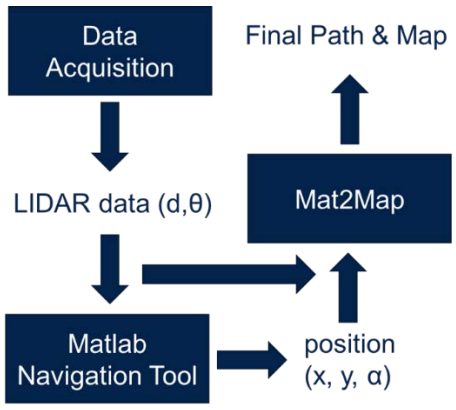


Figure 2: Ground truth generation schema.

2.1. Dataset Generation

The main challenge in deep learning approaches is the acquisition of large amounts of data, to allow the neural network to work fine in any real scenario. Our generated dataset consists in about 51,000 samples, which we tested to be enough in our experiments for training the proposed neural network. Each sample is composed by two subsequent scans acquired by Lidar. Each scan is expressed by multiple couples (d, θ), that is the distance and the angle from the nearest found object.

Data acquisition cannot be used as it is. We instead need to encode acquired data into a panoramic like depth image, so we need a sort of binarization of each scan before to be paired with the following one. Data binarization is inspired by a previous work [21] to encode laser scans into a 1D vector. To obtain a 1D vector, in each scan, all distances are binned into angle bins, according to the chosen angle resolution. In this way, we stored all depth values inside a 1D vector, where all the possible depths are represented (from 0° to 360°). As soon as two subsequent scans are binarized, we can couple them to be used as neural network input.

Laser range scanners usually give (scan by scan) distances from the identified nearest object for constant angles, so binarization is simple, because we have a fixed number of possible angles to be considered into the 1D vector. Instead, in the chosen laser range scanner A2 RPLidar in each scan the angles can vary, from 0° to 360°. For this reason, it is not possible to fix we the angle resolution as in previous works, e.g., in [22] the authors use 0.10° and in [21] the authors use 0.25°, so a preliminary investigation to find optimal angle resolution is needed. In this research, we tried at the same time to maximize array density and to minimize collisions per frame. Array density is for each scan the number of non-zero value bins, while collisions per frame is the total number of data ranges which are in the same bin.

Table 1 shows the impact of the chosen β (angle resolution) on N (total number of bins) and then on mean collisions per frame and mean array density. Since A2 RPLidar have got a 360 degrees view, the laser data is separated into β degree bins, for a total amount of N=360°/β° bins. Of course, increasing β (and then decreasing the total number of N bins) the array density increases and of course collisions per frame will become bigger. In our experiments, we tried different angle resolutions β to make at the end the proper decision about which configuration to use in the proposed neural network.

As indicated in Table 1, particularly at higher angle resolutions β, collision is an important aspect to solve to guarantee the neural network to work property. In [22] authors chose to take the mean of all distances are in the same bin, probably because in their experiments the collision occurred rarely and distances at the same degree was similar. In our experiments, we consider two main aspects: laser range scanner is more precise for lower distances and average of two different distances at the same degree can introduce false objects distances. For these reasons, we choose to take the minimum distance (instead of average distance) for distances falling in the same degree.

Table 1: Dataset Binarization

β	N	Collisions per frame	Array Density
0.10	3600	0.05	9%
0.25	1440	0.13	22%
0.50	720	0.48	46%
1.0	360	2.54	89%

2.2. Neural Network

The network we propose in this work, from a consecutive pair of two scans (s_{t-1},s_t) done by the chosen Lidar, obtains the robot displacement between them, trying the estimation of their relative pose transformation:

$$T = [\Delta x, \Delta y, \Delta \alpha] \quad (1)$$

where Δx and Δy are respectively the horizontal and vertical translations and Δα is the rotation angle between the two scans (s_{t-1},s_t). We can only estimate the displacement of the robot in two dimensions, since we choose to use just a two dimension sensor.

The final objective of the neural network is to learn the unknown function g(), which, at time t, maps (s_{t-1},s_t) to the pose transformation T :

$$T_t = g(s_{t-1},s_t) \quad (2)$$

In the training step the unknown function g() is learned. Moreover, thanks to the accumulation of the estimated local poses from the starting of the process up to time t, we obtain the robot global position at time t. The chosen loss functions are: mean absolute error (MAE) and mean square error (MSE), which are commonly used in deep learning regression problems.

The strategy we tried to implement here is to fit our regression problem with standard deep learning 2D image matching problems. The difference is that, in our case, instead of 2D images, we have obtained (thanks to dataset binarization) 1D panoramic depth images. In this way, as in the case of images, we can use consecutive CNN to extract spatial features obtained by the sensor in the tested conditions. After that, additional fully connected layers (dense layers) allow to the neural network to understand patterns within extracted spatial features to provide the matching and then the current robot position estimation in the unknown environment.

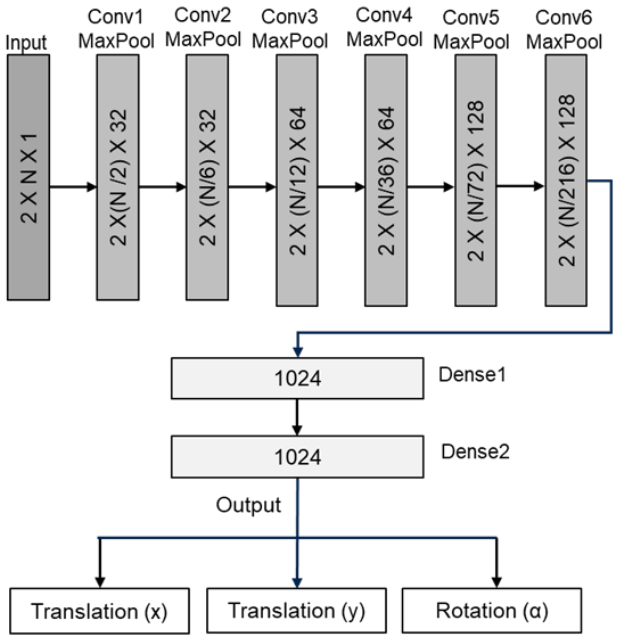


Figure 3: Proposed deep learning approach.

Table 2: Parameters of Neural Network

N	NN Model	NN Parameters
3600	CNN+LSTM	21,078,563
3600	CNN+Dense	5,343,779
1440	CNN+LSTM	15,835,683
1440	CNN+Dense	2,722,339
720	CNN+LSTM	14,262,819
720	CNN+Dense	1,935,907
360	CNN+LSTM	13,214,243
360	CNN+Dense	1,411,619

The proposed neural network is shown in Figure 3. The suggested neural network has been inspired by the one indicated in [22], but some differences should be evidenced. Firstly, we make lots of trials to obtain a low complexity neural network, without to lose in quality performance. Moreover, we used max pool layers (instead of average pool layers) reducing complexity and extracting only most important features spatially distributed. Better results were obtained using max pooling, since as aforementioned (see Table 1), the binarized laser scans tends to be sparse, when the angle resolutions β tends to increase. For the same reason, we also eliminate the stride parameter of the various neural network convolutional filters, which can eliminate useful information for sparse binarized laser scans. To maintain the same dimension, instead of using stride parameter in the convolutional filters, we used other max pooling before applying these filters.

As aforementioned, we experimented different neural network configurations, in particular varying:

- β (angle resolution) and then N (number of bins), to define the correct dimension of input data;
- last two network layers, trying both dense layers [21] and long short term memory (LSTM) layers [22].

The calculation of the NN parameters is shown in Table 2. This number is impacted by N (number of bins) and by the neural network model used. As indicated, when N is increased and when LSTM layers are used, the total number of NN parameters and then the time to be executed will increase too. It is to note that the changes made compared to [22], that is max pool in replacement of strides and average pool, do not impact the overall NN parameters.

2.3. Training

For the sake of clearness, now we indicate in detail the input and ground thru output of the proposed neural network. Input for the network is composed by composed by a set of couples of consecutive Lidar scans (θ, d). These scans are preprocessed allowing to binarize them into sets of 2 X N matrixes, as indicated in Section 2.1. In these matrixes, N depends on β (angle resolution), like indicated in Table 1.

About the reference (ground thru) output of the proposed NN, it is composed by a set of vectors T = [Δx,Δy,Δα], which are the ground thru positions of the moving object, obtained by applying Navigation Tool (MATLAB) to the input Lidar scans (θ, d).

The various experiments were performed in a Python environment with the use of TensorFlow framework and Keras wrapper. Moreover, a workstation was used for training execution, that is a Xeon ES-2630 (Intel) octacore machine with 62 GB of RAM and a Titan X (Pascal) GPU produced by Nvidia. The chosen GPU has got 12 GB of RAM and it is equipped by 3584 CUDA cores, to allow lots of parallelization in training step for faster execution.

The details for training step are the following: 0.0001 function cost minimization learning rate, 500 epochs for training the neural network, 32 batch size and Adam training optimizer used. We tried other kinds of training optimizers, but we did not notice any significant difference in regression performances.

3. Experimental Results

Lots of tests have been executed with different neural network configurations, as indicated in Table 2, and with different indoor environments. As expected, it is important to note that classical Convolutional Neural Networks (CNNs) work better in the case of dense input datasets. For this reason, we used max pooling instead of average pooling in final tests. For the same reason, even if, at the beginning, we tried in our experiments all the different neural network configurations, final research was focused on β (angle resolution) set to one degree and then N (number of bins) set to 360. This choice also allows us to reduce the total neural network parameters and the overall complexity.

Figures in this Section are representative of a particular testing to underline how (depending on scaling applied) the neural network tends to converge (generalizing the regression problem) or not and to underline how the final suggested neural network fits our needs (lightness and precision). In particular, in the X axis the evolution of the network in various epochs (trials) is represented and in the Y axis the loss in precision is represented (first trials in mean square error, after we used mean absolute error). When train and validation curves are similar with low loss the neural network works properly, while when they are different a problem occurs. In this Section we try to explain a particular reason (scale) of this problem.

Table 3 shows the results obtained using input distances and output positions expressed in millimeters. Results are really bad: the proposed network seems to make a sufficient regression work for training set, but for validation and test set it does not reach good performance at all. In general, as expected, max pooling strategy reaches better performance than average pooling and reducing the angle resolution and then the N dimension of the input binarized scans we obtain better loss values (MSE).

To better understand the evolution of this first experiment done on the proposed neural network epoch by epoch, the first 150 epochs are displayed in Figure 4. This graph is referred to the case N = 1440 with max pooling (train loss = 2.48 MSE; validation loss = 216 MSE; test loss = 233 MSE), but similar considerations can be done on the other different configurations. It is easy to note that while the curve for train decreases, the curve for validation is flat, so the proposed network, in this case, is not able to solve the overexposed regression problem and to generalize it.

Table 3: Test Results (MSE) with Input Dataset (Millimeters)

N	Model	Train Loss	Validation Loss	Test Loss
3600	AvePool	2.78	250	258
3600	MaxPool	2.75	248	250
1440	AvePool	2.30	228	236
1440	MaxPool	2.48	216	233
720	AvePool	2.36	221	240
720	MaxPool	3.27	220	238
360	AvePool	7.54	220	235
360	MaxPool	7.02	218	230

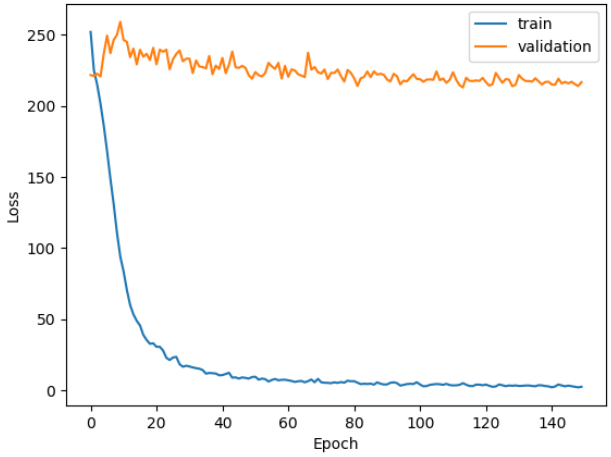


Figure 4. Results obtained with input dataset (millimeter).

In the used approach, the main issue is the scale discrepancy in the input variables (x and y expressed in millimeters and α expressed in degree), which often increases the difficulty to correctly model the neural network to solve the regression problem. A common trick used in these cases is to pre-process the input variables before they are fed to the neural network [26]. In the same way, also the outputs of the network (d expressed in millimeters and θ expressed in degree) should be processed to obtain the correct output values.

A commonly used pre-processing step is just a simple linear scaling of network variables [26], so we just changed the distance measure passing from millimeter to centimeter and the related results are shown in Table 4. As indicated, better results are obtained compared to the first tentative with input distances and output positions expressed in millimeters. Even if results are better, we must again to note that they are not good enough: again, the proposed network seems to make a sufficient regression work for training set, but for validation and test set it does not reach similar good performance. Moreover, as expected, max pooling strategy reaches better performance than average pooling and reducing the angle resolution and then the N dimension of the input binarized scans we obtain better loss values (MSE).

Table 4: Test Results (MSE) with Input Dataset (Centimeters)

N	Model	Train Loss	Validation Loss	Test Loss
3600	AvePool	0.09	4.12	3.90
3600	MaxPool	0.03	3.98	3.83
1440	AvePool	0.28	3.97	3.77

1440	MaxPool	0.11	3.87	3.68
720	AvePool	0.18	3.86	3.76
720	MaxPool	0.13	3.84	3.67
360	AvePool	0.28	3.75	3.70
360	MaxPool	0.19	3.67	3.64

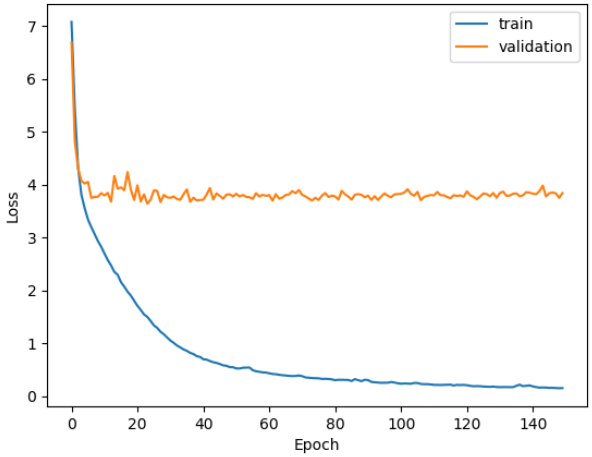


Figure 5. Results obtained with input dataset (centimeter).

To better understand the evolution of the second experiment done on the proposed neural network epoch by epoch, the first 150 epochs are displayed in Figure 5. To allow a visual comparison with Figure 4, also this graph is referred to the case N = 1440 with max pooling (train loss = 0.13 MSE; validation loss = 3.84 MSE; test loss = 3.683 MSE). Of course, similar considerations can be done on the other different configurations. It is important to put into evidence that MSE is not a linear measure, but quadratic. This is the reason why the loss is drastically reduced in comparison with the previous experiment. Moreover, the train curve correctly decreases, as in previous case, while in the current test the validation curve is not completely flat, but it starts to go down. As first experiment, in the current test the proposed network cannot model the neural network to generalize and correctly solve the regression problem, but improvements noticed give us an important hint to work with: to obtain the best regression results, firstly we must find the correct rescaling to apply to input and output data measures (distance and angle).

At this point, we make some rescaling experiments. To maintain similar scale in both input dataset and output variables, we tried to scale translation data by one thousand (in this way we use meters, instead of millimeters as measure) and rotation data by one hundred. This rescaling configuration gives us best results. Furthermore, we decide to make use of Mean Absolute Error (MAE), instead of Mean Squared Error (MSE). In this way, we obtained similar loss curves, but with results simpler to understand and comment, because MAE is a linear measure, while MSE is quadratic. Figure 6 shows that good results are finally reached (train loss: 0.011 MAE; validation loss: 0.011 MAE; test loss: 0.010 MAE), using the new scaling factors to be applied to input and output data with a very simple configuration (only 1,411,619 network parameters). Like other experiments, the train curve correctly decreases, but this time the validation curve also decreases with similar slope. This indicates that finally we reached our main goal: the neural network can now correctly generalize and solve the proposed regression problem.

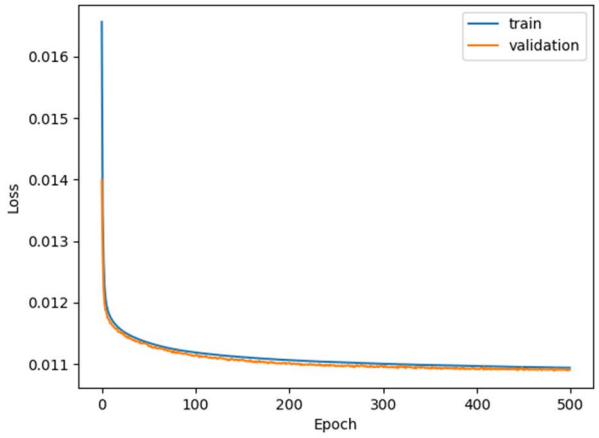


Figure 6: Results obtained with input dataset (meter) with 1,411,619 parameters

In our research, we tried to further reduce the parameters in the proposed neural network for lighter solutions, to be implemented in microcontrollers with low resources and in particular with memory (RAM and FLASH). To reach this goal, we tried to reduce the elements in the last two layers (dense layers). After several trials, we realized that the results are still good also deleting last two layers, obtaining a very low neural network parameters (96,547). Figure 7 shows that in this experiment, even if at the beginning loss values are higher than previous test because the neural network is simpler, at the end of the epochs, train and validation curves are very similar and this network reaches similar results. In fact, Table 5 shows that numerical results between the prosed full neural network with 1,411,619 parameters and the proposed reduced neural network with 96,547 reaches similar performances in terms of MAE regression loss.

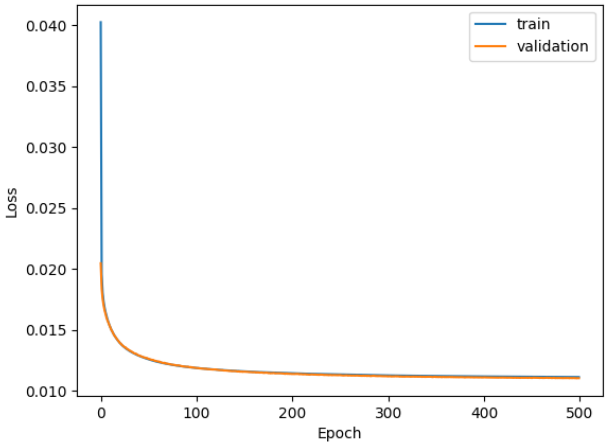


Figure 7: Results obtained with input dataset (meter) with 96,547 parameters

Table 5: Test Results (MAE) with Input Dataset (Meters) (N = 360)

P	Model	Train Loss	Validation Loss	Test Loss
1.4M	MaxPool	0.011	0.011	0.010
96K	MaxPool	0.011	0.011	0.010

Table 6: Test Results (MAE) For (x,y) Only (N=360)

P	Model	Train Loss	Validation Loss	Test Loss
1.4M	MaxPool	0.009	0.009	0.010
1.4M	MaxPool Augmentation	0.009	0.009	0.010
96K	MaxPool	0.010	0.010	0.010
96K	MaxPool Augmentation	0.010	0.009	0.009

Further investigations to increase performance of the proposed neural network have also been carried out. In particular, different normalization techniques (scaling) have been tested in our testing environment, like MinMaxScaler() and StandardScaler(), but no substantial improvement has been obtained. The same behavior (no improvement) has been also noticed changing the function loss (Euclidean distance) and extending the neural network using LSTM layers in replacement of Dense layers.

At this point, we tried to reduce the problem limiting the output of the neural network to only spatial components, i.e., to a vector $T' = [\Delta x, \Delta y]$. Moreover, in this simplified problem version, we also tried an intensive data augmentation, obtained with different strategies. In particular, we take input scans in reverse order, just odds and even scans and finally odds and even scans in reverse order. In this way, we obtained about 204,000 samples, that is about four times the original dataset.

Table 6 summarizes the results obtained by the proposed neural network in the case of simplified problem. Comparing the results with the ones in Table 5, we can note a negligible improvement in train and validation loss, but not in the test loss, so results are almost the same. Even with the intensive augmentation, we obtain a slight but not significant improvement for the simpler neural network with 96,547 parameters.

4. Conclusions

The problem of estimating the moving robot localization, with the only usage of data coming from 2D laser scanner, has been addressed by this paper using a simple deep learning approach. For the dataset used, in the final version composed by 204,000 samples with only 96,547 parameters, the proposed neural network achieved good precision performance in terms of Mean Absolute Error (MAE): one centimeter in translation and one degree in rotation. We also tried to increase performance of the proposed network using different strategies and parameters, but no significant improvements have been obtained.

Even if the encouraging results presented here are almost comparable with classical localization estimation approaches, at moment the approaches based on deep learning could be used in replacement of other state-of-the-art algorithms, since the latter ones are more flexible and can potentially reach better performances. Anyway, the overexposed approach remains a good proof of concept and it can be used for future explorations.

Furthermore, the proposed neural network can be used as interesting complement or can be integrate in classic localization methods, since it works in real-time, taking less than 130 μs to elaborate each estimation on Titan X (Pascal) GPU produced by Nvidia.

Conflict of Interest

None of the authors have any kind of conflict of interest related to the publication of the proposed research.

References

[1] G. Spampinato, A. Bruna, I. Guarneri, D. Giacalone, "Deep Learning Localization with 2D Range Scanner," International Conference on Automation, Robotics and Applications (ICARA), 206-210, 2021, DOI: 10.1109/ICARA51699.2021.9376424.

[2] G. Spampinato, A. Bruna, D. Giacalone, G. Messina, "Low Cost Point to Point Navigation System," International Conference on Automation, Robotics and Applications (ICARA), 195-199, 2021, DOI: 10.1109/ICARA51699.2021.9376545.

[3] T. Sattler, B. Leibe, L. Kobbelt, "Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization," IEEE Transaction on Pattern Analysis and Machine Intelligence, 1744-1756, 2017, DOI: 10.1109/TPAMI.2016.2611662.

[4] R. Mur-Artal, J. M. M. Montiel, J. D. Tardos, "ORB-SLAM: a versatile and accurate monocular SLAM system," IEEE Transactions on Robotics, 31(5), 1147-1163, 2015, DOI: 10.1109/TRO.2015.2463671.

[5] C. Forster, M. Pizzoli, D. Scaramuzza. "SVO: Fast semi-direct monocular visual odometry," IEEE International Conference on Robotics and Automation (ICRA), 15-22, 2014, DOI: 10.1109/ICRA.2014.6906584.

[6] B. Zeisl, T. Sattler, M. Pollefeys, "Camera Pose Voting for Large-Scale Image-Based Localization," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2704-2712, 2015, DOI: 10.1109/ICCV.2015.310.

[7] H. Kim, D. Lee, T. Oh, H. Myung, "A Probabilistic Feature Map-Based Localization System Using a Monocular Camera," Sensors, **15**(9), 21636-21659, 2015, DOI: 10.3390/s150921636.

[8] E. Olson, "M3rsm: Many-to-many multi-resolution scan matching," IEEE International Conference on Robotics and Automation (ICRA), 5815-5821, 2015, DOI: 10.1109/ICRA.2015.7140013.

[9] G. D. Tipaldi, K. O. Arras, "Flirt-interest regions for 2d range data," IEEE International Conference on Robotics and Automation (ICRA), 3619-3622, 2010, DOI: 10.1109/ROBOT.2010.5509864.

[10] S.I. Roumeliotis, G. A. Bekey, W. Burgard, S. Thrun, "Bayesian estimation and Kalman filtering: A unified framework for mobile robot localization," Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2985-2992, 2000, DOI: 10.1109/ROBOT.2000.846481.

[11] S. Park, K. S. Roh, "Coarse-to-Fine Localization for a Mobile Robot Based on Place Learning With a 2-D Range Scan," IEEE Transactions on Robotics, 528-544, 2016, DOI: 10.1109/TRO.2016.2544301.

[12] P. Besl, H.D. McKay, "Method for registration of 3-D shapes," Sensor Fusion IV: Control Paradigms and Data Structures, International Society for Optics and Photonics, 239-256, 1992, DOI: 10.1109/34.121791.

[13] F. Pomerleau, F. Colas, R. Siegwart, S. Magnenat, "Comparing icp variants on real-world data sets," Autonomous Robots, **34**(3), 133–148, 2013, DOI: 10.1007/s10514-013-9327-2.

[14] C. Friedman, I. Chopra, O. Rand, "Perimeter-based polar scan matching (PB-PSM) for 2D laser odometry," Journal of Intelligent and Robotic Systems: Theory and Applications, **80**(2), 231–254, 2015, DOI: 10.1007/s10846-014-0158-y.

[15] J. Zhang, S. Singh, "LOAM: Lidar Odometry and Mapping in Real-time," Robotics: Science and Systems, 109-111, 2014, DOI: 10.15607/RSS.2014.X.007.

[16] A. Kendall, M. Grimes, R. Cipolla. "Posenet: A convolutional network for real-time 6-dof camera relocalization," Proceedings of the IEEE International Conference on Computer Vision, 2938-2946, 2015, DOI: 10.1109/ICCV.2015.336.

[17] S. Wang, R. Clark, H. Wen, N. Trigoni, "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," IEEE International Conference on Robotics and Automation (ICRA), 2043-2050, 2017, DOI: 10.1109/ICRA.2017.7989236.

[18] R. Li, S. Wang, Z. Long, D. Gu, "Undeepvo: Monocular visual odometry through unsupervised deep learning," IEEE International Conference on

Robotics and Automation (ICRA), 7286-7291, 2018, DOI: 10.1109/ICRA.2018.8461251.

[19] H. M. Cho, H. Jo, S. Lee, E. Kim, "Odometry Estimation via CNN using Sparse LiDAR Data," International Conference on Ubiquitous Robots (UR), 124-127, 2019, DOI: 10.1109/URAI.2019.8768571.

[20] M. Pfeiffer, M. Schaeuble, J. Nieto, R. Siegwart, C. Cadena, "From perception to decision: A data-driven approach to end-toend motion planning for autonomous ground robots," IEEE International Conference on Robotics and Automation (ICRA), 1527-1533, 2017, DOI: 10.1109/ICRA.2017.7989182.

[21] J. Li, H. Zhan, B. M. Chen, I. Reid, G. H. Lee, "Deep learning for 2D scan matching and loop closure," International Conference on Intelligent Robots and Systems (IROS), 763-768, 2017, DOI: 10.1109/IROS.2017.8202236.

[22] M. Valente, C. Joly, A. de La Fortelle, "An LSTM Network for Real-Time Odometry Estimation," IEEE Intelligent Vehicles Symposium (IV), 1434-1440, 2019, DOI: 10.1109/IVS.2019.8814133.

[23] M. Velas, M. Spanel, M. Hradis, A. Herout, "CNN for IMU assisted odometry estimation using velodyne LiDAR," IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), 71-77, 2018, DOI: 10.1109/ICARSC.2018.8374163.

[24] S. Xu, W. Chou, H. Dong, "A Robust Indoor Localization System Integrating Visual Localization Aided by CNN-Based Image Retrieval with Monte Carlo Localization," Sensors, **19**(2), 249, 2019, DOI: 10.3390/s19020249.

[25] W. Hess, D. Kohler, H. Rapp, D. Andor, "Real-time loop closure in 2d lidar slam," IEEE International Conference on Robotics and Automation (ICRA), 1271-1278, 2016, DOI: 10.1109/ICRA.2016.7487258.

[26] C. M. Bishop, "Neural Networks for Pattern Recognition," Oxford University Press, 296-298, 1996, ISBN:978-0198538646.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 4, 355-367 (2021)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Enhance Student Learning Experience in Cybersecurity Education by Designing Hands-on Labs on Stepping-stone Intrusion Detection

Jianhua Yang[1], Lixin Wang[*,1], Yien Wang[2]

[1]*TSYS School of Computer Science, Columbus State University, Columbus, GA 31907, USA*

[2]*College of Engineering, Auburn University, Auburn, AL 30060, USA*

ARTICLE INFO

Article history:
Received: 07 June, 2021
Accepted: 09 August, 2021
Online: 26 August, 2021

Keywords:
Stepping-stone
Intrusion Detection
Cybersecurity Curriculum
Ethical Hacking
Hands-on Experience

ABSTRACT

Stepping-stone intrusion has been widely used by professional hackers to launch their attacks. Unfortunately, this important and typical offensive skill has not been taught in most colleges and universities. In this paper, after surveying the most popular detection techniques in stepping-stone intrusion, we develop 10 hands-on labs to enhance student-learning experience in cybersecurity education. The goal is not only to teach students offensive skills and the techniques to detect and prevent stepping-stone intrusion, but also to train them to be successfully adaptive to the fast-changing dynamic cybersecurity world.

1. Introduction

1.1. Cybersecurity Significance

We live in a world where digital technologies are needed for various daily activities. The Internet has revolutionized data communications and significantly changed our daily lives. However, hackers can now easily launch cyberattacks using the Internet. As cyberattacks continue to grow, it is important to secure our critical infrastructures, organizations, business and networks.

1.2. The Importance of Stepping-stone Intrusion Detection

Intrusion techniques are widely used by intruders to invade a computing system. Intrusion detection systems (IDS) are installed on a lot of computer and network systems. Intruders tend to use several compromised hosts, called stepping-stones, to send attacking commands to a remote target host, in order to avoid being detected. Attacks that are launched through a chain of stepping-stone host are called stepping-stone intrusion. With a stepping-stone attack, intruders remotely login to such stepping-stones using tools such as SSH, rlogin, or telnet, and then send the attacking packets to the remote target host.

In this paper, after the survey of many known detection techniques for the stepping-stone intrusion, we propose ten hands-on labs which are developed based on the cutting-edge techniques in stepping-stone intrusion detection. The goal is to help students to learn the techniques of stepping-stone intrusion detection. We aim at educating learners to be qualified professionals in cybersecurity in order to defend various digital data and resources. It is also expected to enhance students' learning in cybersecurity education by conducting the hands-on labs designed.

2. Key Challenges

Before designing the hands-on labs on stepping-stone intrusion and its detection, we discuss how challenge the known detection approaches for stepping-stone intrusion are integrated into cybersecurity curricula. In order to educate learners to be qualified professionals in cybersecurity, it is necessary to teach offensive skills in college cybersecurity major curriculum.

Integrating stepping-stone intrusion and its detection techniques into cybersecurity curriculum can make us move forward a big step to achieve this goal. Although a great number of detection approaches for stepping-stone intrusion have been proposed since the emerging of the Internet, there are still a lot of challenges to integrate these detection approaches into cybersecurity curricula at the college level. The first challenge is why we need to teach college students ethical hacking skills. Would it be possible educate our students to become a hacker against us, not for us? The second challenge is that, since there are

[*]Corresponding Author: Lixin Wang, 4225 University Ave., Columbus, GA 31907, USA. Contact No: 001-706-507-8190. Wang_Lixin@ColumbusState.edu

too many algorithms for stepping-stone intrusion detection proposed in the literature, which approaches among them are suitable to our college students as learning materials? The third challenge is what hands-on labs can be developed and integrated into cybersecurity curriculum. We all know that the difficulty in teaching cybersecurity is not at the delivery of the theory and techniques; it is at the development of hands-on labs for students to practice hacking and defensive skills. Considering the limited budget in each four-year college, the cost is an important factor when designing these hands-on labs. However, we still want to motivate our students to learn cybersecurity skills via hands-on learning experience.

2.1. The Rationale to Teach Ethical Hacking Skills

Should we teach ethical hacking skills to cybersecurity major students? To the best of our knowledge, even though some four-year institutions have included ethical hacking skills as part of their cybersecurity curriculum, there are still some concerns and doubts from students' parents and local communities about the possibility that teaching ethical hacking skills would make their kids to conduct some malicious activities, and commit crimes. We must convince students' parents as well as the local communities with the following advice: 1) the word 'hacker' has long been understood negatively. Hacking actually involves computing skills to find vulnerabilities of a system, penetrate a system, and be able to remove evidence of accessing to a system [1]. Similar to the case that doctors who might criminally abuse their medical skills to hurt humans, a hacker who knows some special offensive hacking skills might also misuse their techniques. However, we should not define the term hacking by its misuse; 2) cybersecurity is a two-edged sword: offensive and defensive. To be effective at defence, students must fully understand the capabilities of hackers and the way how hackers perform cyberattacks; 3) it is widely believed that including both perspectives of "defender" and "attacker" and the related skills could make the cybersecurity curriculum more meaningful and practical [2]. On the other hand, teaching hacking skills can make cybersecurity professionals be equipped with offensive techniques, and well prepared to defend their computing and network system; 4) regardless of teaching hacking skills or not, hackers were out there, and will still be out there. Should hacking skills be integrated into cybersecurity curricula, it would be possible to promote conscious ethical practices and minimize the likelihood that students would misuse the skills.

2.2. Challenging to Integrate the Techniques to a 3-Credit Hours Course

What techniques should be selected to train our students with cybersecurity skills, as there are tons of approaches that have been proposed to detect stepping-stone intrusion since 1995? In a regular course with 48 academic credit hours, it is infeasible to cover all the techniques developed so far, but we do want to train our students not only to have an overall picture of the techniques on stepping-stone intrusion detection, but also to deeply understand some specific and typical intrusion detection approaches. The challenge is to develop contents modules and design hands-on lab exercises. In this paper, we only focus on the designing the hands-on labs on stepping-stone intrusion and its detection. Refer to our prior work [3] for the course modules we developed for integration of detection techniques for stepping-stone intrusion into cybersecurity curricula.

2.3. Challenge on Developing Hands-on Labs of Stepping-stone Intrusion and its Detection

The most difficult part of teaching cybersecurity courses is to design appropriate hands-on labs. We all know the importance of hands-on labs in cybersecurity education. Without the practicing of the techniques covered in cybersecurity class, it is hard to make our students to digest the cybersecurity skills. Conducting cybersecurity hands-on labs needs hardware and software that are more likely not free. Most colleges are equipped with good hardware, such as computers, routers, switches, and different type of servers, but lack of appropriate software. One reason is that some software helping students to practice cybersecurity skills are usually not free, and may be extremely expensive, such as Cyber-range, its price can be as high as more than one million dollars. Therefore, the challenge is how to design appropriate hands-on labs not only can help students to practice stepping-stone intrusion and its detection techniques, but also can reduce the cost to make labs affordable to most colleges.

3. Survey of the Techniques on Stepping-stone Intrusion and its Detection

Many methods have been proposed to detect stepping-stone intrusion. In [4], the authors proposed a thumbprint method to detect stepping-stone intrusion in 1995. This method was developed to compare the contents of TCP/IP packets from the incoming and outgoing sessions of a computer that is chosen to be the sensor for detection. In [5], the authors proposed a detection approach for stepping-stone intrusion by considering the time gaps between the packets captured from the outgoing connection and the incoming connection from a host. In [6], the authors proposed another method for stepping-stone intrusion detection. Their method did not follow the idea of using time-based thumbprints. Instead, the authors in [6] used the deviation between the incoming and outgoing sessions of a computer.

After 2000, a lot more methods were proposed for stepping-stone intrusion detection. One popular approach is to compare the number of packets from the incoming connection with that from the outgoing connection. For the details of this type of approach, please refer to the references [7-9]. A watermark correlation technique was proposed for stepping-stone intrusion detection [10-12]. The idea of using a watermark in stepping-stone intrusion detection is to insert a watermark in the incoming connection of a detection sensor, and then pay attention to the outgoing connections to see if the same watermark can be found in any of these outgoing connections. The rationale used in the papers [10-12] is to analyse and compare the incoming and outgoing connections of a sensor to see if there is any relayed pair. A sensor is defined as a computer host in which all the packets are captured and a detection program runs. If an incoming connection of a sensor is relayed with an outgoing connection, the sensor is considered as a stepping-stone host. However, a user might sometimes use a host as a stepping-stone legitimately due to some special applications. If so, the watermark approach discussed in [10-12] for stepping-stone intrusion detection may produce false positive errors, since this method simply compares an incoming connection with an outgoing one.

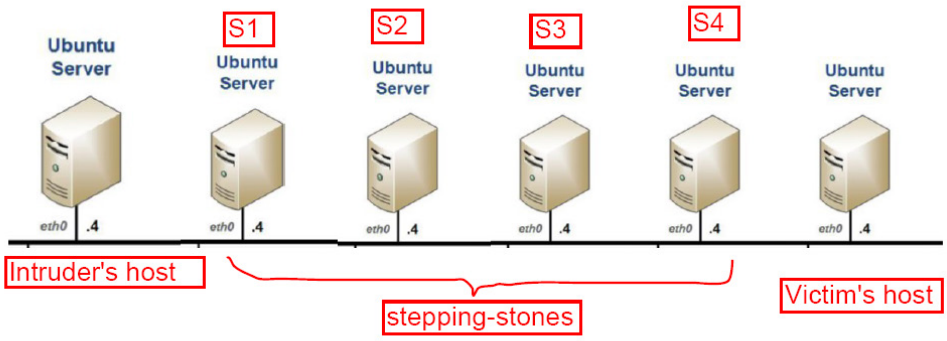


Figure 1: Four Stepping-stone Network Topology

A significant research conducted in [5] has shown that very few professional software employs three or more stepping-stones to access a remote server, although certain legal applications may utilize one or two stepping-stones to access a remote server. Therefore, in order to produce smaller false-positive errors to detect stepping-stone intrusion, an effective method is to estimate the length of a connection chain of stepping-stones. It is extremely challenging to estimate the length of an upper stream connection chain (from the attacker's host to the sensor in the connection chain). Thus, it is impossible to estimate the length of a whole connection chain. By far, most proposed approaches in the literature could only calculate the length of the downstream connection chain (from the sensor to the victim host). This approach to estimate the length of a downstream detection chain was investigated first in [13].

In [13], the authors studied the ratio between the Ack-RTT value and the Echo-RTT. Ack-RTT is defined as the gap between the time to send a packet out and the time to receive its corresponding acknowledgement packet. Echo-RTT is defined as the gap between the time to send a packet out and the time to receive its echo packet. In this way, the length of a downstream connection chain can be approximately estimated. However, this approach could incur false-negative errors.

In [14], the authors proposed a step-function approach motivated by the work that was done in [13] with the purpose of more accurately calculate the length of a downstream connection chain. In [15], the authors proposed another approach by mining network traffic to estimate the number of stepping-stones of a downstream connection chain in 2007. A couple of other methods were also developed in recent years for stepping-stone intrusion detection, including the method using the RTT-based random walk [16], and the method using the idea of RTT Cross-Matching [17].

The stepping-stone intrusion detection approaches have been investigated for about twenty-five years since 1995, unfortunately by far, these important methods have not yet been integrated into cybersecurity curricula at the college level in the U.S. It is vital to educate learners about the known detection approaches for stepping-stone intrusion as more and more professional attackers tend to launch their cyberattacks by using a chain of stepping-stones. Most universities/colleges' professors support to teach the skills and topics of ethical hacking and integrate them into the cybersecurity curricula due to two reasons. First, as far as we know, very few well-educated college students became malicious intruders; second, teaching offensive skills of ethical hacking for college students may produce more and more well-qualified professionals of cybersecurity workforce [18]. We propose ten hands-on labs that allow students to practice in various stepping-stone intrusion detection topics and help them better understand the topics included in the well-designed cybersecurity modules. These hands-on labs will also help enhance students' learning engagement significantly and greatly improve their hands-on experience in cybersecurity.

4. Hands-on Lab Development

Five modules for students to study stepping-stone intrusion and its detection techniques have been proposed and integrated into cybersecurity curriculum [3]. In these five modules, the most popular and the most recently developed techniques have been included. In order to help students to digest the detection and prevention techniques included in the five modules quickly and thoroughly, we design ten hands-on labs as the following,

1) setting up a stepping-stone intrusion connection chain;

2) capturing network traffic;

3) make C# code to capture network traffic;

4) content-based thumbprint detection;

5) time-based thumbprint detection;

6) step-function detection;

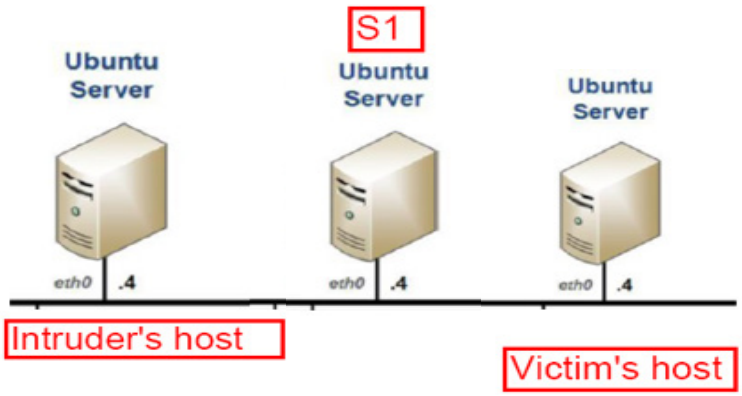
7) packet matching;

8) RTT-based random-walk detection;

9) estimating the length of a long connection chain;

10) intrusion detection using crossover packets.

We apply two rules including relevance and affordability to examine each hands-on lab developed. Relevance means if the lab is closely tied to the modules developed. Affordability means all the labs designed do not use expensive hardware and software. An ideal scenario is that students only need to use the Internet, and free download software to conduct the labs designed.



This designing rule can make it possible for most teaching-focus colleges/universities to offer the labs to cybersecurity majors. Depending on the curriculum design in different institutions, it is not necessary to adopt all the ten labs. However, Lab 1 and Lab 2 are not optional. All the computer hosts used in each lab must be connected in a local area network (LAN). Student must have login credential for each host. All the following labs share the same lab setup as below,

Hardware:

- Each computer must have minimally 4G memory and 500G hard drive capacity.
- Wired or Wireless computer network connection.

Software:

- Ubuntu server or any other type of Linux/Unix installed in each host.
- SSH/OpenSSH client side tool must be installed.
- Each host must have SSH server installed.
- Wireshark, or TcpDump

Login Credentials:

- User Name: Student (Assumed)
- Password: cpsc4166 (Assumed)

All the labs proposed in this paper need students to make a connection chain and to capture TCP/IP packets. A connection chain can be established using OpenSSH under Linux OS which can be a physically installed, or virtual one, such as an OS from VirtualBox, or VMware. It does not need too much memory and second storage. We tried computers with different memory sizes and storage capacity, and found that 4G memory and 500G storage are the minimized requirements. As for the software, TcpDump/Wireshark, SSH client and SSH server package are required minimally.

4.1. Setting up a Stepping-stone Intrusion Connection Chain

4.1.1 Lab objectives

1. Understand TCP/IP protocol; 2. Know how to establish a long interactive connection chain spanning multiple hosts; 3. Understand the concept of Stepping-stones; 4. Obtain the knowledge how an intruder lunches attacks over stepping-stones.

4.1.2 Network topology

It is the same topology as shown in Figure 1.

4.1.3 Lab instructions

1) Start up from any computer in the LAN, and login to a computer that is assumed the Intruder's host with the above credentials.
2) Please open a terminal at the Intruder's host.
3) Browse the current folder, and take a screenshot for the files in the folder.
4) Run SSH to connect to a local host S1: ssh Student@S1 (this can also be the IP address of S1 if host name S1 is not known) in the LAN.
5) As long as connecting to S1, you are prompted to input the password for the user.
6) If connected to S1 successfully, please browse the current folder, and take a screenshot including the folder's name, and all the files in the current folder. Run "ifconfig" to show the IP address and other network related information of S1. Take a screenshot of "ifconfig" results.
7) Compare the screenshot taken at the Intruder's host with the one taken at S1 to see if they are the same.
8) Repeat steps 4), 5), 6) 7) to connect to the computer hosts S2, S3, S4, and the last one respectively. The last host connected is called Victim's host.
9) So far you have locally connected to Victim's host via the hosts S1, S2, S3, and S4. Hosts S1, S2, S3, and S4 are used as stepping-stones.
10) If sniffing the packets at Victim's host, we can see all of the packets are from host S4 other than Intruder's host even though we know all the packets come from the Intruder's host originally. So in this way, intruders can protect themselves via the compromised hosts, such as the hosts S1, S2, S3 and S4.
11) Logout from Victim's host to S4 by typing "Exit" at Victim's

host.

12) Browse the current folder and compare with the screenshot taken at host S4 to see if it is disconnected from Victim's host.

13) Repeat steps 11) and 12) until come back to Intruder's host.

4.1.4 Critical Thinking Practice

1) Ethical Issue Discussion:
Please discuss if there are any ethical issues by making a connection chain across the Internet using legal credentials. How about is it by using illegal credentials?
2) Why do intruders make use of stepping-stones?
3) An interactive session can be encrypted by using SSH. Is it possible to get source IP and destination IP if a TCP/IP packet is captured from such a session? If yes, please tell how? If No, please tell why?
4) Compare to directly access a victim host, is it efficient to access the victim host via some compromised hosts?
5) In the lab, it has five connections in the long interactive session from Intruder's Host to Victim's Host. Each connection is encrypted and set up by using SSH/OpenSSH. Is the encryption key used for the connection from Intruder's Host to S1 the same as the encryption key used for the connection from S1 to S2? Why?

4.2. Capturing Network Traffic

4.2.1 Lab objectives

1. Understand the meaning of each field of a TCP/IP packet header; 2. Know how to store captured packets into different files; 3. Understand the features of TCP, UDP, IP, and ICMP packets; 4. Learn how to use Wireshark to capture network traffic.

4.2.2 Network topology

Refer to Figure 2.

4.2.3 Lab instructions

1) Select any three computer hosts in your local area network, and login to each host with the credentials given.

2) Run "ifconfig" to get the IP address at the three computers respectively and take a screenshot at each host.

3) Follow the instructions in Lab 1 to set up a connection chain as shown in Figure 2. This connection chain spans three computer hosts including Intruder's host, S1, and Victim's host.

4) Type some Linux/Unix commands at Intruder's host to make network traffic from Intruder's host to Victim's host via S1.

5) At S1, run Wireshark to capture TCP packets coming from Intruder's host and leaving to Victim's host only.

6) Store all the packet in Step 5) to a readable file (text file) including timestamp, source IP, destination IP, source Port number, destination Port number, Sequence number, Acknowledgement number, Flag, and Length.

7) At S1, run Wireshark to capture TCP packets coming from Victim's host and going to Intruder's host only. Repeat Step 6).

8) Repeat Steps 5), 6) and 7), but capture UDP packets.

9) Repeat Steps 5), 6) and 7), but capture ICMP packets.

4.2.4 Critical Thinking Practice

1) Ethical Issue Discussion:
2) Would it trigger any ethical issue to capture other users' network traffic under a host with legal login?
3) What is the difference between Display filter and Capture filter in Wireshark?
4) Give the display filter to find the packets of three-way handshake for a connection from host 192. 168.0.1.
5) What is a TCP Send packet?

4.3. Making a Code to Capture Network Traffic

4.3.1 Lab objectives

1. Understand LibPacp package for Linux server; 2. Learn the algorithms to capture computer network traffic; 3. Be able to make C code to capture TCP/IP Packets; 4. Obtain the knowledge to detect network adapters, and open an adapter; 5. Understand the techniques to set up and compile a packet-capturing filter.

4.3.2 Network topology

It has the same network topology as Figure 2 in Lab 4.2.

4.3.3 Mechanism on making the code to sniff network traffic

In order to make a code to capture network packets like what Wireshark does, Libpcap package must be installed in the Ubuntu server. If Windows server is used, please install WinPcap. The way to make a code to sniff computer network traffic is to call the functions built in Libpcap (packet capture) package. Libpcap provides an application-programming interface (API) for capturing network traffic.

We take an example, capturing raw IP packets, to examine the steps to sniff packets by making a program under Linx/Unix system. For the details of the code, please refer to the reference [19]. It has four steps to sniff computer network packets: 1) open a packet capture socket; 2) start packet capture loop; 3) parse and display packets; 4) Terminate capture program.

Open a packet capture socket: A socket is an endpoint for network communication that is identified in a program with a socket descriptor. Opening a packet capture socket involves a series of Libpcap calls that are encapsulated in open_pcap_socket() function. There are a couple of steps needed to open a packet capture socket. The first step is to select a network device using function pcap_lookupdev(). The second step is to open the network device selected for live capture using function pcap_open_live(). The third step is to call function pcap_lookupnet() to get the network address and subnet mask. The fourth step is to compile a packet capture filter by calling function pcap_compile(). The last step is to install the compiled packet filter program into the packet capture device. This causes

```c
pcap_t* open_pcap_socket(char* device, const char* bpfstr)
{
  char errbuf[PCAP_ERRBUF_SIZE];
  pcap_t* pd;
  uint32_t srcip, netmask;
  struct bpf_program bpf;

  // If no network interface (device) is specified, get the first one.
  if (!*device && !(device = pcap_lookupdev(errbuf)))
  {
    printf("pcap_lookupdev(): %s\n", errbuf);
    return NULL;
  }
  // Open the device for live capture, as opposed to reading a packet
  // capture file.
  if ((pd = pcap_open_live(device, BUFSIZ, 1, 0, errbuf)) == NULL)
  {
    printf("pcap_open_live(): %s\n", errbuf);
    return NULL;
  }
  // Get network device source IP address and netmask.
  if (pcap_lookupnet(device, &srcip, &amp;netmask, errbuf) < 0)
  {
    printf("pcap_lookupnet: %s\n", errbuf);
    return NULL;
  }
  // Convert the packet filter expression into a packet filter binary.
  if (pcap_compile(pd, &bpf, (char*)bpfstr, 0, netmask))
  {
    printf("pcap_compile(): %s\n", pcap_geterr(pd));
    return NULL;
  }
  // Assign the packet filter to the given libpcap socket.
  if (pcap_setfilter(pd, &bpf) < 0)
  {
    printf("pcap_setfilter(): %s\n", pcap_geterr(pd));
    return NULL;
  }
  return pd;
}
```

(a)

```c
void capture_loop(pcap_t* pd, int packets, pcap_handler func)
{
  int linktype;
  // Determine the datalink layer type.
  if ((linktype = pcap_datalink(pd)) < 0)
  {
    printf("pcap_datalink(): %s\n", pcap_geterr(pd));
    return;
  }
  // Set the datalink layer header size.
  switch (linktype)
  {
  case DLT_NULL:
    linkhdrlen = 4;
    break;
  case DLT_EN10MB:
    linkhdrlen = 14;
    break;
  case DLT_SLIP:
  case DLT_PPP:
    linkhdrlen = 24;
    break;
  default:
    printf("Unsupported datalink (%d)\n", linktype);
    return;
  }
  // Start capturing packets.
  if (pcap_loop(pd, packets, func, 0) < 0)
    printf("pcap_loop failed: %s\n", pcap_geterr(pd));
}
```

(b)

```c
void parse_packet(u_char *user, struct pcap_pkthdr *packethdr,
          u_char *packetptr)
{
  struct ip* iphdr;
  struct icmphdr* icmphdr;
  struct tcphdr* tcphdr;
  struct udphdr* udphdr;
  char iphdrInfo[256], srcip[256], dstip[256];
  unsigned short id, seq;
  // Skip the datalink layer header and get the IP header fields.
  packetptr += linkhdrlen;
  iphdr = (struct ip*)packetptr;
  strcpy(srcip, inet_ntoa(iphdr->ip_src));
  strcpy(dstip, inet_ntoa(iphdr->ip_dst));
  sprintf(iphdrInfo, "ID:%d TOS:0x%x, TTL:%d IpLen:%d DgLen:%d",
      ntohs(iphdr->ip_id), iphdr->ip_tos, iphdr->ip_ttl,
      4*iphdr->ip_hl, ntohs(iphdr->ip_len));
  packetptr += 4*iphdr->ip_hl;
  switch (iphdr->ip_p)
  {
  case IPPROTO_TCP:
    tcphdr = (struct tcphdr*)packetptr;
    printf("TCP  %s:%d -> %s:%d\n", srcip, ntohs(tcphdr->source),
        dstip, ntohs(tcphdr->dest));
    printf("%s\n", iphdrInfo);
    printf("%c%c%c%c%c%c Seq: 0x%x Ack: 0x%x Win: 0x%x TcpLen: %d\n",
        (tcphdr->urg ? 'U' : '*'),
        (tcphdr->ack ? 'A' : '*'),
        (tcphdr->psh ? 'P' : '*'),
        (tcphdr->rst ? 'R' : '*'),
        (tcphdr->syn ? 'S' : '*'),
        (tcphdr->fin ? 'F' : '*'),
        ntohl(tcphdr->seq), ntohl(tcphdr->ack_seq),
        ntohs(tcphdr->window), 4*tcphdr->doff);
    break;
```

(c)

Figure 3: Packet Capture Sample Code

Libpcap to start collecting the packets with selected filter. The sample code in Figure 3-(a) shows the four steps in opening a packet capture socket.

Start packet capture loop: Libpcap provides three functions to capture packets: pcap_next(), pcap_dispatch(), and pcap_loop(). Since function pcap_next() can only grab one packet at the time to be called. So the program must call this function in a loop to receive multiple packets. The other two functions pcap_loop and pcap_dispatch() can loop automatically to receive multiple packets. Datalink type can be determined by calling pcap_datalink(), and then start packet capture. The sample program shown in Figure 3-(b) uses pcap_loop() to sniff multiple packets. In this code, first to determine the datalink type by calling pcap_datalink(), and then start packet capture loop.

Parse and display packets: The general technique for parsing packets is to set a character pointer to the beginning of the packet buffer then advance this pointer to a particular protocol header by the size in bytes of the header that precede it in the packet. The header can then be mapped to an IP, TCP, UDP, and ICMP header structure by casting the character pointer to a protocol specific structure pointer. A parse_packet() function starts off by defining pointers to IP, TCP, UDP and ICMP header structures. The packet pointer is advanced past the datalink header by the number of bytes corresponding to the datalink type determined in capture_loop(). Casting the packet pointer to struct tcphdr and struct udphdr pointers gives us access to TCP and UDP header fields respectively. The struct icmphdr pointer enables us to display ICMP packet type and code along with the source and destination IP addresses. The sample code in Figure 3-(c) shows the steps to parse and display packets, such as TCP packets that are used to detect stepping-stone intrusion.

Terminate Capturing: The last step is to terminate the packet capture by interrupt signals SIGNIT, SIGTERM, and SIGQUIT through calling function bailout() which displays the packet count, closes the packet capture socket then exits the program.

4.3.5 Lab instructions

1) Start up running your code, and select the interface to sniff
2) Click “Start” button to start packet sniffing
3) Display the following information for each packet captured: source/destination IP address, source/destination port number, packet type, sequence number, acknowledge number, TCP flags, fragmentation information, checksum, receive window, TTL, upper layer protocol, timestamps in format of mm/dd/yy.
4) Click one TCP/IP packet captured to show the details in each of its header field. Take a screenshot for the header details.
5) Store captured packet in a .txt file that can be opened by WordPad, or any other text editor tool.

4.3.4 Critical Thinking Practice

1) Ethical Issue Discussion:
   Would it trigger any ethical issue to capture other users’ network traffic using self-made code under a host with legal login?
2) What is the difference between Winpcap and Libpcap?

3) What functions are called in order to open a packet capture socket?

4) What is the purpose to call pcap_compile()?
5) What is the function of pact_next()?
6) Which function is called to determine the datalink type of a packet?

4.4. Content-based Thumbprint Detection

4.4.1 Lab objectives

1. Understand TCP/IP protocols and network traffic behaviour; 2. Know how to establish an interactive TCP session; 3. Understand using Thumbprint to detect Stepping-stone intrusion; 4. To be familiar with TcpDump and Wireshark.

4.4.2 Network topology

The network topology used in this lab is the same as Figure 2 in Lab 4.2.

4.4.3 Lab instructions

1) Select any three computers in your local area network and name them to be Intruder's host, S1, and Victim's host.

2) Start up the computers in Linux and login to each host with given credentials. Open a terminal in each host.

3) Run "ifconfig" to get the IP address for each host, and take a screenshot from each host.

4) Run SSH from Intruder's host to connect to S1, then to Victim's host just as shown in Figure 2. An interactive session is set up spanning three hosts with S1 working as a Stepping-stone.

5) Students will monitor the traffic of the incoming connection from Intruder's host, and the traffic of the outgoing connection to Victim's host from S1. Here we use the number of TCP packets to represent the corresponding network traffic.

6) Run TcpDump at host S1 to monitor the TCP packets coming to/from Intruder's host but to S1 with destination/source port 22 and store all the packets in IncomingTCP.txt, and also monitor the TCP packets going to Victim's host or come back to S1 with destination/source port 22, and store all the collected packets to OutgoingTCP.txt.

7) In either IncomingTCP.txt or OutgoingTCP.txt, each packet is stored in one row including the following fields separated by ";": Packet Order number; Timestamp; Source IP; Destination IP; Source Port; Destination Port; Flag; Sequence Number; Acknowledge Number; Packet Length

8) Keep operating at Intruder's host for about 15 minutes to make network traffic to Victim's host via S1.

9) Count the number of packets in the two files respectively by counting the number of rows, or just simply check the last row "Packet Order number" field.

10) Compare the two number to see if they are close enough.

11) Identify the Send and Echo packets in the two files. Count the number of Send and Echo packets from IncomingTCP.txt, and denote them as In-S and In-E respectively. Similarly count the number of Send and Echo packets from OutgoingTCP.txt, and denote them as Out-S and Out-E respectively.

12) The rules to determine Send or Echo packet at S1 are as the following,

    a. Send packet is a packet in the incoming link that comes to S1 with Flag.P set up, but in the outgoing link that leaves S1 to Victim's host with Flag.P set up;

    b. Echo packet is a packet in the incoming link that leaves S1 to Intruder's host with Flag.P set up, but in the outgoing link that comes to S1 with Flag.P set up.

13) Compare if the following relation maintains,

    a. In-S is close to Out-S, and

    b. In-E is close to Out-E, and

    c. The sum of In-S and In-E is close to the sum of Out-S and Out-E

14) Please draw your conclusion based on the results from Steps 10) and 13).

4.4.4 Critical Thinking Practice

1) Ethical Issue Discussion:

   If a user has a legal login to a host, captures network packets, and obtains the contents of each packet, would the user's action result in an ethical issue?

2) What is the TcpDump command to sniff the packets in the incoming link?

3) What is the TcpDump command to sniff the packets in the outgoing link?

4) What conclusion you can make based on the information you have in step 10) of the Lab Instructions above? Why?

5) What conclusion you can make based on the information you have in step 13) of the Lab Instructions above? Why?

6) Write a TcpDump command to sniff the packets only acknowledge the requests from Intruders' Host at S1.

4.5. Time-based Thumbprint Detection

4.5.1 Lab objectives

1. Understand using time-based thumbprint to detect stepping-stone intrusion; 2. Learn how to generate time-based thumbprint; 3. Know how to compare time-based thumbprint; 4. Understand the efficiency of thumbprint comparison algorithm.

4.5.2 Network topology

The network topology used in this lab is the same as Figure 2 in Lab 4.2.

4.5.3 Lab instructions

1) Refer to Lab 1 to make an interactive TCP session with at least one host in between attacker and victim machines.
2) On either of the machine of your choice except the target, filter the network capture & save the incoming and outgoing packets including timestamp information for each packet through TcpDump.
3) Examine the packets for the incoming connection and look for the timestamp there and list those timestamps in a sequence.
4) Repeat Step 3 but for the outgoing connection
5) For the incoming connection sequence (list) of timestamps, find the difference in neighboring timestamps and list them in a sequence. This can give a sequence of time gaps for this connection. Find difference using the equation: $|p_i - p_{(i+1)}|$, here p_i is the timestamps of i^{th} packet captured.
6) Repeat Step 5 but for the outgoing connection.
7) Compare the two sequences to get a similarity. If the similarity is larger than a predefined threshold, the host is used as a stepping-stone. Otherwise, not.

4.5.4 Critical Thinking Practice

1) Ethical Issue Discussion:
   If a user has a legal login to a host, captures network packets, and but could not obtain the contents of each packet due to encryption, would the user's action result in an ethical issue?
2) Please describe what a time session-based thumbprint is in your own words.
3) Why would an individual want to perform this method to detect a stepping-stone over other methods?
4) Why do we compare the two sequences of time gaps in our own algorithm as oppose to the Longest Common Subsequence algorithm which can also help to measure similarity?
5) Do you have a better method of comparing the sequences' similarity?
6) Would a time session-based thumbprint be effective with an encrypted connection? If yes, explain why.

4.6. Step-function Detection

4.6.1 Lab objectives

1. Understand packet matching algorithm: First-Match; 2. Learn how to use matched Send and Echo packets to determine the number of compromised hosts; 3. Demonstrate Step-Function algorithm; 4. Illustrate the limits of Step-function detection.

4.6.2 Network topology

The network topology used in this lab is the same as the one shown in Figure 1 of Lab 4.1.

4.6.3 Lab instructions

1) Start up with any computers in the LAN, and login to the Intruder's host, Victim's host, S1, S2, S3, and S4 with the appropriate credentials to make a connection chain.
2) Open a terminal on Intruder's host and S1.
3) On desired sensor host (S1 for initial run), start TcpDump to dump captured packets to a file along with any further options
   a) ###.###.###.###.X is Sensor's IP Address and X is a port number
   b) sudo TcpDump 'tcp[tcpflags] & tcp-push != 0 and host ###.###.###.###.X' -n --number > capturedFile
4) On Intruder's host, Run SSH to connect to a remote host S1: ssh Student@S1 (this can also be the IP address of S1 if host name S1 is not known).
5) As long as S1 is reachable, you will be prompted to input the password for the user "Student".
6) On Intruder's host, repeat steps 4 and 5 replacing S1 with S2, S3, S4, and Victim's host, respectively, to login to further hosts as needed.
7) Interact with Victim's host: browse directories, manipulate files, check available interfaces, etc.
8) End current SSH session and stop TcpDump on the sensor host.
9) Repeat steps 3-8 for multiple setups; such as two/three stepping-stones chains with the sensor on different steps each time
10) You may want to use grep to create two files: one for Send packets and one for Echo. Consider that [^\2\]{2,} matches 22 for SSH
    a) (grep -E '>/b###.###.###.###.[^\2\]{2,}'/bcapturedFile) > downEchoFile
    b) (grep -E '###.###.###.###.[^\2\]{2,}/b>'/bcapturedFile) > downSendFile
11) Use First-Match Algorithm to match Send/Echo Packets:

a) Iterate through both lists, starting with the lowest sequence numbered Send Packet

b) If the current packet is a Send, add it to a list of unmatched Send packets

c) If it is an Echo and there is at least one unmatched Send Packet, Search the list of unmatched Send packets from the beginning. Find the first send packet with an appropriate acknowledgement number [Echo.Seq == Send.Ack].

d) Use the absolute difference between the correct Echo's and Send's timestamps to determine the round trip time (RTT) of the request [RTT = |Echo.Timestamp – Send.Timestamp|]

e) Save RTT to a list of RTTs

f) If it is an Echo and all preceding Send packets have been matched, the algorithm fails. Check if a packet was missed, then try to determine what may have occurred.

12) Sketch the graph of RTT vs. Number of matched Packets

a) RTT in whatever unit of time (typically ms or μs);

b) Number of matched packets indexed from 1 to the number of matches.

4.6.4 Critical Thinking Practice

1) Ethical Issue Discussion:

If a user has a legal login to a host, captures network packets, obtains the round-trip time between matched Send and Echo packets, but could not identify the contents of each packet due to encryption, would the user's action result in an ethical issue?

2) What is the purpose of tcp-push != 0 in the above capture?

3) Explain the difference in the grep statements listed above. Why does the first point to Send packets, while the second points to Echo packets?

4) Did you notice any effects to performance (positive/negative) as more links were introduced to the connection chain? Explain.

5) Would there be any difference to this analysis if the data were clear text, sent using Telnet, or encrypted like in SSH? Justify.

6) Can you determine the length of the entire connection chain with this method? If so, explain why. If not, which portion can you determine the length?

4.7. Packet Matching

4.7.1 Lab objectives

1. Understand the significance of packet matching; 2. Determine the differences in the different packet matching algorithms; 3. Learn how to apply packet matching to detect stepping-stone intrusion; 4. Distinguish the limits of different packet matching algorithms.

4.7.2 Network topology

The network topology used in this lab is the same as the one shown in Figure 2 of Lab 4.2.

4.7.3 Lab instructions

1) Start up any computers in the LAN, and login to the computer, which assumes to be called Intruder's host with the above credentials.

2) On desired sensor host (S1 for initial run), start TcpDump to dump captured packets to a file along with any further options

a) ###.###.###.###.X is Sensor IP Address and X is port number

b) sudo TcpDump 'tcp[tcpflags] & tcp-push != 0 and host ###.###.###.###.X' -n --number > capturedFile

3) Make an SSH connection chain from Intruder's host through any stepping-stone saying host S1 (sensor) to Victim's host.

4) Interact with Victim's host from Intruder's host via the connection chain: browse directories, manipulate files, check available interfaces, etc.

5) Terminate the SSH chain by using the 'exit' command on each of the stepping-stones and Victim's host from the shell of Intruder's host

6) You may want to use grep to create two files: one for Send packets and one for Echo

a) Upstream
   i. (grep '###.###.###.###.X/b>'/b</bcapturedFile) > upEchoFile
   ii. (grep '>/b###.###.###.###.X'/b</bcapturedFile) > upSendFile

b) Downstream – consider that [^\2\]{2,} matches 22 for SSH
   i. (grep -E '>/b###.###.###.###.[^\2\]{2,}'/bcapturedFile) > downEchoFile
   ii. (grep -E '###.###.###.###.[^\2\]{2,}/b>'/bcapturedFile) > downSendFile

7) Use First-Match Algorithm to match Send/Echo Packets:

a) Iterate through both lists, starting with the lowest sequence numbered Send Packet

b) If the current packet is a Send, add it to a list of unmatched Send packets

c) If it is an Echo and there is at least one unmatched Send Packet, Search the list of unmatched Send packets from the beginning. Find the first send packet with an appropriate acknowledgement number [Echo.Seq == Send.Ack].

d) Use the absolute difference between the correct Echo's and Send's timestamps to determine the round trip time (RTT) of the request [RTT = $|Echo.Timestamp - Send.TimeStamp|$]

i. Save RTT to a list of RTTs

e) If it is an Echo and all preceding Send packets have been matched, the algorithm fails. Check if a packet was missed, then try to determine what may have occurred.

8) Use the Conservative Algorithm to match Send/Echo Packets:

a) Iterate through both lists, starting with the lowest sequence numbered Send Packet

b) If the current packet is a Send:

i. If previous packet was Send and time gap was 1 second or more, clear the sendQ and make a note of match-flag = true

ii. Otherwise, add it to a list of unmatched Send packets

c) If it is an Echo:

i. If there is at least one unmatched Send Packet and match-flag = true, search the list of unmatched Send packets from the beginning. Find the first send packet with an appropriate acknowledgement/sequence number [Echo.Seq == Send.Ack && Echo.Ack > Send.Seq].

1. Use the absolute difference between the correct Echo's and Send's timestamps to determine the round trip time (RTT) of the request [RTT = $|Echo.Timestamp - Send.TimeStamp|$]

a. Save RTT to a list of RTTs

ii. Otherwise, set match-flag = false

9) Use the Greedy Heuristic Algorithm to match Send/Echo Packets:

a) Iterate through both lists, starting with the lowest sequence numbered Send Packet

b) If the current packet is a Send:

i. If previous packet was Send and time gap was 1 second or more, clear the sendQ

ii. Otherwise, add it to a list of unmatched Send packets

c) If it is an Echo:

i. If there is at least one unmatched Send Packet, search the list of unmatched Send packets from the beginning. Find the first send packet with an appropriate acknowledgement/sequence number [Echo.Seq == Send.Ack && Echo.Ack > Send.Seq].

1. Use the absolute difference between the correct Echo's and Send's timestamps to determine the round trip time (RTT) of the request [RTT = $|Echo.Timestamp - Send.TimeStamp|$]

a. Save RTT to a list of RTTs

ii. Otherwise, no match detected.

4.7.4 Critical Thinking Practice

1) Ethical Issue Discussion:
If a user has a legal login to a host, captures network packets, matches each Send packet with its corresponding Echo, but could not identify the contents of each packet due to encryption, would the user's action result in an ethical issue?

2) Which two TCP packet types can we exploit to properly match packets within a connection?

3) Explain why Echo.Seq == Send.Ack is used.
4) Explain why Echo.Ack > Send.Seq is used.
5) Why does Conservative Algorithm clear the Send Queue?
6) Looking at the results of running through the algorithms, what differences do you see between them? Explain why that might be.

4.8. RTT-based Random-walk Detection

4.8.1 Lab objectives

1. Understand random-walk model; 2. Learn how to apply random-walk model to detect stepping-stone intrusion; 3. Be familiar with the techniques to evade detection; 4. Demonstrate using RTT to resist intruders' evasion.

4.8.2 Network topology

The network topology used in this lab is the same as the one shown in Figure 2 of Lab 4.2.

4.8.3 Lab instructions

1) Refer to Lab 1 to make an interactive TCP session including at least one stepping –stone host that is used as a sensor.

2) On the sensor, filter the network capture & save the incoming and outgoing packets through TcpDump.

3) Examine the packets for the incoming connection, and match the Send & Echo packets using conservative packet matching algorithm from Lab 4.7, and obtain the number of RTTs from matched packets for this connection, N^{RTT}_{in}.

4) Repeat Step 3) for the packets collected from the outgoing connection, and obtain N^{RTT}_{out}.

5) Take the difference of N^{RTT}_{in} and N^{RTT}_{out}. $N^{RTT}_{in\text{-}out} = |N^{RTT}_{in} - N^{RTT}_{out}|$

6) Compare $N^{RTT}_{in\text{-}out}$ to a predefined upper bound. If it is less than the upper bound, then the incoming & outgoing connections are a relayed pair. The sensor is used as a stepping-stone. If not then, the machine is not used as a stepping-stone.

4.8.4 Critical Thinking Practice

1) Ethical Issue Discussion:
If a user has a legal login to a host, captures network packets, obtains the round-trip time between matched Send and Echo packets, but could not identify the contents of each packet due to encryption, would the user's action result in an ethical issue?
2) Please describe how a RTT-based Random-Walk Detection works in your own words.
3) Why would an individual want to perform this method to detect a stepping-stone over other methods?
4) Could an intruder manipulate this approach to give a false negative?
5) Would this method be effective with an encrypted connection? If yes, explain why.

6) Perform a network capture by following the above instructions with the predefined threshold, Γ, being equal 30. From the results, is the machine a stepping-stone?

4.9. Detection by Estimating the Length of a Long Connection Chain

4.9.1 Lab objectives

1. Understand the RTTs of the packets from the same connection chain can be mined to the same cluster; 2. Learn the number of compromised hosts is equal to the number of outstanding clusters; 3. Demonstrate the approach to estimate the length of a connection chain; 4. Obtain the knowledge on how clustering-partitioning algorithm can resist intruders' evasion.

4.9.2 Network topology

The network topology used in this lab is the same as the one shown in Figure 1 of Lab 4.1.

4.9.3 Lab instructions

1) Start up any computers in the LAN, and login to the computer that assumes to be called Intruder's host with the above credentials.
2) We will use at least 5 hosts in this connection chain. Decide which 5 hosts you want to use, and designate the 2nd host as a sensor host
3) On the sensor host, begin packet capture prior to making any of the connections.
4) Please open a terminal at Intruder's host.
5) Run SSH to connect to a remote host S1 (sensor host): ssh Student@S1 (this can also be the IP address of S1 if host name S1 is not known).
6) As long as connected to S1, you must be prompted to input the password for the user.
7) Repeat steps 4), 5), to connect to computer hosts S2, S3, S4, and the last one respectively. The last host you connect to remotely is called Victim's host.
8) So far you have remotely connected to Victim's host spanning hosts S1, S2, S3, and S4. Hosts S1, S2, S3, and S4 are used as stepping-stones in this lab.
9) Generate traffic to be captured by sensor. (ls, pwd, etc.)
10) After complete the packet capture, analyse the packets captured using clustering-partitioning algorithm. For the algorithm details, please refer to the reference [20].

4.9.4 Critical Thinking Practice

1) Ethical Issue Discussion:
   If a user has a legal login to a host, captures network packets, obtains the round-trip time between matched Send and Echo packets, and can estimate how many connections between the current host and the end of the connection chain. If the user could not identify the contents of each packet due to encryption, would the user's action result in an ethical issue?
2) Why is it important to begin packet capture before you initiate the connection chain? Please explain.
3) What results are we looking for after completing the clustering-partitioning algorithm? Why do these results indicate connections?
4) What is the maximum theoretical complexity of the partitioning clustering algorithm? Why is this algorithm likely never reach this complexity level? Please explain.
5) What percentage of the RTTs should be within a cluster to be considered a valid cluster?
6) If we collected 720 send packets and 810 echo packets, at most, how many comparisons would be necessary for partitioning-clustering algorithm?

4.10. Detection Using Crossover Packets

4.10.1 Lab objectives

1. Understand crossover packets; 2. Know the reason of generating crossover packets; 3. Obtain the relation between the length of a connection chain and the number of crossover packets; 4. Learn how to identify crossover packets.

4.10.2 Network topology

The network topology used in this lab is the same as the one shown in Figure 1 of Lab 4.1.

4.10.3 Lab instructions

We assume Intruder's Host is called iHost, and Victim's host is called vHost. After a connection chain is established, please type the following information at iHost to make some network traffic for each of the following: "This is s test from Hands-on lab 10. Please discard all the wrong messages!"

1) Make a connection chain from iHost to vHost via S1 only. Type the above information at iHost and capture Send and Echo packets at S1 from its outgoing connection. Store the packets to PacketFile1.
2) Make another connection chain from iHost to vHost, but via S1 and S2. Type the above information at iHost and capture Send and Echo packets at S1 from its outgoing connection. Store the packets to PacketFile2.
3) Make the third connection chain from iHost to vHost, but via S1, S2, and S3. Type the above information at iHost and capture Send and Echo packets at S1 from its outgoing connection. Store the packets to PacketFile3.
4) Make the fourth connection chain from iHost to vHost, but via S1, S2, S3, and S4. Type the above information at iHost and capture Send and Echo packets at S1 from its outgoing connection. Store the packets to PacketFile4.
5) Count the number Crossover packets in each file and compare them. Please conclude what you would find from the comparing the results.

4.10.4 Critical Thinking Practice

1) Ethical Issue Discussion:
   If a user has a legal login to a host, captures network packets, obtains the crossover packets, but could not identify the contents of each packet due to encryption, would the user's action result in an ethical issue?
2) Why is it unlikely that you will observe much, if any, Crossover in a LAN environment?
3) Does increasing the connection chain length increase or decrease the likelihood of observing packet Crossover? Why or why not?
4) Does packet Crossover help or hinder packet matching? Why?
5) Why are you more likely to observe packet Crossover in a WAN environment?
6) What information about a connection chain can you gather from detecting many packet Crossovers?

5. Discussion on the Labs Designed

In this session, we will discuss the innovation, contribution, and the effectiveness of the proposed work.

All the hands-on labs were designed based on some research papers. To the best of our knowledge, the is the first time that stepping-stone intrusion detection techniques are integrated into cybersecurity curriculum. The contribution is that college students can learn complex stepping-stone intrusion detection techniques and enhance their experience by conducting the hands-on labs. The labs designed are suitable for teaching-focus colleges who may have limited budget for their cybersecurity curriculum.

Each lab proposed has a critical thinking practice component including discussions about ethical issues, and the questions to train students to be qualified professionals of cybersecurity workforce. Most of the labs proposed were adopted in the course of "Intrusion Detection and Prevention" at Columbus State University, GA from 2018 to 2019. The instructors did class survey to ask the students if they agree with the labs adopted for the class. The survey results are shown in Table 1.

Table 1: Lab Survey Results

Item / Semester	Strongly Agree	Agree	Neutral	Disagree	Agree and Neutral Rate	Attending Rate
Spring 2018	5	4	3	1	92.3%	13 out of 15 = 86.7%
Spring 2019	11	9	6	0	100%	26 out of 28 = 92.9%
Spring 2020	9	5	2	2	88.88%	18 out of 19 = 94.7%
Spring 2021	11	11	4	2	92.9%	28 out of 29 = 96.6%
Average Rate					93.52%	92.73%

From the survey results, we can see that over four years, more than 90 percent of the students like the labs. Their comments and feedback are positive. There are also some negative comments and feedback. The following are some negative feedback extracted from the surveys:1) the time given to finish the labs are not enough; 2) most students prefer to use a physically installed Linux system to conduct the lab, other than a virtual Linux system because it is hard to copy the results out; 3) too many packets are required to capture which costs their too much time; 4) some students expect to have the first lab to refresh the Linux command, other than to make a connection chain.

6. Summary

In order to help college students to learn stepping-stone intrusion detection and prevention techniques and enhance their hands-on learning experience, we developed ten hands-on labs based on the significant results published in the area of stepping-stone intrusion detection since 1995. For making these hands-on labs be easily adopted by university professors in undergraduate cybersecurity courses, we used the following strategies while designing these hands-on labs: 1) save budgets for learners; 2) simplify the requirements for required hardware and software; 3) clear step-by-step instructions; 4) easy assessments by evaluators; 5) easy adoption by instructors.

Most of the hands-on labs we designed in this paper have been adopted in the undergraduate course of Intrusion Detection and Prevention at Columbus State University for four years. The average survey result shows that more than 90% of the students liked the labs and enjoyed the hand-on activities involved in the labs. The rate of disagreement/dislike is less than 10%. All the hands-on labs have been shared within the USA via the Clark system managed by Towson University, MD, USA. Records show that at least six colleges/universities downloaded the hands-on labs. We highly believe that our proposed hands-on labs in stepping-stone intrusion detection will help building the nation's cybersecurity workforce.

Cybersecurity is a rapidly changing and expending field. In order to make our students to be adaptable with fast changing cybersecurity techniques quickly after graduation, in the future, we will improve the proposed hands-on labs following NICE cybersecurity workforce framework initiated by NIST. In this framework, there are seven categories and each category contains one or more specialty areas. Each cybersecurity specialty area is composed of multiple work roles. Each work role includes Knowledge, Skills and Abilities (KSAs) and Tasks. The future hands-on labs will help our students to achieve three targets. First, they will obtain a body of information, which can be directly applied to the performance of a function. Second, they will enhance their skills needed for cybersecurity. Third, they will improve their competence to perform an observable behavior, which can result in an observable product.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work of Drs. Lixin Wang and Jianhua Yang is supported by National Security Agency (NSA) NCAE-C research grant H98230-20-1-0293 with Columbus State University, Columbus GA 31907, USA.

References

[1] P. Logan, A. Clarkson, "Teaching students to hack: curriculum issues in information security," Special Interest Group on Computer Science Education Symposium, St. Louis, MO USA, 2005.

[2] S. Bratus, A. Shubina, M.E. Lacasto, "Teaching the principles of the hacker curriculum to undergraduates," SIGCSE' 10, Milwaukee, Wisconsin USA, 2010.

[3] J. Yang, Y. Zhang, G. Zhao, "Integrate stepping-stone intrusion technique into cybersecurity curriculum," the Proceedings of 31st IEEE International Conference on Advanced Information Networking and Applications, Taipei, Taiwan, published in IEEE proceedings and Digital Library, 1-6, 2017, doi: 10.1109/WAINA.2017.29.

[4] S. Staniford-Chen, L.T. Heberlein, "Holding intruders accountable on the Internet," Proceedings of IEEE Symposium on Security and Privacy, Oakland, CA USA, 39-49, 1995, doi: 10.1109/SECPRI.1995.398921.

[5] Y. Zhang, V. Paxson, "Detecting stepping-stones," Proceedings of the 9th USENIX Security Symposium, Denver, CO USA, 67-81, 2000.

[6] K. Yoda, H. Etoh, "Finding connection chain for tracing intruders," Proceedings of 6th European Symposium on Research in Computer Security, Toulouse, France, Lecture Notes in Computer Science, 31-42, 2000.

[7] A. Blum, D. Song, S. Venkataraman, "Detection of interactive stepping-stones: algorithms and confidence bounds," Proceedings of International Symposium on Recent Advance in Intrusion Detection, Sophia Antipolis, France, 20-35, 2004.

[8] D.L. Donoho, "Detecting pairs of jittered interactive streams by exploiting maximum tolerable delay," Proceedings of 5th International Symposium on Recent Advances in Intrusion Detection, Zurich, Switzerland, 45-59, 2002.

[9] T. He, L. Tong, "Detecting encrypted stepping-stone connections," Proceedings of IEEE Transaction on signal processing, **55**(5), 1612-1623, 2007, doi: 10.1109/TSP.2006.890881.

[10] X. Wang, D.S. Reeves, S.F. Wu, J. Yuill, "Sleepy watermark tracing: an active network-based intrusion response framework," Proceedings of 16th International Conference on Information Security (IFIP/Sec'01), 369-384, 2001.

[11] X. Wang, D.S. Reeves, "Robust correlation of encrypted attack traffic through stepping-stones by manipulation of interpacket delays," Proceedings of ACM CCS '03, 2003.

[12] X. Wang, "The loop fallacy and serialization in tracing intrusion connections through stepping-stones," Proceedings of the 2004 ACM Symposium on Applied Computing, ACM Press, 2004.

[13] K.H. Yung, "Detecting long connecting chains of interactive terminal sessions," Proceedings of International Symposium on Recent Advance in Intrusion Detection (RAID), Zurich, Switzerland, 1-16, 2002.

[14] J. Yang, S.-H.S. Huang, "A real-time algorithm to detect long connection chains of interactive terminal sessions," Proceedings of 3rd ACM International Conference on Information Security (Infosecu'04), Shanghai, China, 198-203, 2004.

[15] J. Yang, S.-H.S. Huang, "Mining TCP/IP packets to detect stepping-stone intrusion," Journal of Computers and Security, Elsevier Ltd., **26**, 479-484, 2007, doi: 10.1016/j.cose.2007.07.001.

[16] J. Yang, Y. Zhang, "RTT-based random walk approach to detect stepping-stone intrusion," Proc. of 29th IEEE International Conference on Advanced Information Networking and Applications, Gwangju, South Korea, 558-563, 2015, doi: 10.1109/AINA.2015.236.

[17] J. Yang, "Resistance to chaff attack through TCP/IP packet cross-matching and RTT-based random walk," Proceedings of 30th IEEE International Conference on Advanced Information Networking and Applications, Crans-Montana, Switzerland, IEEE proceedings and Digital Library, 784-789, 2016, doi: 10.1109/AINA.2016.17.

[18] Z. Trabelsi, W. Ibrahim, "A hands-on approach for teaching denial of service attacks: a case study," Journal of information technology education: Innovations in Proactive, **12**, 299-319, 2013.

[19] J. Yang, L. Wang, B. Lockerbie, A. Lesh, "Manipulating network traffic to evade stepping-stone intrusion detection," Internet of Things, Elsevier, **3**(4), 34-45, 2018, doi: 10.1016/j.iot.2018.08.011.

[20] J. Yang, S.-H.S. Huang, M.D. Wan, "A clustering-partitioning algorithm to find TCP packet round-trip time for intrusion detection," Proceedings of 20th IEEE International Conference on Advanced Information Networking and Applications (AINA 2006), Vienna, Austria, **1**, 231-236, 2006, doi: 10.1109/AINA.2006.13.