

ASTES

Advances in Science, Technology & Engineering Systems Journal

Special Issue

Computing, Engineering and Multidisciplinary Sciences

2024

www.astesj.com
ISSN: 2415-6698

EDITORIAL BOARD
(Special Issue)

Editor-in-Chief

Prof. Passerini Kazmerski
Pritzker School of Molecular Engineering, University of Chicago, USA

Guest Editors

Prof. Wang Xiu Ying
Chongqing University, China

Prof. Yu Xiao Yan
Chongqing Normal University, China

Prof. María Jesús Espinosa Trujillo
Universidad Tecnológica Metropolitana, Mexico

Prof. Ahmad Yusairi Bani Hashim
Universiti Teknikal Malaysia Melaka, Malaysia

Prof. Mohamed Abdelaziz Hassan Eleiwa
University of Hail, KSA

Prof. Nicolae Tudoroiu
John Abbott College, Canada

Editorial

The Special Issue on Computing, Engineering and Multidisciplinary Sciences (2024) in the *Advances in Science, Technology and Engineering Systems Journal (ASTES Journal)* reflects the accelerating convergence of computational intelligence, engineering innovation, and cross-disciplinary collaboration in addressing the complexities of a rapidly evolving technological landscape. As global challenges become increasingly interconnected, the integration of diverse scientific perspectives has emerged as a critical enabler of robust, scalable, and forward-looking solutions. This issue brings together a collection of contributions that demonstrate how computing and engineering, enriched by multidisciplinary approaches, are shaping the next generation of scientific and technological advancement.

A central theme of this issue is the pervasive influence of advanced computing paradigms in transforming engineering systems and multidisciplinary applications. Contributions explore the deployment of artificial intelligence, machine learning, data analytics, and intelligent automation to enhance system performance, optimize resource utilization, and support real-time, data-driven decision-making. These technologies are applied across a broad spectrum of domains, including smart cities, healthcare systems, industrial automation, communication networks, and energy management, highlighting their versatility and transformative potential.

The multidisciplinary dimension of this issue is further reflected in its focus on sustainability, resilience, and innovation. Several papers investigate environmentally responsible engineering practices, renewable energy integration, and efficient resource management, aligning with global priorities for sustainable development. Others examine emerging challenges in digital ecosystems, infrastructure resilience, and human-centered technologies, emphasizing the importance of integrating technical expertise with societal considerations. This convergence of perspectives underscores the value of collaborative research in addressing both technical and socio-economic dimensions of contemporary problems.

Methodological diversity continues to be a defining strength of this special issue. The selected studies employ a wide range of approaches, including theoretical modeling, computational simulations, experimental research, and real-world case analyses. This integrative methodology ensures that the research is both academically rigorous and practically relevant. Many contributions emphasize scalability, adaptability, and robustness, offering solutions that can be effectively implemented across diverse operational contexts and industries.

The 2024 timeframe provides an important context characterized by rapid digital transformation, increasing reliance on intelligent systems, and heightened awareness of sustainability and resilience. The research presented in this issue reflects these trends, showcasing innovations that support automation, connectivity, and sustainable growth while addressing emerging global challenges.

The editorial team expresses its sincere appreciation to the authors for their high-quality contributions and to the reviewers for their thoughtful and constructive evaluations. Their collective efforts have ensured the integrity and scholarly impact of this special issue, reinforcing the journal's commitment to excellence in multidisciplinary research.

This special issue highlights the transformative potential of integrating computing, engineering, and multidisciplinary sciences. By presenting diverse perspectives and innovative research, it

provides a valuable platform for advancing knowledge, fostering collaboration, and supporting the development of solutions that address the complex demands of an increasingly interconnected world.

Guest Editor

Prof. Mohamed Abdelaziz Hassan Eleiwa

ADVANCES IN SCIENCE, TECHNOLOGY AND ENGINEERING SYSTEMS JOURNAL

Special Issue **March 2024**

CONTENTS

Proposal and Implementation of Seawater Temperature Prediction Model using Transfer Learning Considering Water Depth Differences
by Haruki Murakami, Takuma Miwa, Kosuke Shima and Takanobu Otsuka

Web Application Interface Data Collector for Issue Reporting
by Diego Costa, Gabriel Matos, Anderson Lins, Leon Barroso, Carlos Aguiar and Erick Bezerra

Evaluation of Physicochemical Stability in Extemporaneous Omeprazole-Based Preparations
by Ezri Cruz-Pérez, José Locia-Espinoza, Joel Jahaziel Díaz-Vallejo, Magda Olivia Pérez-Vásquez, Luis Morales de la Vega and Luz Irene Pascual-Mathey

Advancements in Explainable Artificial Intelligence for Enhanced Transparency and Interpretability across Business Applications
by Maikel Leon and Hanna DeSimone

Evaluation of Various Deep Learning Models for Short-Term Solar Forecasting in the Arctic using a Distributed Sensor Network
by Henry Toal, Michelle Wilber, Getu Hailu and Arghya Kusum Das

GPT-Enhanced Hierarchical Deep Learning Model for Automated ICD Coding
by Joshua Carberry and Haiping Xu

Spatial Distribution Patterns of the Royal Development Projects Initiated by King Rama 9th of Thailand
by Puntip Jongkroy and Ponthip Limlahapun

Visualization of the Effect of Additional Fertilization on Paddy Rice by Time-Series Analysis of Vegetation Indices using UAV and Minimizing the Number of Monitoring Days for its Workload Reduction
by Taichi Ito, Ken'ichi Minamino and Shintaro Umeki

Early Detection of SMPS Electromagnetic Interference Failures Using Fuzzy Multi-Task Functional Fusion Prediction
by Declan Mallamo, Michael Azarian and Michael Pecht

Exploring Current Challenges on Security and Privacy in an Operational eHealth Information System
by Viktor Denkovski, Irena Stojmenovska, Goce Gavrilov, Vladimir Radevski and Vladimir Trajkovik

Integrating Speech and Gesture for Generating Reliable Robotic Task Configuration
by Shuvo Kumar Paul, Mircea Nicolescu and Monica Nicolescu

Leveraging Machine Learning for a Comprehensive Assessment of PFAS Nephrotoxicity
by Anirudh Mazumder and Kapil Panda

Malaysia's Renewable Energy Policy and its Impact on ASEAN Countries
by WaiYie Leong, LeeSun Heng and YuanZhi Leong

Deploying Trusted and Immutable Predictive Models on a Public Blockchain Network
by Brandon Wetzel and Haiping Xu

Tracing the Evolution of Machine Translation: A Journey through the Myanmar (Burmese)-Wa (sub-group of the Austro-Asiatic language) Corpus
by Florance Yune and Khin Mar Soe

Automated Performance analysis E-services by AES-Based Hybrid Cryptosystems with RSA, ElGamal, and ECC
by Rebwar Khalid Muhammed, Kamaran Hama Ali Faraj, Jaza Faiq Gul-Mohammed, Tara Nawzad Ahmad Al Attar, Shaida Jumaah Saydah and Dlsoz Abdalkarim Rashid

Development and Usability Evaluation of Mobile Augmented Reality Contents for Railway Vehicle Maintenance Training: Air Compressor Case
by Gil Hyun Kang, Hwi Jin Kwon, In Soo Chung and Chul Su Kim

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 4, 01-06 (2024)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Proposal and Implementation of Seawater Temperature Prediction Model using Transfer Learning Considering Water Depth Differences

Haruki Murakami, Takuma Miwa, Kosuke Shima, Takanobu Otsuka

Department of Computer Science, Nagoya Institute of Technology, Nagoya, 23107, Japan

ARTICLE INFO

Article history:
Received: 26 April, 2024
Revised: 24 June, 2024
Accepted: 25 June, 2024
Online: 9 July, 2024

Keywords:
Transfer learning
Seawater temperature prediction
Water depth differences
Aquaculture
Machine learning
Time-series forecasting

ABSTRACT

Aquaculture is one of the most important industries worldwide, and most marine products are produced by aquaculture. On the other hand, the aquaculture farmers are faced on the challenge of damage to marine products due to abnormal seawater temperatures. Research on seawater temperature prediction have been conducted, but many of them require a large amount of training data. Collecting seawater temperature data is not easy, and it takes an enormous time to introduce in new aquaculture farms. Therefore, the purpose of this study is to predict seawater temperature even with a small amount of training data for about one year. In this paper, we propose a seawater temperature prediction model using transition learning. The proposed model also attempts to improve the prediction accuracy by considering the difference in water depth between observation points. The results of the evaluation experiment showed that the prediction accuracy can be improved by transfer learning when learning with a small amount of data. In addition, we also confirmed that adding water depth values to the input layer may not lead to improved prediction accuracy for transfer learning.

1. Introduction

Aquaculture is one of the most important industries supporting the world's food supply. Since around 1990, fisheries production has been on a flat trend, while aquaculture production has continued to increase. As a result, the share of aquaculture in the total production of the seafood industry has increased. In 2021, the production from mariculture was approximately 70 million tons, accounting for about 32% of the total production of the fisheries and aquaculture industry [1]. This is due to growing demands for edible seafood in the world. Global consumption of edible seafood is increasing and has almost doubled in the last 50 years. Also in Japan, aquaculture is an important industry. The followings are statistical data on the aquaculture industry in Japan. In 2022, the production from mariculture was about 910,000 tons, accounting for about 23% of the total production of the seafood industry [2]. Its production value was 521.1-billion-yen, accounting for about 33% of the total. These indicate that aquaculture is an indispensable industry in Japan.

Under these circumstances, one of the problems facing aquaculture farmers is the damage caused by abnormal seawater temperatures. If seawater temperature is not properly controlled, it can increase the risk of fish disease infection [3] and decrease feed efficiency [4]. To prevent such damage, aquaculture farmers need to accurately predict seawater temperatures and protect marine products in advance by moving rafts. However, temperature prediction requires years of experience and a wealth of knowledge. Furthermore, it's getting more difficult to predict seawater temperatures due to global warming and severe weather changes. From these, water temperature management is a heavy burden for aquaculture farmers. To make these temperature management more sustainable in the future, it is necessary to provide seawater temperature predictions based on collected temperature data that can be used by anyone.

Although seawater temperature prediction is an active area of research, this is currently limited to prediction of sea surface temperature (SST) [5, 6]. One of the factors that SST prediction

*Corresponding Author: Murakami Haruki, Nagoya Institute of Technology, Japan, murakami.haruki@otsukalab.nitech.ac.jp

https://dx.doi.org/10.25046/aj090401

has been actively conducted is the abundance of SST datasets provided by satellite imagery [7]. SST prediction is used in many fields, such as marine meteorology and weather forecasting, but it is not suitable for the seawater temperature prediction required by aquaculture farmers. In aquaculture, the depth at which fish are reared depends on the type of product and the size of the fish tank. It may be appropriate to keep at a depth of several 10 meters. As seawater temperature changes with depth, it is necessary to predict using seawater temperature data from each depth. This means that abundant sea surface temperature datasets are not available, and it is necessary to measure temperatures at several depths at each location. However, since these data are measured using dedicated observation equipment, it is not easy to collect sufficiently at each aquaculture farm. Therefore, in this study, we propose and implement a seawater temperature prediction model that provides sufficient prediction accuracy even with a small amount of data for about one year.

2. Related Works

In this chapter, we introduce related studies on the prediction of seawater using measured water temperature.

In [8], the authors proposed a method based on an autoregressive model. In this method, each observation point is classified by cluster analysis, and a principal component analysis is conducted using the water temperature anomalies within each group. By autoregressive predicting using the calculated first principal component, the method enables prediction up to three months ahead. However, because they focus on a wide area such as the entire coastal area, it is difficult to predict water temperatures in a narrow area within an aquaculture farm.

In [9], the authors proposed a method for predicting water temperature in aquaculture areas. In this method, water temperature in the aquaculture area is obtained in real time by installing several small and inexpensive buoys, called ubiquitous buoys [10]. Nevertheless, it is necessary to install a large number of buoys because the prediction range covers a wide area of several kilometers.

In [11], the authors proposed a method that uses meteorological data in addition to measured water temperature data. In this method, sea temperature at multiple depths and data provided by the Japan Meteorological Agency are collected. Using these data as features, a random forest is used to make predictions for each depth. In the evaluation experiment, it was shown that using not only temperature but also wind speed improves the accuracy of the prediction. However, they did not predict for the next day or later. According to authors in [12], aquaculture farmers need to predict water temperatures up to one week ahead within an error margin of 1℃.

The authors in [13, 14] proposed a method that predicts over multiple time periods. In this method, in addition to a model that predicts up to the next day, a model that predicts daily mean water temperatures up to one week ahead is provided. These models are capable of long-term prediction using Gated Recurrent Unit (GRU) [15]. Evaluation experiments showed that the models were able to prediction with higher accuracy than existing methods. However, they required a large amount of training data, about 9 years. It is not easy to prepare such data, and it will take time to introduce them to a new farm. To solve the above issues, we develop a seawater temperature prediction method that meets the needs of aquaculture farmers even with a short period of data.

3. Proposed Method

3.1. Subject Data

In this study, we use measured sea water temperature data and meteorological data to create a prediction model. From previous studies [11, 13, 14], it is clear that in addition to seawater temperature data, it is effective to combine this data with meteorological data for learning. In this section, as an example of the seawater temperature and meteorological data used in the proposed model, details of the data collected in this study are described.

First, the seawater temperature data is described. we prepare a dataset of seawater temperature collected at five points throughout Japan. Figure 1 shows the observation points of seawater temperature in this study. The latitude and longitude of Gokasho is (34.3461°N, 136.7050°E), Matoya is (34.8807°N, 136.8807°E), Ago is (34.3071°N, 136.8038°E), Goshoura is (32.2910°N, 130.2370°E), and Otaru is (43.1904°N, 140.9951°E). As Japan is located in the Northern Hemisphere, it is generally hot in the south and cold in the north throughout the year. Similar trends are observed not only with air temperature but also with seawater temperature. To demonstrate that the proposed model can be used universally throughout Japan despite these trends, data from various locations are used.

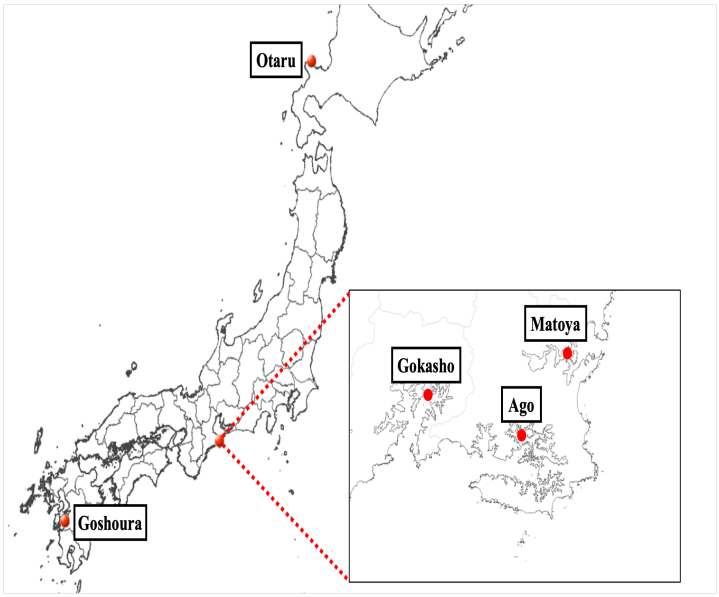


Figure 1: Five observation points of seawater temperature [16]

Table 1 shows the water depths and periods for which seawater temperatures are measured at each point. Under the conditions shown in Table 1, seawater temperatures are measured hourly at each point. At Gokasho, Ago, and Matoya, data are available for about 16 years from 2007 to 2022. On the other hand, Goshoura and Otaru have only about 8 years of data from 2012 to 2020, which means that the amount of data at each point varies. Seawater temperatures are measured simultaneously at three or four depths, which vary point site to point.

Table 1: Measurement conditions for seawater temperature

Point	Depths (m)	Period
Gokasho	0.5, 2, 5, 8	07/02/28 ~ 22/02/28
Ago	0.5, 2, 5, 8	07/03/20 ~ 22/02/28
Matoya	0.5, 2, 5, 8	07/03/20 ~ 22/02/28
Goshoura	1, 3, 10	12/04/01 ~ 20/01/20
Otaru	1, 10, 20	12/04/01 ~ 20/03/02

Next, the meteorological data is described. We prepare temperature and wind speed data from the meteorological database provided by the Japan Meteorological Agency [17]. We use these data measured at the nearest observation stations to the points shown in Figure 1. The periods of these data are set to be the same as the periods in which the seawater temperatures are measured at each point.

3.2. Effects of Depth on Seawater Temperature

In this section, we discuss the effects of different water depths on seawater temperature. Figure 2 shows the daily mean seawater temperature at each depth in Ago in 2015. From this graph, the difference in water temperature for each depth in Ago during the winter (December to February) is small. On the contrary, during the summer (June to August), the difference in water temperature for each depth is large. Seawater temperature tends to decrease with depth in summer. Specifically, the water temperature difference in winter is within 2℃, while that in summer ranges from 2 to 9℃.

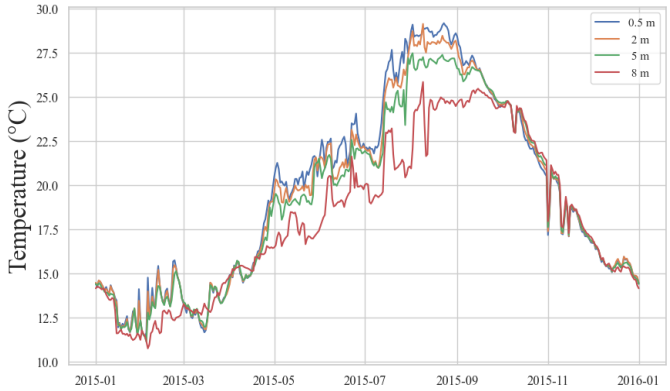


Figure 2: Mean daily seawater temperature at each water depth in Ago (2015).

In summary, the water temperature difference with depth in summer is larger than that in winter, and the larger the water depth, the lower the water temperature in Ago. This tendency is also observed at the other four points. For example, in Gokasho, the water temperature difference in winter is within 2℃, while the water temperature difference in summer ranges from 1℃ to 8℃. In Otaru, the water temperature difference in winter is within 2℃, while the water temperature difference in summer can be as high as 5℃.

As shown in the previous section, seawater temperatures were measured at different depths at each point. This is because the appropriate water depth for aquaculture is different at each point. The depth depends on the type of marine products to be raised and the surrounding climate.

From the above, it is necessary to account for the difference in seawater temperature associated with the water depth between points. With this in mind, the proposed model is presented in the next section.

3.3. Seawater Temperature Prediction Model

The seawater temperature prediction model proposed in this study is an extension of the long-term prediction model proposed in [13,14] by using transfer learning [18], in which water depth values are included in the input layer. This approach aims to solve the problem of the huge amount of train data required by conventional models. In addition, it aims to consider differences in water temperature due to differences in water depth.

Transfer learning is a machine learning approach which applies the knowledge obtained in the source domain to learn in the target domain [19]. In this study, the domain refers to the observation points of seawater temperature. A series of learning is performed in the source domain, and then the learned model is re-learned in the target domain. For this reason, transfer learning has the advantage that it can learn efficiently even if the amount of data in the target domain is small.

Figure 3 shows an overview of the proposed seawater temperature prediction model. The structure of the proposed model is based on Recurrent Neural Network (RNN) [20], which is suitable for time-series forecasting. The proposed model has a three-layer structure consisting of an input layer, a hidden layer, and an output layer, each of which is described separately below.

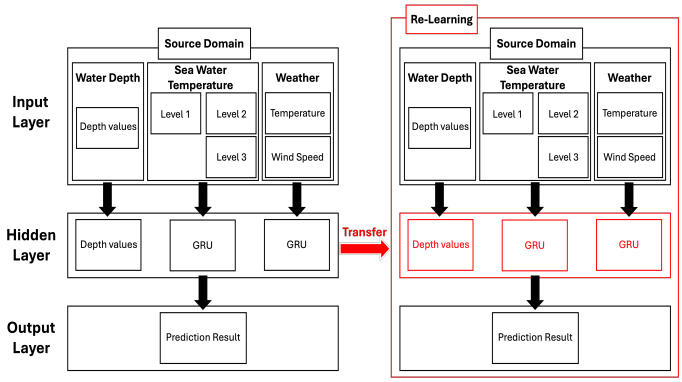


Figure 3: Overall diagram of the proposed model.

In the input layer, two types of time-series data, seawater temperature data and meteorological data, and the water depth values at which seawater temperature measured are input. Since seawater temperature does not change rapidly in a short period of time, the most recent data for the prediction target date is important. Therefore, we use the daily mean seawater temperature and meteorological data for the last seven days. As meteorological data, we use daily mean air temperature and maximum and minimum wind speeds. To account for the effects of multiple depths, we also included seawater temperatures at depths other than the prediction target as an input. As mentioned in the previous section, the depths measured are different at each point, and water temperatures vary depending on the water depth. This means that when performing transfer learning, it is necessary to consider the difference in the water depths between the source and target domains. Therefore, in

the proposed model, the water depth values are added to the input layer.

In the hidden layer, we use GRU, which was also used in the model by the authors in [13,14] Conventional RNNs have two problems: gradient vanishing problem and weight collision, which make it difficult to learn long-term features. On the other hand, GRU has a reset gate and an update gate, and can select a choice of information. This makes it possible to store old information, and thus it can be applied to problems that requires long-term dependence to be considered. In the proposed model, seawater temperature data and weather data are input into another GRUs to be learned separately. The results processed by each GRU and the water depth values are combined, and then passed to the output layer.

In the output layer, the results from the hidden layer are converted to prediction results. As a prediction result, seven days of daily mean seawater temperatures up to one week ahead are output. In the proposed model, this series of learning is performed at points with sufficient seawater temperature data, and then the learned model is used to re-learn at the prediction target points with only a small amount of data.

4. Experiment

4.1. Common experimental setup

In this study, we conducted two evaluation experiments to demonstrate the effectiveness of the proposed model. The first was a comparison experiment of accuracy with and without transfer learning (Experiment 1). The second was a comparison experiment of accuracy with and without the input of water depth values (Experiment 2). This section describes the experimental setup common to both experiments.

Table 2 summarizes the details of the data used in the two experiments. Of the five observation points, we set Gokasho, Matoya, and Ago as the source domain, and Goshoura and Otaru as the target domain. Three different water depth values were set at each point. The period of train data for the source domain was eight years, the target domain was one year, and the test data was one year. Because the daily mean seawater temperature changes with a cycle of one year, we chose one year for the period of train data for the target domain. In addition, by setting the period of train data for source domain to eight years, the amount of train data used at the target domain reaches the amount of train data used by the authors in [13,14].

Table 2: Data used in evaluation experiments.

Point	Depth values [m]	Period of train data	Period of test data
Gokasho	0.5, 2, 5	10/01/01 ~ 17/12/31	
Ago	0.5, 2, 5	10/01/01 ~ 17/12/31	
Matoya	0.5, 2, 5	10/01/01 ~ 17/12/31	
Goshoura	1, 3, 10	18/01/01 ~ 18/12/31	19/01/01 ~ 19/12/31
Otaru	1, 10, 20	18/01/01 ~ 18/12/31	19/01/01 ~ 19/12/31

As evaluation items in the two experiments, we calculated Mean Absolute Error (MAE) and the percentage of predictions with errors more than 1℃. In this study, the standard value was set to 1℃ to meet the needs of aquaculture farmers for an error less than 1℃.

4.2. Experiment 1

In this section, Experiment 1 concerning transfer learning is presented. The purpose of Experiment 1 is to evaluate whether transfer learning is valid for the seawater temperature prediction. Therefore, the proposed model with transfer learning compared with a model learned only with the train data for the prediction target points without transfer learning. The results of Experiment 1 are shown in Tables 3 and 4. Table 3 shows the results when the prediction target point is Goshoura, and Table 4 shows the results when the prediction target point is Otaru. The vertical axis of the table represents the source domain, and the horizontal axis represents the prediction target water depth. The values on the left of the table represent MAE [℃] and the values on the right represent the rate of errors above 1℃ [%].

Table 3: Results of Experiment 1 in Goshoura

Source domain	1m	3m	10m
none	0.535℃, 14.5%	0.669℃, 23.3%	0.605℃, 21.1%
Gokasho	0.293℃, 3.3%	0.256℃, 2.5%	0.258℃, 2.7%
Ago	0.426℃, 7.7%	0.345℃, 4.4%	0.296℃, 3.6%
Matoya	0.330℃, 3.3%	0.280℃, 3.3%	0.272℃, 3.3%

Table 4: Results of Experiment 1 in Otaru

Source domain	1m	10m	20m
none	0.725℃, 27.1%	0.799℃, 31.8%	0.941℃, 38.6%
Gokasho	0.608℃, 17.5%	0.493℃, 10.1%	0.497℃, 9.9%
Ago	0.482℃, 9.3%	0.416℃, 6.8%	0.473℃, 10.1%
Matoya	0.495℃, 10.1%	0.534℃, 12.9%	0.483℃, 10.4%

First, Table 3 shows that when the source domain was 'none', meaning without transfer learning, MAE ranged from 0.535°C to 0.669°C and the rate of errors above 1°C ranged from 14.5% to 23.3%. In contrast, with transfer learning, MAE ranged from 0.256°C to 0.426°C and the rate of errors above 1°C ranged from 2.5% to 7.7%. These results indicate that the proposed model has better prediction accuracy than the model without transfer learning for both evaluation items. The average MAE for each depth was roughly halved for all source domains, and the average rate of errors above 1°C for each water depth was less than one-third for all source domains.

Next, Table 4 shows that without transfer learning, MAE ranged from 0.725°C to 0.941°C and the rate of errors above 1°C ranged from 27.1% to 38.6%. In contrast, with transfer learning, MAE ranged from 0.416°C to 0.608°C and the rate of errors above 1°C ranged from 6.8% to 17.5%. Compared with the model without transfer learning, the average MAE for each depth was about one-half for all source domains, and the average rate of errors above 1°C for each water depth was approximately one-third for all source domains.

In conclusion, Experiment 1 indicated that the prediction accuracy can be improved by transfer learning, regardless of

whether the prediction target point is Goshoura or Otaru, where only about one year of seawater temperature data is available. In transfer learning, prediction accuracy is improved when data from the source and target domains have similar characteristics. As the accuracy increased in transfers to various regions of Japan, the proposed model has the potential to be used in a wide range of aquaculture farms of the country.

4.3. Experiment 2

In this section, Experiment 2 on inputting water depth values is presented. The purpose of Experiment 2 is to evaluate whether water depth values are effective features for improving accuracy in transfer learning. Therefore, the proposed model adding water depth values to the input layer compared with a model that does not use water depth values as input. The results of Experiment 2 were shown in Tables 5 and 6. Table 5 shows the results when the prediction target point is Goshoura, and Table 6 shows the results when the prediction target point is Otaru. The vertical axis of the table represents the source domain and whether or not a bathymetric value was input, and the horizontal axis represents the prediction target water depth. The values on the left of the table represent MAE [℃] and the values on the right represent the rate of errors above 1℃ [%].

Table 5: Results of Experiment 2 in Goshoura

Depth values	Source domain	1m	3m	10m
none	Gokasho	0.312℃, 3.3%	0.260℃, 2.7%	0.276℃, 3.0%
	Ago	0.426℃, 9.3%	0.347℃, 4.9%	0.278℃, 3.6%
	Matoya	0.340℃, 3.8%	0.275℃, 3.0%	0.274℃, 3.3%
input	Gokasho	0.293℃, 3.3%	0.256℃, 2.5%	0.258℃, 2.7%
	Ago	0.426℃, 7.7%	0.345℃, 4.4%	0.296℃, 3.6%
	Matoya	0.330℃, 3.3%	0.280℃, 3.3%	0.272℃, 3.3%

Table 6: Results of Experiment 2 in Otaru

Depth values	Source domain	1m	10m	20m
none	Gokasho	0.518℃, 13.2%	0.448℃, 8.8%	0.485℃, 9.9%
	Ago	0.467℃, 9.3%	0.504℃, 11.2%	0.577℃, 14.5%
	Matoya	0.422℃, 7.7%	0.429℃, 9.0%	0.545℃, 11.5%
input	Gokasho	0.608℃, 17.5%	0.493℃, 10.1%	0.497℃, 9.9%
	Ago	0.482℃, 9.3%	0.416℃, 6.8%	0.473℃, 10.1%
	Matoya	0.495℃, 10.1%	0.534℃, 12.9%	0.483℃, 10.4%

First, Table 5 showed that when the depth values were 'none', meaning without water depth values input, MAE ranged from 0.260°C to 0.426°C and the rate of errors above 1°C ranged from 2.7% to 9.3%. On the other hand, with depth values input, MAE ranged from 0.256°C to 0.426°C and the rate of errors above 1°C ranged from 2.5% to 7.7%. These results indicated that adding water depth values to the input layer led to almost no change in MAE and a slight improvement in the rate of errors above 1°C.

Next, Table 6 showed that without water depth values input, MAE ranged from 0.422°C to 0.577°C and the rate of errors above 1°C ranged from 7.7% to 14.5%. On the contrary, when depth values were input, MAE ranged from 0.416°C to 0.608°C and the rate of errors above 1°C ranged from 6.8% to 17.5%. Adding water depth values to the input layer resulted that the average MAE for each depth was smaller when the source domain was Ago, while larger when it was Goshoura or Matoya. However, whether these prediction accuracies improved or declined, the changes were slight.

Experiment 2 indicated that some combinations improved prediction accuracy, adding water depth values to the input layer. However, regularity between source domains and the prediction target water depth could not be confirmed. Moreover, the degree of changes in accuracy was also marginal. From these results, it was not sufficient to add water depth values to the model input to learn the water depth differences between the source and target domains.

At this point, Table 5 for Goshoura and Table 6 for Otaru were compared in the proposed model. Looking at MAE at the prediction target water depth of one meter, Goshoura is lower than Otaru by 0.056°C to 0.315°C. At other depths, similar results were found. These means that the prediction accuracy is higher when the target domain is Goshoura than Otaru. Focusing on the latitude of the five observation points, Gokasho, Ago and Matoya are about 34°C and Goshoura is 32.2910°N, while Otaru is at 43.1904°N, which is significantly different from the other points. In addition, looking at the currents in the surrounding sea, Otaru is on Tsushima Current, whereas the other points are on Kuroshio Current. Thus, it was found that the two differences mentioned above changed the similarity of the data in the source and target domains. These differences in latitude and nearshore currents could have changed the similarity of the data in the source and target domains. The reason why the prediction accuracy did not improve with the addition of water depth input was thought to be due to the above two differences. Therefore, it is necessary to consider not only depth differences, but also differences in latitude and ocean currents between points in order to further improve prediction accuracy.

5. Conclusion

What is needed in the aquaculture industry is seawater temperature prediction from several meters down a few dozen meters. However, these data are not sufficient because they are not easy to collect. The objective of this study is to enable highly accurate prediction of seawater temperature even for points with a small amount of data. Therefore, we proposed and implemented a prediction model using transfer learning, in which a model that has been learned with data from other points is re-learned with data from the target point. In addition, to account for the depth differences between the two points used in the transfer learning, we added water depth values at which seawater temperature was measured to the input layer.

In the evaluation experiment, it was shown that transfer learning improves the prediction accuracy even for points with only about one year of seawater temperature data. We also showed that the accuracy of transfer learning was not improved by simply adding water depth values to the input layer.

In the future, to solve the above issue, we aim to improve prediction accuracy by considering factors such as differences in latitude and ocean currents. Then, we aim to improve the generalization performance of the model to provide seawater

temperature predictions that meet the demands of more types of aquaculture farmers in more regions.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP21K17803.

References

[1] White Paper on Fisheries, Fisheries Agency, 2022.

[2] Statistical data on the fisheries industry, Ministry of Agriculture, Forestry and Fisheries, 2024.

[3] AnssiKarvonen, PäiviRintamäki, JukkaJokela, ETellervoValtonen, "Increasing water temperature and disease risks in aquatic systems: Climate change increases the risk of some, but not all, diseases," International Journal for Parasitology, **40**(13), 1483–1488, 2010, https://doi.org/10.1016/j.ijpara.2010.04.015.

[4] Helene Volkoff, Ivar Rønnestad, "Effects of temperature on feeding and digestive processes in fish," Temperature, **7**(4), 307-320, 2020, https://doi.org/10.1080/23328940.2020.1765950.

[5] Q. Zhang, H. Wang, J. Dong, G. Zhong, X. Sun, "Prediction of Sea Surface Temperature Using Long Short-Term Memory," IEEE Geoscience and Remote Sensing Letters, **14**(10), 1745-1749, 2017, https://doi.org/10.1109/LGRS.2017.2733548.

[6] Ham, YG., Kim, JH., Luo, JJ., "Deep learning for multi-year ENSO forecasts," Nature, **573**, 568–572, 2019, https://doi.org/10.1109/LGRS.2017.2733548.

[7] Boyin Huang, Chunying Liu, Eric Freeman, Garrett Graham, Tom Smith, Huai-Min Zhang, "Assessment and intercomparison of NOAA daily optimum interpolation sea surface temperature (DOISST) version 2.1," Journal of Climate, **34**(18), 7421–7441, 2021, https://doi.org/10.1175/JCLI-D-21-0001.1.

[8] Tadahiro Saotome, Shinichi Ito, "海洋観測データを用いた福島県沿岸海域の水温予測手法の検討," **11**, 27-, 2004, https://cir.nii.ac.jp/crid/1571417125955513088.

[9] Masaaki Wada, Katsumori Hatanaka, Masashi Toda, "Development of personal ocean observation buoy for scallop cultivation," IPSJ Journal(article in Japanese with an abstract in English), **2006**(14), 387-392 , 2006, https://cir.nii.ac.jp/crid/1050292572092667264.

[10] Kesuke Abe, Masaaki Wada, "A Study of Forecasting the Seawater Temperature in Japanese Scallop Aquaculture Sea Area by Using Ubiquitous Buoy System," The Japanese Society of Fisheries Engineering, **47**(1), 43-46 , 2010, https://doi.org/10.18903/fisheng.47.1_43.

[11] Takanobu Otsuka, Yuji Kitazawa, Takayuki Ito, "Seawater Temperature Prediction Method for Sustainable Marine Aquaculture," Preprints, 2017, https://doi.org/10.20944/preprints201709.0114.v1.

[12] Kenichi Nakagawa, Yoshikazu Fukuda, Hideki Kaneko, Hiroshi Nakamura, Tatsuo Nakamura, "東北地方の養殖漁業のための沿岸水温予測方法の紹介," Meteorological Time Report (Japan Meteorological Agency), **85**, 13-29 , 2018.

[13] Masahito Okuno, Takanobu Otsuka, "How to Predict Seawater Temperature for Sustainable Marine Aquaculture (Student Abstract)," In Proceedings of the AAAI Conference on Artificial Intelligence, **34**(10), 13887–13888, 2020, https://doi.org/10.1609/aaai.v34i10.7216.

[14] Masahito Okuno, Takanobu Otsuka, "Proposal and Implementation of Multiple Term Seawater Temperature Prediction Algorithm for Marine Aquaculture," IPSJ Journal(article in Japanese with an abstract in English), **61**(3), 687–694, 2020, http://id.nii.ac.jp/1001/00204182.

[15] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1724–1734, 2014, https://doi.org/10.3115/v1/D14-1179.

[16] Geographical Survey Institute website, Ministry of Land, Infrastructure, Transport and Tourism, 2024

[17] Historical meteorological data, Japan Meteorological Agency, 2024

[18] S. J. Pan and Q. Yang, " A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering, **22**(10), 1345-1359, 2010, https://doi.org/10.1109/TKDE.2009.191.

[19] Weiss, K., Khoshgoftaar, T.M. & Wang, D, " A survey of transfer learning," Jaurnal of Big Data, **3**(9), 2016, https://doi.org/10.1186/s40537-016-0043-6.

[20] Wim De Mulder, Steven Bethard, Marie-Francine Moens, "A survey on the application of recurrent neural networks to statistical language modeling," Computer Speech & Language, **30**(1), 61–98, 2015, https://doi.org/10.1016/j.csl.2014.09.005.

Copyright: This article is anap open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 9, No. 5, 1-8 (2024)
www.astesj.com
Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTES Journal
ISSN: 2415-6698

Web Application Interface Data Collector for Issue Reporting

Diego Costa*, Gabriel Matos, Anderson Lins, Leon Barroso, Carlos Aguiar, Erick Bezerra

SIDIA Institute of Science and Technology, Manaus, Brazil

ARTICLE INFO

Article history:
Received: 24 April, 2024
Revised: 01 August, 2024
Accepted: 03 August, 2024
Online: 14 September, 2024

Keywords:
Bug Reporting
Software Management
Web Application
Browser API

ABSTRACT

Insufficient information is often pointed out as one of the main problems with bug reports as most bugs are reported manually, they lack detailed information describing steps to reproduce the unexpected behavior, leading to increased time and effort for developers to reproduce and fix bugs. Current bug reporting systems lack support for self-hosted systems that cannot access third-party cloud environments or Application Programming Interfaces due to confidentiality concerns. To address this, we propose Watson, a Typescript framework with a minimalist User Interface developed in Vue.js. The objectives are to minimize the user's effort to report bugs, simplify the bug reporting process, and provide relevant information for developers to solve it. Watson was designed to capture user's interactions, network logs, screen recording, and seamlessly integration with issue tracker systems in self-hosted systems that cannot share their data to external Application Programming Interfaces or cloud services. Watson also can be installed via Node Package Manager and integrated into most JavaScript or TypeScript web projects. To evaluate Watson, we developed an Angular-based application along with two usage scenarios. First, the users experimented the application without using Watson and once they found a bug, they reported it manually on GitLab. Later, they used the same application, but this time, whenever they detect another bug, they reported it through Watson User Interface. Watson, as stated by the experiment participants and the evidences, is useful and helpful for development teams to report issues and provide relevant information for tracking bugs. The identification of bug root causes was almost three times more effective with Watson than manual reporting.

1. Introduction

This paper is an extension of work originally presented at the 2023 IEEE 30th Annual Software Technology Conference (STC) [1]. Bug reports play a crucial role in software maintenance, enabling developers to prioritize, reproduce, identify, and resolve defects [2], [3]. Detailed information is expected from the reports, such as the unexpected behavior, the steps to reproduce it, logs, or screenshots, and others, so developers may recreate it to find a solution [2, 4, 5].

Insufficient information is often pointed out as one of the main problems with bug reports, generally most bugs are reported manually by end-users or testers, the reports lack of details and sufficient information describing the steps to reproduce the unexpected behavior to allow the developers to find a solution [2, 6].

The most common way to report bugs is through issue tracking systems, but in the majority cases, there is not any standard, which causes misinformation for the development team due the unclear or insufficient data [7]. Steps to reproduce the bug, stack trace errors, test case scenarios, logs, and images are factors that impact the quality of the bug reports [6, 8, 9, 10, 11, 12].

In this paper, we introduce Watson, a framework developed in Typescript[1] with an User Interface (UI) developed in Vue.js[2]. The objective is to save time and effort for the person who is going to report the bug, and standardize bug reporting by collecting fundamental information that will aid the developers to reproduce the undesired behavior, as such as: user interaction on the page, network requests, and screen video. The main points of Watson are that it is a framework that can be installed via Node Package Manager (NPM), it can be integrated into most Javascript[3] or Typescript web projects, and it is designed for self-hosted systems that cannot share its data to external Application Programming Interfaces (APIs) or cloud services.

*Corresponding Author: Diego Costa, SIDIA Institute of Science and Technology, Manaus, Brazil, diego.costa@sidia.com
[1]https://www.typescriptlang.org/
[2]https://vuejs.org/
[3]https://developer.mozilla.org/en-US/docs/Web/JavaScript

Watson was evaluated empirically, using a test application developed in Angular[4] and inviting developers and testers to use it with Watson. The goal was to evaluate the participant's perception of Watson usefulness in comparison with manually reporting a bug on GitLab[5]. Experienced web developers an testers judged Watson as useful and Watson's features to collect information proved helpful in identifying the root cause of bugs reported.

The rest of this work is organized as follows: Section 2 provides details about the problem and related work. Section 3 provides details about the proposed software. Section 4 presents the results of the use of Watson. Section 5 concludes by discussing the main points found.

2. Related Works

Our research focused on bug reporting tools that provided relevant and effective information for software maintenance, mainly for web applications, but due to the lack of recent works, we expanded the search for Android applications and approaches that are emerging, such as machine learning.

The authors of [13] propose an automated bug reporting system which serves as a foundation for testing frameworks in web applications, generating failure reports based on the test cases. The generated reports consist of the number of test cases executed, the number of failed, passed and skipped tests, and the time it took to perform the tests. When the report is ready, it is checked if it is a duplicated bug report, then it is mailed to the development team. The tool was used by the authors to simulate regression test cases, resulting in an 8% reduction in test execution time. Additionally, it summarized the bug report, reducing human effort and time spent filtering duplicated bug reports.

In the work [14], they created a tool, Euler, that automatically analyzes the written description of a bug report, evaluates the quality of reproduction steps, and provides feedback to users about ambiguous or missing steps. Neural sequence labeling combined with discourse patterns and dependency parsing identifies sentences and individual steps to reproduce. It matches these steps to program state and Graphical User Interface (GUI) in a graph-based execution model. An empirical evaluation was conducted to determine the accuracy of Euler in identifying and assessing the quality of reproduction steps for bug reports. The results indicated that Euler correctly identified 98% of the existing steps to reproduce and 58% of the missing ones, and 73% of its quality annotations being accurate.

The work [2] presents Bee, an open-source tool that can be integrated with GitHub[6] and automatically analyzes user-written bug reports using machine learning textual classification. It offers insights into the system's observed comportment, expected comportment, and reproduction steps for the reported bugs. As result they achieved 87% recall, indicating the ability to correctly detect and classify the described sentences.

In order to generate better bug reports for Android, CrashScope [15] was created. It works by collecting system version and hardware information, the application state, the user entry text description to reproduce the bug, the GUI events, and app's stack trace error. The purpose was to assess the reliability and comprehensibility of reports generated by CrashScope relative to five current tools. To evaluate this work, they used 8 real world open source applications bug reports extracted from their corresponding issue trackers. They invited 16 users to reproduce 4 bugs reported using CrashScope and 4 bugs reported manually. It was discovered that reports produced by CrashScope were equally reproducible compared to those from other tools, although it yielded more comprehensible and beneficial reports for developers.

In [16], the author presented a chatbot that designed for Android, combines dynamic software analysis, natural language processing, and automated report quality assessment to assist users in writing better descriptions and receiving issue reports. By inviting 18 end-users to identify 12 bugs across 6 Android apps, we found that Burt provides more accurate and complete reproduction steps than Itrac, a template-based bug reporting system used by another 18 users.

The work [17] proposes an automated tool for integrating user feedback into the testing process. In order to achieve this, they collected datasets of mobile application issues reviews along with user feedbacks to train a machine learning algorithm that would be capable to link the user's feedback to stack traces with the objective to relate a feedback that might describe the cause of a failure to a bug. By following this process, they concluded user feedback is highly promising to integrate into the testing process of mobile apps as it complements the capabilities of testing tools identifying bugs that are not reveled in this phase of software development, facilitating the diagnosis of bugs or crashes.

In the study [18], the authors explored automating Android bug replaying using Large Language Models (LLMs). Motivated by the success of these models, they proposed AdbGPT, a method for reproducing bugs from bug reports via prompt engineering. By following this approach, they demonstrated 81.3% effectiveness and efficiency to reproduce bugs from users' reports, and in terms of average time to reproduce bug reports, the AdbGPT outperformed the average time of experiment participants.

Other studies on bug reporting tools for web applications [19, 20] and commercial tools like Usersnap[7], BugHerd[8], and Bird Eats Bugs[9] utilize browser extensions or embedded scripts to capture Document Object Model (DOM) events, stack trace errors, and screenshots. Data is sent to cloud services but lacks integration with self-hosted systems that restrict sending its private data to external APIs. This issue is addressed by Watson, a framework installable via NPM for JavaScript or TypeScript web projects. It provides flexibility for development teams to integrate and utilize self-hosted systems, enabling configuration of Watson to use entirely their own systems while collecting crash report data without external environments.

[4] https://angularjs.org/
[5] https://about.gitlab.com/
[6] https://github.com
[7] https://usersnap.com/
[8] https://bugherd.com/
[9] https://birdeatsbug.com/

3. Watson Framework

Watson was developed in Typescript, offering an API to help collect important information. Watson collects information such as DOM events related to the user's interaction, when a user interacts with the application, Watson interceptors will collect the web page events, such as mouse clicks, network requests and screen video to capture what the user sees on the page. This information will be passed to the reporter class that implements the interface WatsonReporter to attach the information to a bug report with a description provided by the user and the Watson collected data, then it will send to the chosen issue tracking system.

As presented in Figure 1, Watson acts in the native browser API, injecting a collection of interceptors that will intercept data during web application usage. When the user initiates the recording process through the UI, Watson begins collecting data. Upon completion of the recording, the gathered information can be transmitted to an issue tracking system based on the preconfigured reporter implementation. This may involve utilizing a built-in reporter such as GitLabReporter or implementing a customized reporter to facilitate integration with alternative issue trackers or systems.

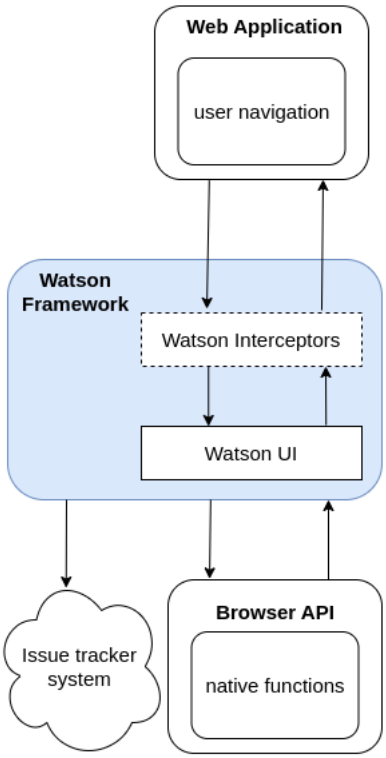
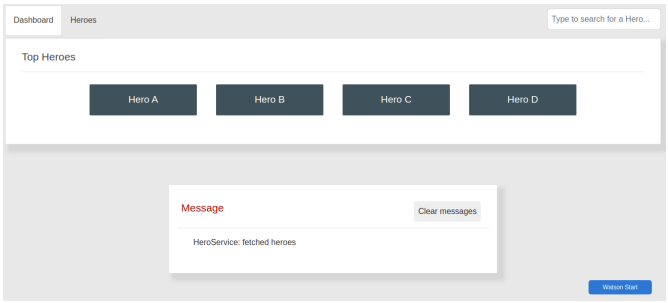
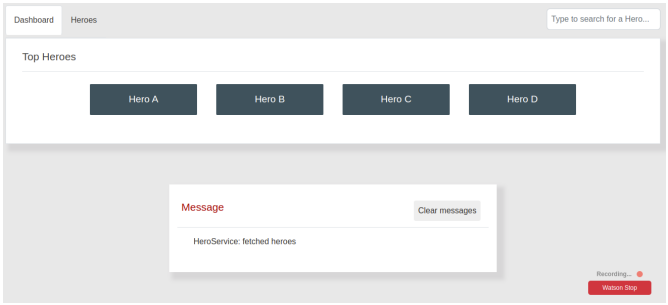


Figure 1: Watson Architecture

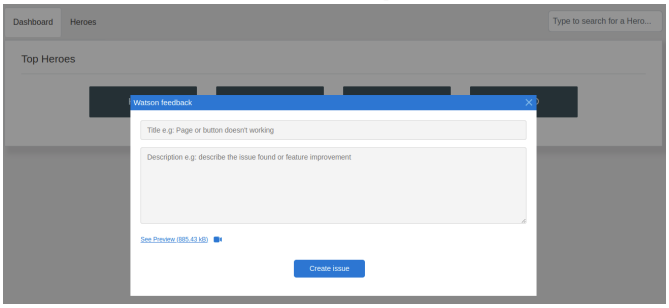
The UI workflow used to report bugs with Watson is shown in Figure 2. Starting with the Watson UI start button (Figure 2a), users can proceed as normal in their testing scenario or regular application usage. Once Watson is running, it attaches interceptors to the browser's native API, listens for native events, collects data, and redirects event parameters to the original event calls as proxies, allowing it to intercept and save the user interaction with the web page.



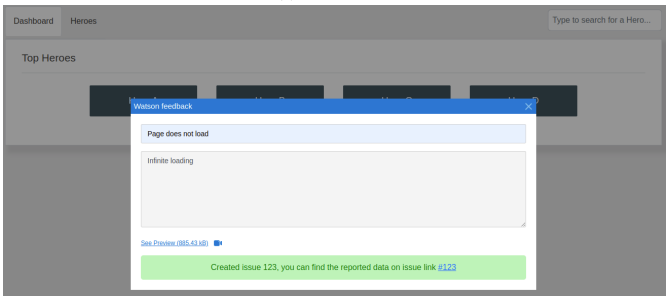
(a) Watson start



(b) Watson stop



(c) Create issue



(d) Issue created

Figure 2: Watson UI

After the data is intercepted, it gets transferred to the initial web browser function API to perform its regular actions. To stop the data collection, the user can use the stop button on Watson UI (Figure 2b), then, as shown in Figure 2c, a dialog message opens for users to provide additional description about the bug and a title for the issue, which will be sent to an external issue tracking system such as GitHub[10], Jira[11] or another one which the developers may have

[10]https://github.com
[11]https://www.atlassian.com/software/jira

previously integrated Watson to it. As in this case, it was integrated with GitLab, it returned the issue link in Figure 2d.

3.1. Implementation details

Figure 3 presents Watson's class diagram. It is composed of four components: **Watson**, **WatsonConfig**, **WatsonReporter**, and **Watson interceptors**.

Watson can be installed on a web project with NPM. After installation, the Watson instance is a singleton and must be initialized at application start-up, initializing and attaching its injectors to the browser's native API, functioning as proxies. The Watson framework must be initialized and configured by adding the code **Watson.getInstance(config)** on application startup code. The config paramter is object of type **WatsonConfig** that contains which interceptors should be active and the reporter implementation of **WatsonReporter** interface that has the responsibility to connect Watson to an external issue tracking system. This enables reporting of collected information alongside user description and issue title.

3.1.1. Watson

The Watson class connects all parts of the system and ensures the data capture functions properly. It manages the start and stop of data capture, stores the collected data, controls the event interceptors activating and deactivating the capture and also uses the reporter instance to send the collected data to issue tracker. With the config parameter, it enables which event interceptors that will be attached to native API and the event types that should be captured, and also defines the WatsonReporter instance to send the collected data.

3.1.2. WatsonConfig

The configuration interface determines which event interceptors should be activated. Currently, there are 3 event interceptors to capture user interaction data: network interceptor, DOM events interceptor, and the screen recorder. By default, all interceptors are enabled, but the development team can define which of them are active according to their requirements. For example, to disable screen recorder by setting **screenRecorder** option to false. Additionally, the implementation instance of the WatsonReporter interface must be specified in the config. This can utilize either a built-in implementation such as GitlabReporter or a custom report class that implements the interface.

3.1.3. WatsonReporter

This is the interface implementation required for reporting to external issue tracking system. A method must be implemented to send the collected data to an external issue tracking system like Jira, GitLab, or others. The development team can write its code to connect to the external system and implement this interface to send data. For this experiment, Figure 3 presents the GitLabReporter class designed to send the collected data to GitLab.

To implement this interface, the development team must first obtain access to the required issue tracking system API and follow its documentation to consume its API and have information about it such as API key, authentication, and available endpoints. Following this step, the developers can implement a class with this interface containing the methods to communicate with the issue tracking system and receive the data collected by Watson to finally transmit them.

The WatsonReporter interface has a method called **reportData**, accepting two parameters. The first parameter consists of the issue's basic information, including its title and description, while the second one is the Watson collected data. This function is responsible for generating issues within the issue tracker and returning an object containing the issue's ID and corresponding link, or null in case of failure.

3.1.4. Watson Interceptors

Watson interceptors are built-in functions that will be attached on native API, in order to listen network requests and DOM events

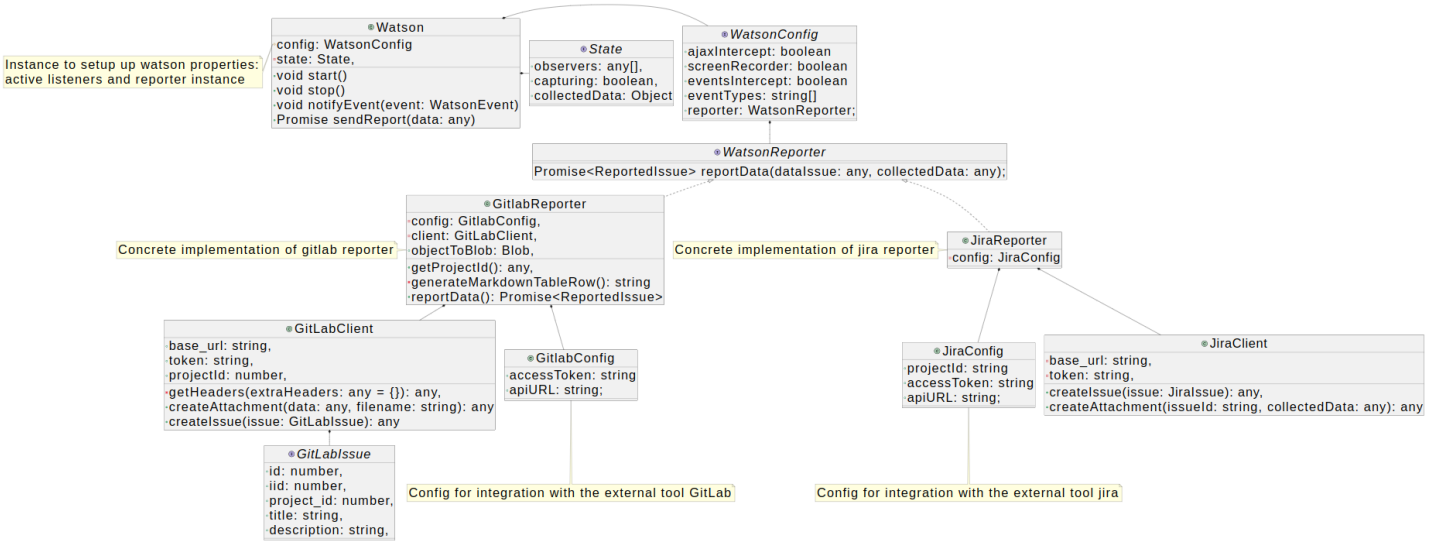


Figure 3: Watson Class Diagram

specified in the event types list, according to active event interceptors specified in the Watsonconfig when Watson is instantiated. These functions will be added to a list of observables at Watson states. Watson has three types of interceptors:

- **Video recording**: Watson utilizes the MediaStream API[12] to capture the user's screen activity. By clicking the start button within the Watson UI, the user initiates the recording process; while stopping it requires clicking the stop button. Besides bug reporting, this can be used to record test scenarios. The recorded video will be attached in the issue report to be sent along with the logs.
- **Network requests**: As represented in Figure 4, Watson injects the ajax interceptor to the XMLHttpRequests API[13] for collecting vital data about network requests. As soon the application is in use, it sends Hypertext Transfer Protocol (HTTP) requests based on user interactions. The native API is proxied, allowing the interceptor to gather information about the network requests such as headers, start and end time, status codes, sent and received data, and request duration. All collected information is stored and used to generate a JSON log file, which will be attached on the issue report.

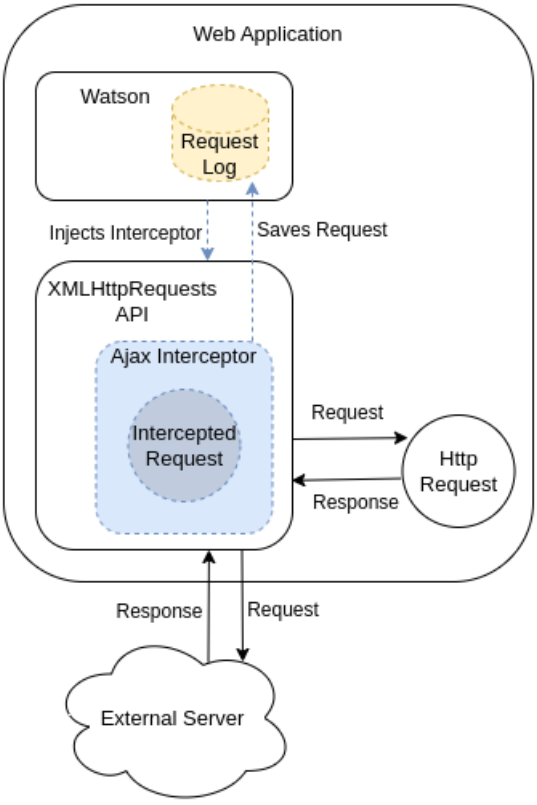


Figure 4: Watson - Network Interceptor

- **Capturing DOM events**: In order to collect the user's interaction with the web application, such as mouse and keyboard actions, Watson injects its interceptor into the native API[14] of AddEventListener, gathering data on events such as clicks, double clicks, mouse enter, mouse leave along with element details including HyperText Markup Language (HTML) tag name, node xpath, text content, as shown be seen in Figure 5.

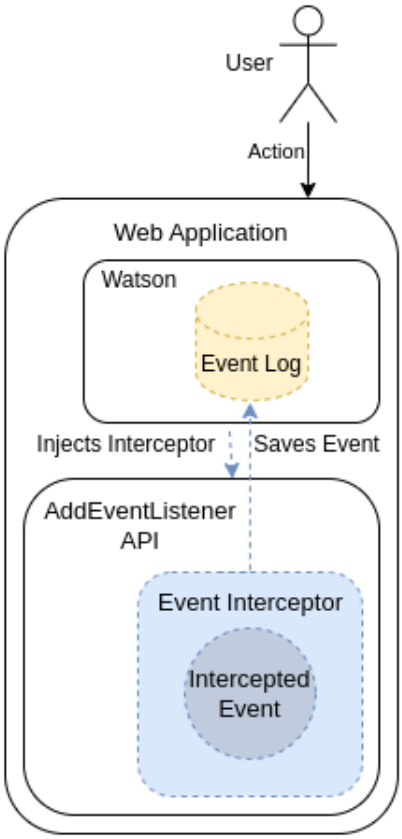


Figure 5: Watson - Event Interceptor

3.1.5. *GitlabReporter*

It is a concrete implementation of interface WatsonReporter, it can create an issue on Gitlab through its Web API and upload the collected information to the created issue.

The components of the reporting system: GitLabClient, GitLabConfig, and GitLabIssue. Each component is detailed below:

GitlabConfig serves as the interface to represent the minimum information necessary to identify the project and access it through the Gitlab API. It requires the server API url, as it may be a self-hosted server, and a valid access token registered to a Gitlab bot authorized to create issues within the project. These information are necessary during the configuring of the GitlabReporter as a WatsonReporter instance at application launch, all those configuration must be provided and setting up by the development team.

GitlabClient is responsible to use the Gitlab Web API use the GitlabConfig information. It is able to create issues and upload attachments to an issue.

[12]https://developer.mozilla.org/en-US/docs/Web/API/MediaStream
[13]https://developer.mozilla.org/en-US/docs/Web/API/XMLHttpRequest
[14]https://developer.mozilla.org/en-US/docs/Web/API/EventTarget/addEventListener

GitlabIssue Is a representation of represents a created issue on Gitlab. It contains basic issue details like issue ID, project ID, title and description.

4. Test Case

A web application was developed with Angular to evaluate the Watson framework. Watson was set up through NPM, configured and incorporated into the application accordingly. As previously mentioned, a GitLab reporter class that implemented the interface for Watson reporter was created to integrate and report the issue to Gitlab.

As in the works of [3, 14, 15, 16, 21], an empirical experiment was conducted in two parts to evaluate if the users enjoyed Watson's functionalities and if these would be relevant for debugging and solving issues.

As described in Table 1, an online questionnaire was created for the users to answer after completing the tests. This questionnaire aimed to assess user perceptions of Watson's features and effectiveness in resolving bugs relative to manual reporting. Quantitative and qualitative questions were included, with participants able to respond on a scale of 0 to 5, providing feedback on areas for improvement and suggestions.

Table 1: Questionnaire

1	Considering the manual method of creating an issue, how error-prone do you think this method is?
2	Considering the Watson method of creating an issue, how error-prone do you think this method is?
3	What would be the main reasons to use the manual method?
4	What would be the main reasons to use Watson?
5	Considering the method using Watson, how much would the network log be relevant to debug?
6	Considering the method using Watson, how much would DOM events be relevant to debug?
7	Considering the method using Watson, how much would the video be relevant to debug?
8	How relevant would Watson be to your project's issue report?
9	If you use or know other issue trackers. How much value would Watson add as a library compared to other issue trackers?
10	Do you have any suggestions for improvement? If yes, which one?

Six web developers and six testers were invited to try the application, discover, and report bugs. Some bugs were found in the web application, including issues with the network connection to database, visual elements, and behavior. The participants were not told about which bugs were incorporated into the application, and Watson was installed and configured by us.

In the fist part of the experiment, the users tried the web application without Watson installed, and once they noticed a bug, they manually reported it on GitLab, describing and attaching anything they judged necessary to solve the bug. In the second part, they tried the application with Watson installed. As in the first part, they explored the system until they found a bug, but this time, with Watson available in the application, they used Watson's UI to report the bug.

Upon completion of testing the application across both scenarios, participants responded to an online survey assessing their satisfaction with Watson's functionalities and the relevance of data gathered by Watson towards addressing reported bugs.

The purpose of the first and second questions was to evaluate how error-prone manual report and the Watson report were. For these two questions, the lower value is better because it indicates less error-prone than a higher value. For Watson method, the average grade was 2.22, and the manual method was 3.11.

The third and fourth questions were about the motives to use one method over another. The distribution about the main reasons the participants considered using the manual method was three votes for duplicate issue report prevention, two votes for organization, one vote for quickly generating evidences, one vote for using a text editor, and one vote for less-error prone, indicating that the participants noticed the prevention of duplicate issues more. For Watson, the results were of 5 votes for taking less time to report an issue, two votes for more bug evidences and two votes for less-error prone, indicating that the participant's preferred Watson because they took less time to report an issue.

The fifth, sixth, and seventh questions are about the relevance of Watson features: network logs, DOM events, and screen records collected by Watson. The average rates for these features were 4.11, 5 and 4.78, respectively.

The eighth question asks about the relevance of Watson for the participant's project issue report. The average rate for it was 4.89. The ninth question asks to compare Watson's reports to other tools if the participant used another one. The average rate for it was 4.22.

The tenth question asks for suggestions. The suggestions the participants cited were to add a text editor for the Watson's description field, an option to download the logs and generated video, an option to choose which project they would like to report the issue and label it, and mainly the history search to avoid duplicate issues before sending the new issue.

To evaluate the efficiency of Watson compared to manual reporting, all the bug reports were evaluated and analyzed for its potential to identify the root cause of the issue. This evaluation involved utilizing videos and user descriptions to recreate the bug scenario, as well as examining the collected stack traces to trace the faulty functionality.

Based on evidence provided by participants through manual bug reporting, it was found the bug root cause on 18.75% of the manually reported issues. In contrast, using Watson's reports containing stack traces and video, the number of issues that could be determined the bug root cause increased to 63.63%.

The recorded video helped to reproduce all of the bug reported with Watson. Additionally, the event and network logs were used to investigate the application's behavior during the occurrence of these bugs.

5. Conclusion

This work presented Watson, a software program developed to standardize the bug report process. Its main purposes are to reduce the

effort to collect bug evidences and provide relevant information for developers to solve it.

Experienced web developers and testers tried Watson and perceived it as useful and relevant for crash reports. Watson scored an average rate of 4.89 out of 5 from the participants questionnaire when asked about its impact on issue reporting in their web projects.

Watson's features for collecting event and network stack traces proved useful in identifying the root cause of bugs in 63.63% of cases when using Watson, against to only 18.75% of the issues reported manually.

This study focused on Watson's ability to gather necessary data for bug reports and make easier for user to report bugs with evidences. Unlike from many other reporting tools, Watson operates independently of any browser plugins and does not require sending data to external servers.

The results of the analysis provided by this article are limited and defined by the test case performed. To interpret the knowledge obtained as a general approach, additional use cases in different projects are needed. Therefore, our proposal suits the need to use a bug reporting tool without using a third-party cloud, which is a requirement for scenarios where highly confidential projects are developed.

As points of improvement mentioned by the participants, most of them were related to the user experience, such as the text editor, options to download collected information, and the recorded screen. It was also suggested to Watson block duplicated bugs.

The main achievement of this project was creating a framework using TypeScript, which can be easily incorporated into web applications and various issue tracking systems. This enables development teams to conveniently obtain standardized bug reports and user feedback. Projects can benefit from this solution as it allows them to bypass third-party servers and incorporate a reporting tool within their self-hosted system, especially when handling sensitive data.

As future work, besides the already mentioned user experience suggestions, we plan to develop a mechanism to prevent duplicate bug reports using Watson and machine learning techniques.

Acknowledgment This work is the result of the R&D project *Projeto de Engenharia de Software e Ciência de Dados aplicados ao Desenvolvimento de Sistemas*, performed by Sidia Instituto de Ciência e Tecnologia in partnership with Samsung Eletrônica da Amazônia Ltda., using resources from Federal Law No. 8.387/1991, and its disclosure and publicity are under the provisions of Article 39 of Decree No. 10.521/2020. Rodrigo José Borba Fernandes, Isabelle Maria Farias de Lima Teixeira, and Thiago Cruz Ferraz former members.

References

[1] G. Matos, D. Costa, A. Lins, E. Bezerra, L. Barroso, C. Aguiar, T. Ferraz, I. Teixeira, "Watson: Web Application Interface Data Collector for Feedback Reporting," in 2023 IEEE 30th Annual Software Technology Conference (STC), 3–6, 2023, doi:10.1109/STC58598.2023.00007.

[2] Y. Song, O. Chaparro, "Bee: A tool for structuring and analyzing bug reports," in Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 1551–1555, 2020, doi:10.1145/3368089.3417928.

[3] K. Moran, "Enhancing android application bug reporting," in Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering, 1045–1047, 2015, doi:10.1145/2786805.2807557.

[4] M. Erfani Joorabchi, M. Mirzaaghaei, A. Mesbah, "Works for me! characterizing non-reproducible bug reports," in Proceedings of the 11th working conference on mining software repositories, 62–71, 2014, doi:10.1145/2597073.2597098.

[5] M. Soltani, F. Hermans, T. Bäck, "The significance of bug report elements," Empirical Software Engineering, **25**, 5255–5294, 2020, doi:10.1007/s10664-020-09882-z.

[6] T. Zimmermann, R. Premraj, N. Bettenburg, S. Just, A. Schroter, C. Weiss, "What makes a good bug report?" IEEE Transactions on Software Engineering, **36**(5), 618–643, 2010, doi:10.1109/TSE.2010.63.

[7] D. Huo, T. Ding, C. McMillan, M. Gethers, "An empirical study of the effects of expert knowledge on bug reports," in 2014 IEEE International Conference on Software Maintenance and Evolution, 1–10, IEEE, 2014, doi:10.1109/ICSME.2014.22.

[8] J. Wang, M. Li, S. Wang, T. Menzies, Q. Wang, "Images don't lie: Duplicate crowdtesting reports detection with screenshot information," Information and Software Technology, **110**, 139–155, 2019, doi:10.1016/j.infsof.2019.03.003.

[9] N. Cooper, C. Bernal-Cárdenas, O. Chaparro, K. Moran, D. Poshyvanyk, "It Takes Two to Tango: Combining Visual and Textual Information for Detecting Duplicate Video-Based Bug Reports," in 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), 957–969, 2021, doi:10.1109/ICSE43902.2021.00091.

[10] M. Bheree, J. Anvik, "Identifying and Detecting Inaccurate Stack Traces in Bug Reports," in 2024 7th International Conference on Software and System Engineering (ICoSSE), 9–14, IEEE Computer Society, Los Alamitos, CA, USA, 2024, doi:10.1109/ICoSSE62619.2024.00010.

[11] Y. Noyori, H. Washizaki, Y. Fukazawa, K. Ooshima, H. Kanuka, S. Nojiri, "Deep learning and gradient-based extraction of bug report features related to bug fixing time," Frontiers in Computer Science, **5**, 1032440, 2023, doi:10.3389/fcomp.2023.1032440.

[12] R. Krasniqi, H. Do, "A multi-model framework for semantically enhancing detection of quality-related bug report descriptions," Empirical Software Engineering, **28**(2), 42, 2023, doi:10.1007/s10664-022-10280-w.

[13] Y. Sharma, A. Dagur, R. Chaturvedi, et al., "Automated bug reporting system in web applications," in 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), 1484–1488, IEEE, 2018, doi:10.1109/ICOEI.2018.8553850.

[14] O. Chaparro, C. Bernal-Cárdenas, J. Lu, K. Moran, A. Marcus, M. Di Penta, D. Poshyvanyk, V. Ng, "Assessing the quality of the steps to reproduce in bug reports," in Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 86–96, 2019, doi:10.1145/3338906.3338947.

[15] K. Moran, M. Linares-Vásquez, C. Bernal-Cárdenas, C. Vendome, D. Poshyvanyk, "Automatically discovering, reporting and reproducing android application crashes," in 2016 IEEE international conference on software testing, verification and validation (icst), 33–44, IEEE, 2016, doi:10.1109/ICST.2016.34.

[16] Y. Song, J. Mahmud, Y. Zhou, O. Chaparro, K. Moran, A. Marcus, D. Poshyvanyk, "Toward interactive bug reporting for (android app) end-users," in Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 344–356, 2022, doi:10.1145/3540250.3549131.

[17] G. Grano, A. Ciurumelea, S. Panichella, F. Palomba, H. C. Gall, "Exploring the integration of user feedback in automated testing of Android applications," in 2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER), 72–83, 2018, doi:10.1109/SANER.2018.8330198.

[18] S. Feng, C. Chen, "Prompting Is All You Need: Automated Android Bug Replay with Large Language Models," in Proceedings of the 46th IEEE/ACM International Conference on Software Engineering, 1–13, 2024, doi:10.1145/3597503.3608137.

[19] B. Burg, R. Bailey, A. J. Ko, M. D. Ernst, "Interactive record/replay for web application debugging," in Proceedings of the 26th annual ACM symposium on User interface software and technology, 473–484, 2013, doi:10.1145/2501988.2502050.

[20] J. Hibschman, H. Zhang, "Unravel: Rapid web application reverse engineering via interaction recording, source tracing, and library detection," in Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology, 270–279, 2015, doi:10.1145/2807442.2807468.

[21] J. Johnson, J. Mahmud, T. Wendland, K. Moran, J. Rubin, M. Fazzini, "An Empirical Investigation into the Reproduction of Bug Reports for Android Apps," in 2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER), 321–322, 2022, doi:10.1109/SANER53432.2022.00048.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 9, No. 5, 9-20 (2024)
www.astesj.com
Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTES Journal
ISSN: 2415-6698

Advancements in Explainable Artificial Intelligence for Enhanced Transparency and Interpretability across Business Applications

Maikel Leon[*,1], Hanna DeSimone[2]

[1]*Department of Business Technology, Miami Herbert Business School, University of Miami, Miami, Florida, USA*

[2]*Miami Herbert Business School, University of Miami, Miami, Florida, USA*

ARTICLE INFO

Article history:
Received: 23 July, 2024
Revised: 14 September, 2024
Accepted: 15 September, 2024
Online: 20 September, 2024

Keywords:
Explainable AI
Transparency
Interpretability

ABSTRACT

This manuscript offers an in-depth analysis of Explainable Artificial Intelligence (XAI), emphasizing its crucial role in developing transparent and ethically compliant AI systems. It traces AI's evolution from basic algorithms to complex systems capable of autonomous decisions with self-explanation. The paper distinguishes between explainability—making AI decision processes understandable to humans—and interpretability, which provides coherent reasons behind these decisions. We explore advanced explanation methodologies, including feature attribution, example-based methods, and rule extraction technologies, emphasizing their importance in high-stakes domains like healthcare and finance. The study also reviews the current regulatory frameworks governing XAI, assessing their effectiveness in keeping pace with AI innovation and societal expectations. For example, rule extraction from artificial neural networks (ANNs) involves deriving explicit, human-understandable rules from complex models to mimic explainability, thereby making the decision-making process of ANNs transparent and accessible. Concluding, the paper forecasts future directions for XAI research and regulation, advocating for innovative and ethically sound advancements. This work enhances the dialogue on responsible AI and establishes a foundation for future research and policy in XAI.

1. Introduction

Artificial Intelligence (AI) has become an increasingly popular topic in recent years. AI is defined as the capability of a machine to replicate cognitive functions associated with the human mind [1]. As new technologies like ChatGPT emerge, uncertainty about the impact of AI technologies on the business world is steadily growing. The complexity of these systems makes it difficult to understand how AI arrives at its conclusions, resulting in a "black box" scenario where the process used to come to a system output is not fully transparent [2]. The black box syndrome in such systems can create problems in critical fields like finance and medical applications. These fields require more transparency and trust when diagnosing or approving a loan. As the use of AI grows, the demand for explainability within knowledge-based systems increases. In the business community, there is worry about human trust in AI recommendations, leading to a desire for transparency in AI systems [3]. The current lack of transparency in AI systems has led to increased focus on the research of explainable AI. There is a clear need for explainability, trust, and transparency in algorithms across various applications. The concept of Explainable AI generalizes new possibilities for AI programs.

The surge in the adoption of AI systems across various sectors necessitates a parallel increase in explainability to ensure these systems are trustworthy, ethical, and accessible. Here are some concise reasons:

- Regulatory Compliance: Increasing global regulations around data privacy and AI transparency demand mechanisms for explaining and justifying automated decisions, especially in critical sectors like healthcare, finance, and legal.
- Ethical Considerations: As AI systems become more prevalent, the ethical implications of their decisions become more significant. Explainable AI facilitates the understanding of automated decisions, supporting ethical auditing and accountability.
- User Trust: Transparency in AI operations fosters user trust and acceptance, crucial for the widespread deployment of AI technologies in sensitive and impactful areas.

These points underscore the essential role of explainability in the responsible scaling of AI technologies. As we delve deeper into the nuances of AI applications, the complexity of these systems grows [4], highlighting the urgent need for advanced research

[*]Corresponding Author: Maikel Leon (mleon@miami.edu)

in explainable AI. Such research not only aids in aligning AI systems with human values and norms but also opens new avenues for innovation in AI governance and policy-making.

2. AI's history highlights

In recent years, the explainability of systems has emerged as a significant factor in adopting AI. It has become essential for practical, social, and legal reasons that users are provided with an explanation of how a system reaches a particular output [5]. Explanations are necessary to understand a system's functions and give users insight into debugging system issues. However, experts have not defined a reason or the qualities it must possess [5]. In early forms of AI, explainability was not prioritized. The origin of AI can be traced back to the 1940s. These first roots of AI were found in the WWII code-breaking machine developed by English mathematician Alan Turing [6]. The technology's ability to outperform humans in decoding caused Turing to question the system's intelligence. In 1950, Turing released a paper discussing how to produce intelligent systems and test their intelligence. In summary, he proposed a test that considered a machine intelligent if a human cannot distinguish between another human and the machine [6]. Today, the Turing Test is still utilized as a benchmark for recognizing the intelligence of a system.

AI foundation traces back to 1956 at Dartmouth College, which kick-started a new era of machine learning research and development. The first hint of explainability can be found in early knowledge-based expert systems in the 1960s. Rule-based expert systems utilize expert human knowledge to solve problems that usually require human-level intelligence [7]. Using expert or domain knowledge, these software systems assist humans in decision-making. Expert systems use an approach of "if-then" statements and have several essential parts, including a knowledge base (usually formatted as a set of rules), an inference engine, and an interface to convey information to a user [6, 7]. Using a top-down approach, expert systems can quickly formalize human intelligence into logical rules that can be followed step-by-step. 1966, at MIT, Joseph Weizenbaum created ELIZA, a natural language processing tool capable of conversing with a human user [6]. ELIZA was one of the first programs to pass the Turing Test. In the early 1970s, governments began to hesitate and pull back funding for AI research, causing a gap in the development of AI.

In the 1980s, Expert Systems, using AI-derived symbolic reasoning techniques to address complex problems, began demonstrating the technology's ability to achieve a firm's goals [8]. However, critics began to argue that overall, expert systems rarely achieved their set goals and, in many cases, could not achieve expert-level performance [8, 9]. These concerns heavily came from the financial sector, as Wall Street did not trust the technology that rarely delivered on its promises.

Due to this suspicion, there was a significant lack of progress in AI initially. There remained a large gap between the expectations and reality of AI capabilities. Expert systems showed impressive potential when attempting problems that can be seamlessly formalized [10]. For example, in 1997, Deep Blue, IBM's chess-playing program, successfully beat Gary Kasparov, the world chess champion, utilizing a tree-search method to evaluate over 200 million potential moves per second [6]. However, this program could not be successfully applied to a problem that is not as quickly standardized, such as face recognition. For a program to accomplish a task like this, the system must correctly interpret data, learn from it, and apply it to various tasks and goals with flexible adaptation [6].

The need for complex decision-making caused an uproar in AI research. While a few machine learning models are labeled interpretable by design - examples include decision trees, rules, and tables- most AI models function as black boxes, meaning the systems do not reveal sufficient details regarding their internal behavior. [5]. The nature of these opaque decision models will be further discussed in the following section. As AI increasingly intertwines with more human-centric applications, the focus has shifted from accuracy to explainability [11].

In the nascent AI development stages, the primary focus was predominantly on enhancing the accuracy and efficiency of AI models:

- Performance Metrics: Early AI research prioritized performance metrics such as precision and recall, with less consideration for how decisions were made within the model.
- Technological Limitations: Limited by the technology of their times, early developers often had to choose between complex, opaque models that offered better performance and simpler, interpretable ones that did not scale.

While this approach was justified in the early days of AI, when the goal was to establish viable, functional AI systems, today's landscape demands a different paradigm [12]. As AI systems increasingly interact with societal and individual decisions, transparency becomes as critical as accuracy. This shift necessitates a robust exploration of XAI, where understanding and clarifying AI processes are not just an academic interest but a societal imperative [13]. The upcoming sections of this paper will delve into the methodologies and impacts of XAI, seeking to bridge the gap between AI capabilities and human-centric values.

3. What Is XAI?

The field of XAI refers to a wide variety of algorithms. These varying algorithms can be grouped by complexity into three main groups: white, gray, and black box models [14]. White-box models are considered systems with full transparency that do not require extra explainability techniques, such as linear regression [14]. Systems that achieve a more advanced performance but lack interpretability, such as neural networks and random forests, are considered black-box models with high accuracy yet lack transparency [11, 14]. These black boxes are considered opaque models, concealing the methods and algorithms mapping inputs to outputs [15]. For example, an opaque system could emerge when an organization licenses closed-source AI to protect its intellectual property and proprietary AI [15]. The "how" and "why" of the system's process are omitted from the output. Finally, gray-box models fall in between, as they are not intrinsically explainable but can be interpretable when explanation techniques are applied [14].

According to the National Institute of Standards and Technology [16], for a system to be considered explainable, it must possess four fundamental properties:

- Explanation: A system must provide accompanying support or evidence with each decision output.
- Meaningful: The system's explanations are understandable to its intended user, considering different user groups' varying knowledge levels and needs.
- Explanation Accuracy: The system's explanation correctly reflects system processes.
- Knowledge Limits: A system only functions within the range of scenarios and conditions it has been trained for. The system can recognize cases that fall outside its scope.

Knowing what "explainability" means is crucial to understanding the importance of explainable AI. The term does not possess an official definition, but experts have culminated several ways to view the concept of explainability. Explainability describes the type of information provided to users through the user interface to allow informed use of a system's output or recommendation [17]. Explainability answers the simple question, "Why did it do that?".

3.1. Explainability, Interpretability, and Transparency

In many cases, explainability and interpretability are used synonymously; however, according to literature on the topic, interpretability and explainability differ slightly. According to Johnson (2020) and Angelov (2021), the definitions of the terms are as follows:

- Explainability: Relates to the concept of explanation as an interface between AI and humans, including AI systems that are comprehensive to humans through explanation [11].
- Interpretability: The ability to determine cause and effect from a machine learning model that is intrinsically understandable to humans [11, 18].

There are notable qualities that explainable and interpretable systems do and do not possess. The terms used are defined as such:

1. Transparency: The quality of AI systems being understandable by themselves, allowing users to comprehend how the system works [11, 19, 20].
2. User Understanding: the ability of human users to immediately make sense of a system's reasoning and behavior without extra explanations or clarifications [21].
3. Comprehensibility: refers to the capacity of a system or a system's explanations to aid a user in task completion [21].
4. Fairness: The goal that explanations should be egalitarian [21].

Systems can be explainable without being interpretable. Explainability considers explanations of the interface between users and an AI system [11]. Explainability is found in AI systems that are accurate and understandable to humans [11]. In addition, explainability works to clarify its internal decision process to users. It emphasizes the ability of parameters, often hidden in deep neural networks, to justify the results [18, 22]. On the other hand, interpretability relates to how accurately a system can link each cause to an effect [18]. Interpretability describes the capacity of a system to give interpretations in formats understandable to humans. Interpretability also includes to what degree users can understand explanations [23]. For example, deep learning models, such as neural networks, tend to perform highly but lack interpretability [14, 24].

In both interpretable and explainable AI systems, fairness is not guaranteed. Although these techniques provide insight into model behavior and reveal biases, achieving fairness requires the consideration of factors including data bias, algorithmic fairness, and ethical considerations [20]. A system's explainability can be determined by several factors, including complexity, transparency, trust, fidelity, accuracy, and comprehensibility [5, 16, 23]. These dimensions of explainability distinguish explainable systems from black-box models and are critical pieces of explainable AI.

One necessary element of explainable AI is transparency. While explainability answers the question "Why did it do that?" transparency addresses "How does it work?" [25]. In summary, transparency is found in systems that have the potential to be understandable by themselves, making transparent systems the opposite of black box models [11]. Transparency helps lift the lid of black box models. This can reveal a model's structural attributes, evaluation metrics, and descriptive properties from training data to users to foster an understanding of a system's underlying logic [5, 25]. Many machine learning models lacked transparency due to a trade-off between explainability and performance [19]. As previous studies focused on performance improvement, transparency was ignored and placed on the back burner.

AI systems' nontransparent nature began to affect human trust and confidence negatively. More specialized knowledge became necessary to understand AI approaches as the complexity increased. Ordinary users with low algorithmic knowledge found it hard to trust AI systems making crucial decisions, and the lack of transparency hindered user understanding of the exact steps of algorithms [26]. This significantly worsened the problem, as user comprehension of why a specific recommendation is made and how their input affects the results is critical to user satisfaction and trust [26]. For example, in a news recommender system, fair and personalized recommendations give users confidence, leading to trust and continued use [26]. Visible transparency improves search performance, as using explanations improves users' overall satisfaction [26]. In recommender systems, personalization has become a determinant of satisfaction and trust [26]. Moreover, the recommendation explanation sets a prerequisite for a relationship of trust between humans and AI [2]. A lack of transparency in medical applications has been identified as a barrier to AI implementation [23]. Trust in medical AI systems is vital, as the recommendations significantly impact patients' health and well-being [23].

The need for transparency has led to a significant interest in XAI. This field ensures that AI benefits rather than harms society by introducing accountability [3]. Systems that lack transparency don't possess this accountability. In some cases, this is not an issue. For instance, in the historic Go game between Lee Sodel, a highly skilled Go player, and AlphaGo, a DeepMind AI system, AlphaGo made an extremely unexpected move [2]. Experts were unsure why the system made this gaming-altering move. In this case, the nontransparent nature of AlphaGo did not matter, as the

application did not drastically affect human well-being. However, in many applications, the opposite is true.

On the other hand, IBM Watson, a supercomputer containing AI and other analytical software, beat the top players at the game show Jeopardy. This software was then marketed to medical facilities as a cancer-detecting system [2]. When providing results, Watson could not display the reasoning for its output, so patients and doctors could not trust the system [2, 3]. IBM Watson's lack of transparency hindered human trust and was not seen as a successful application. The same mindset can be applied to self-driving cars as well. These automated systems did not react efficiently in a new or unfamiliar environment. In 2018, a computerized vehicle owned by Uber crashed, and the operator was charged with negligent homicide [11]. Transparent and explainable systems are necessary, from a public trust perspective and a legal viewpoint, to provide more reliable and safe systems [11].

The capacity of AI-based systems to elucidate their internal decision-making processes is an area ripe for exploration and innovation:

- Model Transparency: Techniques such as model visualization and feature importance metrics provide insights into the working of complex models, enhancing their transparency.
- Decision Justification: Implementing methods that allow AI systems to justify their decisions can facilitate greater understanding and trust among users.

As AI technologies continue to permeate various aspects of personal and professional life, the ability of these systems to offer clear, understandable explanations for their actions becomes crucial. This supports the development of more robust and reliable AI and upholds the user's right to demand transparency [27]. The next section of this paper will discuss methodologies for formulating these explanations, ensuring that AI systems are effective, accountable, and accessible to the users they serve.

4. Explanations

According to [5], explanations can be understood in two ways: as a line of reasoning or as a problem-solving activity. Viewing explanations as a line of reasoning essentially creates understanding by following the path inference rules take to come to a particular decision [5]. The main issue with this approach was the complexity of explanations, as not all users possess the same knowledge to understand the full extent of explanations thoroughly. This idea was re-conceptualized to approach explanation in a different light: explanations as a problem-solving activity. This altered view not only reconstructs the system's reasoning but also considers various degrees of abstraction, meaning different knowledge levels were considered [28].

Post-hoc and model-based explanations are the most prevalent types when categorizing the explanations provided by XAI systems. Post-hoc methods are commonly used on systems that are not intrinsically interpretable to boost their interpretability [29]. Post-hoc methods do not directly reveal the internal workings of a model. Still, they seek to explain behavior to users by studying outputs and factors that contribute to the result [16]. In other words, explanations are derived after a model makes the prediction. The system uses the nature and attributes of results to generate explanations [17].

On the other hand, model-based explanations focus on the mechanical aspect of recommendations and aim to illustrate how an algorithm suggests a distinct output [5]. Model-based explanation strategies use a different model to explain how the task model functions. The levels of soundness and fidelity are particularly essential for assessing model-based explanations [23]. Model-based explanations are strictly based on the system's underlying assumptions and structure [5]. The following subsections briefly overview post-hoc explanations, addressing different techniques and applications. In addition, several other relevant explanation types, such as self-interpretable models, are referenced.

4.1. Post-hoc Explanations

Post-hoc explainability can be applied in two ways: model-specific and model-agnostic approaches. Model-specific methods produce explanations by utilizing the particular system's internal learning process [30]. Since model-specific interpretability is tailored to bring transparency to specific models, the application will not be suitable for other model types [11, 20, 30]. In contrast, model-agnostic methods are independent of the applied system. Model-agnostic methods develop end-user explanations using the inputs and predictions of the model [20, 30]. The lack of specificity of model-agnostic methods allows for wide-scale usage. In addition, the interpretability of post-hoc models can be further divided into local and global methods.

4.1.1. Local Methods

Local methods obtain explainability by segmenting the solution space and providing less intricate explanations that apply to the entire model [29]. A per-decision or single-decision explanation is the most dominant type of local explanation [16]. It provides insight into the aspects that impact the algorithm's decision for a particular input. Local explanations allow for a local approximation of how a black-box model functions [11]. The most well-known example of local methods is LIME (Local Interpretable Model Agnostic Explainer) [16, 17]. LIME functions by taking a decision and creating an interpretable model that illustrates the local decision, which is then used to deliver per-feature explanations [16]. LIME perturbs training data into a new dataset to form a new interpretable model [11]. Another example of local explanations is SHAP (Shapely Additive exPlanations), which uses a mechanism of additive feature attributions to reveal the significance of input factors [14, 17].

4.1.2. Global Methods

Global methods employ interpretable mechanisms, such as decision trees, to extract a simplified version of a complex black box model to supply understandable explanations for each decision made by the model [11]. This makes it possible to comprehend the behavior of the black-box model and how it relates to its trained characteristics [11]. Global explanations can construct post-hoc explanations

on the whole algorithm [16]. Partial Dependence Plots (PDPs) and Testing with Concept Activation Vectors (TCAV) are examples of global explanations. PDPs demonstrate the modification of predicted responses about altered data components. At the same time, TCAVs explain deep neural networks in a more user-friendly manner and have been applied to image classification systems [16].In addition, a global variant of LIME exists, SP-LIME, which uses applicable local LIME outputs as synopsis explanations [16].

4.2. Self-Interpretable Models

Self-interpretable models are intrinsically explainable, meaning humans can directly understand them. The models are the explanation due to a transparent reasoning process [16, 31, 32]. However, many sources claim self-interpretable models are less accurate than post-hoc explanations due to a trade-off between accuracy and interpretability [16, 33]. The most common self-interpretable models include regression models and decision trees [16, 34].

4.3. Other Explanation Models

In addition, several other explanations exist that do not perfectly fit into a category. The most relevant of these explanation models are defined below.

Forms of Model Explanations:

- Introspective Methods: Explanations are formed by connecting inputs to outputs in black-box models. For example, reflective methods can be applied to image classifications with Deep Neural Networks [5, 35, 36] and [37].
- Counterfactual Methods: Explanations provide "what-if" statements regarding how the outputs of a predicted model could be affected by input changes [5, 38, 39, 40] and [41].
- Explanation by Feature Relevance: A method of post-hoc explainability clarifies a model's internal functioning by calculating a relevance score for each variable. The comparison of scores depicts the weight each variable holds [20, 42] and [43].
- Explanation by Simplification: Explanations that use a trained model to formulate a simplified representation to assemble an easily implementable model. These models optimize similarity to the original model while simultaneously decreasing complexity [11, 29] and [44].

AI-based systems must explain their decisions, which may soon transition from a best practice to a mandatory requirement. This shift is driven both by evolving regulatory frameworks aimed at safeguarding consumer rights and by ethical standards that promote transparency and accountability [45]:

- Regulatory Compliance: Legislations such as the EU's General Data Protection Regulation (GDPR) already impose obligations on AI to explain decisions that affect individuals, signaling a broader trend towards legal mandates.
- Ethical Accountability: Beyond compliance, there is a growing recognition of the ethical obligation for AI to be transparent, particularly in systems that impact public welfare and individual freedoms.

This development is poised to significantly benefit numerous business sectors by enhancing consumer trust, facilitating more informed decision-making, and improving the overall user experience with AI technologies.

5. From ANNs (sub-symbolic) to Rules (symbolic)

Extracting rules from ANNs is crucial in demystifying these models' "black-box" nature, making their decisions understandable and interpretable to humans. This process involves translating the intricate, non-linear relationships learned by the network into a set of rules that humans can easily understand. To illustrate this process, we'll explore a detailed example of how rules can be extracted from an ANN trained on a simplified dataset for predicting loan approval based on applicant features.

5.1. Background

Let us use the example of a fictional financial institution that has created an ANN to evaluate loan applications. The ANN considers various applicant features such as Age, Income, Credit Score, and Employment Status and provides a binary decision: Approve or Deny. Despite the ANN's high accuracy, the decision-making process is not transparent. This makes it challenging for loan officers to explain decisions to applicants or to ensure compliance with regulations. The institution aims to derive understandable rules from the ANN to address this.

5.2. ANN Architecture

The ANN in this example is a simple feedforward network with one hidden layer. The input layer has four neurons corresponding to the applicant features. The hidden layer has a few neurons (say five for simplicity) using ReLU (Rectified Linear Unit) as the activation function [41]. The output layer has one neuron and uses a sigmoid activation function to output a probability of loan approval.

5.3. Rule Extraction Process

The rule extraction process involves several steps designed to translate the ANN's learned weights and biases into a set of if-then rules that replicate the network's decision-making process as closely as possible:

- Simplification: The first step involves simplifying the ANN to make the rule extraction more manageable. This could include pruning insignificant weights (shallows values) and neurons that have little impact on the output based on sensitivity analysis.
- Discretization: Since ANNs deal with continuous inputs and hidden layer activations, a discretization process is applied to convert these continuous values into categorical ranges. For

instance, age might be categorized into 'Young', 'Middle-aged', and 'Old'; Income into 'Low', 'Medium', and 'High'; Credit Score into 'Poor', 'Fair', 'Good', and 'Excellent'; and Employment Status into 'Unemployed' and 'Employed'.

- Activation Pattern Analysis: Next, the activation patterns of the neurons in the hidden layer are analyzed for each input pattern. This involves feeding various combinations of the discretized input variables into the simplified network and observing which neurons in the hidden layer are activated for each combination. An activation threshold is defined to determine whether a neuron is considered activated.

- Rule Generation: Based on the activation patterns observed, rules are generated to replicate the ANN's decision process. Each rule corresponds to a path from the input layer through the activated hidden neurons to the output decision. For example:

  - If (Age is Young) and (Income is High) and (Credit Score is Good) and (Employment Status is Employed), then Approve Loan.
  - If (Age is Middle-aged) and (Credit Score is Poor), then Deny Loan.

  This step involves identifying which combinations of input features and hidden neuron activations lead to loan approval or denial, effectively translating the ANN's complex decision boundaries into more interpretable formats.

- Rule Refinement and Validation: The initial set of rules may be too complex or too numerous for practical use. Rule refinement techniques simplify and consolidate the rules without significantly reducing their accuracy in replicating the ANN's decisions. The refined rules are then validated against a test dataset to accurately reflect the ANN's behavior. This may involve adjusting the rules based on misclassifications or applying techniques to handle exceptions and edge cases.

After applying the rule extraction process to our hypothetical ANN, we might end up with a set of simplified, human-readable rules such as:

- Rule 1: If (Income is High) and (Credit Score is Excellent), then Approve Loan.
- Rule 2: If (Employment Status is Unemployed) and (Credit Score is Poor or Fair), then Deny Loan.
- Rule 3: If (Age is Old) and (Income is Low) and (Employment Status is Employed), then Deny Loan.

These rules provide clear criteria derived from the ANN's learned patterns, making the decision-making process transparent and justifiable.

5.4. Advantages and Challenges

Some advantages include:

- Transparency: The extracted rules make the ANN's decisions transparent and understandable to humans.
- Compliance: Clear rules can help ensure compliance with regulatory requirements for explainable AI.
- Trust: Understanding how decisions are made can increase user trust in the AI system.

Some challenges are:

- Complexity: The rule extraction process can be complex, especially for deep or highly non-linear networks [46].
- Approximation: The extracted rules approximate the ANN's decision process and may not capture all nuances.
- Scalability: Extracting rules from large, deep neural networks with many inputs and hidden layers can be challenging and may result in many complex rules [47].

5.5. Summary

Extracting rules from ANNs makes AI decision-making transparent, understandable, and justifiable. Although there are challenges, especially with complex networks, this process is crucial for responsible and ethical AI use. By making AI systems more interpretable, we can establish trust with users, ensure compliance with regulations, and gain valuable insights into decision-making.

6. Fuzzy Cognitive Maps

The pendulum in AI is swinging back from purely statistical approaches toward integrating structured knowledge. FCMs are powerful cognitive tools for modeling and simulating complex systems. They blend elements from artificial neural networks, graph theory, and semantic nets to offer a unique approach to understanding and predicting system behavior. FCMs incorporate the concept of fuzziness from fuzzy logic, enabling them to handle ambiguity and uncertainty inherent in real-world scenarios. This extensive report delves into the origins of FCMs, provides illustrative case studies, and discusses their advantages and disadvantages, with references to their similarities to artificial neural networks, graphs, and semantic nets [48].

6.1. Origins

Bart Kosko introduced the concept of FCMs in the 1980s as an extension of cognitive maps. Cognitive maps, developed by Axelrod, were diagrams that represented beliefs and their interconnections. Kosko's introduction of fuzziness to these maps allowed for the representation of causal reasoning with degrees of truth rather than binary true/false values, thus capturing the uncertain and imprecise nature of human knowledge and decision-making processes. FCMs combine elements from fuzzy logic, introduced by Lotfi A. Zadeh, with the structure of cognitive maps to model complex systems.

6.2. Structure and Functionality

FCMs are graph-based representations where nodes represent concepts or entities within a system, and directed edges depict the causal relationships between these concepts. Each edge is assigned a weight that indicates the relationship's strength and direction (positive or negative). This structure closely mirrors that of artificial neural networks, particularly in how information flows through the network and how activation levels of concepts are updated based on the input they receive, akin to the weighted connections between neurons in neural networks [49].

However, unlike typical neural networks that learn from data through backpropagation or other learning algorithms, the weights in FCMs are often determined by experts or derived from data using specific algorithms designed for FCMs. The concepts in FCMs can be activated like neurons, with their states updated based on fuzzy causal relations, allowing for dynamic modeling of system behavior over time. Integrating structured knowledge graphs with distributed neural network representations offers a promising path to augmented intelligence. We get the flexible statistical power of neural networks that predict, classify, and generate based on patterns—combined with the formalized curated knowledge encoding facts, logic, and semantics via knowledge graphs [50].

6.3. The Inherent Reasoning Mechanism

The primary function of the reasoning rule in FCM models is to update the activation values of concepts iteratively, starting from initial conditions and continuing until a stopping criterion is satisfied. During each iteration, the reasoning rule utilizes three primary components to conduct these calculations: the weight matrix, which signifies the connections between concepts; the activation values of concepts from the previous iteration; and the activation function.

Eq. (1) shows a general rule commonly found in FCMs-related papers:

$$a_i^{(t)} = f\left(\sum_{j=1, i\neq j}^{N} a_j^{(t-1)} w_{ji}\right), \tag{1}$$

Recently, in [51], the author proposed an updated quasi-nonlinear reasoning rule depicted in Eq. (2):

$$a_i^{(t)} = \underbrace{\phi \cdot f\left(\sum_{j=1}^{N} a_j^{(t-1)} w_{ji}\right)}_{\text{nonlinear component}} + \underbrace{(1-\phi)\cdot a_i^{(0)}}_{\text{linear component}}, \tag{2}$$

such that $0 \leq \phi \leq 1$ is the nonlinearity coefficient. When $\phi = 1$, the concept's activation value depends on the activation values of connected concepts in the previous iteration. When $0 < \phi < 1$, we add a linear component to the reasoning rule devoted to preserving the initial activation values of concepts. When $\phi = 0$, the model narrows down to a linear regression where the initial activation values of concepts act as regressors. In their paper, Nápoles et al. [51] used the quasi-nonlinear reasoning rule to quantify implicit bias in pattern classification datasets. In contrast, the authors in [41] resorted to this rule to develop a recurrence-aware FCM-based classifier.

6.4. How Activation Functions Work

The activation function $f : \mathbb{R} \rightarrow I$ is an essential component in the reasoning rule of FCM-based models. This monotonically non-decreasing function keeps the activation value of each concept within the desired image set I, which can be discrete (a finite set) or continuous (a numeric-valued interval). It should be mentioned that I must be bounded; otherwise, the reasoning rule could explode due to the successive additions and multiplications when updating concepts' activation values during reasoning. Table **??** portrays relevant activation functions found in the literature.

6.5. Relevant Case Studies

For illustration purposes, Figure 1 shows an example of an FCM created to model a case of autism [32]. FCMs have been applied across various domains, demonstrating their versatility and effectiveness as a hybrid AI tool:

- Decision Support Systems: FCMs model complex decision-making processes, integrating expert knowledge and data-driven insights to support decisions in healthcare, environmental management, and business strategy.

- Predictive Modeling: In healthcare, FCMs model the progression of diseases or the impact of treatments, incorporating medical expertise and patient data to predict outcomes and support personalized medicine [52].

- System Analysis and Design: FCMs help analyze and design complex systems, such as socio-economic systems or ecosystems, by modeling the interactions between various factors and predicting the impact of changes or interventions.

- Healthcare Management: FCMs have been employed to model and predict patient outcomes in healthcare settings. For example, an FCM can be developed to understand the complex interplay between patient symptoms, treatment options, and possible outcomes, aiding medical professionals in decision-making [53].

- Environmental and Ecological Systems: In environmental studies, FCMs have been used to model the impact of human activities on ecosystems, allowing for the simulation of various scenarios based on different policies or interventions. This application showcases the strength of FCMs in handling systems where data may be scarce or imprecise [54].

- Business and Strategic Planning: FCMs assist in strategic planning and decision-making within business contexts by modeling the relationships between market forces, company policies, and financial outcomes, offering a tool for scenario analysis and strategy development [55].

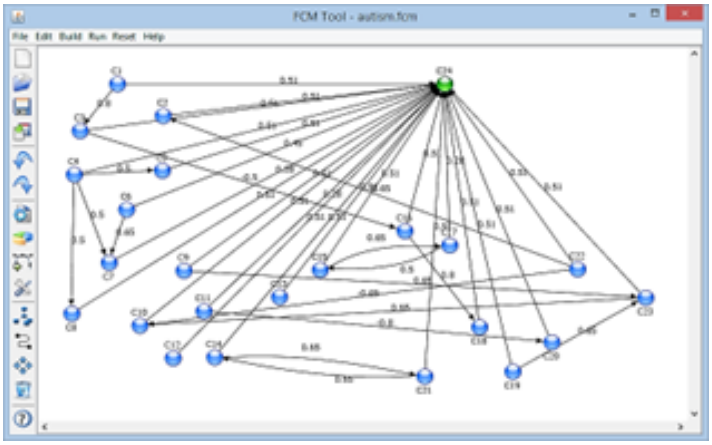


Figure 1: Real example created with FCM Tool.

6.6. Advantages

The hybrid nature of FCMs offers several advantages:

- Interpretability and Transparency: The symbolic representation of concepts and causal relationships in FCMs provides clarity and understandability, facilitating communication with experts and stakeholders and supporting explainable AI.
- Flexibility and Adaptability: FCMs can be easily updated with new knowledge or data, allowing them to adapt to changing conditions or insights. This makes them particularly valuable in fields where knowledge evolves rapidly.
- Handling of Uncertainty: Using fuzzy values to represent causal strengths enables FCMs to deal effectively with uncertainty and ambiguity, providing more nuanced and realistic modeling of complex systems [4].
- Integration of Expert Knowledge and Data-Driven Insights: FCMs uniquely combine expert domain knowledge with learning from data, bridging the gap between purely knowledge-driven and purely data-driven approaches.
- Interpretability: The graphical representation of FCMs, similar to semantic nets, allows for straightforward interpretation and understanding of the modeled system, making it accessible to experts and stakeholders without deep technical knowledge of AI.
- Flexibility: FCMs can incorporate quantitative and qualitative data, effectively handling uncertainty and imprecision through fuzzy logic. This flexibility makes them suitable for a wide range of applications.
- Dynamic Modeling Capability: FCMs can simulate the dynamic behavior of systems over time, providing valuable insights into potential future states based on different inputs or changes in the system [56].

6.7. Limitations

Despite their advantages, FCMs also face several challenges:

- Complexity with Large Maps: As the number of concepts and relationships in an FCM increases, the map can become complex and challenging to manage, analyze, and interpret [57].
- Learning and Optimization: While FCMs can learn from data, adjusting the fuzzy values of causal relationships can be computationally intensive and require sophisticated optimization techniques, especially for large and complex maps [58].
- Quantification of Expert Knowledge: Translating expert knowledge into precise fuzzy values for causal relationships can be challenging and may introduce subjectivity, requiring careful validation and sensitivity analysis [59].
- Subjectivity in Model Construction: The reliance on expert knowledge for constructing FCMs can introduce subjectivity, especially in determining the strength and direction of causal relationships between concepts.
- Complexity with Large Maps: As the number of concepts increases, the FCM can become complex and challenging to manage and interpret, potentially requiring sophisticated computational tools for simulation and analysis.
- Limited Learning Capability: While FCMs can be adjusted or trained based on data to some extent, they lack the deep learning capabilities of more advanced neural networks, which can autonomously learn complex patterns from large datasets [60].

7. Applications

Numerous potential applications exist for XAI techniques and models, including healthcare, law, data science, and business [55]. This section explores the need for explainability in these applications, including their current uses, limitations, and future development.

7.1. Healthcare

In healthcare, there are many applications of XAI such as diagnosis, treatment recommendations, and surgery [23, 61, 62]. For example, an explainable model was proposed for diagnosing skin diseases. Using saliency maps to highlight important parts of the image crucial to diagnosis, dermatologists can easily understand the model's arrival at a diagnosis and then provide a more in-detail diagnosis [61]. According to a survey by [62], LIME is the most commonly used XAI approach in medical applications.

Throughout the COVID-19 pandemic, AI has shown potential in developing solutions to confront the difficulties presented by the virus [61]. However, the lack of transparency in black-box models has hindered their acceptance in clinical practice. With the development in user trust and model performance, XAI can attempt these problems in the future [61]. XAI techniques have been created in the context of medical image analysis to facilitate disease detection and diagnosis through feature visualization [61]. This allows medical professionals and their patients to obtain a deeper insight into the model's process, building confidence in its accuracy. In high-stakes applications, specifically healthcare, there is debate about whether

explainable modeling is necessary. To some, explainability is crucial. On the other hand, some say prioritizing explainability above accuracy in healthcare systems can be unethical [23]. According to [23], the post-hoc explanations can be delusive, but a potential solution is to create post-hoc explanation models with argumentative support.

Suppose the case of an ANN equipped with a rule extraction method can be deployed to diagnose diseases from medical imaging with high accuracy. The ANN processes complex imaging data to identify patterns indicative of specific conditions, such as tumors in MRI scans. A rule extraction technique is integrated into the system to ensure clinicians and patients understand the diagnostic process. This technique translates the ANN's intricate decision-making into simple, interpretable rules, such as the presence of specific shapes or textures associated with malignancy. This not only aids medical professionals in making informed treatment decisions but also enhances patient trust by providing clear explanations for the diagnoses made by the AI system.

7.2. Law

In the context of legal applications, XAI possesses several potential applications. As stated by [61], XAI can be used for legal document analysis, contract review, legal decision-making, and addressing challenges in legal domains. AI can help analyze large volumes of legal documents and sort significant information to facilitate a more accurate analysis, as well as assist in recommending plea bargains or predicting case outcomes [61, 63]. Despite the increasing emphasis on AI in the legal world, systems still struggle to perform at necessary levels due to the precise nature of legal work. Such characteristics include the exact nature of legal jargon, the high level of expertise required, the mass amount of situational exceptions, and the limited tolerance of mistakes [61]. The motivation for interpretable, explainable, and trustworthy systems feeds the recent upsurge of XAI research in legal applications.

In legal applications, an FCM can be a sophisticated tool for modeling and visualizing the intricate dynamics of legal cases and legislative processes. By capturing and representing the causal relationships between various legal factors—such as statutes, precedents, and evidentiary variables—FCMs enable legal professionals to simulate and scrutinize the potential outcomes of different legal strategies in a visually interpretable format. This capability goes beyond basic explainability by showing outcomes and allowing users to interact with the map to adjust variables and immediately see different scenario outcomes. This interactive, interpretable visualization aids in understanding complex legal interdependencies, facilitating more informed decision-making and strategy formulation, especially in cases involving overlapping laws and diverse outcomes.

7.3. Finance

In the financial sector, the applications of XAI can be split into thematic categories. These clusters include financial distress and corporate failure, algorithmic and high-frequency trading, forecasting/predictive analysis, text mining and sentiment analysis, financial fraud, pricing and valuation, scheduling, and investor behavior [64]. In addition, [61] describe the potential applications of XAI in finance as follows:

- **Fraud Detection:** Explain decisions by identifying the reasons behind fraudulent activities and prevent future issues.
- **Credit Scoring:** Allows banks and their customers to understand exactly why a particular credit score was calculated and facilitates lending decisions.
- **Investment Management:** Increased transparency in portfolio management can lead to better performance and more satisfied investors.
- **Compliance:** XAI could assist in mitigating potential biases and avoiding legal issues.
- **Customer Service:** XAI will improve customer service by, for example, including explanations along with loan denials to improve customer understanding and satisfaction.

According to additional literature on the topic, subjects within the finance domain commonly discussed as potential applications of XAI include risk management, portfolio optimization, electronic financial transaction clarification, and anti-money laundering [64]. Due to the high level of regulations in financial domains, XAI is necessary to augment processes to ensure trust and transparency and mitigate risks [65].

Suppose the case of an ANN equipped with a rule extraction method can be effectively used for credit scoring. The ANN analyzes extensive data sets, including transaction history, payment behavior, and credit utilization, to assess the creditworthiness of applicants. By integrating a rule extraction method, the system can transparently generate and provide clear, human-understandable rules that explain its credit-scoring decisions. This transparency not only aids financial analysts in understanding the model's decision-making process but also ensures compliance with regulatory requirements regarding fairness and explainability in credit assessments.

An FCM can model and visualize a client's financial stability or market for the same finance application. By representing elements like market trends, economic indicators, and individual financial behaviors as nodes and their interdependencies as edges, FCMs allow financial analysts to simulate and interpret complex financial scenarios. This method provides a dynamic, interpretable visualization beyond mere explanation, enabling interactive exploration of potential financial outcomes based on varying inputs. Such interpretability is invaluable in strategic financial planning and risk assessment, allowing the decision-makers to foresee and mitigate potential financial instabilities or crises.

8. Future

As complex and human-centric systems become more prevalent, there is a growing need for explainable AI in many applications. Due to the rapid increase in AI, there are currently few regulations and rules governing these systems. However, as the need for trust and transparency continues to rise, regulations are essential to ensure both ethical and accountable AI.

8.1. Current Regulations

Historically, AI-based systems have operated in an environment with minimal regulatory oversight regarding their need to explain internal decision-making processes:

- Early AI Developments: Initially, AI technologies were developed and deployed with a focus on functionality and performance, often at the expense of transparency and accountability [66].

- Regulatory Lag: There has been a significant lag in developing and implementing regulations that require AI systems to be explainable, partly due to the rapid pace of technological advancement outstripping policy development.

However, as the implications of AI technologies have become more apparent, there is a growing consensus among government bodies and policymakers about the necessity of regulatory frameworks that ensure AI systems are transparent and accountable. This shift reflects a broader awareness of AI's potential impacts on society and the need for appropriate safeguards.

The regulation of AI is becoming extremely important in terms of ethics and responsible decision-making. The European Union's General Data Protection Regulation (GDPR) was put into effect in 2018, and the GDPR has raised several legal and ethical questions regarding safety, responsibility, malfunction liability, and the overall trade-offs associated with AI decisions [67]. The GDPR gives citizens a "right to explanation" in algorithmic choices that significantly affect them [68, 69]. Regulations like the GDPR make it nearly impossible to use black-box models in various sectors, emphasizing the growing need for explainability and transparency [70, 71]. Additionally, the EU's intense regulatory actions involving digital markets, including the AI domain, strive to provide an ethical approach to AI applications [72]. Additionally, Hacker (2023) highlights the transformative prospects as well as risks associated with large generative AI models (LGAIMs), such as ChatGPT, and how current regulations are not suited to manage this class of AI [73].

In April 2021, the European Commission proposed a groundbreaking proposal for the first-ever EU regulatory framework for AI. This framework consists of a risk-based classification technique in which the level of risk specifies the regulation applied to a system [74]. The AI Act manages the opacity of particular systems, emphasizing systems classified as high-risk through a focus on transparency [75]. If implemented, the AI Act will represent the world's baseline rules for overseeing AI. Furthermore, generative AI systems such as ChatGPT must follow transparency conditions, such as publishing data synopses for training the system [74].

In summary, AI regulations are developing to address ethical considerations, transparency, and the responsible use of AI across diverse sectors. The GDPR and corresponding endeavors emphasize the demand for transparency and accountability in AI decision-making. At the same time, ongoing discussions in the EU seek to shape AI development in a human-centric and ethical fashion.

8.2. The Future of XAI

The future of XAI holds tremendous promise and challenges. In an increasingly AI-driven world, the possible applications are extensive; however, awareness of the fragile nature and potential biases within AI systems is expanding. As stated previously, global organizations are attempting to craft standards for responsible AI to mitigate concerns. These regulations strive to make AI systems exemplify more transparency and accountability, making the demand for explainable systems higher than ever.

As different organizations and governments pass regulations, the dilemma now shifts: Is regulating the AI available to specific users and not others ethical? When tackling this issue, enforcing rules on AI is essential. Without universal regulations, organizations may pass conflicting laws, which could immensely harm companies attempting to operate systems globally. For example, with search engines experimenting with generative AI systems, such as Google's Bard or Gemini, non-universal regulations would require several system versions to adhere to local regulations, causing unnecessary complexities. Moreover, universal regulations would provide businesses with legal certainty. Ethically, universal regulations will form a standard for ethical AI, assisting in eliminating biased and discriminatory systems. This will also allow users to feel more trust in consistently observed systems, leading to increased adoption of systems.

In conclusion, from a business perspective, the universal enforcement of AI regulations provides many advantages. Companies should prioritize accountable AI and support coordinated regulations to develop ethical, transparent, and innovative AI technologies. Explainable systems are the key to the future of Responsible AI.

References

[1] V. K. Michael Chui, B. McCarthy, "An Executives Guide to AI," McKinsey & Company, 2018.

[2] W. Samek, T. Wiegand, K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," arXiv preprint arXiv:1708.08296, 2017, doi:10.48550/arXiv.1708.08296.

[3] "Unlocking the black box with explainable AI - Infosys," Infosys, 2019.

[4] M. Leon, "Aggregating Procedure for Fuzzy Cognitive Maps," The International FLAIRS Conference Proceedings, **36**(1), 2023, doi:10.32473/flairs.36.133082.

[5] R. Confalonieri, L. Coba, B. Wagner, T. R. Besold, "A historical perspective of explainable Artificial Intelligence," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, **11**(1), e1391, 2021, doi:10.1002/widm.1391.

[6] M. Haenlein, A. Kaplan, "A brief history of artificial intelligence: On the past, present, and future of artificial intelligence," California Management Review, **61**(4), 5–14, 2019, doi:10.1177/0008125619864925.

[7] A. Abraham, "Rule-Based expert systems," Handbook of Measuring System Design, 2005, doi:10.1002/9780470027325.s6405.

[8] T. G. Gill, "Early expert systems: Where are they now?" MIS Quarterly, **19**(1), 51–81, 1995, doi:10.2307/249711.

[9] J. Kastner, S. Hong, "A review of expert systems," European Journal of Operational Research, **18**(3), 285–292, 1984, doi:10.1016/0377-2217(84)90202-0.

[10] E. Struble, M. Leon, E. Skordilis, "Intelligent Prevention of DDoS Attacks using Reinforcement Learning and Smart Contracts," The International FLAIRS Conference Proceedings, **37**(1), 2024, doi:10.32473/flairs.37.1.135349.

[11] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, P. M. Atkinson, "Explainable artificial intelligence: an analytical review," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, **11**(5), e1424, 2021, doi:10.1002/widm.1424.

[12] M. Leon, "Fuzzy Cognitive Maps as a Bridge between Symbolic and Subsymbolic Artificial Intelligence," International Journal on Cybernetics & Informatics (IJCI). 3rd International Conference on Artificial Intelligence Advances (AIAD 2024), **13**(4), 57–75, 2024, doi:10.5121/ijci.2024.130406.

[13] M. Leon, L. Mkrtchyan, B. Depaire, D. Ruan, K. Vanhoof, "Learning and clustering of fuzzy cognitive maps for travel behaviour analysis," Knowledge and Information Systems, **39**(2), 435–462, 2013, doi:10.1007/s10115-013-0616-z.

[14] M. Schemmer, N. Kühl, G. Satzger, "Intelligent decision assistance versus automated decision-making: Enhancing knowledge work through explainable artificial intelligence," arXiv preprint arXiv:2109.13827, 2021, doi:10.48550/arXiv.2109.13827.

[15] D. Doran, S. Schulz, T. R. Besold, "What does explainable AI really mean? A new conceptualization of perspectives," arXiv preprint arXiv:1710.00794, 2017, doi:10.48550/arXiv.1710.00794.

[16] P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski, M. A. Przybocki, "Four principles of explainable artificial intelligence," Gaithersburg, Maryland, 18, 2020, doi:10.6028/NIST.IR.8312.

[17] P. Bhattacharya, N. Ramesh, "Explainable AI: A Practical Perspective," Infosys, 2020.

[18] J. Johnson, "Interpretability vs explainability: The black box of machine learning," BMC Blogs, 2020.

[19] D. Minh, H. X. Wang, Y. F. Li, T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," Artificial Intelligence Review, **55**, 3503–3568, 2022, doi:10.1007/s10462-021-10088-y.

[20] A. Rai, "Explainable AI: From black box to glass box," Journal of the Academy of Marketing Science, **48**, 137–141, 2020, doi:10.1007/s11747-019-00710-5.

[21] C. Meske, E. Bunde, J. Schneider, M. Gersch, "Explainable Artificial Intelligence: Objectives, Stakeholders and Future Research Opportunities," Information Systems Management, 2020, doi:10.1080/10580530.2020.1849465.

[22] S. S. Band, A. Yarahmadi, C.-C. Hsu, M. Biyari, M. Sookhak, R. Ameri, I. Dehzangi, A. T. Chronopoulos, H.-W. Liang, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," Informatics in Medicine Unlocked, **40**, 101286, 2023, doi:10.1016/j.imu.2023.101286.

[23] A. F. Markus, J. A. Kors, P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies," Journal of Biomedical Informatics, **113**, 103655, 2021, doi:10.1016/j.jbi.2020.103655.

[24] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, A. Holzinger, "Explainable AI: the new 42?" in Machine Learning and Knowledge Extraction: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27–30, 2018, Proceedings 2, 295–303, Springer, 2018, doi:10.1007/978-3-319-99740-7_21.

[25] C. Oxborough, E. Cameron, A. Rao, A. Birchall, A. Townsend, C. Westermann, "Explainable AI: Driving business value through greater understanding," Retrieved from PWC website: https://www.pwc.co.uk/audit-assurance/assets/explainable-ai.pdf, 2018.

[26] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI," International Journal of Human-Computer Studies, **146**, 102551, 2021, doi:10.1016/j.ijhcs.2020.102551.

[27] G. Nápoles, M. L. Espinosa, I. Grau, K. Vanhoof, R. Bello, Fuzzy cognitive maps based models for pattern classification: Advances and challenges, volume 360, 83–98, Springer Verlag, 2018.

[28] G. Nápoles, M. Leon, I. Grau, K. Vanhoof, "FCM Expert: Software Tool for Scenario Analysis and Pattern Classification Based on Fuzzy Cognitive Maps," International Journal on Artificial Intelligence Tools, **27**(07), 1860010, 2018, doi:10.1142/S0218213018600102.

[29] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, **58**, 82–115, 2020, doi:10.1016/j.inffus.2019.12.012.

[30] D. Vale, A. El-Sharif, M. Ali, "Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law," AI and Ethics, **2**, 815–826, 2022, doi:10.1007/s43681-022-00142-y.

[31] M. Xue, Q. Huang, H. Zhang, L. Cheng, J. Song, M. Wu, M. Song, "Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition," arXiv preprint arXiv:2208.10431, 2022, doi:10.48550/arXiv.2208.10431.

[32] M. Leon Espinosa, G. Napoles Ruiz, "Modeling and Experimentation Framework for Fuzzy Cognitive Maps," Proceedings of the AAAI Conference on Artificial Intelligence, **30**(1), 2016, doi:10.1609/aaai.v30i1.9841.

[33] A. Adadi, M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," IEEE Access, **6**, 52138–52160, 2018, doi:10.1109/ACCESS.2018.2870052.

[34] V. G. Costa, C. E. Pedreira, "Recent advances in decision trees: An updated survey," Artificial Intelligence Review, **56**(5), 4765–4800, 2023, doi:10.1007/s10462-022-10275-5.

[35] J. F. Allen, S. Schmidt, S. A. Gabriel, "Uncovering Strategies and Commitment Through Machine Learning System Introspection," SN Computer Science, **4**(4), 322, 2023, doi:10.1007/s42979-023-01747-8.

[36] A. Heuillet, F. Couthouis, N. Díaz-Rodríguez, "Explainability in deep reinforcement learning," Knowledge-Based Systems, **214**, 106685, 2021, doi:10.48550/arXiv.2008.06693.

[37] P. Sequeira, E. Yeh, M. T. Gervasio, "Interestingness Elements for Explainable Reinforcement Learning through Introspection." in IUI workshops, volume 1, 2019, doi:10.48550/arXiv.1912.09007.

[38] X. Dai, M. T. Keane, L. Shalloo, E. Ruelle, R. M. Byrne, "Counterfactual explanations for prediction and diagnosis in XAI," in Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 215–226, 2022, doi:10.1145/3514094.3534144.

[39] M. T. Keane, E. M. Kenny, E. Delaney, B. Smyth, "If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques," arXiv preprint arXiv:2103.01035, 2021, doi:10.48550/arXiv.2103.01035.

[40] G. Warren, M. T. Keane, R. M. Byrne, "Features of Explainability: How users understand counterfactual and causal explanations for categorical and continuous features in XAI," arXiv preprint arXiv:2204.10152, 2022, doi:10.48550/arXiv.2204.10152.

[41] G. Nápoles, Y. Salgueiro, I. Grau, M. Leon, "Recurrence-Aware Long-Term Cognitive Network for Explainable Pattern Classification," IEEE Transactions on Cybernetics, **53**(10), 6083–6094, 2023, doi:10.48550/arXiv.2107.03423.

[42] P. A. Moreno-Sanchez, "An automated feature selection and classification pipeline to improve explainability of clinical prediction models," in 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI), 527–534, IEEE, 2021, doi:10.1109/ICHI52183.2021.00100.

[43] J. Tritscher, A. Krause, A. Hotho, "Feature relevance XAI in anomaly detection: Reviewing approaches and challenges," Frontiers in Artificial Intelligence, **6**, 1099521, 2023, doi:10.3389/frai.2023.1099521.

[44] J. Tritscher, M. Ring, D. Schlr, L. Hettinger, A. Hotho, "Evaluation of post-hoc XAI approaches through synthetic tabular data," in Foundations of Intelligent Systems: 25th International Symposium, ISMIS 2020, Graz, Austria, September 23–25, 2020, Proceedings, 422–430, Springer, 2020, doi:10.1007/978-3-030-59491-6_40.

[45] G. Nápoles, F. Hoitsma, A. Knoben, A. Jastrzebska, M. Leon, "Prolog-based agnostic explanation module for structured pattern classification," Information Sciences, **622**, 1196–1227, 2023, doi:10.1016/j.ins.2022.12.012.

[46] Z. Yang, J. Liu, K. Wu, "Learning of Boosting Fuzzy Cognitive Maps Using a Real-coded Genetic Algorithm," in 2019 IEEE Congress on Evolutionary Computation (CEC), 966–973, 2019, doi:10.1109/CEC.2019.8789975.

[47] W. Liang, Y. Zhang, X. Liu, H. Yin, J. Wang, Y. Yang, "Towards improved multifactorial particle swarm optimization learning of fuzzy cognitive maps: A case study on air quality prediction," Applied Soft Computing, **130**, 109708, 2022, doi:10.1016/j.asoc.2022.109708.

[48] Y. Hu, Y. Guo, R. Fu, "A novel wind speed forecasting combined model using variational mode decomposition, sparse auto-encoder and optimized fuzzy cognitive mapping network," Energy, **278**, 127926, 2023, doi:10.1016/j.energy.2023.127926.

[49] W. Hoyos, J. Aguilar, M. Toro, "A clinical decision-support system for dengue based on fuzzy cognitive maps," Health Care Management Science, **25**(4), 666–681, 2022, doi:10.1007/s10729-022-09611-6.

[50] W. Hoyos, J. Aguilar, M. Toro, "PRV-FCM: An extension of fuzzy cognitive maps for prescriptive modeling," Expert Systems with Applications, **231**, 120729, 2023, doi:10.1016/j.eswa.2023.120729.

[51] G. Nápoles, I. Grau, L. Concepción, L. K. Koumeri, J. P. Papa, "Modeling implicit bias with fuzzy cognitive maps," Neurocomputing, **481**, 33–45, 2022.

[52] K. Poczeta, E. I. Papageorgiou, "Energy Use Forecasting with the Use of a Nested Structure Based on Fuzzy Cognitive Maps and Artificial Neural Networks," Energies, **15**(20), 7542, 2022, doi:10.3390/en15207542.

[53] G. D. Karatzinis, N. A. Apostolikas, Y. S. Boutalis, G. A. Papakostas, "Fuzzy Cognitive Networks in Diverse Applications Using Hybrid Representative Structures," International Journal of Fuzzy Systems, **25**(7), 2534–2554, 2023, doi:10.1007/s40815-023-01564-4.

[54] O. Orang, P. C. de Lima e Silva, F. G. Guimarães, "Time series forecasting using fuzzy cognitive maps: a survey," Artificial Intelligence Review, **56**, 7733–7794, 2023, doi:10.1007/s10462-022-10319-w.

[55] M. Leon, "Business Technology and Innovation Through Problem-Based Learning," in Canada International Conference on Education (CICE-2023) and World Congress on Education (WCE-2023), Infonomics Society, 2023, doi:10.20533/cice.2023.0034.

[56] E. Jiya, O. Georgina, A. O., "A Review of Fuzzy Cognitive Maps Extensions and Learning," Journal of Information Systems and Informatics, **5**(1), 300–323, 2023, doi:10.51519/journalisi.v5i1.447.

[57] R. Schuerkamp, P. J. Giabbanelli, "Extensions of Fuzzy Cognitive Maps: A Systematic Review," ACM Comput. Surv., **56**(2), 53:1–53:36, 2023, doi:10.1145/3610771.

[58] S. Yang, J. Liu, "Time-Series Forecasting Based on High-Order Fuzzy Cognitive Maps and Wavelet Transform," IEEE Transactions on Fuzzy Systems, **26**(6), 3391–3402, 2018, doi:10.1109/TFUZZ.2018.2831640.

[59] T. Koutsellis, G. Xexakis, K. Koasidis, N. Frilingou, A. Karamaneas, A. Nikas, H. Doukas, "In-Cognitive: A web-based Python application for fuzzy cognitive map design, simulation, and uncertainty analysis based on the Monte Carlo method," SoftwareX, **23**, 2023, doi:10.1016/j.softx.2023.101513.

[60] D. Qin, Z. Peng, L. Wu, "Deep attention fuzzy cognitive maps for interpretable multivariate time series prediction," Knowledge-Based Systems, **275**, 110700, 2023, doi:10.1016/j.knosys.2023.110700.

[61] G. P. Reddy, Y. P. Kumar, "Explainable AI (XAI): Explained," in 2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), 1–6, IEEE, 2023, doi:10.1109/eStream.2023.00001.

[62] Y. Zhang, Y. Weng, J. Lund, "Applications of explainable artificial intelligence in diagnosis and surgery," Diagnostics, **12**(2), 237, 2022, doi:10.3390/diagnostics12020237.

[63] A. Nielsen, S. Skylaki, M. Norkute, A. Stremitzer, "Effects of XAI on Legal Process," ICAIL '23: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, 2023, doi:10.1145/3593013.3594067.

[64] P. Weber, K. V. Carl, O. Hinz, "Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature," Management Review Quarterly, **74**, 867–907, 2023, doi:10.1007/s11301-023-00320-0.

[65] H. DeSimone, M. Leon, "Explainable AI: The Quest for Transparency in Business and Beyond," in 2024 7th International Conference on Information and Computer Technologies (ICICT), 1–6, IEEE, 2024, doi:10.1109/icict62343.2024.00093.

[66] G. Nápoles, J. L. Salmeron, W. Froelich, R. Falcon, M. Leon, F. Vanhoenshoven, R. Bello, K. Vanhoof, Fuzzy Cognitive Modeling: Theoretical and Practical Considerations, 77–87, Springer Singapore, 2019, doi:10.1007/978-981-13-8311-3_7.

[67] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, A. Holzinger, "Explainable artificial intelligence: Concepts, applications, research challenges and visions," in International Cross-Domain Conference for Machine Learning and Knowledge Extraction, 1–16, Springer, 2020, doi:10.1007/978-3-030-57321-8_1.

[68] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, J. Zhu, "Explainable AI: A brief survey on history, research areas, approaches and challenges," in Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II, 563–574, Springer, 2019, doi:10.1007/978-3-030-32236-6_51.

[69] W. Saeed, C. Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," Knowledge-Based Systems, **263**, 110273, 2023, doi:10.48550/arXiv.2111.06420.

[70] A. Toosi, A. G. Bottino, B. Saboury, E. Siegel, A. Rahmim, "A brief history of AI: how to prevent another winter (a critical review)," PET Clinics, **16**(4), 449–469, 2021, doi:10.1016/j.cpet.2021.07.001.

[71] T. Hulsen, "Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare," AI, **4**(3), 652–666, 2023, doi:10.3390/ai4030034.

[72] R. Justo-Hanani, "The politics of Artificial Intelligence regulation and governance reform in the European Union," Policy Sciences, **55**(1), 137–159, 2022, doi:10.1007/s11077-022-09452-8.

[73] P. Hacker, A. Engel, M. Mauer, "Regulating ChatGPT and other large generative AI models," in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 1112–1123, 2023, doi:10.1145/3593013.3594067.

[74] E. Parliament, "EU AI Act: first regulation on artificial intelligence," 2023.

[75] C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. Soler Garrido, et al., "The role of explainable AI in the context of the AI Act," in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 1139–1150, 2023, doi:10.1145/3593013.3594069.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 9, No. 3, 12-28 (2024)
www.astesj.com
Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTES Journal
ISSN: 2415-6698

Evaluation of Various Deep Learning Models for Short-Term Solar Forecasting in the Arctic using a Distributed Sensor Network

Henry Toal[*,1], Michelle Wilber[1], Getu Hailu[2], Arghya Kusum Das[3]

[1] *University of Alaska Fairbanks, Alaska Center for Energy and Power, Fairbanks, 99775, United States*

[2] *University of Alaska Anchorage, Mechanical Engineering Department, Anchorage, 99508, United States*

[3] *University of Alaska Fairbanks, Department of Computer Science, Fairbanks, 99775, United States*

ARTICLE INFO

Article history:
Received: 07 March, 2024
Revised: 02 May, 2024
Accepted: 03 May, 2024
Online: 22 May, 2024

Keywords:
Solar Photovoltaics
Sensor Network
Machine Learning
Deep Learning

ABSTRACT

The solar photovoltaic (PV) power generation industry has experienced substantial, ongoing growth over the past decades as a clean, cost-effective energy source. As electric grids use ever-larger proportions of solar PV, the technology's inherent variability—primarily due to clouds—poses a challenge to maintaining grid stability. This is especially true for geographically dense, electrically isolated grids common in rural locations which must maintain substantial reserve generation capacity to account for sudden swings in PV power production. Short-term solar PV forecasting emerges as a solution, allowing excess generation to be kept offline until needed, reducing fuel costs and emissions. Recent studies have utilized networks of light sensors deployed around PV arrays which can preemptively detect incoming fluctuations in light levels caused by clouds. This research examines the potential of such a sensor network in providing short-term forecasting for a 575-kW solar PV array in the arctic community of Kotzebue, Alaska. Data from sensors deployed around the array were transformed into a forecast at a 2-minute time horizon using either long short-term memory (LSTM) or gated recurrent unit (GRU) as base models augmented with various combinations of 1-dimensional convolutional (Conv1D) and fully connected (Dense) model layers. These models were evaluated using a novel combination of statistical and event-based error metrics, including Precision, Recall, and F_β. It was found that GRU-based models generally outperformed their LSTM-based counterparts along statistical error metrics while showing lower relative event-based forecasting ability. This research demonstrates the potential efficacy of a novel combination of LSTM/GRU-based deep learning models and a distributed sensor network when forecasting the power generation of an actual solar PV array. Performance across the eight evaluated model combinations was mostly comparable to similar methods in the literature and is expected to improve with additional training data.

1. Introduction

Solar photovoltaic (PV) power generation is an increasingly attractive method for expanding global carbon-free power generation capacity, growing from around 39 GW in 2010 to over 1.2 TW in 2022 [1, 2]. Despite the technology's competitive cost and relative ease of installation, continued solar PV expansion is challenged by the fact that PV systems can experience significant variability in generation potential throughout the course of a day due to shading from weather events such as clouds, with sudden losses or gains in production of over 70% of maximum capacity per minute regularly observed [3, 4, 5]. This variability poses a challenge from a grid-integration standpoint due to the potential for a mismatch between electrical generation and demand [6]. This is especially true for smaller, electrically isolated grids or "microgrids" with high proportions of geographically concentrated solar PV generation since a single cloud event could quickly and dramatically impact total grid production [7].

Communities with microgrids typically rely on diesel generators as a major component of their total generation capacity. These generators also serve as a convenient source of fast-response generation or

[*]Corresponding Author: Henry Toal, 1764 Tanana Loop, Suite 420, Fairbanks, AK 99775, ehtoal@alaska.edu

"spinning reserve" which allows the grid to absorb sudden changes in solar PV generation due to cloud events. Unfortunately, diesel generators are not an ideal solution to this problem [8]. For one, the grid-scale generators used by these communities typically require between 1 to 2 minutes to start and become grid-synchronized, meaning that they cannot be left off until needed to compensate for a loss in PV production. Additionally, these generators typically must be operated at or above 30% capacity. This means that, even if a grid's total PV production exceeds demand, the generators must be kept running at minimum capacity to provide reserve, increasing costs from additional fuel burn and maintenance as well as increasing emissions [9, 10]. The relevance of this problem is not limited to small, rural communities without electrical connection to a wider grid. There is substantial and growing evidence that developing self-sufficient microgrid capabilities within existing grid networks could increase grid resiliency by reducing centralized dependence on electrical generation and also by allowing critical infrastructure to more easily transition to backup generation sources in the event of a grid disruption [11, 12]. Thus, the problem of sudden changes or "ramp" events in solar PV production within high solar penetration grids is likely to only grow in importance over the coming years.

A potential solution to this problem is to forecast short-term ramp events caused by clouds. If grid operators were to have access to accurate and reliable forewarning of incoming sharp ramp events at a 1-to-2-minute horizon, diesel generators could potentially be left off and started only when needed. Additionally, as energy storage systems become more widespread, both in microgrids and beyond, these forecasts have the potential to increase storage efficiency by providing grid control systems with advanced knowledge of when energy storage should start and stop.

A wide range of technologies exist for producing solar PV forecasts. These include numerical weather prediction (NWP) models, satellite imagry, total sky imager (TSI) systems (sometimes called "sky cameras"), and networks of distributed light sensors, with the latter two being most effective for the time horizon presented by this problem [13]. Deep learning neural network models have also seen substantial advancement in solar PV forecasting across a wide range of data collection methods and time horizons. While the majority of studies have utilized TSI systems, there is growing evidence that distributed sensor networks have a number of distinct advantages, including lower cost and improved forecast accuracy [10, 14]. A variety of models have been proposed for use in conjunction with distributed sensor networks, including matching local production peaks and examining covariance between sensors [15, 16], but there has been very limited exploration into sensor networks combined with deep learning models, especially of those deployed around an actual solar PV in the type of environment where this type of forecasting is likely to be particularly valuable. This study aims to fill the gap in the established body of work by developing and deploying a distributed sensor network around a utility-scale solar PV arraying a remote, electrically isolated community and using data from these sources to train a variety of deep learning model architectures. This study also aims to advance the field of solar PV forecasting by evaluating the performance of these models using both standard error metrics such as root-mean-square error (RMSE) but also quantifying their ability to detect individual ramp events and examining the impact of location-specific variability on forecast quality.

This study was part of an ongoing collaboration with the Kotzebue Electric Association (KEA) in Kotzebue, Alaska, USA. Kotzebue, a community of just over three thousand people located just above the arctic circle at 66.9 °N, provides an excellent reference for a study examining the intersection of distributed sensor networks, deep learning models, and harsh, isolated environments. With over 1.2 MW of solar PV capacity across two co-located arrays, Kotzebue is an ideal example of a community which relies heavily on its diesel generators to absorb ramp events from high-penetration solar PV. This study uses high-resolution inverter data collected directly from one 72-kW sub-array of the 575-kW eastern KEA solar PV plant as a target forecast variable along with data from the sensors as model inputs.

In this paper, we review the current state of short-term solar PV forecasting and examine why the combination of deep learning models and distributed sensor network is of particular importance. We go on to provide an in-depth overview of our data collection and processing methodologies, the mechanism behind the various deep learning models used and our specific implementation as well as our chosen performance metrics and their relative advantages. We conclude with an analysis of model performance both in terms of the standard error metrics used throughout the literature as well as relative ramp detection performance across a range of ramp magnitudes.

2. Related Work

2.1. Solar PV Forecasting

The methodologies employed for solar forecasting are highly dependent on the amount of time into the future one wishes to predict (otherwise known as the "forecast horizon") as well as the sources and availability of local weather data. Forecast horizons can generally be thought of in four timescale categories: (1) day-ahead (more than 1 day), (2) intra-day (1-24 hours), (3) intra-hour (10-60 minutes), and (4) very-short-term, sometimes called "nowcasting" (less than 10 minutes) [13]. There are also four primary types of data sources used for solar PV forecasting: NWP models, satellite images, TSI systems, and distributed sensor networks. NWP uses regional weather data and complex weather simulations to forecast cloud cover over wide areas and, due to its computational intensity, is primarily used for day-ahead forecasts. Prediction using satellite images is generally best for intra-day horizons since the images typically lack the spatial and temporal resolution for more refined forecasting [17]. TSI systems use a specialized camera to take hemispheric images of the sky around the solar PV array from which cloud coverage and relative motion information can be transformed into a power production forecast. The majority of the research into intra-hour and nowcasting forecasts uses TSI systems but there is evidence that, for horizons shorter than 5 minutes, these systems may suffer from reduced accuracy due to a protective element inside the imager called a "shadow band" which shields the sensor from direct sunlight but also obscures cloud cover information in the region directly around the sun, the exact zone needed for quality very-short-term forecasting [18, 19]. In an attempt to produce more accurate forecasts in the 1-to-5-minute range, a handful of studies over the

past decade have investigated the use of networks of inexpensive solar irradiance sensors placed far enough around the PV array to detect ramp events before they impact the array [10, 15, 16, 20]. Results from these systems are promising and have the advantage of being more affordable than TSI systems, a particular concern for small communities lacking significant funds for additional power infrastructure. This encouraging performance as well as the lack of studies examining the technique in high-latitude locations is why a distributed sensor network was selected as the basis for this study.

Studies on solar PV nowcasting have used a variety of model structures for generating time series forecasts from input data. These can include peak-matching algorithms, wavelet decomposition models, computer vision algorithms and, more recently, deep learning and/or recurrent neural networks (RNNs) such as long short-term memory (LSTM), gated recurrent unit (GRU), and 1-dimensional convolutional (Conv1D) models which show a substantial ability to learn complex relationships between their input and target data sets.

2.2. Distributed Sensor Networks

Despite their lower cost and potential for improved nowcasting performance over TSI systems [14], distributed sensor networks have seen limited use in the literature. A study presented in [15] state that in 2013 utilized 80 rooftop PV installations across a 2500 km^2 area in Tucson, Arizona as a network of sensors and a centralized PV installation as the target for forecasting. The installations were located at various distances from the central array. Data were recorded at 15-minute intervals and a forecast was generated by calculating the covariance between pairs of installations to determine the speed and direction of clouds as they passed over the array. The measurement installation most closely correlated with the central PV array at any given time step was used as that time step's forecast by lagging the data point by the calculated time-of-arrival. Forecasts were evaluated on horizons of 15 to 90 minutes using mean-square error (MSE). In [16], the authors developed a network of five irradiance sensors located within a large PV plant to extract cloud speed and direction information using the cross-correlation between sensors. This was done by assuming that each cloud had a linear edge and then using geometry to find the orientation of the edge and its estimated speed. Similarly to [15], data from the sensor most correlated to the PV array output was lagged by the calculated time-of-arrival at each time step to create the forecast. Performance was evaluated using RMSE. In [10], the authors used 19 low-cost irradiance sensors localized to the north-west of a PV array in combination with a peak-matching algorithm to find the lag between each sensor and the PV array output, find the most correlated sensor, then time-shift the sensor data stream by the time lag to generate a forecast. The novel peak-matching algorithm was compared to simple autoregressive neural networks with and without inputs from the sensor data. The peak-matching algorithm outperformed both neural network models on an RMSE basis. A 2015 study presented in [20] used the US National Renewable Energy Laboratory's (NREL) Horizontal Irradiance Grid, a network of 17 irradiance sensors distributed over approximately 1 km^2 in Oahu, HI as a data source with a central sensor used as a PV array analog. A regression forecasting model known as Least Absolute Shrinkage and Selection Operator (LASSO) was used with good results when compared to ordinary least squares and auto-regressive models. Models were compared using normalized mean absolute error (nMSE) and a metric known as "forecast skill" which compares model RMSE to the RMSE of a "naïve" or "persistence" forecast (which simply assumes the next data point will be the same as the current).

2.3. Recurrent Neural Networks for Solar Forecasting

A number of studies over the past decade have explored RNNs as a model structure for intra-hour and nowcasting irradiance forecasting using TSI systems, sensor networks, and single-source irradiance combined with local meteorological data [21, 22, 23]. For example, the authors of [24] explored the use of various hybrid LSTM, GRU, and Conv1D models for generating forecasts at horizons between 5 and 60 minutes. The researchers used a large database of solar PV production data from a microgrid at the University of Trieste, Italy and were able to produce very accurate forecasts, particularly at a 1-minute, 1-step-ahead time horizon. The study also demonstrated the feasibility of multi-step ahead forecasting, albeit with slightly lower accuracy. The authors did find potential for further improvements, especially during periods of heavy cloud cover and for multi-step-ahead forecasts. Overall, they determined that their methodology was sufficient to aid in the design of small-scale energy storage systems. In [25], the authors utilized a LSTM-Conv1D hybrid deep learning model to forecast 15-minute resolution data at various horizon using real-world solar data from Rabat, Morocco. Their hybrid model showed improved performance when compared to simpler machine learning techniques and non-hybrid models in terms of mean absolute error (MAE) and RMSE as well as enhanced stability in training. The researchers noted accurate performance up to a horizon of 90 minutes and concluded that their hybrid model could prove useful for real-time microgrid energy management and that there was potential for further forecast integration with wind power and load forecasting. A 2021 study [26] also used a Conv1D-LSTM hybrid model but this time in conjunction with the same NREL Horizontal Irradiance grid data as was used in [20]. The model was used to first extract localized dependencies between sensors which were then compared to long-term patterns learned by the LSTM component. The hybrid model outperformed both simple LSTM and GRU models on a RMSE and MAE basis. The work presented in [26], to the best of these authors' knowledge, is the only example in the literature of an RNN model used to generate a forecast using a distributed sensor network and the lack of an actual solar PV array in the data set represents a gap in the established literature which this paper is intended to fill.

3. Methodology

3.1. Sensor Network

In order to fulfill the research objectives of this study, it was important to consider the extreme environment in which the sensor network would operate. The sensors needed to be survivable in harsh, arctic conditions under high winds and temperatures below -40 °C while also being cost-effective for a small, rural electric utility to install. Additionally, the remote location of the KEA PV array necessitated sensors which could communicate wirelessly and

operate under their own power. To meet these requirements, a simple design was formulated which uses a 1-watt PV panel mounted at a 30-degree tilt angle both as a power source and measurement instrument.

The sensors' data collection system was based around the RocketScream Mini Ultra™ Arduino-compatible microcontroller, a low-power specialty microcontroller based on the Arduino ATmega4808, which used a sensing resistor and its built-in voltage sensor to record the PV panel voltage and translate it into irradiance measurements. For energy storage, the sensors used a 4.7-farad supercapacitor as well as a non-rechargeable 3.6-volt lithium thionyl chloride battery for use during times when PV power was not available. A circuit diagram for the sensors in shown in Figure 1 and an exhaustive list of the materials used in the sensor design is presented in Table 1.

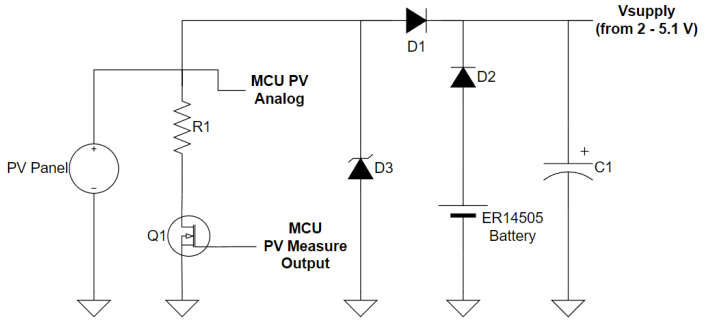


Figure 1: Internal sensor wiring schematic. "MCU" refers to the microcontroller. Component labels are referenced in Table 1.

In total, ten sensors were constructed and deployed in an approximately circular pattern around the 575-kW eastern portion of the KEA PV array. The locations of the sensors were selected based on the minimum distance required to detect a ramp event caused by a cloud 2 minutes before the event would impact the PV array. This was calculated by using historical wind speed data from the site as a surrogate for cloud speed. A cloud moving at the 99.9th percentile wind speed (21.1 m/s) would require a distance of approximately 2500 m to meet the desired 2-minute delay, so each sensor was placed as close to 2500 m from the PV array as possible given the physical accessibility of the site as well as local land use agreements. The locations of the ten sensors in reference to the KEA 575-kW PV array are shown in Figure 2.

3.1.1. Data Transmission and Limitations

Cellular modems are commonly employed for similar remote data collection tasks, however the extremely limited cellular coverage in the area surrounding the KEA PV array made this an infeasible option. While satellite communications do not have this geographic restriction, the high cost and power consumption of satellite modems would have necessitated a much more expensive and cumbersome sensor design. For these reasons, this study made use of short-range radio transmissions using the Long-Range Wide Area Network (LoRaWAN) protocol [27]. LoRaWAN uses a system of centralized nodes, often called "gateways," which serve as connection points between LoRaWAN-capable devices and the wider internet. For this study, one Dragino™ LG16 Indoor LoRaWAN gateway was installed at the KEA PV site (where a wired internet connection was available) and was fitted with a Dragino™ ABL-AN-040 fiberglass extended-range outdoor antenna for improved signal reception. The sensors themselves each utilized a Seeed Grove™ LoRa-E5 LoRaWAN module to interface with the gateway and transmit data.

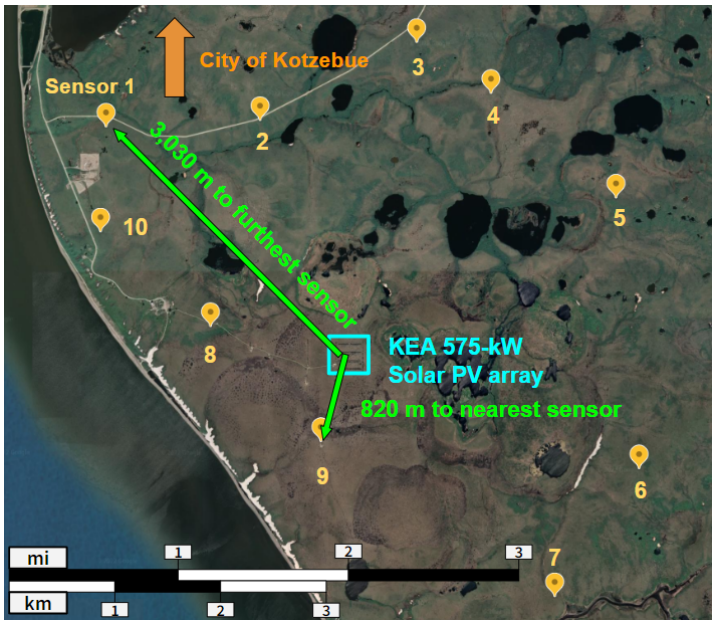


Figure 2: Map showing the locations of the sensors (labeled "1" through "10") relative to the KEA solar PV array.

While the LoRaWAN protocol does generally require much less power than competing technologies, transmitting data still accounted for the vast majority of energy used by the sensors. To ensure a reasonable service life of at least 2 years, limits were placed on the frequency of data collection and transmission. The sensors were programmed to record and store a rolling window of 10 irradiance measurements, each 2 seconds apart but only transmitted the 10 data points currently in memory when certain threshold values were met:

- Absolute mean difference between the last 10 data points is greater than 6.5 W/m^2.
- Absolute median difference between the last 10 data points is greater than 10 W/m^2

These values were determined by calculating a sensor's maximum average energy usage which would still allow it to operate and transmit data using only its PV panel and supercapacitor during the day. A maximum number of transmissions per day was then calculated from this value and found to be approximately 1 transmission every 2 minutes. The maximum measured irradiance ramp rate which would result in an average of 1 ramp every 2 minutes during daylight hours was determined empirically using global horizontal irradiance (GHI) data collected from the KEA PV site. The first 10 data points (20 seconds) of each of these qualifying ramp events as well as the mean and median absolute differences between each of the first 10 data points were recorded for each ramp and the averages of these values were used to set the thresholds listed above. Transmissions were also halted entirely when the mean of the 10-measurement window was less than 15 W/m^2.

Table 1: Sensor components.

Component	Manufacturer	Model
Microcontroller	RocketScream	Mini Ultra™
LoRaWAN module	Seeed	Grove LoRa-E5
Solar panel	Voltaic Systems	1 Watt 6 Volt Solar Panel
Solar panel mounting bracket	Voltaic Systems	Solar Panel Bracket - Small
Solar panel cable extension	Voltaic Systems	Cable Extension with Exposed Leads
Supercapacitor (C1)	Eaton	PHV-5R4V474-R
Diode (D1, D2)	Diodes Incorporated	1N5817-T
Zener diode (D3)	Vishay General Semiconductor	BZX85B5V1-TR
Sensing resistor (R1)	Vishay General Semiconductor	MRS25000C6808FC100
MOSFET transistor	Infineon Technologies	IRLZ44NPBF
ER14505 battery	EVE Energy Co.	ER14505 STD
Antenna for LoRaWAN module	SparkFun Electronics	915MHz LoRa Antenna
LoRaWAN antenna adapter	Padarsey Electronics	RF U.FL(IPEX/IPX) Mini PCI to SMA Female Pigtail
Breadboard (for mounting internal components)	Adafruit	Perma-Proto Quarter-sized Breadboard
Cable tension relief glands	Lokman	PG7
Electronics housing	Polycase	An-16F
Tripod mounting bracket	Liuhe	Universal Stainless Steel Vertical Pole Mount
Mounting tripod	Davis Instruments	Weather Station Mounting Tripod

As a final precaution against excessive power draw, the sensors were programmed to be unable to transmit more than once in any 2-minute period. This impacted the ultimate quality of the data since, if another ramp event were to occur within that 2-minute waiting period, it would not be recorded. It was ultimately decided that the detection of ramp events large enough to warrant the activation of a diesel generator at a resolution greater than 2 minutes would likely not be significantly more valuable to grid operators, since the generators required to ensure grid stability would already be in the process of starting.

3.2. Data

Data used in this research were sourced from two distinct locations: (1) the Kotzebue solar PV site and (2) the NREL 1-second Global Horizontal Irradiance Measurement Grid in Oahu, Hawaii.

The data from the Kotzebue site can be broken down into three categories:

- 1-minute resolution alternating current (AC) power production data measured from the inverter one of eight 72-kW sub-arrays that form the eastern, 575-kW segment of the KEA PV array.
- 1-minute resolution meteorological data recorded from a meteorological station immediately adjacent to the sub-array.
- 2-second, non-continuous irradiance data from each of the 10 sensors.

PV and meteorological data at the Kotzebue site were measured every 5 seconds via the same data logger and averaged over each minute at the time of recording, leaving only the 1-minute resolution data available. Figure 3 shows an example period of approximately 70 minutes of AC solar PV power data as well as several ramp events from sensors 4 and 5, demonstrating the difference in temporal resolution and continuity.

Figure 3 shows that, while the sensors were not able to capture the entire irradiance time series, they do transmit clear upwards or downwards ramp event data in the minutes preceding the ramp experienced by the PV array.

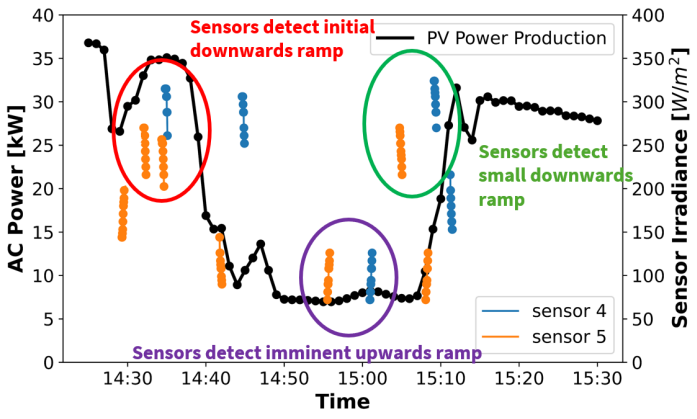


Figure 3: A section of data taken from September 13th, 2022 showing how ramp events detected by upwind sensors correspond to upcoming ramp events in PV power production.

Data from the meteorological station and KEA sub-array were collected from late 2021 through March of 2023 with 486 days of data deemed usable. While the sensor network was installed in July of 2022, only 71 days of data between August and November of 2022 were deemed to be usable due to persistent snow cover of the sensors' 1-watt reference PV panels as well as low light levels during the winter months of the 2022-2023 season.

The 1-Second Global Horizontal Irradiance Grid data set provided by NREL consists of 17 GHI sensors distributed across approximately 1 km^2 in Oahu, Hawaii. Data were collected from these sensors from March of 2010 through October of 2011 [28]. A number of studies have used this data set as training data for various solar nowcasting models [20, 26, 29] and it was used in this

study to validate the results from the Kotzebue site as well as to investigate the impact of differing ramp rate profiles on the efficacy of various models. The NREL Oahu data set is not a perfect analog to the Kotzebue data set as it lacks wind speed and direction data. Additionally, the sensors were placed much closer together than those deployed in Kotzebue for this study.

Due to the different sample rates between the Oahu and Kotzebue data sets, it was necessary to modify the Oahu data set to simulate the limitations of the Kotzebue sensors. First, the data were resampled to a 2-second resolution to match the internal sampling rate of the Kotzebue sensors then the ramp detection thresholds and 2-minute transmission limit described in Section 3.1 were also applied. This resulted in a discontinuous data set comparable to the Kotzebue sensor data. Additionally, since the Oahu sensor network did not include a centrally located PV array, the center-most sensor was chosen as a surrogate and its data were resampled to a 1-minute resolution to match the data from the KEA PV array.

3.2.1. Quality Control

For the solar PV power production data, the capacity of the sub-array inverter (67 kW) was used as a maximum reference value. All data points exceeding 67 kW were discarded.

Data were also removed based on their associated solar elevation (angle of the sun above the horizon). There is evidence to show that irradiance data with solar elevation angles less than 5° can be unreliable due to low incident angles on reference surfaces [30, 31]. In this data set, due to Kotzebue's extreme latitude, solar elevation did not exceed 5° at all from January 1st to January 28th or from November 13th to December 31st but did not overlap with the available data period.

3.2.2. Data Restructuring

Data from the sensors were transmitted at a 2-second resolution with gaps between of at least 2 minutes while data from the PV sub-array and meteorological data were logged as 1-minute averaged data. A truncated example of some of the collated data from a day in September, 2022 is shown in Table 2 below.

Note that data from multiple sensors may overlap but the majority of each sensor column is empty due to the 2-minute transmission restriction. Data from the meteorological station and solar PV sub-array (the examples of power production and ambient air temperature are used in Table 2) is only timestamped each minute, leaving substantial data gaps in this unmodified state.

Neural networks of the kind used in this study cannot be trained with missing data, so data manipulation was required to produce a continuous data set. One solution considered was to resample the solar PV sub-array and meteorological data to a 2-second resolution but this proved impractical due to the gaps in the sensor data caused by the 2-minute transmission limit. It was instead decided to alter the sensor data to fit the 1-minute resolution of the remaining data fields. A diagram of this manipulation is shown in Figure 4 below.

The key insight for this manipulation is that, due to the 2-minute hard cap on sensor transmissions, any given minute can only contain, at most, one transmission per sensor. This allowed for the sensor data to be pivoted, with each of the 10 data points from each transmission assigned their own column. In this structure, assuming an arbitrary sensor transmission is a vector of 10 data points $[t_1, t_2, ..., t_{10}]$, then column "sensor 1, data point 1" contains every t_1 data point from sensor 1, "sensor 1, data point 2" contains every t_2 data point from sensor 1, and so on with column "sensor m, data point n" containing all *nth* (1-10) data points from sensor m (1-10).

Table 2: Example of collected data.

Time	Sensor 1	Sensor 2	...	PV Pwr. (kW)	Air Temp. (C)
0:00	-	-	...	52.55	13.33
0:10	288	-	...	-	-
0:12	225	-	...	-	-
0:14	225	-	...	-	-
0:16	252	-	...	-	-
0:18	243	-	...	-	-
0:20	243	-	...	-	-
0:22	234	378	...	-	-
0:24	225	387	...	-	-
0:26	225	400	...	-	-
0:28	225	414	...	-	-
0:30	-	423	...	-	-
0:32	-	432	...	-	-
0:34	-	432	...	-	-
0:36	-	432	...	-	-
0:38	-	387	...	-	-
0:40	-	378	...	-	-
0:42	-	-	...	-	-
0:44	-	-	...	-	-
0:46	-	-	...	-	-
0:48	-	-	...	-	-
0:50	-	-	...	-	-
0:52	-	-	...	-	-
0:54	-	-	...	-	-
0:56	-	-	...	-	-
0:58	-	-	...	-	-
1:00	-	-	...	54.51	13.21

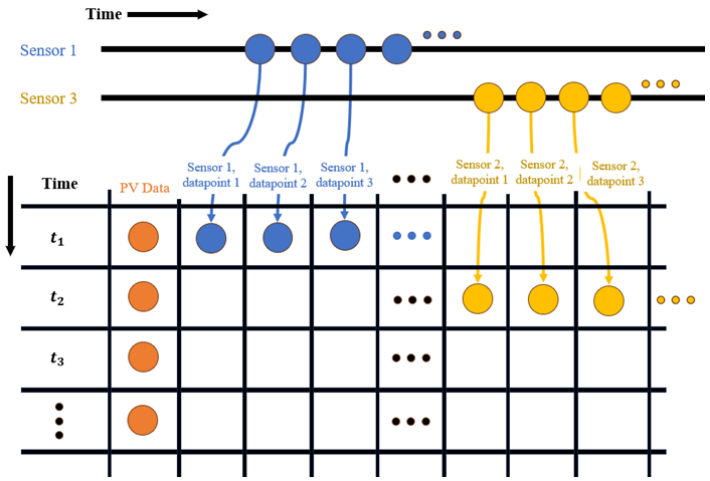


Figure 4: Example of how the sensor data is re-organized to fit into the 1-minute timestamps of the solar PV power production data set. The ramp detected by sensor 1 took place during minute 1 while the ramp detected by sensor 3 occurred in minute 2 and was thus assigned accordingly.

To preserve some additional time resolution information, the time from the beginning of the ramp event until the end of its corresponding minute was recorded and assigned its own data field for each sensor. To aid the models' ability to isolate ramp events, a binary flag field was created for each sensor to indicate if that minute contained a transmission (1 for transmission, 0 for no transmission). If a sensor did not transmit a ramp event in a given minute, these fields for that sensor were set to 0.

In total, 125 fields were used as inputs to the models:

- 10 data-point fields for each of the 10 sensors.
- 1 time-to-end-of-minute field for each sensor.
- 1 ramp-event binary flag for each sensor.
- 2 for wind direction sine and cosine components.
- 1 for wind speed.
- 1 for solar elevation to aid the models in tracking the time of day.
- 1 for the solar PV production.

Of the 71 days of available data, 47 were used for training data, 16 were used for validation, and eight days with a notable diversity of ramp profiles were not used in the training process at all and instead selected for testing.

3.3. Solar Variability

Forecasting weather events in general is a difficult task. Solar irradiance, and thus PV production, undergoes long-term seasonal variation due to the solar cycle and short-term variability from weather effects such as clouds. Figure 5 shows an example of four different days of measured solar PV AC power production under dramatically different cloud coverage conditions.

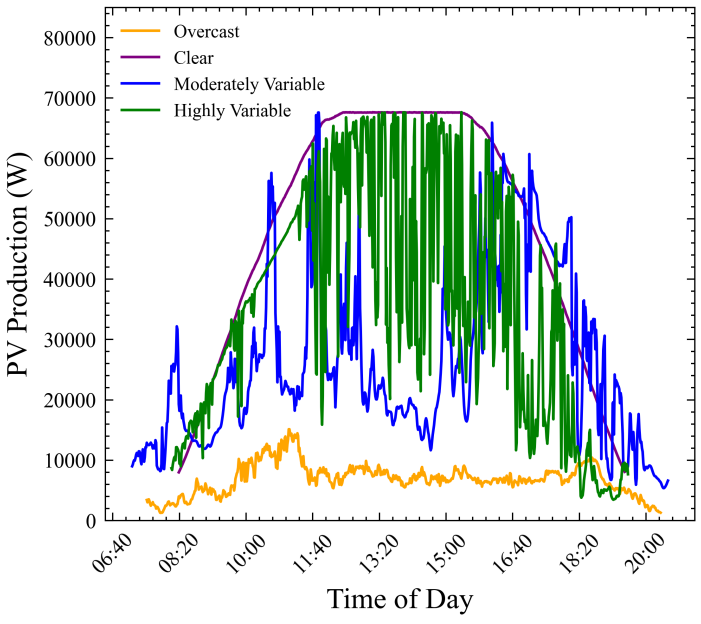


Figure 5: Examples of different daily PV production profiles from Kotzebue showing the clear difference in frequency of sudden ramp events across the four categories [32]

Assessing the variability of solar PV production on both timescales is another common tool apart from deterministic forecasts to which can help improve PV integration efficiency. While a number of metrics exist for evaluating the variability of time series data, a simple one used often when assessing PV production specifically is the Variability Index (VI) [33]. This metric was used in previous work by these authors in assessing the variability of the Kotzebue PV production data set [32].

$$VI = \frac{\sqrt{\sum_{k=2}^{n}(IRR_k - IRR_{k-1})^2 + \Delta t^2}}{\sqrt{\sum_{k=2}^{n}(CSI_k - CSI_{k-1})^2 + \Delta t^2}} \quad (1)$$

VI characterizes the variability of solar PV data over a given period of n time steps by comparing it to a corresponding clear-sky (synthetic, unobstructed PV data) over the same period. A value of 1 indicates a perfectly clear day while larger values correspond with increasing variability. References [33] and [32] used the VI metric to categorize daily variability over the course of a year. In (1) above, IRR_k is the measured irradiance at time step k, CSI_k is the clear-sky irradiance at time step k, n is the number of time steps, and Δt is the length of time between time steps.

This metric is of particular interest in this work because it provides a potential explanation for the differences in forecasting accuracy depending of the data used to train models. As outlined in Section 3.2, a comparison was done between the Kotzebue PV and sensor data and data from the NREL sensor network in Oahu. These two locations have significantly different daily and yearly weather patterns and, since the objective of this research is to develop methodologies for detecting sudden ramp events, it was deemed useful to evaluate the frequency of such events across the two data sets.

3.4. Artificial Neural Networks

Artificial neural networks (ANNs) refer to a wide classification of computational tools which approximate the function of a biological brain. They are composed of complex collections of simple arithmetic operations which can be optimized to produce a desired output from a given input [34]. Because they can be adapted to arbitrary problems, ANNs have found widespread use across many data-driven fields, including solar PV forecasting [35, 36, 37, 38, 39, 40, 41, 42].

ANNs are composed of interconnected structures called "neurons" which intake the weighted sum of arbitrary numerical inputs, add an arbitrary bias value, and apply a activation function (usually a function bounding between 0 and 1 or restricting to positive values). The output $\mathbf{y}$ of a neuron with n inputs can be written as $\mathbf{y} = f(\mathbf{X} \cdot \mathbf{W} + \mathbf{b})$ where $\mathbf{X}$ is the vector of input values to the neuron $[x_1, x_2, x_3, ..., x_n]$, $\mathbf{W}$ is the vector of weights associated with each of the input values $[w_1, w_2, w_3, ..., w_n]$, $\mathbf{b}$ is the bias, and f is the activation function.

The values which a neuron intakes can either be raw input data such as temperature, wind speed, etc. or the output of another neuron. The ability to stack neurons into layers gives ANNs the ability to model complex tasks efficiently [43]. The layers of neurons between the input and output layers are often called "hidden layers" as their inputs and outputs are not directly captured. Figure 6 below

shows a diagram of an arbitrary ANN with its interconnected layers. For solar PV forecasting, examples of input fields might include ambient temperature, time-of-day, current solar irradiance, etc. while the output layer would necessarily be future PV production.

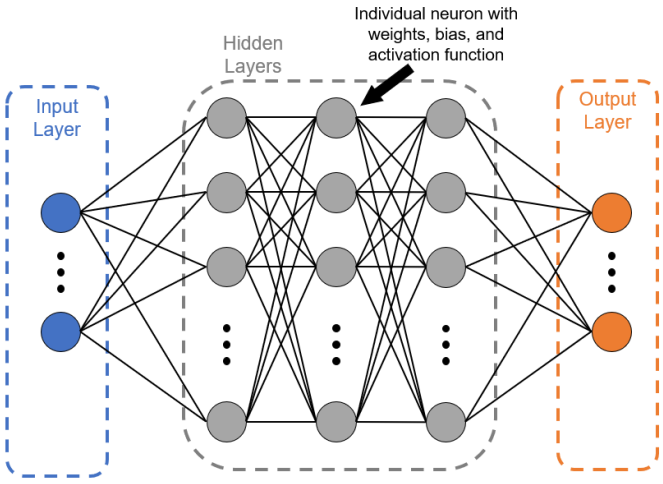


Figure 6: A diagram showing the interconnected structure of an ANN with 3 hidden layers and and an arbitrary number of neurons per layer.

ANNs are optimized to learn input/output relationships in a given data set through a process called "training" during which the weights and bias of each neuron are iteratively adjusted until a stable solution is found. Training is performed through a process called "backpropagation" where, starting from the output end of the network, a gradient descent algorithm is used to find the value of each weight and bias for which some user-defined error function is minimized [44]. This procedure uses a loss function $Q(w)$ of the form:

$$Q(w) = \frac{1}{n}\sum_{i=1}^{n} Q_i(w) \tag{2}$$

where n is the length of the data set and w is the weight or bias value to be optimized. An initial guess for w is selected then iteratively updated using the process: $w = w - \alpha\nabla Q(w)$ where α is a scalar factor often called the "learning rate" which modulates the size of each step down the gradient and ∇ is the gradient operator. This iteration repeats until w stabilizes at some minimum value of $Q(w)$, at which point the next weight or bias is adjusted. This repeats until the entire network is optimized [45].

3.4.1. Long Short-Term Memory

While the general ANN structure (sometimes called a dense ANN, since each neuron is connected to every neuron in the next layer) is still widely used for numerous forecasting applications, other variants exist which are specialized for specific tasks. One such variant is the RNN which is designed for sequential applications, such as time series forecasting. RNNs work by adding the weighted output of each neuron's activation function together with the weighted input from the next time step. Each neuron can be thought of as a chain of "cells" which each share the same weights and biases. Figure 7 below shows the structure of a simple RNN neuron [46].

Since each cell in the chain shares weights and a bias, an RNN which intakes n time steps will have its output values scaled by w_2^n. For even small values of n, this can quickly lead to an unstable gradient and is often referred to as the "exploding/vanishing gradient problem." This flaw makes simple RNN structures like that shown in Figure 7 unfeasible for learning the longer-term, yearly dependencies required for accurately forecasting solar PV data [47].

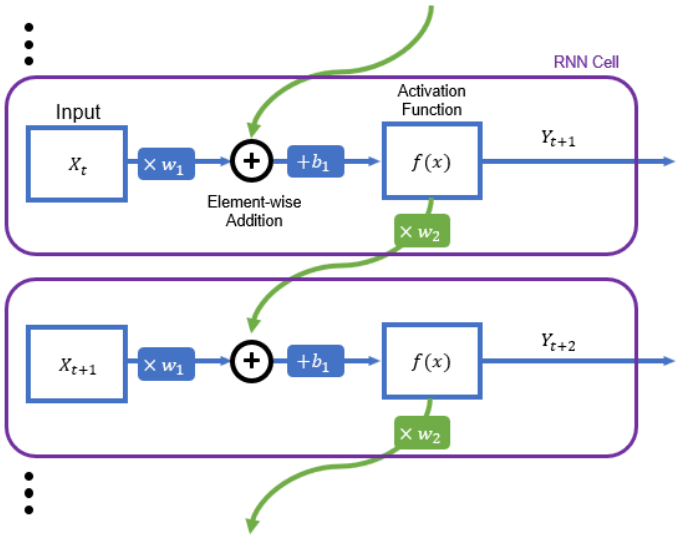


Figure 7: The chain of cells which make up an individual RNN neuron. X is the vector of input values at time t, $w_{1,2}$ are the weights, and b_1 is the bias.

LSTM-based models have become an extremely popular choice for time series forecasting due to its ability to overcome the exploding/vanishing gradient problem [48, 49, 50]. This is achieved by a more complex cell structure which contains three sub-units called "gates," each of which contains either one or two sets of their own weights, biases, and activation functions. This structure allows the LSTM cell to maintain two internal states: the short-term and long-term memory. The short-term memory retains information about the previous time steps in the chain while the long-term memory is used to learn dependencies across much wider timescales, allowing long-term pattern information to be transferred without the gradient growing too large/small [51, 52]. The structure of an LSTM cell can be seen in Figure 8 below.

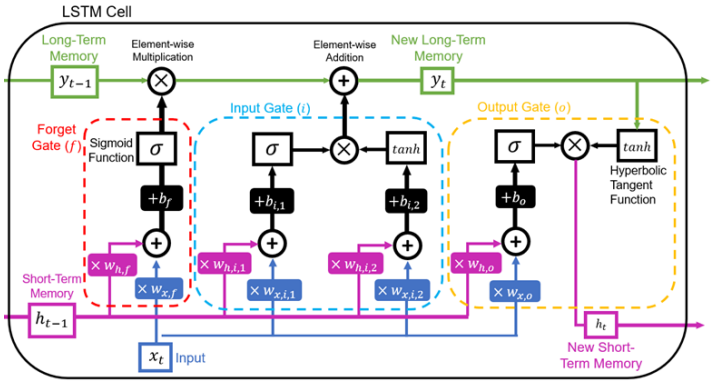


Figure 8: LSTM internal structure showing the three gates, "forget" (f), "input" (i), and "output" (o) as well as their internal weights and biases.

The output of the three gates at time t can be expressed as follows:

$$f_t = \sigma(h_{t-1}w_{h,f} + x_t w_{x,f} + b_f) \tag{3}$$

$$i_t = \sigma(h_{t-1}w_{h,i,1} + x_t w_{i,1} + b_{i,1}) \cdot tanh(h_{t-1}w_{h,i,2} + x_t w_{i,2} + b_{i,2}) \tag{4}$$

$$h_t = \sigma(h_{t-1}w_{h,o} + x_t w_{x,o} + b_o) \cdot tanh(y_t) \tag{5}$$

Here, x_t can either be a scalar (in the case of a single variable input) or a vector of length equal to the number of input variables and the weights w and biases b are the same shape as x_t. All operations are also assumed to be element-wise across the vector. The current timestamp's long-term memory state, y_t, can be expressed as:

$$y_t = y_{t-1} \cdot f_t + i_t \tag{6}$$

Like with simple RNNs, the weights and biases inside each LSTM cell are shared between all cells in the neuron's chain. LSTM cells can also be interconnected into layers, processing multiple time steps simultaneously, and allowing for even deeper pattern recognition (see Figure 9 below).

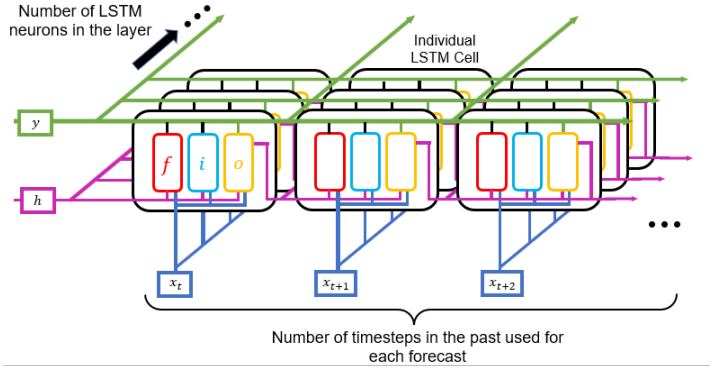


Figure 9: One layer of LSTM of arbitrary size, with the length of the chain (the number of previous time steps to be used in the next time step prediction) extending to the right. Note how the input vector x is connected to each neuron at every time step and likewise with the long-term memory y and short-term memory h.

3.4.2. Gated Recurrent Unit

The Gated Recurrent Unit (GRU) model is an advancement on the LSTM and is fundamentally very similar in its application with some notable differences. As opposed to the three gates present in the LSTM cell, GRU cells only have two. Additionally, the GRU design opts to forgo the two separate memory states of the LSTM in favor of one, long-term memory [53, 54]. The structure of the GRU cell is shown in Figure 10.

Here, the output states of the reset and update gates can be expressed as:

$$r_t = \sigma(x_t w_{x,r} + y_{t-1}w_{y,r} + b_r) \cdot y_{t-1} \tag{7}$$

$$z_t = \sigma(x_t w_{x,z} + y_{t-1}w_{y,z} + b_z) \cdot tanh(r_t w_r + x_t w_x + b) \tag{8}$$

Finally, the new singular long-term state, y_t can be written as:

$$y_t = y_{t-1} \cdot (1 - \sigma(x_t w_{x,z} + y_{t-1}w_{y,z} + b_z)) + z_t \tag{9}$$

This highlights the fundamental differences between the LSTM and GRU models, primarily the fact that the GRU model cannot control the amount the previous long-term state y_{t-1} to intake into the current calculation (which LSTM is able to do with the forget gate). This potentially allows for a more nuanced ability to learn across multiple time steps. There is also the lack of short-term memory state in the GRU formulation. This affords the model notably more computational efficiency, allowing for larger models to be practical while also potentially limiting its ability to extract more detailed patterns from training data [55].

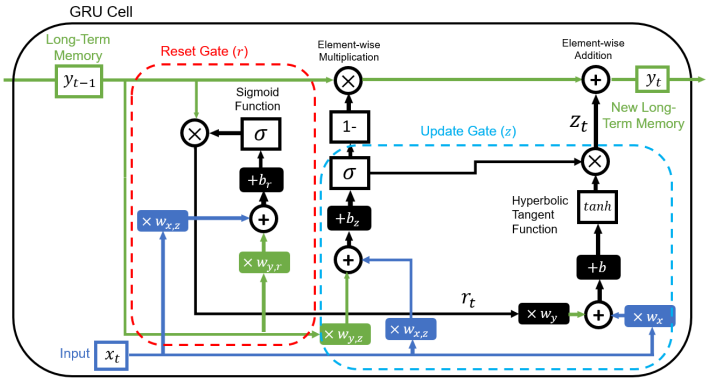


Figure 10: The internal structure of a GRU cell showing the two gates, "reset" (r) and "update" (z) as well as their internal weights and biases.

3.4.3. 1-Dimensional Convolutional

Convolutional Neural Networks (CNNs) are a subset of ANNs most commonly used in the field of image processing. They use a 2-dimensional sliding window over a data set and then feeding that window into a set of connected neurons. This same methodology can be applied, with some adjustment, to 1-dimensional sequential data, such as a time series. These 1-dimensional CNNs are often referred to as Conv1D models when used in this context. Conv1D models are particularly useful for reducing the number of input variables in large, multivariate time series as well as for determining which patterns and combinations of variables have the largest impact of accuracy. As mentioned in Section 2, they are often layered before LSTM and/or GRU layers in order to reduce the input size and improve training efficiency [56]. The structure of a Conv1D neuron is shown

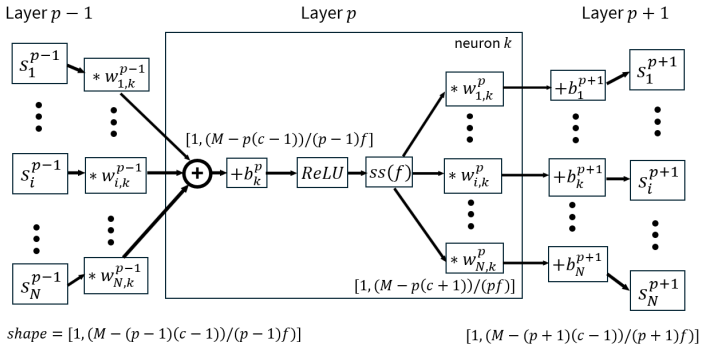


Figure 11: Diagram of three consecutive Conv1D layers showing one neuron (k) of layer p.

Figure 11 shows the Conv1D data flow process over a data set with N data columns and an initial window of size M. The output vectors of the previous layer, s_i^{p-1} convoluted ($*$ being the convolution operator) with "kernels" or vectors of shape $[1, c]$ (denoted by w_i^{p-1}). These resulting vectors are now reduced in size by $c-1$, since the kernel must fit entirely within the rolling window M. A bias b_k^p is then applied followed by a rectified linear unit (ReLU) activation function. The vector is then sub-sampled (denoted by the ss operator) by a factor of f before being convoluted with the kernels of layer p and interconnected with the neurons of layer $p+1$. The output of layer p for each data column i can be expressed as:

$$s_i^p = \sum_{k=1}^{L}(ss_f(ReLU(\sum_{i=1}^{N}(s_i^{p-1} * w_{i,k}^{p-1}) + b_k^p)) * w_{i,k}^p) \qquad (10)$$

where L is the number of neurons in layer p and each subsequent layer reduces the vector size, thus limiting the possible combinations of kernel and window size [57].

3.5. Model Structures and Implementation

For this study, eight combinations of LSTM, GRU, Conv1D, and Dense layers were tested. These include: a base LSTM and GRU model, an LSTM and GRU model proceeded by a Dense model, an LSTM and GRU model preceded by a Conv1D model, and a "hybrid" model, which combines the Conv1D, LSTM/GRU, and Dense models. The naming convention for these eight models is presented in Table 3 below.

Table 3: Base model structure naming conventions.

Model	LSTM	GRU
N/A	LSTM	GRU
Dense	Dense-LSTM	Dense-GRU
Conv1D	Conv1D-LSTM	Conv1D-GRU
Dense and Conv1D	Hybrid-LSTM	Hybrid-GRU

Another useful addition to machine learning models is the "dropout" layer. Not a computational layer like the Conv1D, Dense, or CNN layers, the dropout instead randomly sets a specified fraction of the weights between connected layers in a model to zero on each training iteration. This is highly valuable to combat a phenomenon known as "overfitting" whereby a model will optimize its weights and biases to fit its training data set but will perform poorly on an as-yet unseen (or "validation") data set. This happens because the model learns patterns which are specific to the training data but cannot be generalized to future data. The dropout layer forces the model to adapt to variation in the input data set as, with the dropout layer, the model now cannot rely on any one neural pathway between layers. The generalized model structure with dropout layers is shown in Figure 12. The specific number of Conv1D, LSTM/GRU and Dense layers, as well as the fraction of each dropout layer, is specified in Section 3.7 below.

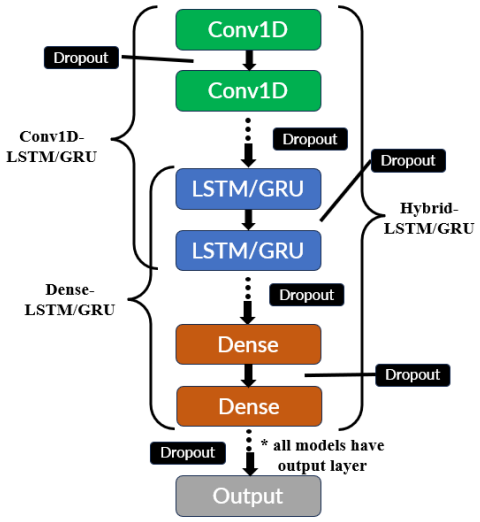


Figure 12: Illustration of the base LSTM/GRU layers and how additional layers are stacked to form a full model.

These eight models were constructed, trained, tuned, and validated using the Python library "Keras" which provides a wide variety of tools for implementing machine learning models of all kinds including neural networks. Keras provides a framework for creating hybrid neural networks in a highly modular fashion and is ideal for quickly prototyping models while also including a robust system for refining and validating model performance.

3.6. Performance Metrics

The metrics used to evaluate the efficacy of the various model structures were divided into two categories: (1) standard error metrics and (2) event-based metrics.

3.6.1. Standard Error Metrics

Previous research into solar PV forecasting across all time horizons used some combination of the standard statistical error metrics MAE, MSE, RMSE, and the coefficient of determination, r^2.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \qquad (11)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{12}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{13}$$

$$r^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \tag{14}$$

In the above equations, y is the actual data set, $\hat{y}$ is the forecast, $\bar{y}$ is the mean of the actual data set, and N is the length of both y and $\hat{y}$. Each of these metrics quantifies the error of a forecast differently. MAE treats all individual errors proportionally while MSE penalized larger absolute differences more severely. RMSE preserves the units of the input data and is also used to calculate another metric ("skill"), which is discussed below. r^2 is bounded between 0 and 1 and shows the degree to which changes in $\hat{y}$ are determined by changed in y. In order to more directly compare results with previous work, variants of MAE and MSE, normalized Mean Absolute Error (nMAE) and normalized Mean Square Error (nMSE) were used. These simply scale the error by the maximum value of the actual data set.

$$nMAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|/max(y) \tag{15}$$

$$nMSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2/max(y) \tag{16}$$

Forecast skill compares the RMSE score of $\hat{y}$ to that of a persistence model. A persistence model, sometimes referred to as a "naive" model, is the simplest prediction algorithm possible. It simply assumes that the next data point will be the same as the previously measured data point.

$$\hat{y}_{t+H} = y_t \tag{17}$$

In (17), H is the forecast horizon, or the number of steps ahead to be forecast. The most commonly used skill definition uses RMSE as its metric.

$$skill = 1 - \frac{RMSE_{forecast}}{RMSE_{persistence}} \tag{18}$$

A score of 0 indicates that the model in question performed no better than a persistence model while negative and positive scores demonstrate worse and better relative performance respectively. While not itself useful for generating meaningful forecasts, persistence models and forecast skill are important tools for establishing baseline model performance. Data sets with large periods of little change, such as extremely sunny or extremely overcast days in the case of solar irradiance data, can lead to surprisingly good performance from a persistence model in terms of the standard error metrics.

3.6.2. Event-based Metrics

While the standard error metrics are useful for understanding the aggregate performance of a model, the ultimate goal of this research is to develop a deterministic forecasting method which can accurately predict individual large ramp events which would threaten the stability of a high-solar-penetration electric grid. In order to more directly measure this capability, the models presented in this paper were also evaluated on an individual event basis. This was accomplished using three standard event-based metrics: Precision (PR), Recall (RE), and F_β.

$$PR = \frac{hits}{hits + false\ alarms} \tag{19}$$

$$RE = \frac{hits}{hits + misses} \tag{20}$$

$$F_\beta = (1 + \beta^2) \cdot \frac{PR}{\beta^2 \cdot PR + RE} \tag{21}$$

Hits, misses, and false alarms are binary classifications of data points. PR represents a model's ability to only forecast events which actually happen while RE reflects its success at actually forecasting events in the first place. In (21), β is a scaling factor used to weight PR and RE differently. These metrics are commonly used in machine learning classification applications, but are rarely applied to time series forecasting. This makes sense for situations in which one is primarily concerned with cumulative performance but, considering the highly event-based nature of this research, these metrics were warranted. Additionally, while it may be that those using such a model would place more importance on either PR or RE in their calculation of F_β, this research evaluated all models at $\beta = 1$, which is written as F_1.

Applying these metrics to a time series data set first required defining hits, misses, and false alarms. Since all ramp events are relative, it was necessary to define a minimum ramp size threshold h. Additionally, while both upward and downward ramp events effect grid stability, upwards ramps can be dealt with more easily through dispatchable loads. For this reason, only downward ramp events were considered in this paper. Misses were calculated by first locating points at time t in the actual data set y where $y_{t-1} - y_t \geq h$ and then evaluating against the forecasted data set $\hat{y}$ to find if $\hat{y}_t - y_t \geq h$. False alarms were calculated in the opposite way, where all of the data points in the forecasted data set $\hat{y}$ where $\hat{y}_{t-1} - \hat{y}_t \geq h$ and then checking if $y_t - \hat{y}_t \geq h$. A hit was defined as a data point for which both $\hat{y}_{t-1} - \hat{y}_t \geq h$ and $y_{t-1} - y_t \geq h$. Figure 13 shows a visual representation of how each event is classified.

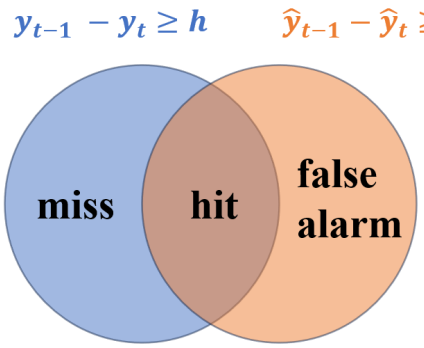


Figure 13: Venn diagram of the definition of misses, hits, and false alarms.

Under real-world conditions, the minimum downward ramp magnitude for which a forecast would be useful to a grid operator is extremely dependent on both their specific grid's configuration and current state. A small grid on a clear day with a large proportion of its power coming from solar PV is able to absorb a different relative ramp magnitude compared to a larger grid with battery storage, alternative generation sources, and low PV penetration. For this reason, PR, RE, and F_1 were evaluated over a range of h values in order to explore how the models performed at predicting different ramp sizes. Plotting RE, PR, and F_1 versus h can give a visual sense of model performance, but it was still deemed useful to select a consistent baseline value of h at which to evaluate the models. The 99th percentile downward ramp rate was selected, which corresponded to approximately 25.3% of maximum PV capacity per minute, and is referred to as h_{99} throughout the paper. Values of RE, PR, and $F1$ evaluated at h_{99} are labeled PR_{h99}, RE_{h99}, and $F_{1,h99}$ respectively.

A version of forecast skill to be used with F_1 was also devised in order to use a persistence model as a baseline. The F_1 forecast skill, F_1^* is presented in (22) below and is similar to the standard RMSE definition shown in (18), although modified to preserve the convention of positive values indicating better performance over the persistence model.

$$F_1^* = \frac{F_{1,model}}{F_{1,persistence}} - 1 \tag{22}$$

3.7. Hyperparameter Tuning

The term "parameter" is often used to refer to the trainable internal weights and biases within a neural network. Similarly, the term "hyperparameter" refers to the values which define the overall architecture of the model itself, including the number of layers of each type, the number of neurons (or "units" as referred to in Keras), and learning rate α. The ideal values for these hyperparameters is highly dependent on the task and there is no set rule for determining them. The model design process often involves significant iterative trial and error where one specifies a range of values for each hyperparameter, models are constructed using various combinations of those values, then tested to determine their relative accuracy. This process, often referred to as "tuning," is easily automated using Keras. The ranges of the hyperparameters tuned for each type of model layer are presented in Table 4 below.

Table 4: Hyperparameter ranges for each layer.

Parameter	Min.	Max	Step
LSTM/GRU Layers	1	5	1
LSTM/GRU Units per Layer	4	128	4
Conv1D Layers	1	5	1
Conv1D Units per Layer	4	128	4
Dense Layers	1	5	1
Dense Units per Layer	4	128	4
Dropout % Between Layers	0.0	0.3	0.05
Learning Rate (α)	1×10^{-5}	1×10^{-2}	10

The results of the hyperparameter tuning, and thus the architectures of the eight models used in this study, are presented in Table 5. The number of values in parentheses denote the number of layers in each category and are listed in order of how the data flows through the layers of the model (i.e. the Dense-GRU model has three GRU layers, the first with 24 units, the second with 24, and the third with 12 a well as 2 Dense layers, the first with 20 units and the second with four). Each layer also has a dropout layer proceeding it, the associated dropout fraction of which is also ordered sequentially in the "Dropout Layers" column.

4. Results and Discussion

4.1. Performance of the Eight Base Models

Models were constructed using Keras based on the hyperparameter tuning results and each was evaluated using the performance metrics outlined in Section 3.6. As with the hyperparameter tuning, the loss function used for all eight models was MSE. To eliminate the effect of random variation in the initial parameter guesses during training, each model was trained a total of 10 times, after which all internal parameters were reset. The models were not restricted in terms of how many full passes through the training data, or "epochs" each training iteration used, but instead a "patience" threshold was set which automatically ended the training process if three epochs occurred without any improvement in the models' MSE score. Once trained, each of the 10 iterations of the eight model architectures was used to create a 2-minute-ahead forecast for the eight test days and the performance of the resulting forecasts was evaluated using the standard and event-based metrics. The mean values of the performance metrics were calculated across the 10 iterations for each model and are presented in Table 6 below with h_{99} again equal to 25.3% of maximum PV capacity per minute.

Overall, it is apparent from Table 6 that the standard error metrics and event based metrics are not strictly linked. The basic LSTM/GRU models performed poorly across all metrics, showing a severe inability to detect ramp events at or above h_{99} and generally having high error scores across the standard metrics. The Dense-LSTM fared little better, although the Dense-GRU did show the highest PR_{h99} of any model and improved standard error scores. The Conv1D-LSTM/GRU models displayed particularly interesting results, with the Conv1D-GRU showing underwhelming event-based scores and middling standard errors but the Conv1D-LSTM architecture having a substantially higher $F^*_{1,h99}$ than any other model,

Table 5: Hyperparameter tuning results.

Model	RNN Layers	Conv1D Layers	Dense Layers	Dropout Layers	Learning Rate (α)
LSTM	(20, 8, 24)	N/A	N/A	(0.15, 0.15, 0.1)	1×10^{-3}
GRU	(16, 8, 8)	N/A	N/A	(0.2, 0.2, 0.25)	1×10^{-3}
Dense-LSTM	(28, 8)	N/A	(16, 16, 16)	(0.15, 0.15, 0.15, 0.2, 0.2)	1×10^{-3}
Dense-GRU	(24, 24, 12)	N/A	(20, 4)	(0.25, 0.1, 0.15, 0.15, 0.15)	1×10^{-3}
Conv1D-LSTM	(28, 24)	(36)	N/A	(0.2, 0.1, 0.2)	1×10^{-4}
Conv1D-GRU	(32, 16, 16)	(64)	N/A	(0.15, 0.2, 0.15, 0.1)	1×10^{-4}
Hybrid-LSTM	(32, 16)	(36)	(8, 8)	(0.1, 0.1, 0.2, 0.1, 0.15)	1×10^{-4}
Hybrid-GRU	(28, 4, 20)	(28)	(12, 4)	(0.1, 0.15, 0.15, 0.25, 0.15, 0.2)	1×10^{-4}

Table 6: Model performance across all metrics over the eight test days.

Model	nMAE	nMSE	r^2	Skill	PR_{h99}	RE_{h99}	$F_{1,h99}$	$F^*_{1,h99}$
LSTM	0.215	0.412	0.825	-0.165	0.000	0.000	0.000	-1.000
GRU	0.211	0.370	0.843	-0.104	0.000	0.000	0.000	-1.000
Dense-LSTM	0.212	0.401	0.842	-0.100	0.000	0.000	0.000	-1.000
Dense-GRU	0.181	0.331	0.869	0.000	0.895	0.009	0.017	-0.630
Conv1D-LSTM	0.210	0.397	0.832	-0.143	0.242	0.069	0.102	1.220
Conv1D-GRU	0.211	0.396	0.872	0.008	0.000	0.000	0.000	-1.000
Hybrid-LSTM	0.186	0.365	0.856	-0.050	0.006	0.028	0.041	-0.107
Hybrid-GRU	0.169	0.330	0.870	0.001	0.030	0.022	0.015	-0.673

although with poor results in other metrics. Finally, the Hybrid-lSTM/GRU models showed improved standard error scores but generally had skill scores around 0, indicating rough parity with persistence models.

The results in Table 6 indicate a rather large gap in performance across the models in terms of the event-based metrics. This was likely due to the relatively small size of the available data set. A downwards ramp event with a magnitude of at least h_{99} only occurred 64 times across the eight test days. This meant that missing only a few more events than another model lead to drastically lower event-based scores. The relative performance of the eight models in terms of F_1 is better illustrated in Figure 14.

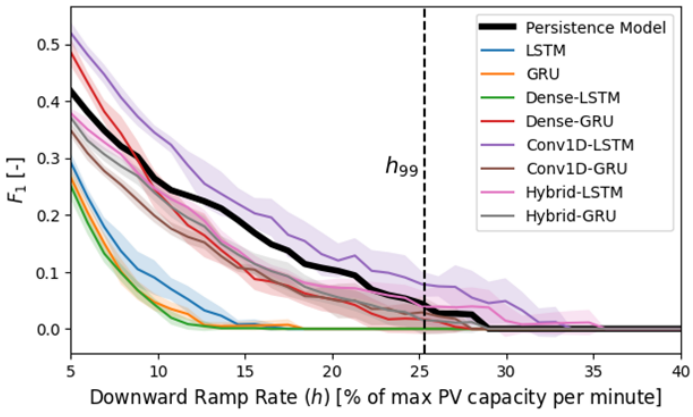


Figure 14: Mean F_1 score versus downward ramp rate magnitude (h) distribution for the eight base models over 10 trials. Shaded region around the mean line represents 1 standard deviation.

Figure 14 shows the mean and variance of each of the 10 models' F_1 score across the 10 trials along with a persistence model for downwards ramp events h of 5% to 40% of the maximum PV array capacity per minute at increments of 1% with h_{99} plotted as a reference. No ramp events over 40% per minute were recorded during any of the eight testing days. This plot shows a number of surprising results. The basic LSTM/GRU as well as the Dense-LSTM did not approach the persistence model's performance for any value of h while the Dense-GRU, Hybrid-LSTM/GRU, and Conv1D-GRU models all performed relatively similarly. Interestingly, the Conv1D-LSTM was able to consistently outperform persistence across all values of h. It is apparent from Figure 14 that there is significant variability within each model's F_1 score, especially for larger values of h where there are fewer ramp events of that size.

The fact that only one of the models was able to barely outperform the persistence model across multiple values of h shows that these models are struggling to accurately forecasting significant ramp events when they happen. The standard error metrics are also somewhat high compared to similar studies in the literature [10, 20, 58]. One potential reason for the former is the use of MSE as the loss function for the model training process. While MSE does penalize larger errors disproportionately, it does not specifically incentivize models to train in a way which produces good event-based metric scores. This could also be impacted by the hard-coded 2-minute window between sensor transmissions, during which ramp events could be missed. This is most likely due to a general lack of quality training data.

4.2. Comparison to Alternative Data Set

While not the best performing model across all metrics, the Conv1D-LSTM's notable improvement over other models in terms of F_1 warranted further testing. As mentioned in Section 3.2, data from

NREL's Global Horizontal Irradiance Grid sensor network in Oahu, HI was used as a comparison to the Kotzebue PV production data. The same Conv1D-LSTM architecture was used with the 570 days of available data from Oahu, with similar proportions of training, validation, and testing data. This is not an ideal comparison due to the Oahu data set (1) not having associated wind speed and direction data, which was an input into the Kotzebue models, and (2) the data from Kotzebue being production data from a PV array compared to the pyranometer measurements from Oahu. Since it takes time for a cloud to move across a PV array, ramp events happen slower when compared to a point-source measurement, like a pyranometer, where the cloud edge impact happens all at once. This spatial smoothing effect has been observed across a wide variety of locations, including Kotzebue [32, 59, 60].

To create a more fair comparison between the Oahu and Kotzebue data sets, the Conv1D-LSTM model architecture was trained on the same Kotzebue data as before minus the wind speed and direction variables. The models were again fully re-trained a total of 10 times and the results across all metrics were averaged. The F_1 plot of these is shown below in Figure 15.

It is apparent in Figure 15 that the Conv1D-LSTM trained on and targeting the Oahu data was much less successful in terms of F_1 compared to its respective persistence model but also showed significantly less variance between trials than the Kotzebue Conv1D-LSTM. The lower variance was likely due to the proportionally larger data set from Oahu while the reduced relative performance can likely be explained using the variability index VI outlined in Section 3.3. The daily VI was calculated for both the Kotzebue PV production and the central Oahu PV-surrogate pyranometer. To obtain a better sense of the variability in Kotzebue, 286 days within calendar year 2022 were used. These included the 71 days for which there was also usable sensor array data.

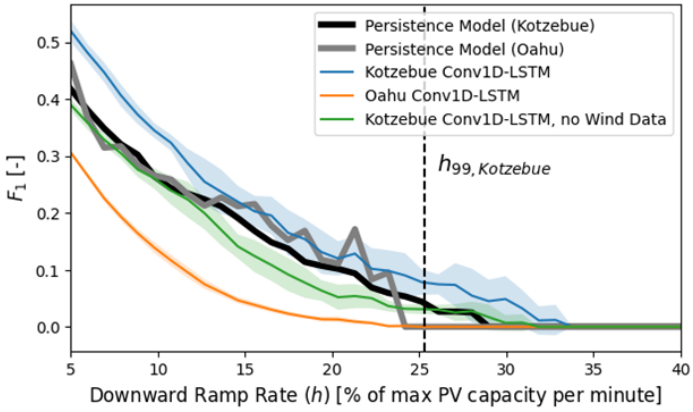


Figure 15: Mean F_1 score versus downward ramp rate magnitude (h) distribution for the base Kotzebue Conv1D-and no-wind Conv1D-LSTM models from both Kotzebue and Oahu. Shaded region around the mean line represents 1 standard deviation.

Figure 16 also includes VI values for a plane-of-array (POA) reference cell located at the Kotzebue PV array over the same 286-day time span. This reference cell, like a pyranometer, takes single point measurements and thus should show more variability than the PV array production due to spatial smoothing. This is evident in Figure 16 but it is also clear that the Oahu data set contains significantly more days of higher variability than either of the Kotzebue data sets. This indicates that the increased variability in Oahu is not simply due to a lack of spatial smoothing and is instead a result of local weather patterns.

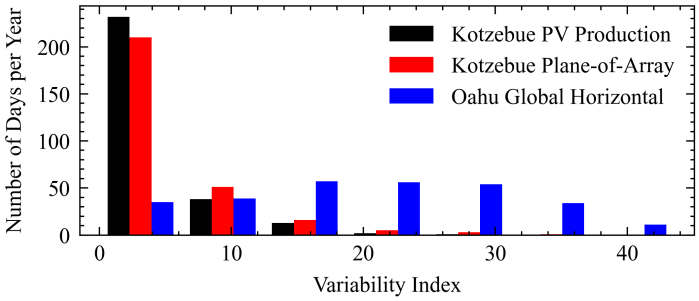


Figure 16: Histogram showing the relative frequency of various VI values between the Kotzebue and Oahu data sets.

It is possible that there are so many sudden ramp events in the Oahu data set that a persistence model is able to achieve a relatively high F_1 score simply because ramp events happen to occur within the 2-minute time horizon of each other very frequently. It is also likely that the Conv1D-LSTM architecture model learned that optimizing for MSE could be achieved without trying to match every local minima/maxima. This is supported by the standard error metrics presented in Table 7 below which show that the Oahu Conv1D-LSTM model performed significantly better than either of the Kotzebue models and even achieved a positive skill.

Table 7: Mean statistical metrics over 10 trials for Kotzebue Conv1D-LSTM and no-wind Conv1D-LSTM models from both Kotzebue and Oahu.

Model	nMAE	nMSE	r^2	Skill
Base Conv1D-LSTM	0.210	0.397	0.832	-0.143
Oahu Conv1D-LSTM	0.115	0.274	0.88	0.142
No-Wind Kotzebue Conv1D-LSTM	0.178	0.348	0.863	-0.025

5. Conclusion

The aim of this research was to develop and test a method for predicting sudden ramp events in solar PV production data at a 2-minute time horizon. To this end, a network of 10 solar irradiance sensors was deployed around a utility-scale solar PV array in Kotzebue, Alaska, just above the Arctic Circle. Data from these sensors was used to detect incoming changes caused by passing clouds via a number of deep neural network models, all of which were evaluated using both standard error metrics such as MSE, MAE, and forecast skill as well as an application of the event-based metrics PR, RE, and F_1. A total of eight models were tested, combining the widely used RNN models LSTM and GRU with 1-dimensional convolutional (Conv1D) and fully connected (dense) networks into various hybrid models. While most of the models showed middling results in terms of the standard error metrics and largely failed to outperform a persistence model in terms of F_1 across a range of minimum ramp magnitudes, the combined Conv1D-LSTM model show notable improvement over the others in the F_1 case, although its

mediocre standard error metric scores are not enough to definitively conclude that it is the best model for this specific application.

It is likely that a model's ability to forecast ramp events is related to the frequency and severity of ramp events. When the *VI* metric was applied to both the Kotzebue and Oahu data sets, it is apparent that the Oahu site experiences significantly more extreme ramp events over a similar time period when compared to the Kotzebue site. This likely both impedes the models' ability to accurately learn what constitutes a ramp event as well as gives the persistence models an advantage since shifting the data set by 2 data points is more likely to coincidentally line up with existing ramp events. This performance difference is reflected in the event-based metrics for the Oahu data set as compared to Kotzebue.

There were two notable limitations with this research: (1) the minimum 2-minute window between sensor transmissions, during which time ramp events could be missed, and (2) the general lack of available training data. It is possible that the models were simply unable to reliably learn patterns between the sensor, meteorological, and PV production data sets but would see improved performance with additional data. It is also likely that the failure of most of the models to outperform persistence on an F_1 basis was due to the use of MSE (or any standard error metric for that matter) as the loss function. MSE and functions like it are not designed to optimize a model for event-based forecasting. Unfortunately, RE, PR, and F_1 use discrete values and as such are non-differentiable, meaning they cannot be used as a loss function.

This research provides two main contributions to the field of solar PV nowcasting. For one, it represents by far the highest-latitude deployment of a distributed irradiance sensor network yet tested. It also marks a step forward in evaluating the ability of nowcasting models to accurately and reliably predict sudden changes which might threaten grid stability. While there has been great success in the past two decades in PV forecasting of all kinds according to the standard error metrics, very little analysis has been presented which quantifies the number of critical events captured by models. The few studies which do either evaluate these metrics over too long of a time horizon for this specific application, only evaluate them at one minimum ramp magnitude h, or both [61, 62, 63].

Lastly, assessing the capabilities of artificial intelligence (AI) across various domains has emerged as a vibrant research area. Particularly in the Arctic, AI is proving to be a vital tool for exploring the region's future and supporting its communities amidst climate change [64, 65]. Further research is essential to fully harness AI's potential in this context. By evaluating multiple AI models towards Arctic energy challenges, this paper has laid the groundwork for numerous research endeavors in this direction, paving the way for a deeper understanding of AI's applications in the Arctic and beyond.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment This work is part of the Arctic Regional Collaboration for Technology Innovation and Commercialization (ARCTIC) Program, an initiative supported by the Office of Naval Research (ONR) Award #N00014-18-S-B001. This paper is also supported in part by U.S. National Science Foundation's Major Research Instrumentation grant (Award #2320196) and EPSCoR Research Infrastructure Improvement Track 4 grant (Award #2327456).

The authors of this paper extend our deepest thanks to the Kikiktagruk Inupiat Corporation who generously allowed the sensor network to be installed on their land, to the Kotzebue Electric Association (KEA) for the use of their PV production data, and to the community of Kotzebue as a whole for making this project possible. We also thank Alan Mitchell of Analysis North for his invaluable help in designing the sensors as well as Christopher Pike of the Alaska Center for Energy and Power (University of Alaska Fairbanks) for installing and managing the KEA PV data collection system.

References

[1] G. Masson, E. Bosch, A. V. Rechem, M. de l'Epine, "Snapshot of Global PV Markets 2023," 2023.

[2] R.-E. Precup, T. Kamal, S. Z. Hassan, editors, Solar Photovoltaic Power Plants: Advanced Control and Optimization Techniques, Power Systems, Springer Singapore, Singapore, 1st edition, 2019, doi:10.1007/978-981-13-6151-7, published: 07 February 2019 (eBook), 20 February 2019 (Hardcover).

[3] J. Marcos, L. Marroyo, E. Lorenzo, D. Alvira, E. Izco, "Power output fluctuations in large scale pv plants: One year observations with one second resolution and a derived analytic model," Progress in Photovoltaics: Research and Applications, **19**(2), 218–227, 2010, doi:10.1002/pip.1016.

[4] S. Abdollahy, A. Mammoli, F. Cheng, A. Ellis, J. Johnson, "Distributed compensation of a large intermittent energy resource in a distribution feeder," in 2013 IEEE PES Innovative Smart Grid Technologies Conference (ISGT), 1–6, 2013, doi:10.1109/ISGT.2013.6497911.

[5] R. van Haaren, M. Morjaria, V. Fthenakis, "Empirical assessment of short-term variability from utility-scale solar PV plants," Progress in Photovoltaics: Research and Applications, **22**(5), 548–559, 2012, doi:10.1002/pip.2302.

[6] H. M. Diagne, P. Lauret, M. David, "Solar irradiation forecasting: state-of-the-art and proposition for future developments for small-scale insular grids," in WREF 2012 - World Renewable Energy Forum, Denver, United States, 2012.

[7] P. Nikolaidis, A. Poullikkas, "A novel cluster-based spinning reserve dynamic model for wind and PV power reinforcement," Energy, **234**, 121270, 2021, doi:https://doi.org/10.1016/j.energy.2021.121270.

[8] N. Green, M. Mueller-Stoffels, E. Whitney, "An Alaska case study: Diesel generator technologies," Journal of Renewable and Sustainable Energy, **9**(6), 061701, 2017, doi:10.1063/1.4986585.

[9] C. S. McCallum, N. Kumar, R. Curry, K. McBride, J. Doran, "Renewable electricity generation for off grid remote communities; Life Cycle Assessment Study in Alaska, USA," Applied Energy, **299**, 117325, 2021, doi: https://doi.org/10.1016/j.apenergy.2021.117325.

[10] S. Achleitner, A. Kamthe, T. Liu, A. E. Cerpa, "SIPS: Solar Irradiance Prediction System," in IPSN-14 Proceedings of the 13th International Symposium on Information Processing in Sensor Networks, 225–236, 2014, doi: 10.1109/IPSN.2014.6846755.

[11] A. Hussain, V.-H. Bui, H.-M. Kim, "Microgrids as a resilience resource and strategies used by microgrids for enhancing resilience," Applied energy, **240**, 56–72, 2019, doi:10.1016/j.apenergy.2019.02.055.

[12] A. Lagrange, M. de Simón-Martín, A. González-Martínez, S. Bracco, E. Rosales-Asensio, "Sustainable microgrids with energy storage as a means to increase power resilience in critical facilities: An application to a hospital," International Journal of Electrical Power & Energy Systems, **119**, 105865, 2020, doi:10.1016/j.ijepes.2020.105865.

[13] D. S. Kumar, G. M. Yagli, M. Kashyap, D. Srinivasan, "Solar irradiance resource and forecasting: a comprehensive review," IET Renewable Power Generation, **14**(10), 1641–1656, 2020, doi:10.1049/iet-rpg.2019.1227.

[14] P. Sukič, G. Štumberger, "Intra-minute cloud passing forecasting based on a low cost IoT sensor—A solution for smoothing the output power of PV power plants," Sensors, **17**(5), 1116, 2017, doi:10.3390/s17051116.

[15] V. P. Lonij, A. E. Brooks, A. D. Cronin, M. Leuthold, K. Koch, "Intra-hour forecasts of solar power production using measurements from a network of irradiance sensors," Solar energy, **97**, 58–66, 2013, doi:10.1016/j.solener.2013.08.002.

[16] J. L. Bosch, J. Kleissl, "Cloud motion vectors from a network of ground sensors in a solar power plant," Solar Energy, **95**, 13–20, 2013, doi:10.1016/j.solener.2013.05.027.

[17] R. H. Inman, H. T. Pedro, C. F. Coimbra, "Solar forecasting methods for renewable energy integration," Progress in Energy and Combustion Science, **39**(6), 535–576, 2013, doi:10.1016/j.pecs.2013.06.002.

[18] C. W. Chow, B. Urquhart, M. Lave, A. Dominguez, J. Kleissl, J. Shields, B. Washom, "Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed," Solar Energy, **85**(11), 2881–2893, 2011, doi:10.1016/j.solener.2011.08.025.

[19] R. Marquez, C. F. Coimbra, "Intra-hour DNI forecasting based on cloud tracking image analysis," Solar Energy, **91**, 327–336, 2013, doi:10.1016/j.solener.2012.09.018.

[20] D. Yang, Z. Ye, L. H. I. Lim, Z. Dong, "Very short term irradiance forecasting using the lasso," Solar Energy, **114**, 314–326, 2015, doi:10.1016/j.solener.2015.01.016.

[21] M. Jaihuni, J. K. Basak, F. Khan, F. G. Okyere, T. Sihalath, A. Bhujel, J. Park, D. H. Lee, H. T. Kim, "A novel recurrent neural network approach in forecasting short term solar irradiance," ISA Transactions, **121**, 63–74, 2022, doi:https://doi.org/10.1016/j.isatra.2021.03.043.

[22] P. Kumari, D. Toshniwal, "Deep learning models for solar irradiance forecasting: A comprehensive review," Journal of Cleaner Production, **318**, 128566, 2021, doi:10.1016/j.jclepro.2021.128566.

[23] S. Mishra, P. Palanisamy, "An Integrated Multi-Time-Scale Modeling for Solar Irradiance Forecasting Using Deep Learning," CoRR, **abs/1905.02616**, 2019.

[24] A. Mellit, A. M. Pavan, V. Lughi, "Deep learning neural networks for short-term photovoltaic power forecasting," Renewable Energy, **172**, 276–288, 2021, doi:10.1016/j.renene.2021.02.166.

[25] A. Bhatt, W. Ongsakul, J. G. Singh, et al., "Sliding window approach with first-order differencing for very short-term solar irradiance forecasting using deep learning models," Sustainable Energy Technologies and Assessments, **50**, 101864, 2022, doi:10.1016/j.seta.2021.101864.

[26] X. Jiao, X. Li, D. Lin, W. Xiao, "A graph neural network based deep learning predictor for spatio-temporal group solar irradiance forecasting," IEEE Transactions on Industrial Informatics, **18**(9), 6142–6149, 2022, doi:10.1109/TII.2021.3133289.

[27] J. Haxhibeqiri, E. De Poorter, I. Moerman, J. Hoebeke, "A survey of LoRaWAN for IoT: From technology to application," Sensors, **18**(11), 3995, 2018, doi:10.3390/s18113995.

[28] M. Sengupta, A. Andreas, "Oahu solar measurement grid (1-year archive): 1-second solar irradiance; Oahu, Hawaii (data)," 2010.

[29] A. W. Aryaputera, D. Yang, L. Zhao, W. M. Walsh, "Very short-term irradiance forecasting at unobserved locations using spatio-temporal kriging," Solar Energy, **122**, 1266–1278, 2015, doi:10.1016/j.solener.2015.10.023.

[30] J. S. Stein, C. W. Hansen, M. J. Reno, "Global horizontal irradiance clear sky models: implementation and analysis." Technical report, Sandia National Laboratories (SNL), Albuquerque, NM, and Livermore, CA, 2012.

[31] V. Lara-Fanego, J. Ruiz-Arias, D. Pozo-Vázquez, F. Santos-Alamillos, J. Tovar-Pescador, "Evaluation of the WRF model solar irradiance forecasts in Andalusia (southern Spain)," Solar Energy, **86**(8), 2200–2217, 2012, doi:10.1016/j.solener.2011.02.014.

[32] H. Toal, A. K. Das, "Variability and Trend Analysis of a Grid-Scale Solar Photovoltaic Array above the Arctic Circle," in 2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI), 242–247, 2023, doi:10.1109/IRI58017.2023.00049.

[33] J. Stein, C. Hansen, M. J. Reno, "The variability index: A new and novel metric for quantifying irradiance and PV output variability." Technical report, Sandia National Laboratories (SNL), Albuquerque, NM, and Livermore, CA, 2012.

[34] B. Yegnanarayana, Artificial neural networks, PHI Learning Pvt. Ltd., 2009.

[35] V. Z. Antonopoulos, D. M. Papamichail, V. G. Aschonitis, A. V. Antonopoulos, "Solar radiation estimation methods using ANN and empirical models," Computers and Electronics in Agriculture, **160**, 160–167, 2019, doi:10.1016/j.compag.2019.03.022.

[36] V. Srikrishnan, G. S. Young, L. T. Witmer, J. R. Brownson, "Using multi-pyranometer arrays and neural networks to estimate direct normal irradiance," Solar Energy, **119**, 531–542, 2015, doi:10.1016/j.solener.2015.06.004.

[37] F.-V. Gutierrez-Corea, M.-A. Manso-Callejo, M.-P. Moreno-Regidor, M.-T. Manrique-Sancho, "Forecasting short-term solar irradiance based on artificial neural networks and data from neighboring meteorological stations," Solar Energy, **134**, 119–131, 2016, doi:10.1016/j.solener.2016.04.020.

[38] H. T. Pedro, C. F. Coimbra, "Assessment of forecasting techniques for solar power production with no exogenous inputs," Solar Energy, **86**(7), 2017–2028, 2012, doi:10.1016/j.solener.2012.04.004.

[39] S. Z. Hassan, H. Li, T. Kamal, M. Nadarajah, F. Mehmood, "Fuzzy embedded MPPT modeling and control of PV system in a hybrid power system," in 2016 International Conference on Emerging Technologies (ICET), 1–6, 2016, doi:10.1109/ICET.2016.7813236.

[40] A. Abbas, N. Mughees, A. Mughees, A. Mughees, S. Yousaf, S. Z. Hassan, F. Sohail, H. Rehman, T. Kamal, M. A. Khan, "Maximum Power Harvesting using Fuzzy Logic MPPT Controller," in 2020 IEEE 23rd International Multitopic Conference (INMIC), 1–6, 2020, doi:10.1109/INMIC50486.2020.9318188.

[41] S. Z. Hassan, H. Li, T. Kamal, U. Arifoğlu, S. Mumtaz, L. Khan, "Neuro-Fuzzy Wavelet Based Adaptive MPPT Algorithm for Photovoltaic Systems," Energies, **10**(3), 2017, doi:10.3390/en10030394.

[42] S. Mumtaz, S. Ahmad, L. Khan, S. Ali, T. Kamal, S. Z. Hassan, "Adaptive Feedback Linearization Based NeuroFuzzy Maximum Power Point Tracking for a Photovoltaic System," Energies, **11**(3), 2018, doi:10.3390/en11030606.

[43] C. M. Bishop, Neural networks for pattern recognition, Oxford university press, 1995.

[44] A. Goh, "Back-propagation neural networks for modeling complex systems," Artificial Intelligence in Engineering, **9**(3), 143–151, 1995, doi:10.1016/0954-1810(94)00011-S.

[45] P. Baldi, "Gradient descent learning algorithm overview: A general dynamical systems perspective," IEEE Transactions on neural networks, **6**(1), 182–195, 1995, doi:10.1109/72.363438.

[46] I. Sutskever, J. Martens, G. E. Hinton, "Generating text with recurrent neural networks," in Proceedings of the 28th international conference on machine learning (ICML-11), 1017–1024, 2011.

[47] Y. Bengio, P. Simard, P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," IEEE transactions on neural networks, **5**(2), 157–166, 1994, doi:10.1109/72.279181.

[48] S. Siami-Namini, N. Tavakoli, A. S. Namin, "The performance of LSTM and BiLSTM in forecasting time series," in 2019 IEEE International conference on big data (Big Data), 3285–3292, IEEE, 2019, doi:10.1109/BigData47090.2019.9005997.

[49] A. Yadav, C. Jha, A. Sharan, "Optimizing LSTM for time series prediction in Indian stock market," Procedia Computer Science, **167**, 2091–2100, 2020, doi:10.1016/j.procs.2020.03.257.

[50] X. Qing, Y. Niu, "Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM," Energy, **148**, 461–468, 2018, doi:10.1016/j.energy.2018.01.177.

[51] S. Hochreiter, J. Schmidhuber, "Long Short-Term Memory," Neural Computation, **9**(8), 1735–1780, 1997, doi:10.1162/neco.1997.9.8.1735.

[52] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," Physica D: Nonlinear Phenomena, **404**, 132306, 2020, doi:10.1016/j.physd.2019.132306.

[53] R. Dey, F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS), 1597–1600, IEEE, 2017, doi:10.1109/MWSCAS.2017.8053243.

[54] R. Zhao, D. Wang, R. Yan, K. Mao, F. Shen, J. Wang, "Machine health monitoring using local feature-based gated recurrent unit networks," IEEE Transactions on Industrial Electronics, **65**(2), 1539–1548, 2018, doi:10.1109/TIE.2017.2733438.

[55] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.

[56] S. Kiranyaz, A. Gastli, L. Ben-Brahim, N. Al-Emadi, M. Gabbouj, "Real-time fault detection and identification for MMC using 1-D convolutional neural networks," IEEE Transactions on Industrial Electronics, **66**(11), 8760–8771, 2019, doi:10.1109/TIE.2018.2833045.

[57] S. Kiranyaz, T. Ince, M. Gabbouj, "Real-time patient-specific ECG classification by 1-D convolutional neural networks," IEEE transactions on biomedical engineering, **63**(3), 664–675, 2016, doi:10.1109/TBME.2015.2468589.

[58] H. Zhou, Y. Zhang, L. Yang, Q. Liu, K. Yan, Y. Du, "Short-Term Photovoltaic Power Forecasting Based on Long Short Term Memory Neural Network and Attention Mechanism," IEEE Access, **7**, 78063–78074, 2019, doi:10.1109/ACCESS.2019.2923006.

[59] R. Van Haaren, M. Morjaria, V. Fthenakis, "Empirical assessment of short-term variability from utility-scale solar PV plants," Progress in Photovoltaics: Research and Applications, **22**(5), 548–559, 2014, doi:10.1002/pip.2302.

[60] R. Perez, M. David, T. E. Hoff, M. Jamaly, S. Kivalov, J. Kleissl, P. Lauret, M. Perez, et al., "Spatial and temporal variability of solar energy," Foundations and Trends® in Renewable Energy, **1**(1), 1–44, 2016, doi:10.1561/2700000006.

[61] M. Abuella, B. Chowdhury, "Forecasting of solar power ramp events: A post-processing approach," Renewable Energy, **133**, 1380–1392, 2019, doi:10.1016/j.renene.2018.09.005.

[62] A. Sanfilippo, L. Martin-Pomares, N. Mohandes, D. Perez-Astudillo, D. Bachour, "An adaptive multi-modeling approach to solar nowcasting," Solar Energy, **125**, 77–85, 2016, doi:10.1016/j.solener.2015.11.041.

[63] M. Abuella, B. Chowdhury, "Forecasting solar power ramp events using machine learning classification techniques," in 2018 9th IEEE International Symposium on Power Electronics for Distributed Generation Systems (PEDG), 1–6, IEEE, 2018, doi:10.1109/PEDG.2018.8447599.

[64] Y. Wang, C. Purev, H. Barndt, H. Toal, J. Kim, L. Underwood, L. Avalo, A. K. Das, "Toward Energy-Efficient Deep Neural Networks for Forest Fire Detection in an Image," The Geographical Bulletin, **64**(2), 13, 2023.

[65] F. Huettmann, P. Andrews, M. Steiner, A. K. Das, J. Philip, C. Mi, N. Bryans, B. Barker, "A super SDM (species distribution model)'in the cloud'for better habitat-association inference with a 'big data'application of the Great Gray Owl for Alaska," Scientific Reports, **14**(1), 7213, 2024, doi:10.1038/s41598-024-57588-9.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 4, 21-34 (2024)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

GPT-Enhanced Hierarchical Deep Learning Model for Automated ICD Coding

Joshua Carberry, Haiping Xu*

Computer and Information Science Department, University of Massachusetts Dartmouth, Dartmouth, MA 02747, USA

ARTICLE INFO

Article history:
Received: 22 April, 2024
Revised: 01 July, 2024
Accepted: 04 July, 2024
Online: 18 July, 2024

Keywords:
ICD coding
Deep learning
GPT-4 prompt
Fine-grained data point
Hierarchical classification

ABSTRACT

In healthcare, accurate communication is critical, and medical coding, especially coding using the ICD (International Classification of Diseases) standards, plays a vital role in achieving this accuracy. Traditionally, ICD coding has been a time-consuming manual process performed by trained professionals, involving the assignment of codes to patient records, such as doctor's notes. In this paper, we present an automated ICD coding approach using deep learning models and demonstrate the feasibility and effectiveness of the approach across subsets of ICD codes. The proposed method employs a fine-grained approach that individually predicts the appropriate medical code for each diagnosis. In order to utilize sufficient evidence to enhance the classification capabilities of our deep leaning models, we integrate GPT-4 to extract semantically related sentences for each diagnosis from doctor's notes. Furthermore, we introduce a hierarchical classifier to handle the large label space and complex classification inherent in the ICD coding task. This hierarchical approach decomposes the ICD coding task into smaller, more manageable subclassification tasks, thereby improving tractability and addressing the challenges posed by the high number of unique labels associated with ICD coding.

1. Introduction

In healthcare, both structured and unstructured information is recorded and stored to ensure that all relevant healthcare processes and patient observations are properly documented. As a key element of the structured data associated with a specific hospital visit, medical codes, such as those in the ICD (International Classification of Diseases) standards [1], are used to accurately describe the visit and indicate the treatment and diagnoses of the patient. ICD coding has long been a labor-intensive manual process performed by trained professionals, requiring meticulous coding of patient records, including doctor's notes. In this paper, we examine how medical codes can be automatically assigned to reflect the different medical diagnoses a patient may receive during a hospital visit. Due to the high complexity of healthcare and medicine, there are thousands of unique ICD codes reflecting a myriad of diagnoses. Among the globally prevalent standards currently used in healthcare and healthcare finance, the ICD codes assigned to a particular hospital visit are highly relevant to the patient's healthcare. This code assignment allows healthcare and finance professionals to accurately communicate hospital visit information and avoid misunderstandings and inaccuracies in billing and treatment [2], [3], [4]. Therefore, medical coding, the process of assigning the appropriate ICD codes for a specific hospital visit, is an important step in healthcare.

Typically, doctors write comprehensive notes that contain important information related to a patient's visit. These notes, while containing crucial observations and diagnoses associated with the visit, are largely written in unstructured natural language. That is, these notes are kept in a traditional note-taking style, which is not conducive to communication when compared to highly specific and universally recognized medical codes. Professionals known as medical coders are employed to process the unstructured doctor's notes into structured lists of medical codes, complete documentation, and improve record keeping and communication within the healthcare ecosystem. However, this task is non-trivial due to the complexity of healthcare, which involves dealing with countless diagnoses, many of which are highly similar and easily confused. Introducing automation into the medical coding process could enhance human performance and help allocate resources to more critical aspects of healthcare. Nonetheless, ICD coding automation faces several challenges [2]. As a classification task, ICD coding requires assigning a unique label to each relevant diagnosis. In practical scenarios, dozens of unique codes may be necessary to describe a particular visit, and appropriate codes must be selected from potentially similar codes that could lead to

*Corresponding Author: Haiping Xu, University of Massachusetts Dartmouth, Dartmouth, MA 02747, Email: hxu@umassd.edu

https://dx.doi.org/10.25046/aj090404

confusion. In addition to the challenges of the classification task itself, doctor's notes as input data typically exhibit a number of characteristics that further complicate the coding task. For example, the notes do not have a prescribed uniform structure and are written in natural language, which may vary significantly from doctor to doctor and from institution to institution. The writing may include jargon, abbreviations, and typographical errors such as misspellings. In addition, the notes often span several pages, listing a variety of information that may or may not be useful in assigning ICD codes. This means that there is often a large amount of raw input data, but much of it is not relevant to the specification of a particular ICD code to a diagnosis.

In this study, we present an automated ICD coding method that employs a fine-grained hierarchical procedure to predict the ICD codes to be assigned to a given instance of doctor's notes. Many existing approaches to automated ICD coding employ the following two main steps: first, a vector representation is generated for the natural language input of the doctor's notes; second, the vector is fed into a multi-label classifier that outputs all predicted ICD codes at once. We refer to methods that employ this popular strategy as *coarse-grained*. Unlike the coarse-grained approaches, which attempt to code the entire doctor's notes document in one shot, producing all predicted codes at once [5], [6], [7], we minimize the complexity of code prediction by performing fine-grained assignments that locate and target diagnoses individually within the notes. Using a fine-grained approach, the various code predictions required to fully code an instance of doctor's notes can be made separately, thus constituting a series of less complex individual classifications [8]. To further support the classification of a given diagnosis, we use GPT (Generative Pre-trained Transformer) to derive related concepts for a diagnosis and identify sentences in doctor's notes that are semantically related to the diagnosis. The diagnosis is then combined with the related sentences to form a fine-grained data point that is now ready for classification. Since the classifier is responsible for classifying only one diagnosis at a time, the complexity of classification is reduced compared to classifying all diagnoses at once with multiple labels. Furthermore, each fine-grained data point is a human-understandable footprint that can be reviewed to determine the evidence used to arrive at the prediction for a particular ICD code. The proposed approach incorporates a hierarchical classifier that can further decompose the classification task of a single diagnosis into multiple steps or subclassifications. For example, the first subclassification can identify the disease family, and subsequent subclassifications can become more and more specific until they reach the ICD code prediction. Furthermore, the design of the hierarchical classifier is analogous to human decision-making and ensures a higher level of understandability and explainability for users in healthcare. The main contributions and novelties of the paper are summarized as follows:

- Implemented a fine-grained ICD coding approach that predicts one ICD code at a time, thus limiting the complexity of classification while improving human comprehensibility.
- Demonstrated the feasibility and effectiveness of generative large language model (LLM) GPT-4 for sentence extraction, which greatly improves the performance of downstream ICD code classification.
- Introduced a modular hierarchical approach that leverages existing ICD code organization to enable high performance of automated coding when many unique ICD codes must be considered and improve the human comprehensibility of the classification results.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 presents the fine-grained ICD coding approach and deep learning architectures used. Section 4 discusses the GPT-powered related sentence extraction technique used in the fine-grained approach. Section 5 details the automated hierarchical classification approach using deep learning. Section 6 presents a case study and the analysis results. Section 7 concludes the paper and mentions future work.

2. Related Work

Prior to the popularization of public datasets such as Medical Information Mart for Intensive Care III (MIMIC-III) [9], researchers achieved automated medical coding primarily through the use of rule-based systems. These systems can leverage expert knowledge and reduce the need for large-scale training data. In [3], the authors introduced a rule-based system that assigns ICD codes by analyzing sentence elements in radiology reports. The system uses production rules that incorporate domain knowledge and simple logic to draw conclusions about the ICD codes that may be indicated by the reports. The increasing volumes of electronic healthcare records (EHR) over time have allowed the application of methods that require larger amounts of training data. In [4], the authors developed a method to automatically generate a set of rules from radiology report data to predict ICD codes. The automatically generated rules were found to be similarly effective to rule sets handmade by domain experts, which is an encouraging sign for medical coding automation. In [5], the authors introduced one of the first machine learning approaches to automated medical coding. They applied Naïve Bayes-based classifier and established the viability of machine learning approaches for the medical coding task.

More recently, the release of large-scale EHR datasets has enabled the use of deep learning models in automated medical coding. This increased data availability has motivated researchers to explore more resource-intensive methods. One generalized two-step deep learning approach has been successful in automated ICD coding. In this generic approach, the language in a doctor's notes document is first vectorized, and then the entire document is used as input by a multi-label classifier that predicts ICD codes from the resulting vector. In [6], the authors introduced a method that vectorizes the entire document, which is then passed to a convolutional neural network that outputs the predicted ICD code. In [10], the authors achieved high performance in the ICD coding task by exploring various options for the initial vectorization step and using a BERT-based deep learning classifier for final classification. However, the large amount of text in doctor's notes poses a unique challenge, as the large input can confuse a deep learning classifier, especially if it is biased toward a limited portion of the input. To address this challenge, some research efforts have focused on developing classifiers that can build more complex representations at multiple levels of analysis, aiming to understand doctor's notes in terms of overall conceptual or even sentence-level complexity [6], [7], [8]. Other research efforts have addressed

this problem by introducing a labeling attention mechanism in deep learning methods to produce different document interpretations for each unique label [11], [12], [13]. While these methods demonstrated effectiveness and enhanced classification performance, they also come with limitations. In particular, these methods employ a two-step strategy known as the *coarse-grained* approach. In other words, given a report or doctor's notes document, they treat the document as a single input and then predict all associated ICD codes together. Consequently, these methods confront a more complicated classification challenge, where all labels must be predicted from a single classifier input, and potentially output results that can be difficult for users to interpret. In contrast, we present a fine-grained ICD coding approach that targets and predicts one ICD code at a time. To achieve this, we decompose the automated coding task for doctor's notes into a series of single-label classifications. This approach reduces the complexity of the medical coding process compared to the multi-label classification method. More specifically, our approach starts with a diagnosis from doctor's notes and then uses GPT-enhanced sentence extraction to identify the sentences that are semantically related to the diagnosis. This information is grouped into a fine-grained data point for ICD code prediction, which can be examined to reveal evidence used by the classifier for medical coding, thereby improving the explainability and confidence of predictions.

A number of existing studies have exploited the inherent relationships between labels, using hierarchical classifiers to make predictions through a series of classification steps. Hierarchical classifiers are characterized by specially designed subcomponents to handle the various decisions required to achieve a complete classification. These involved decisions are arranged into a hierarchical tree with each node representing a decision. One way to implement the hierarchy is to assign a separate subclassifier to each node of the hierarchy to process the decisions, which is known as the per-node local classifier approach. In this approach, the classifiers required for each decision or subclassification can be specially designed for their particular tasks, and the complexity of each subclassification can be much lower compared to the single step in a non-hierarchical classification approach. The per-node local classifier approach has been used with success in a variety of classification tasks. In [14], the authors used a local classifier per node hierarchy for the galaxy morphology classification task. Their approach used an established taxonomy of galaxy morphologies to design the local classifier hierarchy. In [15], the authors proposed a local classifier-based solution for genomic data classification. Their approach applied multi-label annotation of examples along several paths of the hierarchy and the results showed that deeper and more detailed hierarchies could produce better results. In [16], the authors used a local classifier hierarchy for the task of protein function classification. In their approach, the physical and chemical properties of proteins were used to predict their function in an organism. In [17], the authors employed local classifiers to classify natural language documents by topic. Their approach exploited the hierarchical relationships between related topics and used these relationships to design effective local classifier hierarchies. In [18], the authors studied the design of local classifiers at a higher level and investigated the impact of hierarchical design on classification performance. Their results showed that their hierarchical classification approach could significantly improve classification performance, but the performance gains would depend heavily on a good hierarchical design and training parameters. Unlike the above approaches, our approach is to design a classifier hierarchy by using a taxonomy of ICD codes to enable the automated coding of diagnoses. In our hierarchical approach, we first classify the disease type or family for a given diagnosis, and then attempt to classify specific ICD codes using subclassifiers in a hierarchical tree.

There have also been some efforts to implement hierarchical components in a single classifier, which is known as a global hierarchical classifier. In [19], the authors compared the global Naïve Bayesian classification method with local methods and found that the global Naïve Bayesian classification method has a performance advantage in the protein function classification task. In [20], the authors tackled the face recognition task by combining several techniques, including global hierarchical classification. Their models included a convolutional component consisting of layers that could produce successively more complex classification features. Like hierarchical classification with local classifiers, using a global hierarchical classifier allows designers to address specific aspects of classification through hierarchical components. The resulting global classifier is highly cohesive and easy to train and apply. However, a global classifier lacks the modularity provided by the individuality and separability of local classifiers. Therefore, a global classifier can be difficult to design and it is not possible to train their hierarchical components in a highly specific manner to optimize their performance. In contrast, the modular design in our per-node local classifier approach allows local classifiers to be reused, repurposed, or rearranged in the hierarchy or other hierarchies without the need for extensive retraining. Depending on the nature of the classification task, efficiency can also be improved by parallelizing the training of multiple local classifiers. In addition, the use of local classifiers in a hierarchical architecture allows for more direct decision tracking through the explicit hierarchical decision tree, whereas the hierarchical components and decisions of a global classifier reside in the black box of a single classifier, and thus interpreting the behavior of the global classifier may be more difficult.

One of the major performance bottlenecks in ICD coding is extreme labeling bias. That is, some ICD codes are so frequent that thousands of unique examples may appear in a given dataset, while other codes may represent rare diagnoses with only a few examples. Typical deep learning methods require many examples to learn a given label, and are therefore particularly vulnerable to label bias introduced by ICD codes. Several recent studies have specifically addressed the problem of label bias in an attempt to directly address this key challenge. In [21], the authors used a debiasing method that first statistically analyzes the model's performance to detect bias. Once quantified, the model's bias for each class is used to calculate a debiasing factor, which is utilized to adjust the confidence of the model's output for each class before deciding on the final prediction. In addition to deep learning models, some researchers used fuzzy logic and string matching techniques to improve the performance of few-shot and zero-shot ICD coding [22]. After initially identifying the ICD code category through a deep learning classifier, fuzzy string matching was used to compute Levenshtein distances between sentences in the doctor's notes and the various ICD codes included in the predicted category from which the final code predictions were selected. In

[23], the authors explored the use of transfer learning in a related unsupervised learning task to provide additional learning data for the deep learning classifier. In their approach, the classifier is trained not only on labeled examples that can be very scarce especially for underrepresented ICD codes, but also on completely unlabeled clinical texts using token-level similarity. In this paper, we take a hierarchical approach to the ICD coding problem, building labeling relationships (between different ICD codes) into the model architecture. This predefined hierarchical structure leverages the existing knowledge about labels and supports label classification, even if few examples can be trained. However, beyond this hierarchical structure and the application of class weights during training, the problem of label bias is not directly addressed. Therefore, the above strategies for mitigating label bias can be used as a complement to our approach in future research.

While LLMs and generative AI are best known for their generative capabilities through chatbot applications such as ChatGPT and Microsoft Copilot, they can also be used for more straightforward tasks like text classification [24], [25]. Most LLMs have a large internal network whose output is passed through a specially trained "head" to produce the final desired output. One way to adapt an LLM to perform classification is to replace the generative head with a classification head. The network can then be trained holistically to learn the classification task and fine-tuned using the existing knowledge to be applied to that task [26]. This approach has been used in Google's BERT LLM and its many variants [27]. In [28], the authors applied several BERT-based models to influential text classification benchmarks and achieved state-of-the-art results. Some of the benchmarks involved include topic classification, sentiment classification, and goods and services identification. Similar to these methods, we use BERT-based classifiers with a multiclass classification head to perform classification tasks. In addition to being used directly for classification tasks, the generative capabilities of LLM can also be used to augment or enrich the input data points for other classification methods. In [29], the authors introduced GPT3Mix, a method that augments training data by using the generative ability of GPT-3. In GPT3Mix, GPT-3 combined multiple training examples to generate a hybrid synthetic training example, and downstream classifiers trained on GPT3Mix-enhanced data points showed significant improvement over the baseline model. In [30], the authors utilized LLM-based text augmentation to improve classification performance for grant proposal research topics. Their approach enhances imbalanced training data by targeting underrepresented classes and generating new training data points to populate them. In contrast to these methods, our medical coding approach enriches data points indirectly by using generative language models. Based on an initial diagnosis concept, we prompt the model for a set of semantically related terms. We then use these related terms to mine related sentences from the free text of doctor's notes and combine them with the diagnosis to generate an enriched fine-grained data point to improve training and classification performance.

3. Fine-Grained ICD Coding Using Deep Learning

3.1. A Novel Approach for Automated ICD Coding

During a hospital visit, healthcare professionals collect and record various data about the patient. One key record comes in the form of doctor's notes, which are text-based records generated and maintained by hospital staff. Doctor's notes typically cover the entire healthcare process and may involve anything from medical measurements and observations to patient medical histories and miscellaneous comments. Some parts of doctor's notes are loosely structured, presenting information in bulleted or numbered lists; while others can be unstructured, presented in common sentences with agrammatic or misspelled language. Figure 1 shows a text snippet from a randomly sampled example of doctor's notes.

> Cardiovascular: On telemetry, the patient was noted to have multiple premature ventricular contractions. These were asymptomatic and not treated. Due to the sudden episodes of pulmonary edema …

Figure 1: Text snippet from sample doctor's notes.

The free-text format of doctor's notes allows for flexibility and convenience in covering a wide range of information about the patient and their hospitalization, but this flexibility comes at a price. Because of the free-form nature of doctor's notes, much of the data they encode is not sufficiently structured and organized to be used effectively. For example, a doctor may have to flip through pages of irrelevant patient history to find important details related to a specific diagnosis; and an external institution such as an insurance company may not be able to recognize information related to billing due to the peculiarities of the way it is written. For this reason, doctor's notes must be annotated with a set of highly specific codes that show the exact diagnoses and course of treatment. This allows other healthcare professionals and external entities to quickly and directly assess critical information that has previously been obscured by the difficulties associated with doctor's notes. The ICD international standard provides a robust and extensive set of medical codes used to identify a myriad of disease diagnoses in healthcare. A version of the standard, ICD-9, was widely used in modern healthcare, leading to the release of a number of datasets coded using the standard. Recently, the newer ICD-10 has been accepted and frequently used by hospitals in the United States and some other countries. There are several accessible datasets using the ICD-10 standard, which are used for a variety of machine learning tasks, including automated ICD coding. The latest standard, ICD-11, has not yet been widely adopted, thus the available data using this standard are limited. However, ICD-11, like its predecessors, greatly expands the code bases offered by the previous standards and provides an important research motivation for the topic of automated coding using a large number of unique ICD codes. Despite the importance of ICD coding in healthcare, identifying and assigning the appropriate codes for a given instance of doctor's notes has been a non-trivial task. To maintain manageable classification complexity and enhance ICD coding accuracy, we introduce a novel approach to automated ICD coding of doctor's notes using GPT-enhanced text mining. Figure 2 shows an overview of the key components and steps of the proposed automated ICD coding approach. As shown in the figure, we use a fine-grained method, which performs ICD code assignment as a series of single-label classifications rather than a single multi-label classification. This means that the classifier used is only responsible for predicting one ICD code from a given input, thus reducing the complexity of the ICD code prediction process. In order to construct the fine-grained data points used to code a given instance of doctor's notes, a diagnosis

needs to be selected from the doctor's notes and paired with related sentences from the free text of the doctor's notes. To mine sentences related to a specific diagnosis, we derive a set of diagnosis-related concepts using GPT-enhanced text mining. Subsequently, we search the free text of the doctor's notes to mine sentences that contain one or more of the derived related concepts. The extracted related sentences are then combined with the diagnosis to form a fine-grained data point, which is fed into the fine-grained classifier to generate an ICD code prediction.

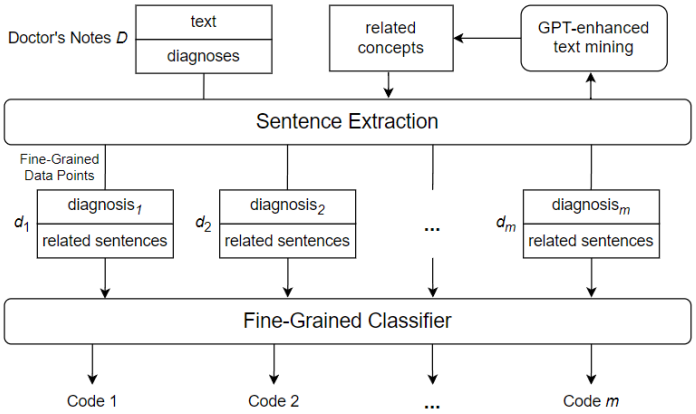


Figure 2: Overview of a fine-grained approach enhanced by GPT-4.

3.2. Fine-Grained ICD Code Assignment

Unlike the *fine-grained* approach used in this paper, existing methods typically use *coarse-grained* techniques for ICD coding. These methods perform ICD coding through multi-label classification. That is, given a single input, which is the entire contents of a doctor's notes instance, the goal is to produce a set of outputs, i.e., all appropriate ICD codes for the doctor's notes instance. The classification procedure performs only one operation to process the entire doctor's notes as a whole and outputs the predicted ICD codes accordingly. Figure 3 shows an overview of a typical coarse-grained classification approach.

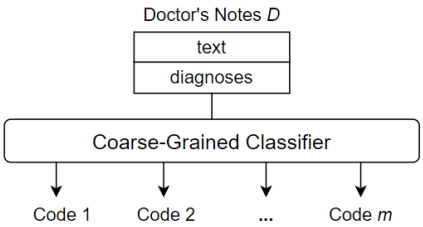


Figure 3: Automated ICD coding using a coarse-grained approach.

One of the main disadvantages of coarse-grained methods is that they can lead to very complex classifications. First, the classifier has to deal with a large amount of raw data, as the contents of doctor's notes can be several pages long. Much of this information is not useful for a specific ICD code assignment, and some may be useless for any code assignments, essentially useless noise. Second, a multi-label classification is inherently difficult, especially when the number of unique labels to be assigned is large. The ICD standards currently in use contain thousands of unique codes, and the code base continues to grow as new versions are released. Therefore, large-scale ICD coding is considered to be a difficult classification task with a very large label space. That is, when the number of unique labels in a multi-label classification task is n, then there exist 2^n unique label combinations. The resulting complexity must be captured by the training data and be tractable for the classifier architecture, which becomes increasingly challenging. In contrast, we adopt a fine-grained approach to ICD code assignment that seeks to avoid the difficulties associated with a coarse-grained strategy. The key to the fine-grained approach is to subdivide the ICD code assignment process down to the individual code level. In other words, rather than predicting all codes at once, we predict one code at a time from some "starting point", namely a diagnosis, in the doctor's notes. This shifts the task from a multi-label classification to single-label classifications, thus limiting the inherent classification complexity. Since we are predicting one code at a time, there is also no need to overwhelm the classifier with the full text of the doctor's notes, instead using only the text that we believe is relevant to the current code classification. Thus, fine-grained classification not only reduces the complexity of the classification step, but also provides the opportunity to trim or enrich the input data used to predict each individual code in the medical coding process.

To facilitate the division of the ICD code assignment task for a doctor's notes instance, we can utilize the diagnoses in the doctor's notes to mark the presence of codes to be identified. A doctor's notes instance typically contains a section of "Discharge Diagnoses," which is a delimited list of the most important diagnoses at the time of a given patient's visit. Figure 4 shows the discharge diagnoses from randomly sampled doctor's notes.

DISCHARGE DIAGNOSES:
- Pulmonary edema
- Congestive heart failure
- Metastatic carcinoma

Figure 4: An example of discharge diagnoses from doctor's notes.

While the diagnoses provide the initial concepts for assigning codes to the doctor's notes, they often lack the specific details required to determine the assignment of individual ICD codes. These diagnoses may contain abbreviations, misspellings, or other incompleteness that preclude precise assignment. Therefore, while they cannot be used alone to draw conclusions about the ICD codes needed for a given doctor's notes instance, they can serve as ideal starting points for fine-grained classifications. For each diagnosis in a discharge diagnoses section, we perform a single-label classification on a different subset of the notes to predict the corresponding code. After assigning a predicted code to each diagnosis, the resulting set constitutes the set of predicted codes for the entire doctor's notes instance. Let F_{fine} be a fine-grained classifier defined as a function that outputs an array of confidences for the output classes. Let doctor's notes D be a 2-tuple (*DIAG*, *FTXT*), where *DIAG* is a list of diagnoses and *FTXT* is the free text of the doctor's notes, respectively. The procedure for classifying an instance D of doctor's notes using a fine-grained classifier F_{fine} is described in Algorithm 1. As shown in the algorithm, each individual diagnosis in the discharge diagnosis section is combined with a set of semantically related sentences in the free-text of the doctor's notes to form a fine-grained data point *dp*. The data point *dp* is then used as an input to the single-label multiclass ICD code classifier F_{fine}, which predicts a suitable ICD code. Once all diagnoses in the doctor's notes have been processed, the generated ICD code set is returned as the predicted code set for the doctor's

notes. Note that in our fine-grained approach, sentences in the doctor's notes are appropriately ignored if they are not relevant for the classification of a diagnosis. Therefore, our approach utilizes targeted and useful information for each prediction, enabling the classifier to predict relevant ICD codes more accurately. In addition, since the amount of raw text in the doctor's notes may pose a problem for a classifier architecture that is limited by the size of the input, sentence extraction provides yet another key benefit. By removing irrelevant information, we also limit the average length of the data points, thereby expanding the scope of applicable classifiers and training techniques, which were previously limited by potentially high data volumes. A key consideration when using a fine-grained approach is how to identify the subset of doctor's notes that are relevant and useful for the classification of a particular diagnosis. Thus, it is critical to develop an effective method for the inclusion of specific text passages in doctor's notes based on their usefulness.

Algorithm 1: Automated ICD Code Assignment

Input: an instance of doctor's notes *dNotes,* a single-label multiclass ICD code classifier F_{fine}
Output: a set of m predicted ICD codes *codeSet*, where $m = |dNotes.DIAG|$

Initialize $codes = \emptyset$
for each diagnosis α **in** $dNotes.DIAG$:
    Extract a set of sentences Ψ from $dNotes.FTXT$ related to α
    Let fine-grained data point dp be (α, Ψ)
    $confidences = F_{\text{fine}}(dp)$
    $code = argmax(confidences)$
    $codeSet = codeSet \cup \{code\}$
end
return *codeSet*

3.3. Transformer-Based Deep Learning Models

Deep learning models are one of the dominant tools in the field of Natural Language Processing (NLP) [31]. Deep learning utilizes deep neural networks to perform tasks that are often difficult to solve programmatically because they are complex and potentially poorly-defined. Neural networks are computational architectures composed of artificial neurons. As shown in Figure 5, an artificial neuron takes one or more values as inputs and transform these values to generate an output value. The output value is either sent to the next neuron or neurons for further computation or as the final output of the network.

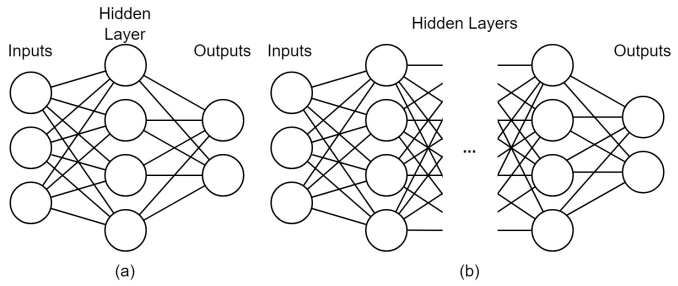


Figure 5: (a) A traditional neural network with one hidden layer. (b) A deep neural network with more than one hidden layer.

During the passage through the neural network, the input values are transformed according to a set of learned parameters called model weights. These weights are learned during the training phase, which optimizes the weights according to some correctness criteria. For example, a neural network classifier is trained to predict the true class of a given example, which means that the network will be optimized towards correct classification. After training, the network should be able to predict classes with reasonable classification performance even for unseen examples. While traditional neural networks contain only one hidden layer of neurons, limiting the complexity of possible transformations and resulting performance, deep neural networks use any number of hidden layers, meaning that arbitrarily complex architectures can be developed to cope with difficult problems. A comparison between a traditional neural network and a deep neural network is illustrated in Figure 5 (a) and (b).

One deep learning architecture that dominates in NLP is the transformer-based architecture. A transformer-based architecture is a deep neural network consisting of modules called encoders and decoders equipped with self-attention mechanisms. These self-attention mechanisms allow the network to learn and express relationships between individual tokens (e.g., words) in a natural language input sequence. At a high level, transformer-based architectures provide the complexity necessary to capture the way language changes based upon its context. One example is homonymy, where the same pronunciation or spelling has different meanings. For example, "saw" can represent one of two meanings depending on context. Take this sentence for example: "Patient reported he saw black spots in his vision," where "saw" comes from the verb "see." In another sentence: "Patient admitted with injuries related to a power saw," "saw" is referring to a power tool. While a more primitive NLP model may assign the same meaning to both usages of "saw," the self-attention mechanisms in a transformer-based deep learning network allow the model to distinguish between the two uses, resulting in a more robust and accurate understanding of the language. In essence, self-attention mechanisms allow the model to process each token (e.g., word) in the input sequence considering its relations to the surrounding tokens. When found nearby the word "vision", "saw" is likely to refer to eyesight. On the other hand, when "power" is located nearby, "saw" is likely to refer to power tools. Other words that may have different meanings include pronouns, i.e., "it," "they," "these," "those," and other non-specific nouns used to denote other nouns. Whereas a more primitive NLP model may make little use out of such words, self-attention models can decipher the meaning of pronouns and give them the proper treatment. Figure 6 shows how an attention-based model can characterize a pronoun that other models may not understand.

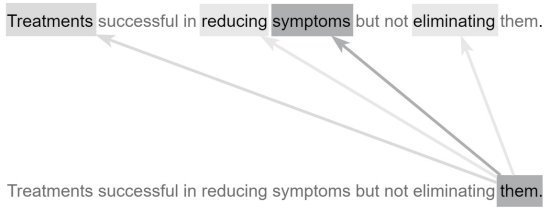


Figure 6: An attention-based model considers a pronoun "them" using the context of the surrounding words.

As shown in Figure 6, the word "them" should not be taken at face value. Instead, the attention mechanism uses the surrounding words to modify its meaning to give it a more specific and useful

characterization. The word "them" attends most strongly to the word "symptoms" because it is the actual word "them" refers to. It also attends weakly to "treatments," a word related to "symptoms" and attends slightly to the verbs "reducing" and "eliminating," both of which act upon the "symptoms" in the sentence. Transformer-based architectures are characterized by the presence of transformer encoders and decoders, which are special architectural modules containing self-attention mechanisms. Encoders are responsible for converting natural language inputs into vectors. These vectors, known as encodings, can be used for a variety of downstream tasks. Decoders work in the opposite direction, taking encoded vectors as input and converting them back into natural language tokens (e.g., words). For example, a question-answer model may first locate the correct answer to a given question in the semantic or meaning space before using decoders to generate the natural-language expression of the answer. In this paper, we explore the automated ICD coding task using the fine-grained approach described in the previous sections. In our approach, a fine-grained data point is formed using a diagnosis and its semantically related sentences. The fine-grained coding task takes one fine-grained data point (natural language input sequence) and outputs one ICD code. To accomplish this task, some essential pre-processing tasks need to be performed, including lowercasing all letters as well as removing specific dates and identifiers as they are not useful for the classification task. Figure 7 shows a fine-grained classifier used to complete the classification step of the fine-grained ICD coding method.

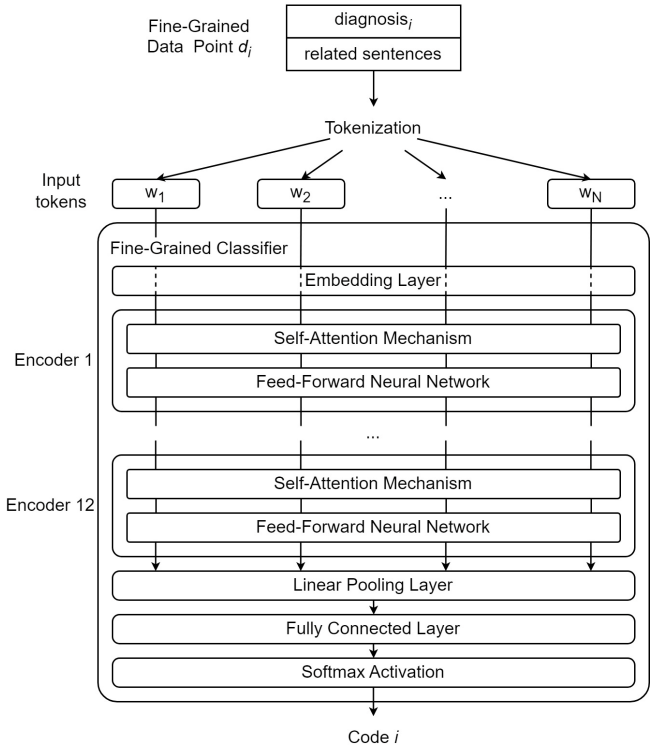


Figure 7: The transformer-based fine-grained deep learning classifier used to generate final ICD code predictions under the fine-grained approach.

As shown in Figure 7, we *tokenize* the text by splitting the data point into a sequence of distinct tokens (e.g., words, suffixes). We then employ a fine-grained classifier that performs the following steps: first, tokens are given their initial vector representations or "embeddings" through an embedding layer. Once the original tokens have been converted to embeddings, they are ready to be encoded. In this case, they are fed into the first encoder, where self-attention is applied, and the resulting vectors are fed into a feed-forward neural network that further transforms the vectors. After passing through the neural network, the vectors are released from the first encoder and go to the next encoder. In this paper, we use a BERT-based model that contains a series of 12 encoders. After the 12 encoders have processed the vectors, the fully encoded vector is pooled and passed through a fully connected linear layer. The obtained values are then passed through a *Softmax* activation layer to output the final prediction.

In addition to the advantages offered by transformer-based architectures, contemporary NLP models such as GPT-4 and BERT benefit greatly from pre-training. That is, they are extensively trained on very large datasets to gain general knowledge of the language and its meaning before being used for a specific application. After pre-training, users can extend or refine the general knowledge of the model for a specific task through further training (i.e., fine-tuning). We use MedBERT [32] as a fine-grained classifier, which is based on Google's BERT architecture, a transformer-based model. In addition to the general pre-training of BERT on BookCorpus and the English Wikipedia, MedBERT was pre-trained on electronic medical record data, which means that it is particularly well suited for ICD coding. We fine-tuned MedBERT's pre-training weights using the ICD coding task to maximize its performance. In addition to using MedBERT for classification, we use another transformer-based pre-trained model, GPT-4, to perform the sentence extraction step of the method. In the next section, we present a GPT-powered concept matching approach to support the extraction of semantically related sentences from doctor's notes to a specific diagnosis.

4. Sentence Extraction Using GPT-4

To extract semantically related sentences from doctor's notes, we introduce a mechanism for reasoning about the potential relationship between a given sentence and a given diagnosis. Briefly, a sentence can be viewed as a sequence of words referring to a set of concepts. In the most straightforward case, a sentence can directly refer to the concept of a diagnosis. In this case, it is easy to determine that the sentence is related to the diagnosis and can be useful for downstream classification. However, a sentence may also be indirectly related to a diagnosis. For example, a sentence talks about a certain symptom such as "runny nose"; although it never explicitly refers to a diagnosis such as "influenza virus", it would certainly be related to the diagnosis *through* the mentioned symptom. Thus, in order to get good coverage of sentences that might be related to a diagnosis, we need to consider not only sentences that talk about the diagnosis, but also sentences that mention other concepts related to the diagnosis. Figure 8 shows the process of sentence extraction using GPT-4. As shown in the figure, the first step in extracting semantically related sentences from doctor's notes is to generate a set of concepts related to the selected diagnosis. These related concepts can be used in a subsequent search for related sentences in the free text portion of doctor's notes. Thus, our approach is to identify a set of sentences related to the original diagnosis by one or more related concepts. In previous work, the sentence extraction step was carried out using a set of related concepts generated through an ontology, a knowledge representation that encodes concepts and

relations in a directed graph [2]. Ontologies offer a number of advantages for this part of the ICD coding procedure. First, ontologies provide a suitable knowledge base for reasoning about medical concepts and their relationships, which is critical to the sentence extraction step. Furthermore, ontologies can be handcrafted by experts and adapted to the ICD coding process. However, the need for manual design of ontologies can be seen as a weakness, as it incurs development costs and requires expert domain knowledge. In this paper, we present an alternative approach that utilizes the Chat Completions API of GPT-4 to generate a list of related concepts to be used in the sentence extraction step. Unlike ontologies, GPT-4 is a pre-trained LLM that works out-of-the-box; while its use and performance may require expert monitoring and validation, it does not incur the initial cost of ontology design specifically for ICD coding tasks.

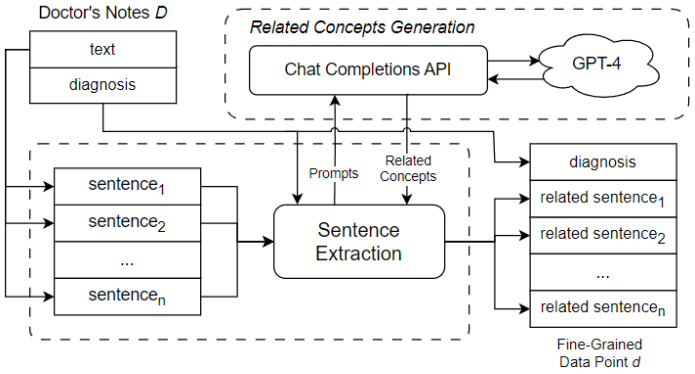


Figure 8: The process of sentence extraction using GPT-4.

GPT-4, the latest available version of OpenAI's LLM, was trained on a large text dataset mined from online documents. It has demonstrated capabilities extending beyond standard language processing tasks. In one study, GPT-4 outperformed expected human scores on the official United States Medical Licensing Examination (USMLE), demonstrating its understanding of healthcare and medical concepts [33]. The approach outlined in this paper utilizes the GPT-4's ability in the healthcare domain to generate a list of related concepts for use in the sentence extraction step. In our approach, we use a diagnosis to generate prompts for GPT-4 designed to elicit a list of related concepts. The prompts are then fed into GPT-4 through its Chat Completions API, which is an interface for the generative features of OpenAI models. Similar to a typical chatroom scenario, the interface accepts an incomplete chat log as an input, from which GPT-4 predicts how the chat will continue by generating the subsequent message.

Since the performance of our approach depends on the quality of the related concepts used to perform sentence extraction, we focused a great deal of attention on designing Chat Completions prompts to elicit appropriate behaviors and concepts from GPT-4. Two main factors were considered when designing the prompts. First, some prompt needs to be designed to direct GPT-4 to properly organize the outputs because we need a list of text strings that correspond to the concepts associated with a given diagnosis. For this reason, we want our Chat Completions prompts to elicit text outputs that are listed, bulleted, or numbered so that we can easily separate the different concepts. To this end, we devise a "system" message specifying that GPT-4 should respond with a dashed list. In this way, the text output by GPT-4 can be reliably separated into individual concepts, which we can then use to extract related sentences. In addition to this generic "system" message, we must design multiple prompts, each including a "user message" representing a query for each subset of related concepts. Table 1 shows several examples of user messages used to elicit concepts related to "Asthma". Another important requirement for the Chat Completions prompt is that it only elicit concepts related to the diagnosis. Therefore, when extracting related sentences from the free text of doctor's notes using the related concepts, they can provide useful information or context for predicting the corresponding ICD code for the given diagnosis.

Table 1: Examples of Prompts for Deriving Related Concepts to "Asthma."

User Message (Prompt)	Related Concepts (Output)
"List the different ways a doctor may indicate the diagnosis 'Asthma' in healthcare documentation."	asthmatic condition, chronic asthmatic, reactive airway disease, asthmatic disorder, …
"List the treatments associated with the diagnosis 'Asthma'."	inhalers, steroids, bronchodilators, leukotriene modifiers, …
"List the symptoms associated with the diagnosis 'Asthma'."	shortness of breath, chest tightness, wheezing, coughing, …
"List the body parts and organs that may be affected by the diagnosis 'Asthma'."	lungs, air passages, bronchial tubes, respiratory tract, …

Since the purpose of sentence extraction is to generate a fine-grained data point for the classification step, it is desirable to derive related concepts and related sentences to support optimal classification results. To this end, we prompt the LLM to generate several classes of concepts related to the diagnosis in different ways (e.g., synonymy, treatment, symptom, etc.). Algorithm 2 describes the detailed steps to extract a set of semantically related sentences Ψ from free text ξ of doctor's notes *dNotes* for a given diagnosis α.

Algorithm 2: Extract Related Sentences for a Given Diagnosis

Input: a given diagnosis α, free text ξ of doctor's notes *dNotes*.
Output: a set of related sentences Ψ

Initialize a set of related concepts $\Gamma_\alpha = \{ \alpha \}$
Initialize a set of related sentences $\Psi = \emptyset$
Define a list of prompts Π to elicit related concepts for α
Let *delimiter* be delimiting symbols (e.g., ',', ';', '\n', …)
for each *prompt* **in** Π:
  Receive a response σ from GPT-4 Chat Completions API
  Derive a set of concepts Γ_σ by tokenizing σ on *delimiter*
  $\Gamma_\alpha = \Gamma_\alpha \cup \Gamma_\sigma$
end
Split the free text ξ into a list of sentences Σ
for each sentence β **in** Σ:
  Derive a set of concepts Γ_β mentioned in β
  if $\Gamma_\beta \cap \Gamma_\alpha \neq \emptyset$:
    $\Psi = \Psi \cup \{ \beta \}$
  end
end
return Ψ

As shown in Algorithm 2, we first initialize the set of related concepts Γ_α so that it contains the original diagnosis concept α, which serves as the starting point for deriving related concepts. The set of related sentences Ψ is also initialized to an empty set. We then define a list of prompts to elicit concepts related to α and receive responses using GPT-4 Chat Completions API. These

responses are tokenized into related concepts and included in Γ_α. After deriving the related concepts using GPT-4, we divide the free text ξ into a list of sentences, tokenizing each sentence to derive the set of concepts discussed in the sentence. If any concept discussed in a sentence also appears in Γ_α, the sentence is considered relevant and is therefore included in Ψ. Finally, Ψ is returned and combined with the diagnosis text to form a fine-grained data point.

5. Automated Hierarchical Classification Using BERT

Traditionally, classification problems are solved by monolithic or "flat" classifiers, which process inputs and make predictions in a linear path through a single classifier or module. Monolithic classifiers are easy to implement and provide good performance in a variety of classification tasks, especially if the number of unique labels is kept reasonable. However, monolithic classification methods are not scalable within a given classifier architecture; as more unique labels are added to the classification task, the complexity required to model the extended label space eventually becomes too great, compromising classification performance. In this paper, we introduce a hierarchical classification approach that predicts the corresponding ICD code for a fine-grained data point through a series of classification steps of increasing specificity. Depending on the design, a hierarchical classifier can first identify the general type of disease and then narrow it down for more precise classification until a specific ICD code classification is derived. In our hierarchical classification approach, additional subclassifiers or classification modules are added, each of which is responsible for a certain step in the overall classification process. We refer to these classification steps as *subclassifications.* For example, the first subclassification of a given instance of doctor's notes might be a binary determination of whether it is a respiratory disease or a circulatory disease, while the second subclassification, which is more specific, might be to determine which individual disease to code for. Figure 9 compares a monolithic classification architecture with a hierarchical classification architecture that consists of three subclassifiers A, B and C, where A is the root subclassifier for the hierarchical classification.

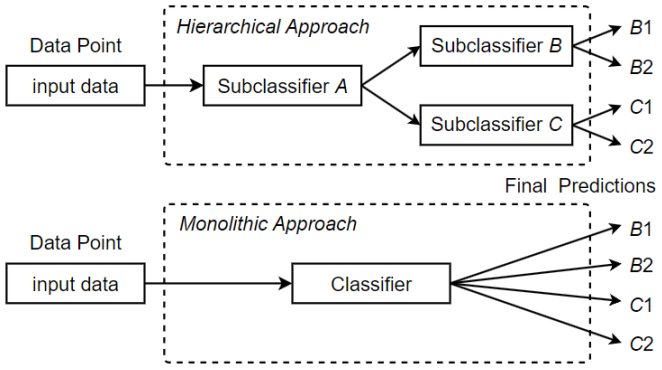


Figure 9: Comparison of hierarchical and monolithic classification approaches.

In a hierarchical classification architecture, the organization of subclassifiers forms a classification hierarchy, where data points flow from the root subclassifier to a leaf subclassifier. The main benefit of using a hierarchy of subclassifiers is that it scales to the large label spaces associated with the ICD coding task. When greater code coverage is required, the classification can be split into individual subclassifications of lower complexity. Each subclassification can then be covered by a separate subclassifier. As long as the subclassifications are sufficiently tractable, classification performance remains high at every step, and the overall performance can be maintained despite the additional complexity introduced by the ever-expanding label space. We define a single-label multiclass subclassifier F_{sub} as a 2-tuple (*FSUB*, *CHID*), where *FSUB* is the prediction function of the subclassifier, which outputs an array of confidences for the output classes, and *CHID* is an array of children subclassifiers, or the empty set $\emptyset$ if the classifier has no further children. A special subclassifier F_{root} is defined as the root subclassifier for hierarchical classification. Algorithm 3 describes the detailed procedure of hierarchical classification. As shown in the algorithm, using a hierarchical classification approach, a fine-grained data point *dp* is first processed by the root subclassifier F_{root}. Based on the confidence level of the output classes, the child subclassifier corresponding to the class with the highest confidence level is selected for further subclassification. This process is repeated until a leaf subclassifier with no child subclassifiers is reached. In this case, the ICD code corresponding to the class predicted by the leaf subclassifier is returned as the matching code for *dp*.

Algorithm 3: Hierarchical Classification

Input: a fine-grained data point *dp,* the root subclassifier F_{root} for the hierarchical classification
Output: the predicted ICD code *code*

Let $F_{current}$ be the current single-label multiclass subclassifier
Initialize $F_{current} = F_{root}$
while $F_{current} \neq null$:
    $confidences = F_{current}.FSUB(dp)$
    $class = argmax(confidences)$
    if $F_{current}.CHID \neq \emptyset$ **then** $F_{current} = F_{current}.CHID[class]$
    else $F_{current} = null$
End
Set *code* to the ICD code corresponding to the predicted *class*
return *code*

It is worth noting that the hierarchical implementation also allows for a degree of modularity in the classification methodology, which can be used for multiple purposes in ICD auto-coding applications. The introduction of hierarchical organization divides the overall classification into discrete subclassification steps designed by domain experts. These divisions and the resulting subclassifications necessarily have human-understandable meanings. Thus, our hierarchical approach has the advantage that even users unfamiliar with the details of deep learning can have some understanding of the various decisions made by the final classification. In other words, it is possible to examine and study the order of the subclassifications or the decision path that led to the final classification. This enhances the trustability that code assignments are made in a coherent and consistent manner and adds to the overall interpretability and credibility of the method. Finally, the modular design of the hierarchical classification facilitates performance analysis, and in cases where classification is difficult or unclear, users or developers can track the subclassifications involved to gain clarification or identify erroneous features. During training and testing, the errors caused by each subclassifier can be examined on a case-by-case basis to identify areas where the

problem is particularly severe and where there is significant potential for improvement of system performance.

6. Case Study

In this section, we conduct a case study to demonstrate the feasibility and effectiveness of our approach using the MIMIC-III dataset, a publicly available healthcare dataset containing medical data from over 50,000 hospital visits [9]. We conducted a smaller experiment with 7 classes (ICD codes) and a larger experiment with 40 classes (ICD codes) to emphasize the advantages of the method with different numbers of classes. For our fine-grained classifier, we adopt MedBERT, a variant of BERT trained on EHR data, previously introduced in Section 3.3. Table 2 presents the complexity matrix and architecture parameters of the MedBERT classifier. For both experiments, we split the collected data into 80% training dataset and 20% test dataset. In the training data, we use 5-fold cross-validation to track model performance and select the best performing checkpoints. To avoid overfitting the classifier and take full advantage of MedBERT's pre-trained knowledge, classifiers were trained with a low learning rate of 5e-5 for 5 epochs, after which the best performing model is selected. In addition, the classifiers were trained using cross-entropy loss. All training and testing processes were performed on a machine with 16 GB of main memory, an Intel Core i7-9700 CPU, and an NVIDIA GeForce RTX 2060 SUPER (8 GB VRAM) GPU.

Table 2: Complexity matrix and parameters of the MedBERT classifier.

Layer #	*Layer Name*	*Input Size*	*Output Size*
1	Embedding	512	(512, 768)
2	Encoder 1 self-attention	(512, 768)	(512, 768)
3	Encoder 1 feed-forward net	(512, 768)	(512, 768)
4-21	Encoders 2-12	(512, 768)	(512, 768)
22	Linear pooling layer	(512, 768)	768
23	Dropout (p = 0.1)	768	768
24	Fully connected linear	768	# of ICD codes
25	Softmax activation	# of ICD codes	# of ICD codes

6.1. Automated Medical Coding Process

Our fine-grained approach performs the medical coding task for one diagnosis at a time. In this experiment, we examine the complete automated medical coding process for predicting the ICD code for a single diagnosis. Suppose the diagnosis to be considered is "Asthma." The first step in the automated medical coding process is to generate a set of concepts that are semantically related to the diagnosis. In our approach, we generate a list of related concepts by prompting GPT-4 through its Chat Completions API. We send multiple prompts to GPT-4 to guide it in generating categories of related concepts for the diagnosis "Asthma." Examples of prompts for the diagnosis "Asthma" can be found in Table 1. Figure 10 shows the procedure for promoting GPT-4 and collecting the related concepts for sentence extraction. The prompts listed in Table 1 are sent through the Chat Completions API, which communicates with and receives responses from GPT-4. Each response contains a partial list of related concepts (e.g., symptoms, treatments, etc.). These outputs from GPT-4 are parsed and collected into a complete list of related concepts, which is then passed to the *Sentence Extraction* module (as shown in Figure 10). The *Sentence Extraction* module segments the free text in doctor's notes into sentences and iterates over each sentence. If the current sentence mentions one or more related concepts, the sentence is extracted; otherwise, the sentence is discarded.

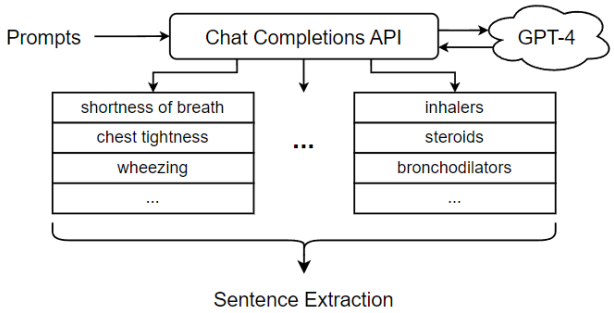


Figure 10: Procedure for prompting GPT-4 and collecting the responses.

After extracting all related sentences from the doctor's notes, they are combined with the original diagnosis "Asthma" to form a fine-grained data point. Figure 11 shows a portion of the fine-grained data point generated for the "Asthma" diagnosis.

DISCHARGE DIAGNOSIS: -Asthma
Pulmonary/Asthma/OSA: The patient inially had a 2L O2 requirement and was weaned to room air after fluid removal at … Pulmonary: The patient was initially admitted with a chronic obstructive pulmonary disease exacerbation andinitially treated with …

Figure 11: Example fine-grained data point generated for diagnosis "Asthma".

In addition to sentences directly referring to asthma, sentences discussing related terms such as "pulmonary" (lung-related) conditions and "O2" (oxygen) requirements, should also be used to enrich the fine-grained data point, providing additional details that can help with classification. With the generated fine-grained data point, we can now pass it to the hierarchical classifier for ICD code prediction. The hierarchical classifier does not immediately classify the data point, but instead generates the final prediction through a series of subclassifications. Figure 12 shows a hierarchical classification pathway that leads to a medical code for the "Asthma" diagnosis.

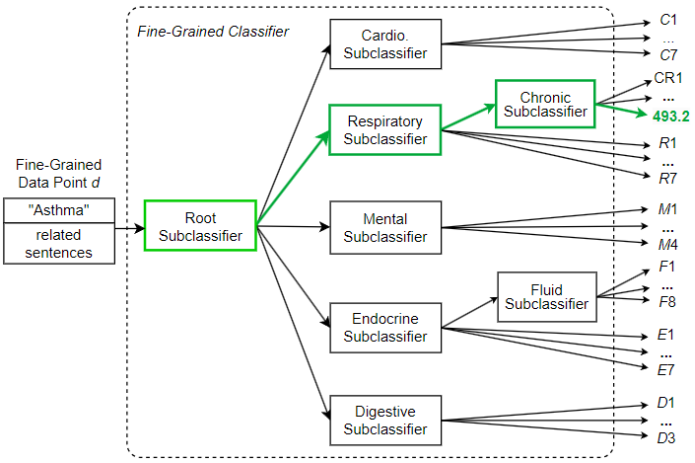


Figure 12: Classification pathway for the diagnosis "Asthma."

As shown in Figure 12, the fine-grained data point *d* created for diagnosis "Asthma" is sent to the root subclassifier. This root subclassifier predicts the broad disease category to which the data point belongs. In this example, the root subclassifier recognizes data point *d* as belonging to the respiratory (breathing-related)

disease category and sends *d* to the *Respiratory* subclassifier. Then the *Respiratory* subclassifier further recognizes *d* as belonging to the chronic disease category and sends *d* to the *Chronic* subclassifier. Finally, the leaf *Chronic* subclassifier recognizes *d* as ICD code 493.2 (chronic obstructive asthma).

Unlike conventional two-step coarse-grained approaches, the fine-grained approach presented in this section follows several explicit human-understandable steps, including diagnosis selection, sentence extraction, and stepwise hierarchical classification. This information can be provided to the user at any point in the medical coding process to illustrate and justify the method's medical coding decisions. This additional information maintains a high level of explainability, thereby eliminating the "black box effect" that occurs when a system with a complex architecture focuses solely on deep learning, thus depriving the regular user of important decision-making information.

6.2. Medical Coding with a Small Code Set

In this section, we establish the viability of the approach and examine its performance with a reduced set of medical codes. For this purpose, we employ a set of 7 ICD codes corresponding to closely related heart diseases. In addition, a comparison with a two-step coarse-grained approach was made to demonstrate the improved performance of the proposed method. Although the code set selected contains only a small number of unique codes, the similarities between the various heart diseases poses difficulties for automated ICD coding. In a typical classification task, similar classes are difficult to distinguish, which greatly increases the overall difficulty of classification. Due to class similarities, the classifiers are more likely to confuse the classes, which reduces performance. This is common in ICD coding, as many of the unique codes involved often refer to variants of the same disease. Among the 7 codes examined in this experiment, two of them refer to hypertension, but they do not refer to the same type of hypertension. Code 401.1 refers to benign hypertension, which can be determined by measurement or testing, but without any apparent problematic symptoms. On the other hand, code 401.9 refers to essential hypertension, which can also be determined by measurement and testing, but may be a dangerous condition that requires some form of medical treatment or lifestyle adjustment. Clearly, keeping these two medical codes separable and correctly identifying each is critical to accurate record keeping and effective patient care. Given this particular difficulty, in order to improve coding performance, we opt for a hierarchical classifier design despite the small number of unique codes. That is, we include an additional classification step responsible for separating potentially difficult instances that fall into one of the hypertension classes. Figure 13 shows the hierarchical design of the fine-grained classifier to predict one of the 7 unique ICD codes. As shown in the figure, going through the root subclassifier, a fine-grained data point *d* may be immediately assigned a final classification and receive the label corresponding to one of the 5 non-hypertension codes in the set. Otherwise, it is assigned to the *Hypertension* subclassifier in order to differentiate whether the hypertension is benign (code 401.1, "Benign hypertension") or essential (code 401.9, "Unspecified essential hypertension"). Since the root subclassifier does not need to distinguish the types of suspected hypertension, its classification task becomes simpler and can identify non-hypertension classes more effectively. On the other hand, the *Hypertension* subclassifier is defined as a dedicated subclassifier trained specifically for separating the two hypertension codes, so it is more capable of predicting one of the two hypertension classes. Under this approach, we split a potentially complex classification task into a series of two less complex subclassification tasks. As a result, each subclassification task has low complexity and high performance, helping to improve the overall classification performance.

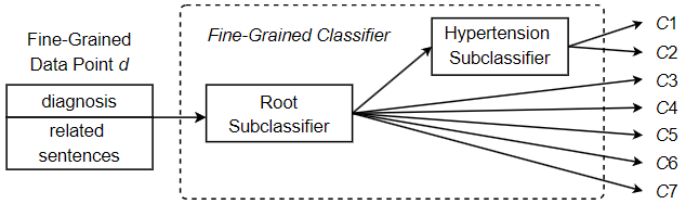


Figure 13: Hierarchical classification with a small code set.

Table 3 shows the performance metrics for processing a dataset with 7 ICD codes using the monolithic and hierarchical approaches. As shown in the table, for this smaller code set, both the monolithic and hierarchical approaches are viable, with an accuracy of over 94.9% and a macro average F1-score of over 87.6%. However, the hierarchical design gives the classifier a slight advantage as it is better able to distinguish between two highly similar hypertension codes.

Table 3: Performance metrics for processing a dataset with 7 ICD codes.

Method	***Accuracy***	***F1-score***	***Precision***	***Recall***
Monolithic	0.949	0.876	0.938	0.851
Hierarchical	0.956	0.894	0.972	0.856

The ROC (Receiver Operating Characteristic) curves and epoch *vs.* validation accuracy plots are presented in Figure 14. The ROC curves in the figure show that the AUC (Area Under the Curve) for both methods is very high, approaching 1.0. As shown in the epoch vs. validation accuracy graphs, both models start out with relatively high performance (thanks to effective pre-training with MedBERT). With fine-tuning, both models showed moderate improvement in accuracy, but neither model demonstrated a significant advantage in validation accuracy.

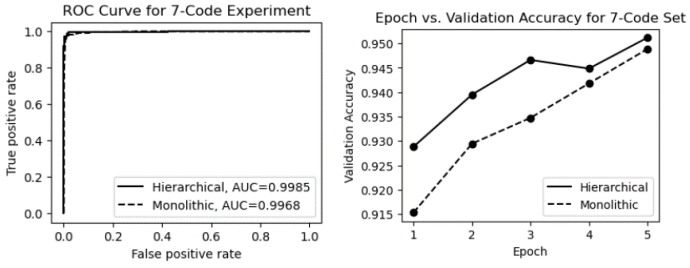


Figure 14: ROC curves and training epoch vs. validation accuracy plots for hierarchical and monolithic classification approaches on the 7-code set.

In addition to these models, we trained and evaluated an approach that employs the conventional two-step or coarse-grained strategy of first vectorizing doctor's notes documents and then feeding the vectors into a coarse-grained multi-label classifier for ICD code prediction. For comparison, we implemented a coarse-grained multi-labeling approach without a sentence extraction step and fine-grained data point formation to exemplify the advantages offered by these novel components of our proposed approach. We

used the same deep learning architecture (MedBERT) to implement this coarse-grained solution with the performance metrics shown in Table 4.

Table 4: Performance metrics for the conventional two-step coarse-grained approach with 7 ICD codes.

Method	*Accuracy*	*F1-score*	*Precision*	*Recall*
Conventional	0.539	0.460	0.375	0.644

Due to the different nature of multi-label and single-label classification and classification metrics, accurate comparison of specific values is not feasible. However, it is clear from the metrics that there is a significant performance difference between the coarse-grained approach and the two fine-grained approaches. The coarse-grained approach demonstrates more difficult and lower performance classification tasks while the deep learning architecture remains the same. On the other hand, the same deep learning architecture was able to keep up and provide excellent performance when working on the same task in a fine-grained manner. In the next section, we will investigate a larger code set where the performance difference between the two fine-grained approaches (monolithic and hierarchical) is demonstrated.

6.3. Medical Coding with a Large Code Set

Since modern standards such as ICD-9 and ICD-10 contain thousands of unique medical codes, useful automated coding tools must be scalable and able to handle large numbers of codes. In this section, we explore how to apply our fine-grained hierarchical ICD coding approach to a dataset containing 40 unique ICD codes. The expanded code set leads to a more complex label space and a more complex hierarchical classification. Despite the difficulties, an effective automated ICD coding approach should remain robust and usable in terms of classification performance. Our approach employes a hierarchical classification method that divides the complex classifications into multiple subclassifications. Figure 12 shows the design of a hierarchical fine-grained classifier for processing a dataset containing 40 ICD codes. As shown in the figure, predictions are made only after two or three steps of subclassifications. The root subclassifier first accepts the fine-grained data point *d* and decides which disease family it may belong to. In this experiment, the 40 codes belong to five families, each affecting a different organ system. These five families are *cardiovascular*, *respiratory*, *endocrine*, *digestive*, and *mental*. Once the root subclassifier predicts the family of a fine-grained data point, that data point is passed down through the predicted branch. For example, if a data point is classified as an endocrine disease by the root subclassifier, it is passed down to the *Endocrine* subclassifier for further classification. At this point, some classes can be predicated immediately, resulting in a final ICD code prediction in code set {*E*1, *E*2, …, *E*7}. However, for endocrine diseases associated with body fluids, the data point is sent to the *Fluid* subclassifier for further classification before an ICD code in code set {*F*1, *F*2, …, *F*8} can be predicted. Table 5 shows the performance metrics for processing a dataset with 40 medical codes using the monolithic and hierarchical approaches. As shown in the table, the hierarchical classification approach remains effective even when the number of unique labels increases significantly. In the case where 40 unique ICD codes need to be predicted, the hierarchical classification method achieved an accuracy and a macro average F1-score of 91.8% and 88.9%, respectively. This successful classification performance provides support for the proposed GPT-enhanced automated coding approach as a potentially useful tool for ICD coding. On the other hand, the performance metrics of the monolithic classifier decreased significantly, with accuracy and macro average F1-score dropping to 73.8% and 70.2%, respectively. This is due to the fact that the subclassifiers of the hierarchical approach have much lower subclassification complexity and correspondingly higher accuracy, whereas the flat or single classifier of the monolithic approach eventually becomes overwhelmed by the increased complexity of the ICD code label space, leading to a degradation of its classification performance.

Table 5: Performance metrics for processing a dataset with 40 ICD codes.

Method	*Accuracy*	*F1-score*	*Precision*	*Recall*
Monolithic	0.738	0.702	0.733	0.696
Hierarchical	0.918	0.889	0.905	0.876

Figure 15 shows the ROC curves and epoch vs. validation accuracy plots for both models. As shown in the figure, the same advantages are demonstrated by the ROC curves and the plots of epoch vs. validation accuracy. Both models gained the base performance from pre-training with MedBERT; however, unlike the previous experiments with a small code set, the hierarchical classifier had a significant performance advantage after the first training epoch (validation accuracy ~0.82 vs. ~0.61). In addition, the ROC curves and AUCs are quite different from previous experiments, which further demonstrates the clear advantage of the hierarchical approach as the number of unique ICD codes (number of classes) increases.

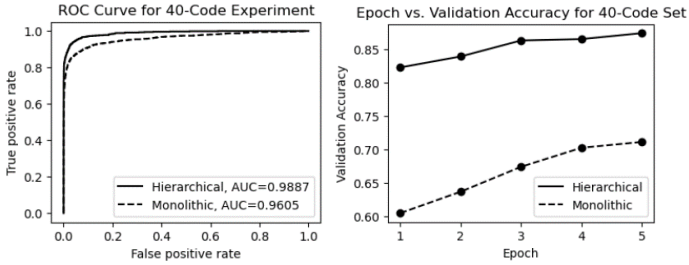


Figure 15: ROC curves and training epoch vs. validation accuracy plots for hierarchical and monolithic classification approaches on the 40-code set.

7. Conclusions and Future Work

Despite the many advances in language processing and classification techniques, medical coding remains a challenging task in healthcare. As a classification task, medical coding poses a number of unique challenges that are addressed by the various techniques presented in this paper. Since medical coding is a multi-label classification task, processing doctor's notes using a fine-grained code assignment method helps limit the label space for individual classifications while still producing an appropriate set of codes. In our approach, we classify only one diagnosis at a time; consequently, individual classifications are simpler and performance is correspondingly higher, even if a large number of unique ICD codes must be considered. Furthermore, we split the prediction process of ICD codes for a given diagnosis into a hierarchical procedure consisting of multiple subclassification steps, thus limiting the complexity of the classification process. Improvements in the hierarchical classifier suggest that

classification tasks like ICD coding with related labels (for example, multiple diseases belonging to a shared family) may benefit from having these relations built into the model architecture, as in ICD code hierarchies. To support a fine-grained approach and provide more informative input to the hierarchical classifier used, we have implemented GPT-enhanced sentence extraction, which prompts GPT-4 for related concepts to be used for locating related sentences in doctor's notes. Unlike previous approaches that use ontologies to generate the related terms used in the sentence extraction step, an approach using GPT-4 provides a useful solution to avoid the additional overhead required to develop a formal medical ontology. The proposed fine-grained hierarchical approach to automated ICD code assignment has yielded promising results in the experiments and provides a basis for effective classification using GPT-enhanced text mining. In addition, the flexibility and effectiveness of GPT-4 in extracting semantically related sentences suggests that there could be further unexplored uses for ICD coding and feature engineering beyond sentence extraction.

While GPT-4 has been used to provide a list of concepts required for our sentence extraction step, the LLM is best known for its wide range of capabilities in a variety of complex tasks. Future work may explore the capabilities of the LLM in automating the various steps of ICD code assignment. For example, GPT-4 could be responsible for the entire sentence extraction step, rather than just generating related concepts. GPT-4 could even be used to perform the entire ICD code prediction, although its performance would need to be carefully examined and characterized. Another potential usage of the generative model is to provide a human-friendly interface for medical coders to answer questions and resolve queries related to predicted codes. Future work could also explore the design and definition of more complex and modular hierarchical classifiers. In particular, improved hierarchical classifiers could employ a variety of decision processes, including decision trees, rule-based reasoning, and deep learning, to produce final classification results. Heterogeneous hierarchies of this type can be designed to be more specific in order to deal with each step of the classification according to the most efficient method. For example, it is usually simple to distinguish between heart disease and eye disease. In most cases, a relatively simple decision-making process can handle this distinction. On the other hand, distinguishing between many highly correlated eye diseases is much more difficult and may require a more powerful decision process, such as the deep learning classifier used in this paper. Since any classification task involving labels can somehow be meaningfully arranged into a hierarchy or taxonomy, future work may explore the application of this hierarchical classification approach not only in ICD coding, but also in different classification tasks such as object recognition, anomaly detection, and topic classification.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We thank the editors and all anonymous referees for spending their valuable time carefully reviewing this paper and making many suggestions for improvement. We also thank the University of Massachusetts Dartmouth for providing financial support to the first author to complete this work.

References

[1] WHO, "International statistical classification of diseases and related health problems (ICD)," Health Topics, World Health Organization (WHO), January 1, 2022. Retrieved on December 12, 2023 from: https://www.who.int /standards/classifications/classification-of-diseases

[2] J. Carberry, H. Xu, "Fine-grained ICD code assignment using ontology-based classification," In Proceedings of the 2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI), 228-233, San Diego, CA, USA, August 2022, doi: 10.1109/ IRI54793.2022.00058.

[3] I. Goldstein, A. Arzumtsyan, O. Uzuner, "Three approaches to automatic assignment of ICD-9-CM codes to radiology reports," AMIA Annual Symposium Proceedings, (1), 279-283, 2007.

[4] R. Farkas, G. Szarvas, "Automatic construction of rule-based ICD-9-CM coding systems," BMC Bioinformatics, **9**(suppl 3, S10), April 2008, doi: 10.1186/1471-2105-9-S3-S10.

[5] J. Medori, C. Fairon, "Machine learning and features selection for semi-automatic ICD-9-CM encoding," In Proceedings of the NAACL HLT 2nd Louhi Workshop Text Data Mining Health Documents, 84-89, Los Angeles, June 2010.

[6] M. Li, Z. Fei, F. Wu, Y. Li, Y. Pan, J. Wang, "Automated ICD-9 coding via a deep learning approach," IEEE/ACM Transactions on Computational Biology and Bioinformatics, **16**(4), 1193-1202, July-August 2019, doi: 10.1109/TCBB.2018.2817488.

[7] T. S. Heo, Y. Yoo, Y. Park, B. Jo, K. Lee, K. Kim, "Medical code prediction from discharge summary: document to sequence BERT using sequence attention," In Proceedings of the 20th IEEE International Conference on Machine Learning and Applications (ICMLA), 1239-1244, Pasadena, CA, USA, 2021, doi: 10.1109/ICMLA52953.2021.00201.

[8] J. Carberry, H. Xu, "A hierarchical fine-grained deep learning model for automated medical coding," In Proceedings of the 3rd International Conference on Computing and Machine Intelligence (ICMI 2024), Central Michigan University, Michigan, USA, April 13-14, 2024.

[9] A. E. W. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, "MIMIC-III, a freely accessible critical care database," Scientific Data, **3**(1), 160035, 2016, doi: doi.org/10.1038/sdata.2016.35.

[10] P. Chen, S. Wang, W. Laio, L. Kuo, K. Chen, Y. Lin, C. Yang, C. Chiu, S. Chang, F. Lai, "Automatic ICD-10 Coding and Training System: Deep Neural Network Based on Supervised Learning," JMIR Med Inform, **9**(8) :e23230, August 2021, doi: 10.2196/23230.

[11] Y. Wu, Z. Chen, X. Yao, X. Chen, Z. Zhou, J. Xue, "JAN: Joint attention networks for automatic ICD coding," IEEE Journal of Biomedical and Health Informatics, **26**(10), 5235-5246, October 2022, doi: 10.1109/JBHI.2022.3189404.

[12] V. Mayya, S. S. Kamath, V. Sugumaran, "LATA - Label attention transformer architectures for ICD-10 coding of unstructured clinical notes," In Proceedings of the 2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 1-7, Melbourne, Australia, 2021, doi: 10.1109/CIBCB49929.2021.9562815.

[13] Y. Dong, "A deep learning-driven disease classification method using MIMIC-III database and natural language processing," In Proceedings of the 2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA), 1068-1072, Dalian, China, October 2023, doi: 10.1109/ICDSCA59871.2023.10392478.

[14] M. Marin, L. E. Sucar, J. A. Gonzales, R. Diaz, "A hierarchical model for morphological galaxy classification," In Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2013), 438-443, St. Pete Beach, FL, USA, May 22-24, 2013.

[15] M. Ramírez-Corona, L. E. Sucar, E. F. Morales, "Hierarchical multi-label classification based on path evaluation," International Journal of Approximate Reasoning, **68**, 179-183, January 2016.

[16] A. D. Secker, M. N. Davies, A. A. Freitas, J. Timmis, M. Mendao, D. R. Flower, "An experimental comparison of classification algorithms for the

hierarchical prediction of protein function," Expert Update (Magazine of the British Computer Society's Specialist Group on AI), **9**(3), 17-22, 2007.

[17] K. Wang, S. Zhou, Y. He, "Hierarchical classification of real life documents," In Proceedings of the 2001 SIAM International Conference on Data Mining (SDM01), 1-16, Chicago, IL, USA, April 5-7, 2001, doi: 10.1137/1.9781611972719.22.

[18] K. Daisey, S. D. Brown, "Effects of the hierarchy in hierarchical, multi-label classification," Chemometrics and Intelligent Laboratory Systems, **207**, 104177, December 2020, doi: 10.1016/j.chemolab.2020.104177.

[19] C. N. Silla Jr., A. A. Freitas, "A global-model naive bayes approach to the hierarchical prediction of protein functions," In Proceedings of the 9th IEEE International Conference on Data Mining, 992-997, December 6-9, 2009, Miami Beach, FL, USA, doi: 10.1109/ICDM.2009.85.

[20] S. Lawrence, C. L. Giles, A. C. Tsoi, A. D. Back, "Face recognition: a convolutional neural-network approach," IEEE Transactions on Neural Networks, **8**(1), 98-113, January 1997, doi: 10.1109/72.554195.

[21] X. Peng, T. Tan, T. Fan, "Automatic ICD coding based on bias removal," In Proceedings of the 2024 IEEE 2nd International Conference on Control, Electronics and Computer Technology (ICCECT), 53-57, Jilin, China, 2024, doi: 10.1109/ICCECT60629.2024.10546138.

[22] T. Chomutare, A. Budrionis, H. Dalianis, "Combining deep learning and fuzzy logic to predict rare ICD-10 codes from clinical notes," In Proceedings of the 2022 IEEE International Conference on Digital Health (ICDH), 163-168, Barcelona, Spain, July 2022, doi: 10.1109/ICDH55609.2022.00033.

[23] A. Kumar, S. Das, S. Roy, "Transfer learning improves unsupervised assignment of ICD codes with clinical notes," In Proceedings of the 2023 IEEE International Conference on Digital Health (ICDH), 278-287, Chicago, IL, USA, July 2023, doi: 10.1109/ICDH60066.2023.00047.

[24] N. T. K. Le, N. Hadiprodjo, H. El-Alfy, A. Kerimzhanov, A. Teshebaev, "The recent large language models in NLP," In Proceedings of the 2023 22nd International Symposium on Communications and Information Technologies (ISCIT), 1-6, Sydney, Australia, October 2023, doi: 10.1109/ISCIT57293.2023.10376050.

[25] J. Fields, K. Chovanec, P. Madiraju, "A survey of text classification with transformers: how wide? how large? how long? how accurate? how expensive? how safe?" IEEE Access, **12**, 6518-6531, 2024, doi: 10.1109/ACCESS.2024.3349952.

[26] C. Sun, X. Qiu, Y. Xu, X. Huang, "How to fine-tune BERT for text classification?" arXiv:1905.05583 [cs.CL], February 2020, Verson 3, doi: 10.48550/arXiv.1905.05583.

[27] J. Devlin, M. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv: 1810.04805 [cs.CL], May 2019, Version 2, doi: 10.48550/arXiv.1810.04805.

[28] F. Karl, A. Scherp, "Transformers are short text classifiers: a study of inductive short text classifiers on benchmarks and real-world datasets," arXiv:2211.16878v3 [cs.CL], 2023, doi: 10.48550/ arXiv.2211.16878.

[29] K. M. Yoo, D. Park, J. Kang, S. Lee, W. Park, "GPT3Mix: leveraging large-scale language models for text augmentation," arXiv:2104.08826 [cs.CL], April 2021, doi: 10.48550/arXiv.2104.08826.

[30] X. Cai, M. Xiao, Z. Ning, Y. Zhou, "Resolving the imbalance issue in hierarchical disciplinary topic inference via LLM-based data augmentation," In Proceedings of the 2023 IEEE International Conference on Data Mining (ICDM), 956-961, Shanghai, China, December 2023, doi: 10.1109/ ICDM58522.2023.00107.

[31] D. W. Otter, J. R. Medina, J. K. Kalita, "A survey of the usages of deep learning for natural language processing," IEEE Transactions on Neural Networks and Learning Systems, **32**(2), 604-624, February 2021, doi: 10.1109/TNNLS.2020.2979670.

[32] C. Vasantharajan, K. Z. Tun, H. Thi-Nga, S. Jain, T. Rong, C. E. Siong, "MedBERT: a pre-trained language model for biomedical named entity recognition," In Proceedings of the 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 1482-1488, Chiang Mai, Thailand, November 2022, doi: 10.23919/APSIPAASC55919.2022.9980157.

[33] D. Brin, V. Sorin, A. Vaid, B. S. Glicksberg, A. W. Charney, G. Nadkarni, E. Klang, "Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments," Scientific Reports, **13**, 16492, 2023, doi: 10.1038/s41598-023-43436-9.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 2, 26-32 (2024)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Spatial Distribution Patterns of the Royal Development Projects Initiated by King Rama 9th of Thailand

Puntip Jongkroy, Ponthip Limlahapun*

Geography Department, Kasetsart University, Bangkok, 10900, Thailand

ARTICLE INFO

Article history:
Received: 05 January, 2024
Revised: 15 March, 2024
Accepted: 16 March, 2024
Online: 31 March, 2024

Keywords:
King Rama IX
Spatial distribution patterns
Spatial web-based
The Royal development projects
Virtual platform learning

ABSTRACT

The study aimed to create a chronological overview of the royal development projects initiated by King Rama IX and to analyze their spatial distribution patterns. The research used a mixed-methods approach, combining quantitative and qualitative data collection methods such as obtaining data from relevant offices, internet research, and field observations. Data analysis involved descriptive statistics and content analysis across three dimensions as temporal, spatial, and disciplinary. The findings were visualized using digital mapping, tables, and diagrams. The analysis revealed that over King Rama IX's seven-decade reign, there was a strong focus on improving the lives of marginalized populations in areas with limited development opportunities and services. The research specifically aimed to 1) present the timeline of changes in the number of royal development projects and 2) analyze the spatial distribution patterns of these projects. The minor objectives included analyzing the patterns of the projects and their spatial distribution. The spatial distribution of the royal development projects was found to be extensive, covering various regions of the country from north to south. However, while there was comprehensive information on the projects, there was no centralized spatial database. Online data accessibility provided a flexible way for users to access project information based on their interests. Additionally, a virtual learning platform was developed to engage younger generations and present the research findings in a more engaging and accessible manner.

1. Introduction

The current polarization of ideas in society, exacerbated by political differences, has widened the generation gap in thinking. This divide has expanded beyond political matters to include challenges to the monarchy, often seen on platforms like Twitter, Facebook, and Instagram. While there is a growing demand for democratic change, actions against the monarchy, especially considering King Rama IX's contributions, based on distorted data from unreliable sources, unfairly undermine one of the country's key institutions. This research focuses on analyzing data related to King Rama IX's royal duties over his seven-decade reign, particularly emphasizing the Royal Development Projects initiated. Quantitative methods were used to analyze the spatial distribution patterns of these projects, with the findings presented through maps, tables, graphs, and diagrams. This study, along with a related project titled "Development Web Spatial Database System of the Royal Projects under the Inspiration of King Rama 9th" [1], which employs web-based GIS for easy retrieval and comprehension, aims to accurately showcase the late King's contributions. The data were categorized into two main types: those with identifiable locations (e.g., reservoir development projects, road construction, schools, hospitals) and those without (e.g., scholarships, public health initiatives, facilities, utilities, conservation and reforestation efforts). To process the data, three approaches were employed which are geocoding to reference coordinate locations using Python in Jupyter and Anaconda, creating project locations as points using QGIS (an open-source GIS software) and storing geospatial data using GeoServer and Openlayers. Additionally, a website was developed using HTML and PHP to integrate the web-based system, enabling users to visualize project information on a map [2] and access detailed project data without needing to install software.

2. Research Materials and Methods

2.1. Scope of Study

Quantitative and qualitative data collection and analysis methods.

*Corresponding Author: Ponthip Limlahapun, thip.limlahapun@gmail.com

https://dx.doi.org/10.25046/aj090204

- Spatial dimension: Thailand and analyzed at the provincial and the regional levels. The six geographic regions and their respective provinces under study are 21 provinces and 1 metropolis in the Central Region, 9 provinces in the Northern Region, 7 provinces in the Eastern Region, 20 provinces in the Northeastern Region, 14 provinces in the Southern Region, and 5 provinces in the Western Region.
- Temporal dimension : The changes in the number of the royal development projects throughout the 70 years of the late King's reign from 1946 when he ascended to the throne to 2016, the year of his passing. The changes were divided into 7 periods of 10 years each: from 1946-1956 through to 2007-2016, and illustrated with diagrams.
- Disciplinary dimension : The disciplinary focus was divided into two major issues in harmony with the research objectives as timeline of the project development based on changes in the number of the royal projects in each decade of the reign classified provincially by the project type. The spatial distribution of the projects in each province and each region was analyzed by comparing the number of the projects to 100,000 population on the provincial and the regional. For the demographic and urbanization perspective, the researchers analyzed the level of urbanization, population density, and the average population growth rate. Finally, the constraint to development from the physical perspective was scrutinized through topography and infrastructures.

These concepts were grouped into eight categories by the Office of the Royal Development Project Board (ORDPB) to be summarized into "the project pattern" in three major groups, as follows. Development projects on resources and the environment (water supply source projects and other environmental and resource projects). Sectoral development projects (agricultural projects; socio-economic project consisting of occupational support, public health, social and educational projects; and infrastructure projects, namely, those involving transportation and communication). Other development projects (projects with integrated concepts or agricultural projects based on the late King's New Theory) [3].

Software and techniques utilized primarily on spatial web-based application. During the web development phase, the focus was on effectively managing data, handling backend operations, and ensuring a user-friendly interface. The system preparation was divided into four main components, namely geocoding, geospatial server setup, database management, and web development.

Geocoding, a notable aspect of this process included restructuring more than 4,500 project records in MS Excel to facilitate streamlined location searches [4]. This required the conversion of address text into geographic coordinates (latitude and longitude) [5], facilitating accurate mapping of locations. While users typically engaged through a Google web browser without the requirement for additional software installation or programming proficiency, the scale of data prompted the adoption of a scripted approach to simultaneously locate multiple locations. The Anaconda and Jupyter environment were set up to execute Python programming for acquiring geocoordinates [6].

GeoServer, a freely available geospatial server environment, is utilized for storing spatial data in diverse formats such as geopackage, shapefile, image, PostGIS, among others. It operates seamlessly on the Windows operating system or functions as a web server service.

General databases such as MySQL and MS Access lack spatial capabilities. Instead, project information was stored using PostGIS, PostgreSQL, and PgAdmin software [7-9]. The developed system incorporates interactive time-series data on an annual basis. Hence, project information was loaded into PostgreSQL and organized based on the dates when the projects were initiated.

Geoserver and web elements were incorporated to provide users with project information presented in spatial and non-spatial formats [10]. Additionally, interactive time-series maps and virtual learning platforms were integrated to enrich user engagement and enhance comprehension of the data.

2.2. Literature Reviews

Analysis on the spatial distribution pattern is part of the spatial process analyses that focuses on the dynamic locational decision-making for land use activities. In [11], the authors highlighted human geographer's interests in people and places that are key to understanding of natural and human-made geographic phenomenon. In this regard, distribution, diffusion, and interaction are the main subsections of the concepts of interaction, communication, and movement on the earth surface. Similarly, in [12], the authors described spatial distribution as the starting point for analyzing the structured content of place that distinguishes geography from other sciences. Common elements to analyze spatial distributions are density, dispersion, and pattern. In [13], the author aims to describe the regular distribution of objects in space. By observing spatial arrangement of research objects across the earth the three properties of distribution: density, concentration, and pattern can be analyzed. For the purpose of the analysis of spatial distribution pattern, the spatial geometric arrangement of interested objects involved two features: 1) the data for the number of projects in each administrative division are depicted by polygons; and 2) quantitative data across an area or surface are illustrated with various shades of color. The research conceptual framework is presented in Figure 1.

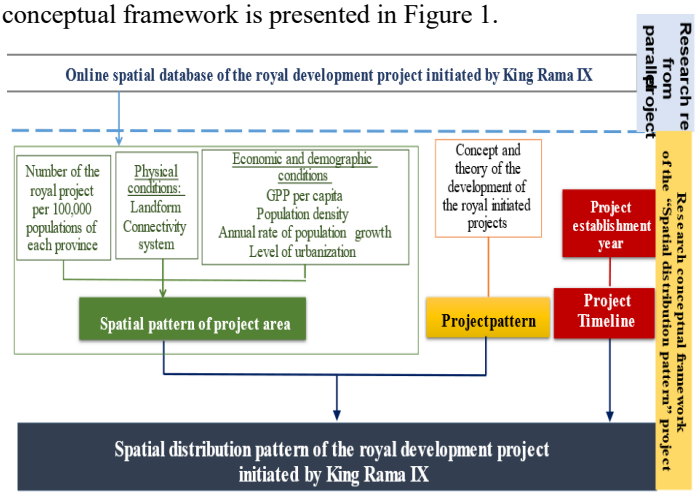


Figure 1: The Research Conceptual Framework (Source: by Authors)

The challenge with spatial data lies in accurately locating all the projects. Traditional techniques for converting address text to geographic coordinates using ArcGIS software are not suitable for Thai addresses. Instead, the Geocoding Application Interface (Geocoding API) from Google is utilized to convert text addresses into geographic coordinates, represented as latitude and longitude [14]. To streamline this process, a custom-written program is employed to handle multiple records simultaneously, eliminating the need for manual searches through Google. This automated approach improves efficiency and accuracy in geocoding a large volume of project records.

In summary, this research project was designed to transform the number of projects and the spatial data on provinces and regions into the ordinal scale. Indices and criteria used in the spatial ordering were set for five issues. They were the spatial ordering of project distribution per 100,000 population in each province of each region; the ordering by economic index (per capital income); the ordering of population and urbanization (annual average of population growth rate, level of urbanization, and population density). Five levels of ordering and interpretation were set including the summary of the overall development level in each province, as shown in Table 1.

Table 1: Development Level

Number of projects per capital	Economic index	Population and urbanization indices			Level
	GPP per capita (USD/ year) [1]	Average annual growth rate (%)[2]	% of urban population[3]	Population density (per sq.km.) [4]	Interpretation
Very few (0.08-5.69)	Low (<1,045)	Contracted (<0.00)	Low-very low (<30%)	Low-very low (<100)	Low
Few (5.70-11.30)	Middle-low (1,046-4,125)	Low (0.00-0.34)	Moderate (30.1-40%)	Low (101-250)	Slightly low
Moderate (11.31-16.91)	Middle (4,126-8,431)	Moderate (0.35-1.00)	Slightly high (40.1-60%)	Moderate (201-500)	Moderate
Slightly large (16.92-22.52)	Middle-high (8,432-12,735)	Slightly high (1.01-2.00)	High (60.1-80%)	Slightly high (501-1,000)	Slightly high
Large (22.53-28.13)	High (>12,735)	High (>2.00)	Very high (>80%)	High-very high (>1,000)	High

Remarks: [1]Applied from the criteria of national development level according to the World Bank's average income per capita [15]..; [2]Classified by the maximum and minimum values from research results; [3]Applied from the United Nations' guideline on urbanization classification [16].;[4]Applied from CIESIN's criteria (Center for International Earth Science Information Network [17].

The developed website integrates virtual learning platforms [18] aimed at involving younger generations in comprehending the vision and concepts driving project development. This includes knowledge about the philosophy and theories of His Majesty King Bhumibol Adulyadej's reign, focusing on projects like the Sufficiency Economy, the Chai Pattana Aerator, a floating buoy that rotates slowly to pump oxygen into wastewater, and Soil Development. These initiatives aim to address issues such as soil acidity, saline soil, vetiver grass, and the prevention of soil erosion and degradation.

3. Results

3.1. Timeline of Changes in the Number of the Royal Development Projects

Quantitative data analysis together with the temporal dimension indicated that 42% of the royal development projects were carried out in the fourth decade of the late King's reign (1977-1986). Only a small number of these projects (0.3%-20%) took place in other decades. It can be argued that the royal development projects were concentrated mostly Table 1. Criteria to summarize "level of area" for spatial pattern analysis of projects with multi-faceted factor displayed with the radar graph. in the fourth decade, especially the water supply source development projects—when as many as 93% of them were implemented. In the fourth decade of the late King's reign, Thailand was in the 4th and 5th National Socio-economic Development Plans. It was also the period of economic transition for Thailand, from an agricultural country to one of the newly industrialized countries. Discovery of natural gas in the Gulf of Thailand had spurred economic growth especially in manufacturing and urban infrastructure system. At the same time, remote rural areas had no accessibility to such development (Figure 2). Moreover, the 4th decade of the late King's reign saw the world in turmoil from the oil crisis that negatively impacted people's lives. Development of the royal development projects, then, put the emphasis primarily on water supply source for the agricultural sector followed by the promotions of occupation, society, and education, as well as other projects such as the New Theory for agriculture or those with integrated activities within the same project. These projects were found mainly outside the metropolitan area.

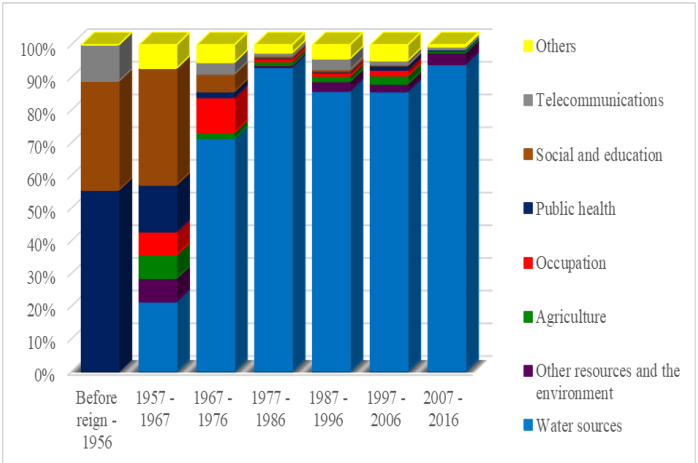


Figure 2: Comparing the Proportion of the Royal Development Projects by Decades of His Reign (Source: by Authors)

Towards the end of the 4th decade, the "Royal Development Center" had been inaugurated in every region. Each center is the hub of efficient coordination and dis-semination of all the royal development projects throughout the regions, supported by numerous experimental projects King Rama IX had established on the grounds of the Chitralada Rahotan Royal Residence. The inauguration of these regional centers and the acceleration of the experimental pilot projects truly laid the strong foundation and the

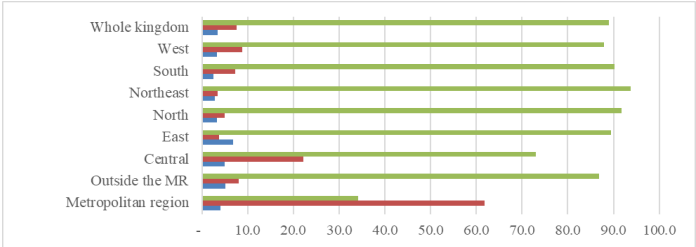
manifestation of sustainable progress of the royal development projects. Features of the above-mentioned development projects were different from the previous three decades that mainly involved projects focusing on people's quality of life, first in the capital and eventually diffused to other regions, specifically to the West and the North. However, the issues regarding the King's health particularly in the last three decades of His reign (5th-7th decades) saw a sharp drop in the number of new projects. These mainly involved transportation and communication, resources and the environment, and public health. Such projects were the direct act in response to the dilapidated natural resources and the environment from industrial development and urbanization. It should be pointed out that development projects concerning transportation and communication had markedly increased in the 5th decade of His reign (1987-1996) with the continuation of infrastructure system development to support urbanization in Thailand from the previous decades. In the latter part of His reign, the area of the royal development projects had been scattering in almost every province of the country, with the metropolitan area taking the backseat. Projects concerning flood mitigation in the metropolis were initiated in the 6th decade but none in the 7th decade. Overall, it is clear that in all the decades of King Rama IX's reign, the royal development projects all aimed at fighting poverty in both urban and rural areas, with great emphasis on the most underprivileged groups in remote rural areas.

3.2. Spatial distribution pattern of the royal development project

- Pattern of the royal development project.

From the three groups of the royal development projects, namely, projects on re-sources and the environment, regional-based development projects, and other projects, it was found that 88.9% of them involved resources and the environment. Among these were mainly for water supply source--the major pattern of the royal development project that could be found in every region, especially in the North. Overall, the distribution of the royal development projects, as illustrated in Figure 3, could be summarized in two patterns.

The first pattern was based on the development concept for resources and the environment. Almost all of the projects in this pattern involved water supply source development and were established in almost every region in Thailand, except in the metropolitan area known as Bangkok Metropolitan Region (BMR) (comprising of Bangkok Metropolis or Krung Thep Maha Nakhon and its 5 vicinity provinces: Nonthaburi, Pathum Thani, Samut Prakarn, Samut Sakhon, and Nakhon Pathom provinces). Nevertheless, differences in the pattern of regional development concept were discovered. In this regard, the North, the Northeast, and the South had well over 90% of all the projects whereas the East, the West, and the Central, with the exception of BMR had 80.1-90%. Overall, only 40-80% of all the projects in the Central Region (including BMR) involved resources and the environment. The second pattern was based on various sectoral project development concepts and was found to cover as much as 60% of all the projects. In this pattern, the development project for resources and the environment was not the major concept as found in other regions of the country. This second pattern was found specifically in the metropolitan area, consisting of Bangkok Metropolis and the surrounding five provinces.



	Metropolitan region	Outside the MR	Central	East	North	Northeast	South	West	Whole kingdom
Resources and the environment	34.2	86.8	73.0	89.4	91.7	93.8	90.2	87.9	88.9
Sectoral development projects	61.8	8.0	22.1	3.7	5.0	3.4	7.3	8.8	7.6
Other concepts	4.0	5.2	4.9	6.9	3.3	2.8	2.5	3.3	3.5

Figure 3: Percentages of the Royal Development Project, Categorized by the Major Development Concept in Each Region in Thailand (Source: by Authors)

- Spatial pattern of the royal development project

Results from spatial distribution analysis of the 2,738 royal development projects in Thailand on the regional basis albeit the total number of projects, the project density per area, or the number of projects per capita—clearly showed the dominance of the North as a major development target with 838 projects or 31% of all the projects in the country. The project density in the North was 9 per 1,000 square kilometers (higher than the national average of 5:1,000 sq.km.) with the highest project numbers per capita of 13.25 per 100,000 population whereas other regions had less than 6 projects per capita. When the data on the number of projects per 100,000 population in each province were depicted in the spatial distribution map of the royal development project, it was found that the province with the highest number of projects per capita was Mae Hong Son in the North (28.11 projects), followed by Narathiwat in the South (25.98 projects). These two provinces had the highest project density, as shown in Figure 4.

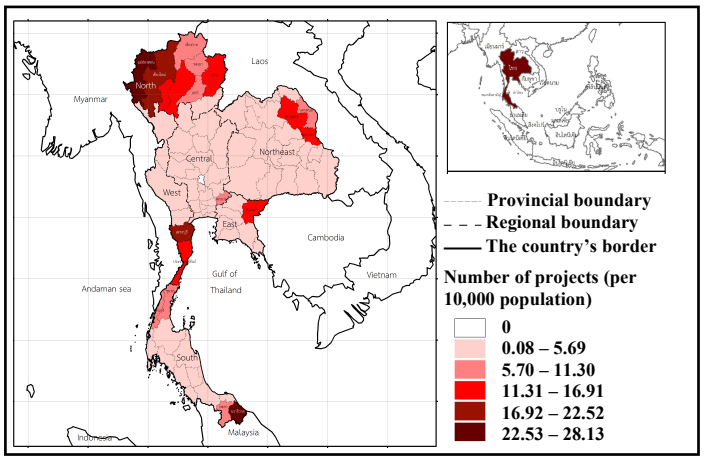


Figure 4: The Spatial Distribution of the Royal Development Project in Thailand, by Province (Source: by Authors)

Results of the spatial pattern analysis in terms of the economic, demographic, and urbanization perspectives together with the number of projects per 100,000 population as illustrated in the radar chart (Figures 5 and 6) could be interpreted into the findings on the regional and provincial spatial pattern (showing the selected provinces with the highest number of project per capita in each region only), as follows. The regional perspective, it was found that the North, a target area for the royal development project, is mostly mountainous interspersed with hills, valleys, and

intermontane basins. Steep areas are headwaters of numerous rivers and creeks. On the socio-economic condition, its population had "middle to low income" which was the lowest income level found in Thailand; its economic base was mixed, its population density was "low", its average annual population growth rate was "low", and its urbanization was "moderate". In fact, the North is the only region experiencing population contraction with six out of nine provinces having lost their population in recent years which was similar or lower than almost every region in Thailand. The provincial perspective, the two provinces with the highest number of projects per capita in the country, classified as the "high" level were Mae Hong Son (28.11 projects) and Narathiwat (25.98 projects). The socio-economic conditions in almost all dimensions of these two provinces were in the "low-lowest" levels in comparison to other provinces in their respective region. Direct fieldwork survey in Mae Hong Son, with the highest number of projects in the country, indicated that most of the royal development projects were in remote areas and far from the provincial center. Its population had limited infrastructure system in both public utility and public facility.

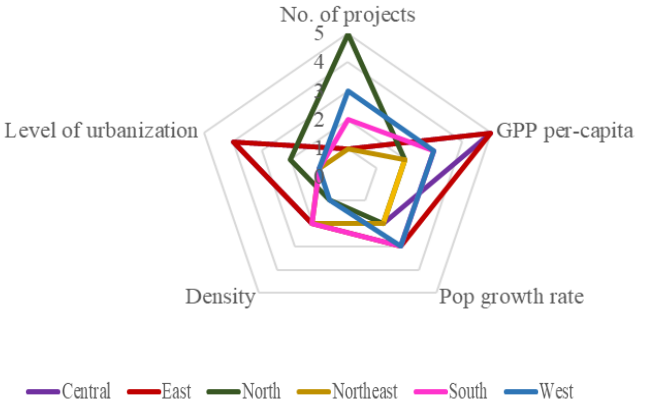


Figure 5: Economic, Demographic, and Urbanization Conditions Scrutinized Together with the Number of Projects Per Capita at the Regional Level.

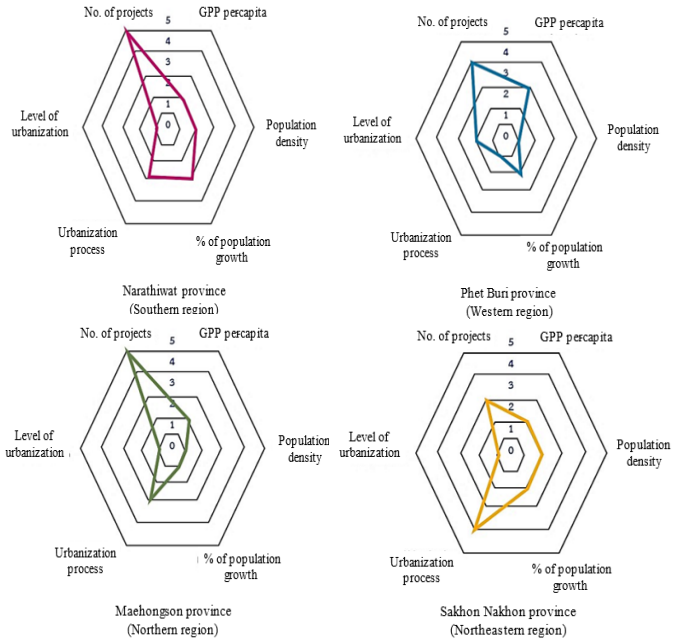


Figure 6: Economic, Demographic, and Urbanization Conditions Scrutinized Together with the Number of Projects Per Capita at the Provincial Level.

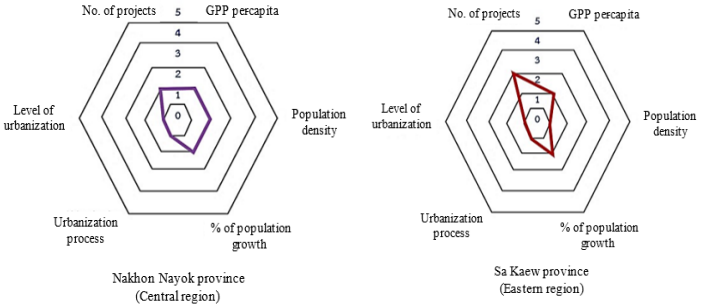


Figure 6 (Cont.): Economic, Demographic, and Urbanization Conditions Scrutinized Together with the Number of Projects Per Capita at the Provincial Level.

3.3. Web development

The web component is intended to enable the sharing of data and information. Data is organized by project types and provinces. The web development is dynamic and comprehensive, incorporating complex web applications that present detailed project information in an interactive and user-friendly manner. An example of this is the virtual learning platform depicted in Figure 7. The website can be accessed at [19].



Figure 7: Virtual Learning Platform (a Sample of Methods to Address Soil Acidity

4. Discussions

The spatial distribution of the royal development projects in the seven decades of His reign could be summarized into two major parts: timeline, and spatial patterns of the distribution of royal development projects in all regions of Thailand during the past seven decades of His reign.

4.1. Timeline of the Royal Development Projects

The distribution of the 2,738 projects were mostly initiated in the fourth decade of His reign (1977-1986) when Thailand was in the process of laying down the infrastructure that was the foundation of national development to become a newly industrialized country. The fourth decade saw a complete inauguration of all types of development project, but the focus was on water supply source development. This focus - in reality, was the core pattern that could be found in every decade of His reign. Moreover, it was also found that most of the projects were distributed in remote regions, far from the capital city. The overall findings in terms of timeline of the royal development projects throughout seven decades of His reign indicated that His Majesty

aimed his efforts at alleviating poverty and hardship in remote areas where the poorest poor reside and at the water source development that benefits particularly those in agricultural sector. This was likely the same as a couple of Ayoo's suggestions on poverty reduction strategies in developing countries. To alleviate poverty, the author [20] recommended stimulating inclusive economic growth, economic and institutional reforms, promoting microfinance institutions and programs, improving the marketing systems, and cash and income transfer programs.

4.2. Spatial Distribution Pattern of the Royal Development Project

Analysis based on the regional perspective indicated that pattern of most of the royal development projects focused on the development of 'resources and the environment' particularly water supply source development which accounted for more than 80% of all the projects. The largest concentration of these projects was found in the North-the only region where the number of projects per capita was classified as "high". The findings noted his focus on water source development which aligns with the concepts explored in the study on water allocation management during scarcity [21]. Setting aside the search for feasible study of applying a bankruptcy approach to water allocation management under scarcity, the authors also pointed out the prominent of the scarcity of water resource that decision makers should focus. This is because it dealt closely with the most vulnerable households, especially those in remote areas.

Regarding the spatial pattern of the North as a development target, it was found that this region had the lowest average income per capita, low population density, and low population growth rate in comparison to other regions. At the same time, its level of urbanization was "moderate" (lower than the Central and the East that were at "high level", but higher than the Northeast, the South, and the West which were at the "low level"). Thus, it could be argued that King Rama IX placed the first priority on the neediest area – where the majority of the most vulnerable group resided.

However, projects concerning rural livelihood and agricultural society such as those for water supply source, agriculture, the environment, occupational development, and other aspects had a random spatial pattern. That is to say that although the projects were distributed in every region but in each region the royal development projects would be concentrating only in selected provinces where poverty and hardship in life were their common problem issues.

From the provincial perspective, the royal development projects could be found primarily in provinces with very low level of urbanization but contained the highest number of projects per capita in each region. As seen in the provinces with the highest number of projects per capital of each region, all had the lowest level of development in almost all aspects (income level, average annual population growth rate, population density, or level of urbanization).

In the end, it can be concluded that targeting the royal development projects on water supply sources and putting focus on remote areas highlighted the uniqueness of what was known as the King's Philosophy. In reference [22], the authors observed that the unique philosophy was a practice that took into account the geographical, economical, and social contexts of each area inclusively. Problem solving strategies were designed not only in a holistic way but also with the focus on building the sense of belonging of local people in order to achieve three ultimate goals: survival, self-reliance, and sustainability.

5. Conclusions

The research findings on the timeline of the royal development projects and the above-mentioned spatial distribution patterns clearly indicated that almost half of the royal projects initiated in the fourth decade of His reign (1977-1986) when Thailand was in the process of the infrastructure installation that was the foundation of national development to become a newly industrialized country. Regardless of the year, King Rama IX placed the highest emphasis on alleviating poverty and raising marginal people's livelihood through the development of water supply sources.

In relation to the spatial distribution patterns, it was found that most of the royal development projects focused on the development of 'resources and the environment' particularly water supply source development which accounted for more than 80% of all the projects. The largest concentration of these projects was found in Northern Thailand where not only the number of projects per capita was classified as "high" but also where the most vulnerable inhabitants lived. Apart from the problems of poverty, destituteness, and disparity in the chance for development; it should be pointed out that hardship in everyday life and inadequate infrastructure system all acted as the pull factor for the initiation of the royal development project as well.

6. Recommendations

The above-mentioned findings clearly echo the royal grace of King Rama IX whose primary intention was to raise life quality of the impoverished. His effort as well as those of the Royal Family resulted in almost 5,000 royal development projects 2,738 of which were initiated within the seven decades of His reign, averaging 40 projects annually that are unparalleled amongst monarchies the world over. His royal grace be-stowed upon the nation should be appreciated and commemorated. Relevant agencies should expedite dissemination of correct information both academically and in-depth in response to false accusation and fake news appearing in Thai society lately that had reached parts of society especially among the youths.

Furthermore, the findings clearly pointed out that most of the royal development projects were found in provinces with low to slightly low level of economic, demographic, and urbanization conditions. Such projects have benefitted communities in remote areas with poor accessibility. Field surveys conducted in various provinces still found the royal development projects in remote areas that had there been without them they would have been in a much worse condition than at present. The social gap and economic disparity between urban areas and remote rural areas would have been far greater than what it is today. Nonetheless, many royal development project areas visited on field surveys in 2022 were found to be dilapidated and poorly maintained. Therefore, relevant agencies should expedite the rejuvenation of these project areas so that they will continue to effectively raise the quality of life of the underprivileged and impoverished.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The researchers wish to thank the Faculty of Social Sciences for the funding and the Department of Geography for the research equipment. Special appreciation is given to the Office of the Royal Development Projects Board for providing the information concerning the royal development projects that were the basis for analyzing the project spatial distribution pattern. The researchers are also indebted to the Department of Provincial Administration, the Ministry of Interior as well as the Office of the National Economic and Social Development Council for invaluable data on population and gross provincial product.

References

[1] P. Limlahapun, P. Jongkroy, "Development web spatial database system of the Royal Projects under the Inspiration of King Rama 9th," Association for Computing Machinery, In Proceedings of the 8th International Conference on Frontiers of Educational Technologies (ICFET '22), New York, NY, USA, 201–209, 2022. DOI: 10.1145/3545862.3545894.

[2] P. Limlahapun, H. Fukui, W. Yan, T. Ichinose, "Integration of flood forecasting model with the web-based system for improving flood monitoring, and the alert system," in International Conference on Control, Automation and Systems, 767-772. 2011.

[3] Office of the Royal Development Projects board (in Thai), http://www.rdpb.go.th.

[4] K. A. Kupzyk, M. Z. Cohe, "Data validation and other strategies for data entry," Western Journal of Nursing Research, 37 (4), 546-556, 2014. DOI: 10.1177/0193945914532550

[5] C. Owusu, Y. Lan, M. Zheng, W. Tang, E. Delmelle, "Geocoding fundamentals and associated challenges," in Geospatial Data Science Techniques and Applications, 2017. doi: 10.1201/9781315228396-3.

[6] Anaconda, https://www.anaconda.com/products/individual.

[7] PostGIS, https://postgis.net/.

[8] PostgreSQL, https://www.postgresql.org/.

[9] PgAdmin, https://pgadmin.org/.

[10] Geoserver, https://geoserver.org/.

[11] W. Norton, M. Mercier, Human geography, Tenth edition, Oxford University Press, Don Mills, Ontario, 2019.

[12] J.D. Fellmann, M. Bjelland, A. Getis, J. Getis, Human Geography: Landscapes of Human Activities, Eleventh edtion, McGraw-Hill, 2010.

[13] J.M. Rubenstein, Contemporary Human Geography, Pearson Education, Inc., 2016.

[14] H. M. Habeeb and N. Al-A'araji, "Extracting implicit geolocation based on google maps geocoding API of social media networks," Research Journal of Applied Sciences, 11(10), 2016. doi: 10.3923/rjasci.2016.1083.1088.

[15] World Bank, World Bank countries classifications by income level, https://blogs.worldbank.org/opendata/new-world-bank-country-classifications-income-level-2021-2022, 2021.

[16] United Nations, World Urban prospects 2018, Population Division, https://population.un.org/wup/Maps/World Urbanization Prospects, 2018.

[17] Center for International Earth Science Information Network, Mapping urban and rural population, In FAO, 2021. https://www.fao.org/3/a0310e/A0310E06.htm.

[18] J.M. Rubenstein, Contemporary Human Geography, Pearson Education, Inc., 2016.

[19] Royal Majesty Project, Web-based development, 2023. http://geosocial-soc.ku.ac.th:81/king_projects/en/index.html.

[20] C. Ayoo, "Poverty reduction strategies in developing countries", Rural Development - Education, Sustainability, Multifunctionality, Intech, 2022. http://dx.doi.org/10.5772/intechopen.101472.

[21] Y. Zheng, X. Sang, Z. Liu, S. Zhang, P. Liu, "Water allocation management under scarcity: a bankruptcy approach," Water Resources Management, 36, 2891–2912, 2022. Doi: 10.1007/s11269-022-03098-0

[22] P. Disayathat, S. Chitsuchon, S. Chandrarat, D. Latthapipat, P. Siriprapanukul, A. Muthitacharoen, P. Luengaram, P. Luengnarumitchai, S. Manprasert, "King Rama IX and Economics" (in Thai)", 2016. https://www.pier.or.th/abridged/2016/23/.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 3, 29-40 (2024)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Visualization of the Effect of Additional Fertilization on Paddy Rice by Time-Series Analysis of Vegetation Indices using UAV and Minimizing the Number of Monitoring Days for its Workload Reduction

Taichi Ito[1,*], Ken'ichi Minamino[1], Shintaro Umeki[2]

[1]*Graduate School of Software and Information Science, Iwate Prefectural University, Takizawa, 0200611, Japan*

[2]*Hanamaki Satellite, Research Center for Industrial Science and Technology, Iwate University, Hanamaki, 0250312, Japan*

ARTICLE INFO

Article history:
Received: 14 March, 2024
Revised: 2 May, 2024
Accepted: 4 May, 2024
Online: 25 May, 2024

Keywords:
Additional Fertilization
Machine Learning
Paddy Rice
Time-Series Clustering
Unmanned Aerial Vehicle
Vegetation Index

ABSTRACT

This research is an extension of the research (ISEEIE 2023), which dealt with Time-Series Clustering (TSC) of Vegetation Index (VI) for paddy rice. The novelty of this research is "Visualization of growth changes before and after additional fertilization," "Analyzing the appropriate amount of additional fertilizer," and "Optimization of monitoring period to minimize the number of monitoring days for its workload reduction using Unmanned Aerial Vehicle (UAV)." For visualization of growth changes before and after fertilization, a UAV was used to obtain VIs for each mesh and make its time-series data during the monitoring period. Then, TSC was performed on the data. As a result of clustering, NDVI and NDRE increased with additional fertilizer, making it possible to visualize the fertilizer effects. For analyzing the appropriate amount of fertilizer, the its amount applied was changed for each paddy field (2.8, 3.5, 4.2g/m^2). In a field experiment conducted, both the TSC results and the crop estimates by unit acreage sampling for each paddy field revealed no difference in yield among fields, indicating that the paddy field with the least fertilizer amount (2.8g/m^2) is optimal. It was estimated that this would reduce nitrate nitrogen, which is harmful to soil and the human body, by 0.070mg/L. In addition, for optimization of the monitoring period, the importance of each independent variable outputted by Random Forest (RF) was used to find a subset of monitoring dates. In any VI, there is a period, determined by the range of effective accumulated temperature, when the clustering result does not change even if the number of monitoring dates is reduced (The period could be reduced from 30 to 40 days, which is particularly important for three vegetation indices). From these results, the technologies can help reduce fertilizer costs, excessive fertilization and environmental impacts, and promote the use of UAV.

1. Introduction

This paper is an extension of work originally presented at the 3rd International Symposium on Electrical, Electronics, and Information Engineering (ISEEIE 2023) [1].

A long-standing problem in Japanese agriculture is the aging and decline of farmers. According to a survey, from 2015 to 2020, the number of farmers decreased by approximately 22% to 1,363,000, and the percentage of those aged 65 years or older increased by 5% to 69.6% [2]. The main causes of this are that appropriate farm work requires tacit knowledge and much human labor. Regarding the former, appropriate work is based on the heuristics of experts; therefore, it is difficult to share. Regarding the latter, Japan's agricultural fields are small, and mechanization has not yet progressed. These problems create barriers for new farmers and lead older farmers to quit farming. To solve this problem, "Smart Agriculture" is being promoted. Specifically, production management systems and the automatic operation of agricultural machinery are expected to clarify and save labor in farm work. However, these technologies are expensive and difficult to implement by farmers. Therefore, in this study, crop monitoring using unmanned aerial vehicles (UAV) was adopted, which is more reasonable than other technologies.

Corresponding Author: Taichi Ito, Iwate Prefectural University, Takizawa,0200611, Japan, s236w001@s.iwate-pu.ac.jp

In UAV monitoring, multispectral cameras are often used to obtain the sunlight reflectance of crop leaves in various bands. Generally, healthy crops have high reflectance in the near-infrared band and low reflectance in the red band. However, in stressed crops, the difference in reflectance between the two bands decreases, as shown in Figure 1 [3], [4]. Based on this reflection characteristic, many vegetation indices (VI) have been developed using calculation formulas to understand vegetation conditions [5]. For example, using VIs, studies have created rice yield maps using the Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Red Edge Index (NDRE) [6], [7]. These studies have shown that the VI value in the young panicle formation stage and the heading stage are correlated with yield; therefore, yield prediction is possible. However, these predictive models cannot be applied everywhere but within the research environment. Furthermore, despite multiple monitoring sessions, the model predicted using only the VI value input during the heading stage. In Japan, yield is important for planning rice demand and supply. Therefore, early and highly accurate predictions with numerous features are required.

This study focused on additional fertilization during farming. This study is important because it directly influences growth and yield. Generally, this work is conducted between the young panicle-forming and heading stages to supply nitrogen and potassium and can be expected to increase yield by enriching the contents of ears [8]. However, excessive fertilizer causes lodging owing to overgrowth, releasing water from the roots due to osmotic pressure, and health hazards owing to nitrate-nitrogen dissolution into groundwater [9], [10]. For nitrate-nitrogen, the Japanese standard value for water pollution is set at 10 mg/L [11]. Based on these negative effects, it is necessary to determine the appropriate amount of fertilizer and apply variable fertilization. Because the effect of farm work appears gradually, it is important to compare it before and after work. Therefore, in this study, we use a new method of time-series analysis of VIs obtained from a UAV to visualize the vegetation before and after additional fertilization and analyze the appropriate fertilization amount. Furthermore, to reduce the workload of UAV monitoring, we use random forest (RF) to minimize the monitoring period.

2. Method

2.1. Monitoring

The monitoring targeted paddy fields owned by private farmers in Hanamaki City in Iwate Prefecture, Japan, and yielded 10 analysis results for six sites from 2021 to 2023 [1], [12], [13]. Several sites were used in the experiment, and the additional fertilizer amount varied for each field. This is because growth differences were observed for each fertilizer amount.

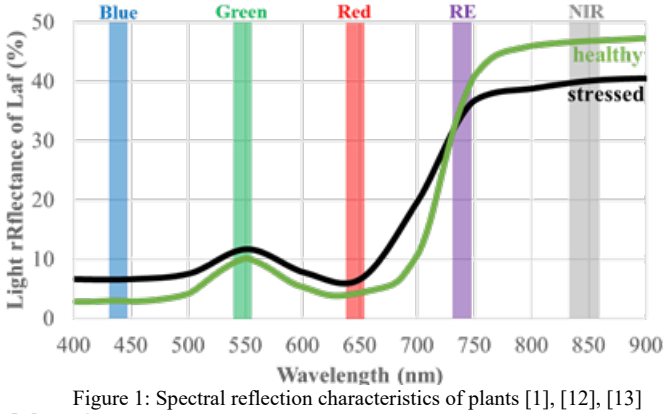


Figure 1: Spectral reflection characteristics of plants [1], [12], [13]

Table 1: Site and Cropping Information

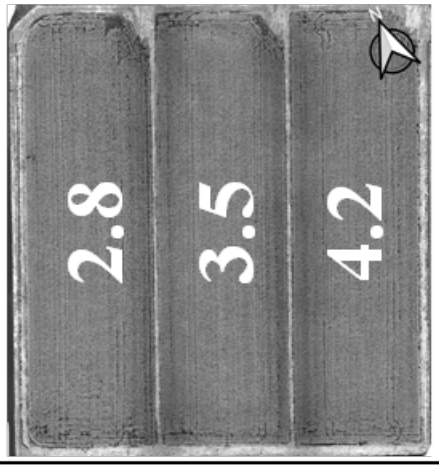
		Yuguchi-8 (2022)	Yuguchi (2023)
Variety		Yumi-azusa	Homusume-mai
Date	Number of Meshes	819	857
	Transplanting	May 7	May 5
	Additional Fertilization	Jul. 20	Middle Jul.
	Heading	Aug. 3	Late Jul.
	Harvest	Oct. 6	Late Sep.
Aerial photo. & Fertilizer Amount $[gN/m^2]$			



Figure 2: UAV used for monitoring (Phantom 4 Multispectral)

Table 1 shows the sites and cropping information where the comparative fertilizer application experiments were conducted (Yuguchi-8 in 2022 and Yuguchi in 2023). The distance between the two fields was approximately 1 km; therefore, the weather conditions were similar. Both fields are adjacent to each other and can be divided into North, Center, and South in Yuguchi-8, and West, Center, and East in Yuguchi. Fertilizer amounting to 2.8, 3.5, and 4.2 gN/m^2 was applied in the middle of July using a pesticide spraying UAV (AGRAS MG-1, DJI Co., Ltd. [14]).

To calculate the VI, a UAV equipped with a multispectral camera is required. Thus, a UAV (Phantom 4 Multispectral, DJI Co., Ltd.), shown in Figure 2 was used. It weighs 1,487 g, has a maximum flight time of 27 min, and has a multispectral camera with one RGB sensor for visible light and five monochrome sensors for multispectral light. The bands of the monochrome sensor were 450 nm (blue), 560 nm (green), 650 nm (red), 730 nm (red edge: RE), and 840 nm (near-infrared: NIR) [15]. The flight altitude was set to 30 m, and orthogonal images of each band were created after monitoring. The VI values were then calculated from these orthoimages. The VIs calculated in this study were (1), (2), and (3).

$$NDVI = (R_{NIR} - R_{Red})/(R_{NIR} + R_{Red}) \quad (1)$$

$$NDRE = (R_{NIR} - R_{RE})/(R_{NIR} + R_{RE}) \quad (2)$$

$$sNDRE = (NDRE - \mu^d_{NDRE})/\sigma^d_{NDRE} \quad (3)$$

where

R_b : *Reflectance of band "b"*
μ^d_{NDRE} : *Average of NDRE on date "d"*
σ^d_{NDRE} : *Standard Deviation of NDRE on date "d"*

The NDVI was used for vegetation diagnosis. This was calculated by normalizing the NIR and Red bands. It is widely used; however, it has been reported that the soil easily affects it, and the index becomes saturated during the peak vegetation season [6], [16], [17]. The NDRE was used to diagnose crop stress and nitrogen content. The RE band reaches deeper into the canopy than the Red band; therefore, stress can be detected earlier. Additionally, it is not easily saturated; therefore, depending on the season, it may be more appropriate than NDVI for diagnosing vegetation and yield [7], [17], [18]. A Standardized Normalized Difference Red Edge Index (sNDRE) was developed to ensure that the NDRE did not change unless it was under significant stress. This is the relative value for each date and is used for a relative comparison within a field [1], [12], [13]. In addition, it uses only NDRE for calculations,

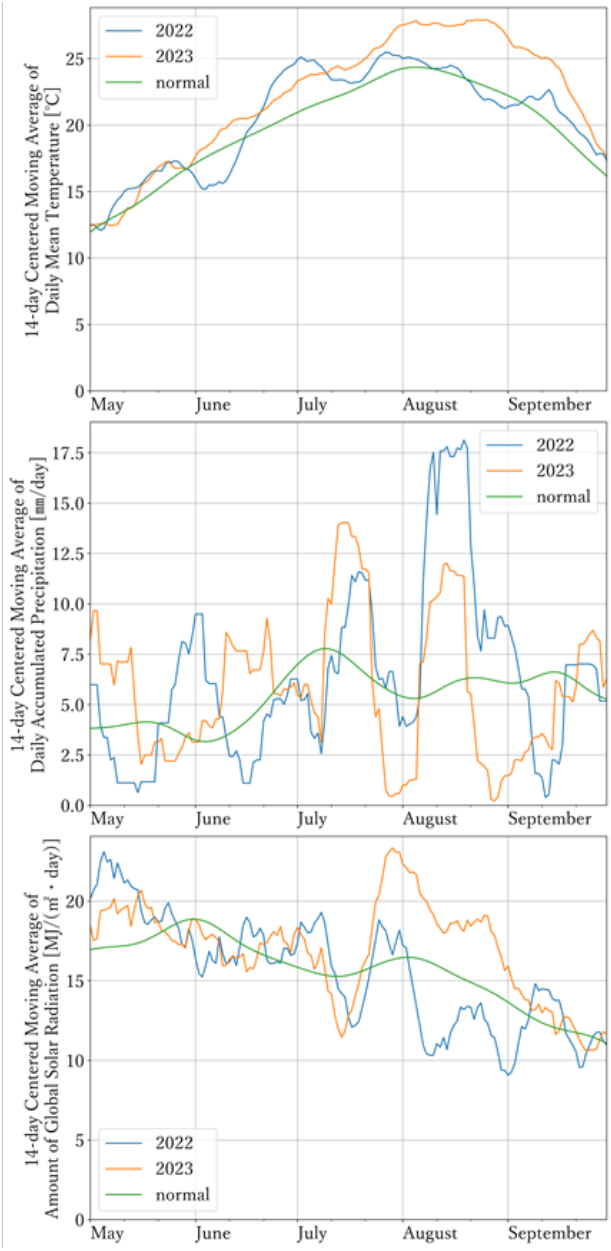


Figure 3: Meteorological data at the experimental field
(Daily mean temperature, Precipitation, and Solar radiation)

and it is possible to diagnose the relative stress and nitrogen content levels in the field.

After computing these three VIs, meshes were created in the rice field portion of the orthoimage. The mesh size was 3 m square.

The VI of each mesh was determined as the average VI value of pixels within the mesh. After computing the VIs of all meshes on all monitoring dates, VI time-series data were generated for each mesh.

Figure 3 shows the 14-day centered moving average of the daily mean temperature, accumulated precipitation, and global solar radiation at Yuguchi during the rice-growing period (May to September) in 2022, 2023, and a normal year. As a characteristic point, there was heavy rainfall in August 2022 and high

temperatures in 2023. To adjust for these weather differences, monitoring dates were converted to growth stages using the Effective Accumulated Temperature (EAT) from the transplantation date and paddy rice growth model [19], [20]. Table 2 shows the correspondence between the EAT and growth stages, and Figure 4 shows the relationship between the growth stages and dates for both years.

2.2. Evaluation of Additional Fertilization

To visualize the fertilization effect, this research used time-series clustering (TSC) with the K-Means++ method on the VI time-series data presented in Section 2.1. Clustering usually classifies 2 or 3-dimensional data which can be shown in scatter diagrams [21]. However, this analysis aimed to classify time-series data. Therefore, a program was developed to handle time-series data with several missing monitoring dates [1]. This program is an improved version of the function "TimeSeriesKMeans" [22] in the Python library "tslearn." This program was used in this research to conduct TSC. The elbow method [1], [23] was used to determine the number of clusters. From the TSC results, the VI time-series transition of each cluster's centroid, cluster distribution, and cluster ratio for each paddy field were obtained. Based on these results, we determined the growth differences according to the fertilizer amount and the appropriate fertilizer amount. Furthermore, a crop estimate by unit acreage sampling was conducted immediately before harvest to confirm the consistency between the appropriate fertilizer amount determined by TSC analysis and the actual yield. The sampling method was to select five meshes per field so that the heading-stage NDVI was scattered, and five stubbles per mesh were collected. The selected meshes are listed in Table 3.

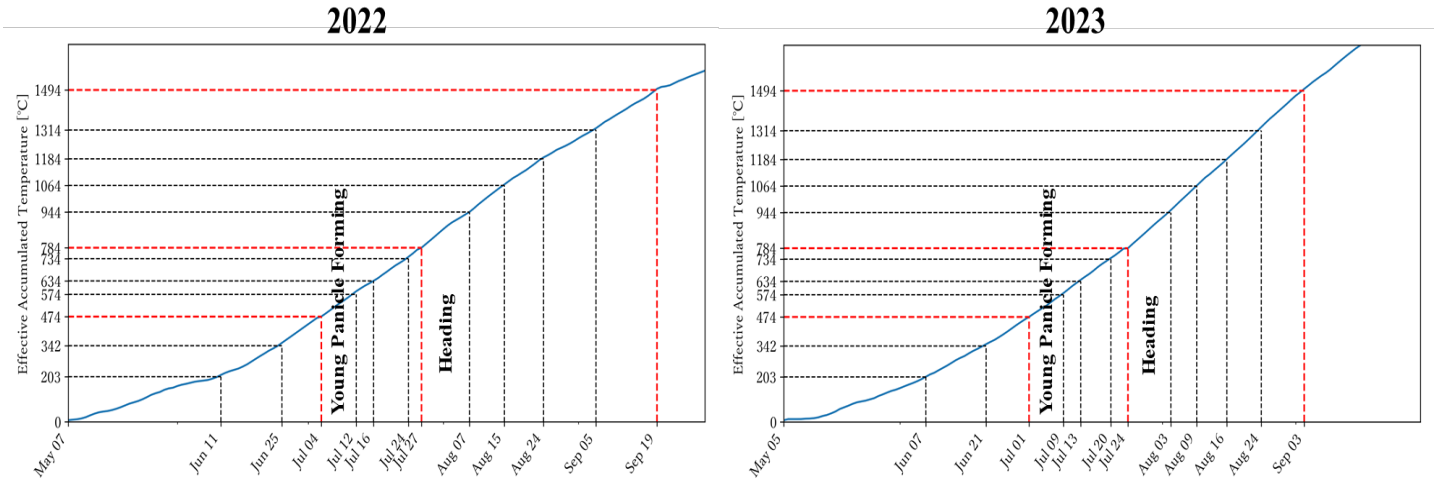


Figure 4: Relationship between the growth stages and dates in 2022 and 2023

Table 2: Correspondence between growth stage and EAT

Model	Effective & Upper limit temperature	Growth Stage	Required EAT [°C]
Leaf number on the main culm	Effective: 10°C Upper limit: 24°C	Slow leaf emergence	0
		Turning leaf emergence rate	203
		Fast leaf emergence	342
Young panicle development	Effective: 10°C	Young panicle forming	474
		Pollen mother cell differentiation	574
		Meiosis	634
		Pollen accumulation	734
Brown Rice Development	Effective: 10℃	Heading	784
		Milk Ripe	944
		Dough Ripe	1064
		Yellow Ripe	1184
		Full Ripe	1314
		Harvest	1494

2.3. Optimization of Monitoring Period

Previous studies have analyzed healthy vegetation growth, stress, etc. in paddy rice using the TSC of VI time-series data [12], [13]. However, UAV monitoring, such as moving to sites, preparing for flight, and computing VI values from aerial photographs, takes time. Therefore, a survey was conducted to identify the subset of monitoring dates for which the feature quantity was not reduced significantly [1]. RF was used in this survey. This was adopted because it can handle the categorical values that are the output results of TSC (i.e., cluster ID) and it outputs the importance of each independent variable (monitoring date).The execution conditions for RF are listed in Table 4 [1]. From the RF, a classifier model was created that outputs the cluster ID when given the VI time-series data of each mesh. The optimization method is shown in Figure 5. First, a classifier was created using the training data of the VI time series over the entire period (Entire RF). Second, the optimized period was determined based on the importance of the test data given to the classifier. Third, another classifier was created using the training data of the VI time series values in the optimized period (Optimized RF). Fourth, the cluster ID match percentages were calculated when classifying the test data between the TSC and Entire RF ($ppppp_{RRReee}$) and between the TSC and Optimized RF ($ppppp_{ooooee}$). Finally, the McNemar's test was performed with the null hypothesis as "$pp_{RRReeee} = ppoooee$" and it was evaluated that optimization was successful if the null hypothesis was not rejected. Using this method, the optimized period was determined from the monitoring data of eight sites (excluding the two sites in Table 1 from the monitoring data of all ten sites). This confirmed the consistency between the appropriate fertilizer amount using TSC analysis results in the optimized period and the actual yield.

3. Results

3.1. Comparison of Additional Fertilizer Amounts

Table 5 presents the yields of the five sampled meshes at both sites. Sample 5 from West of Yuguchi was missing because the survey was discontinued. The sampled yield was calculated as the weight of five stubbles. These units were converted to SI units, and the average for each rice field is shown in the second line from the bottom. The bottom line shows the p-value of the Kruskal-Wallis test (non-parametric, unpaired, differences between three or more groups) with the null hypothesis set as "the average of sampled yields of all rice fields are equal." As a result, the null hypothesis was adopted even at the 10% significance level, as the p-value for both sites significantly exceeded 0.1.

For the entire TSC period, the time-series transitions in each cluster centroid, cluster ID distribution map, and percentage of

Table 3: Sampled meshes

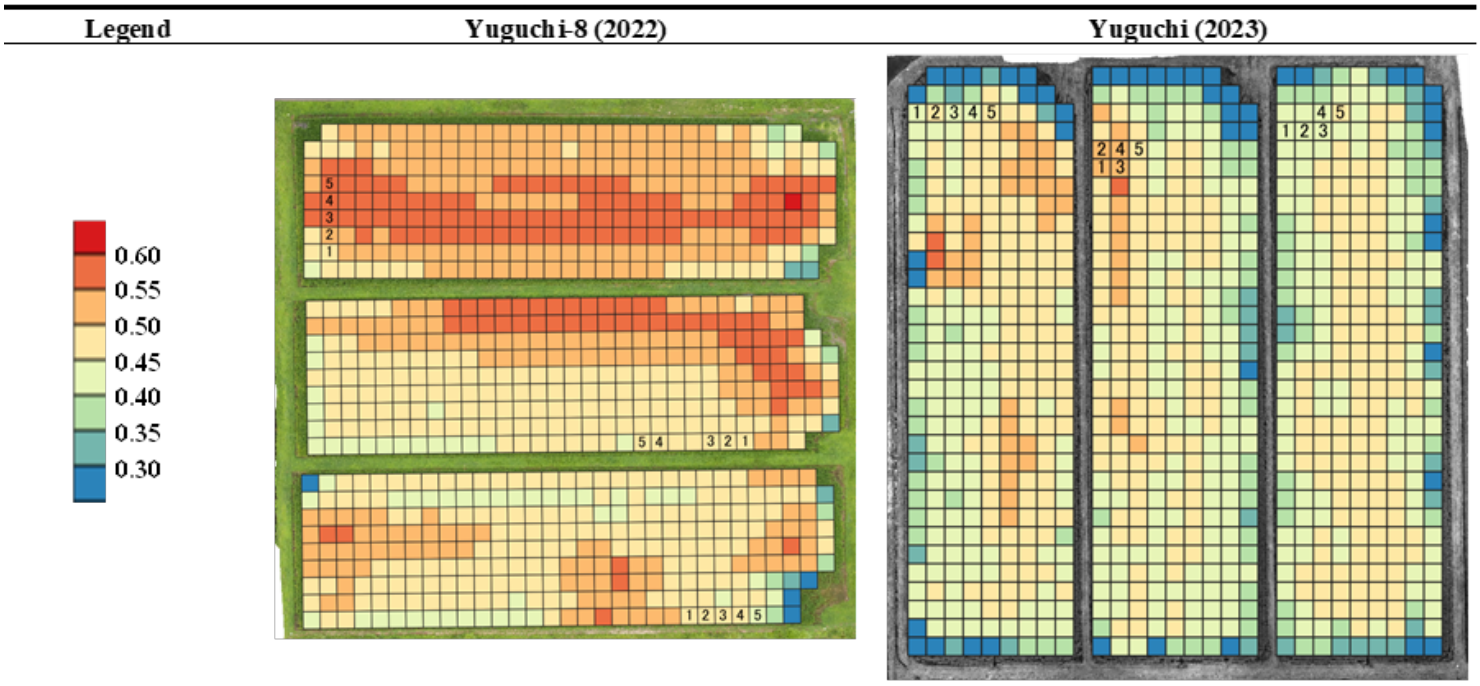


Table 4: Execution conditions of RF

Item	Condition
Independent Valuable	Time-series VI value per monitoring date of each mesh
Dependent Valuable	Cluster ID of each mesh
Training Data	Sampling by Stratified Sampling with Cluster ID as a layer (25% of total meshes)
Test Data	Remaining 75% of total meshes

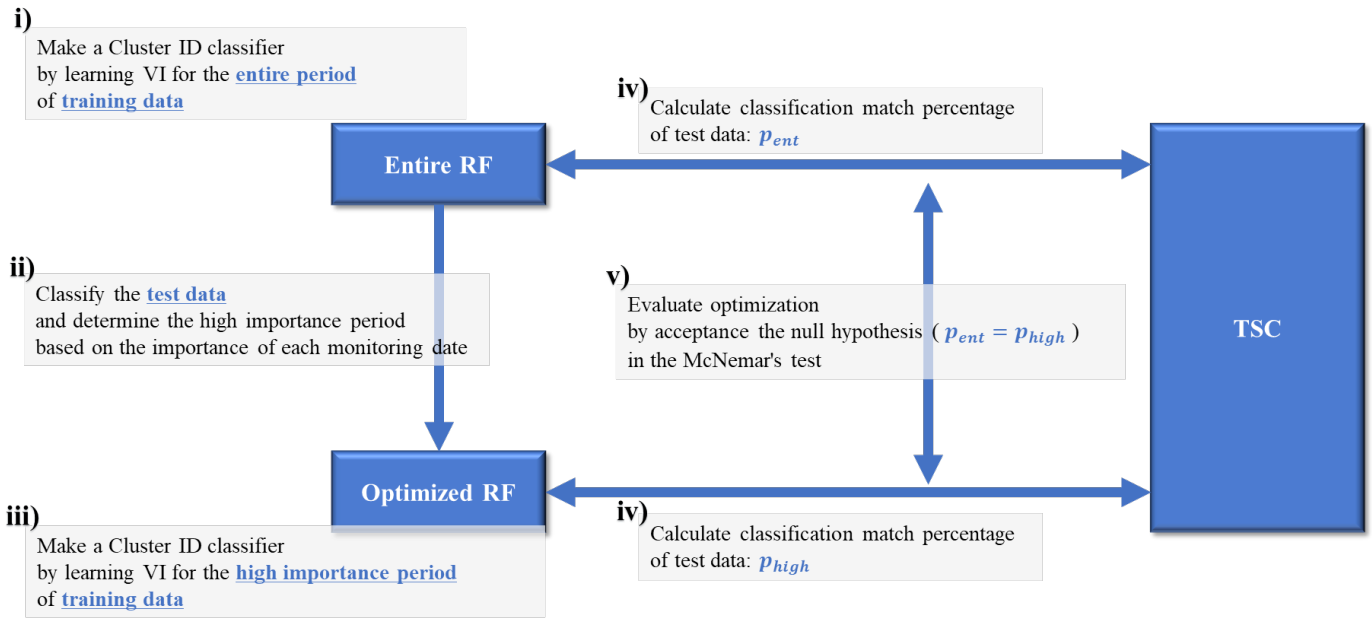


Figure 5: Method of monitoring period optimization

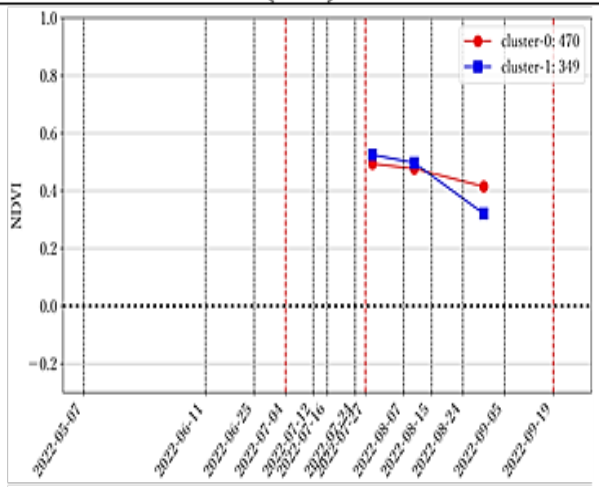
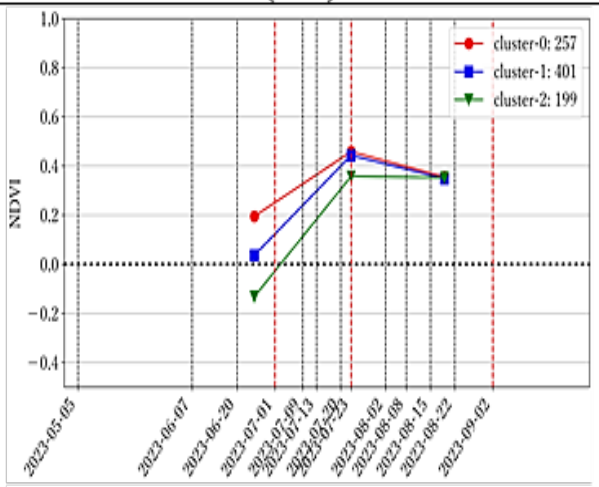
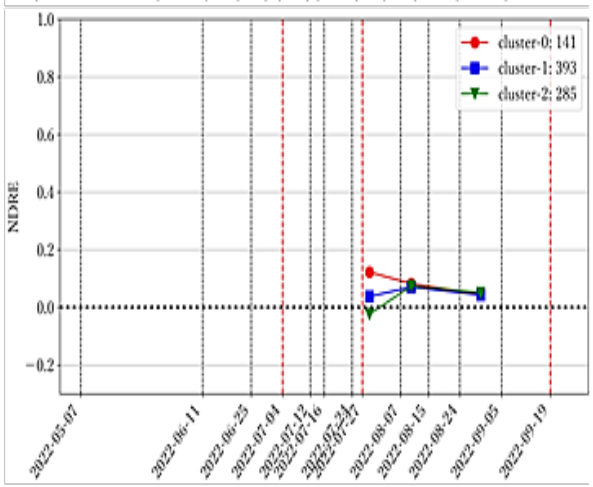
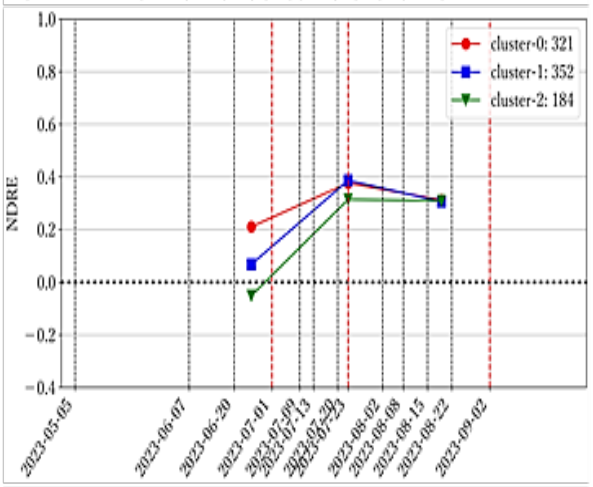
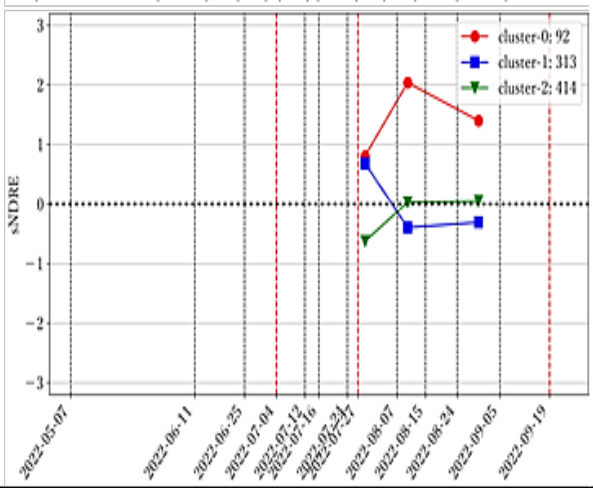
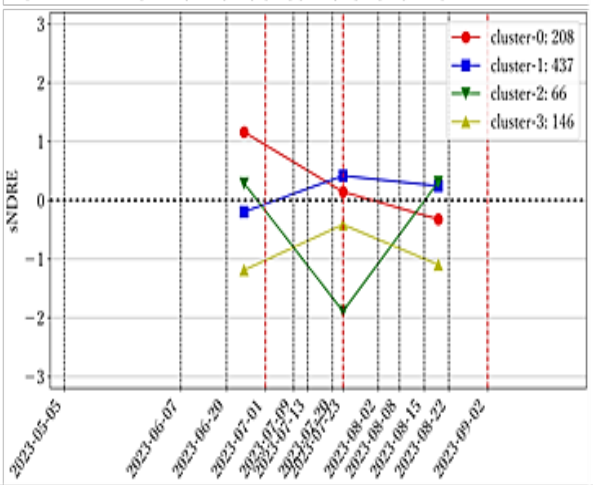
Table 5: Yields of 5 sampled meshes and results of Kruskal-Wallis test

g / 5stubbles	Yuguchi-8 (2022)			Yuguchi (2023)		
	North	Center	South	West	Center	East
Sample1	176.0	180.8	220.6	136.4	161.8	221.9
Sample2	239.0	172.2	188.4	203.0	66.6	175.7
Sample3	189.6	192.0	164.0	190.5	204.8	142.0
Sample4	180.2	178.2	155.6	177.6	214.3	206.9
Sample5	213.0	175.8	139.2	N/A	181.5	178.0
Average	199.6	179.8	173.6	176.9	165.8	184.9
Standard Deviation	26.3	7.5	31.7	28.9	59.1	30.9
Average [g/㎜²]	724	653	630	642	602	671
p-value of Kruskal-Wallis Test	0.249			0.873		

each cluster are shown in Table 6, Table 7, and Table 8. For Yuguchi-8 in 2022, the meshes were classified into two, three, and three clusters for NDVI, NDRE, and sNDRE, respectively. For the NDVI, there was no difference between clusters in the heading stage, and the yields were expected to be similar in all rice fields. In the NDRE, differences were observed at the heading stage, but no differences were observed thereafter. Cluster-0, which had relatively low stress, was found near the ridges. In sNDRE, Cluster-0 was high and often observed near the ridge. The central meshes of each paddy field were many Clusters-1 and Cluster-2, but after heading, the value did not differ by more than 0.5, with no difference in the stress level. The same can be said for all three rice fields, and no growth differences were observed between them. Therefore, the appropriate fertilizer amount was analyzed as the north field (2.8 gN/m^2), where the amount is the lowest, considering fertilizer costs etc.

At Yuguchi in 2023, the meshes were classified into three, three, and four clusters for NDVI, NDRE, and sNDRE, respectively. In the NDVI, there was a difference of approximately 0.15 between clusters at the first monitoring. However, the low value (Cluster-2) improved because of the fertilizer. Thus, it was concluded that there was no difference in growth between the fields because the value of each cluster was similar at the heading stage. There was a large difference in sNDRE between the four clusters. Similar to other VI, high-value (Cluster-0 and Cluster-1) were observed at the center of the field. These two clusters appeared slightly more frequently in the West and Center. Similar

Table 6: Centroid of each cluster (entire period)

Centroid of each cluster	Yuguchi-8 (2022)	Yuguchi (2023)
NDVI		
NDRE		
sNDRE		

to Yuguchi-8, the same was observed for all three fields, but no growth differences were observed. Therefore, the appropriate amount of fertilizer was analyzed in the western field (2.8 gN/m^2).

3.2. Optimization of Monitoring Period

The importance results for the six sites (excluding Yuguchi-8 in 2022 and all sites in 2023) are shown in Figure 6. The vertical axis in this figure represents the Relative Importance. The Importance output by the RF is the percentage contribution [1], [24]. That is, the fewer the monitoring dates, the higher the value. To adjust for this, Relative Importance was calculated by dividing the usual output importance by the reciprocal of the number of monitoring dates [1]

Table 7: Cluster ID distribution map (entire period)

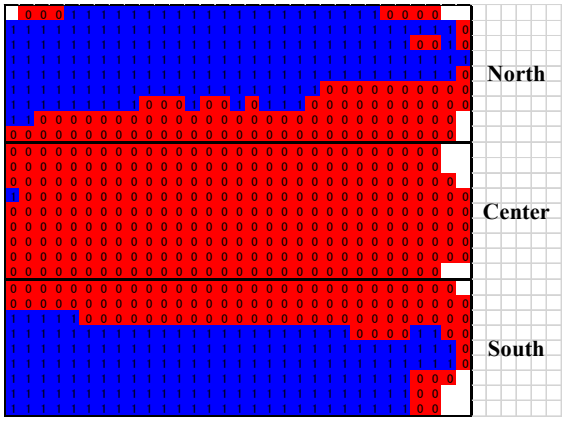
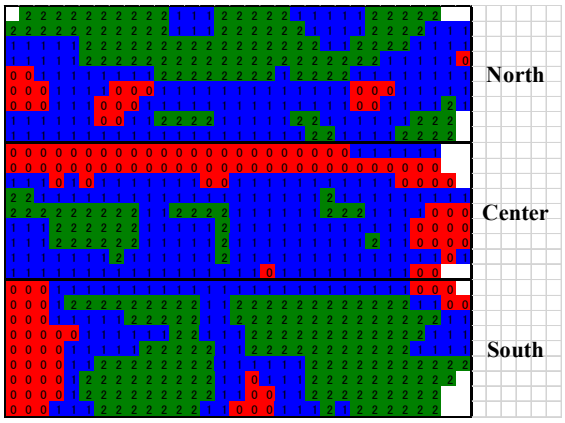
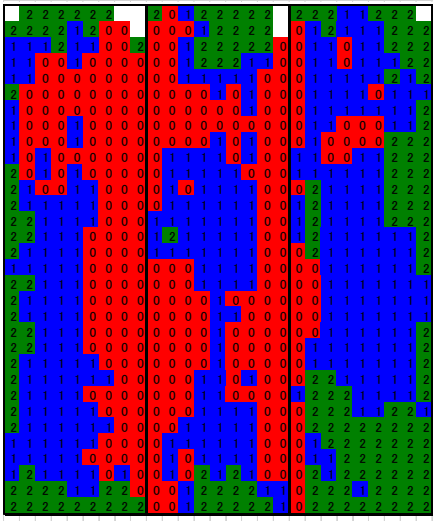
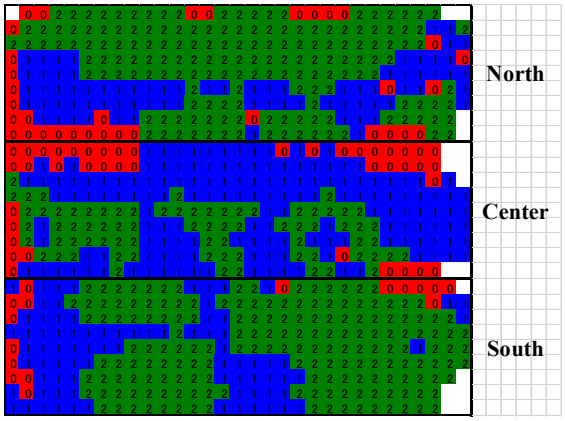
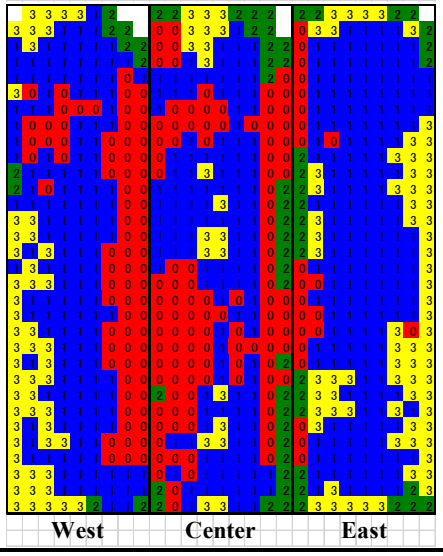
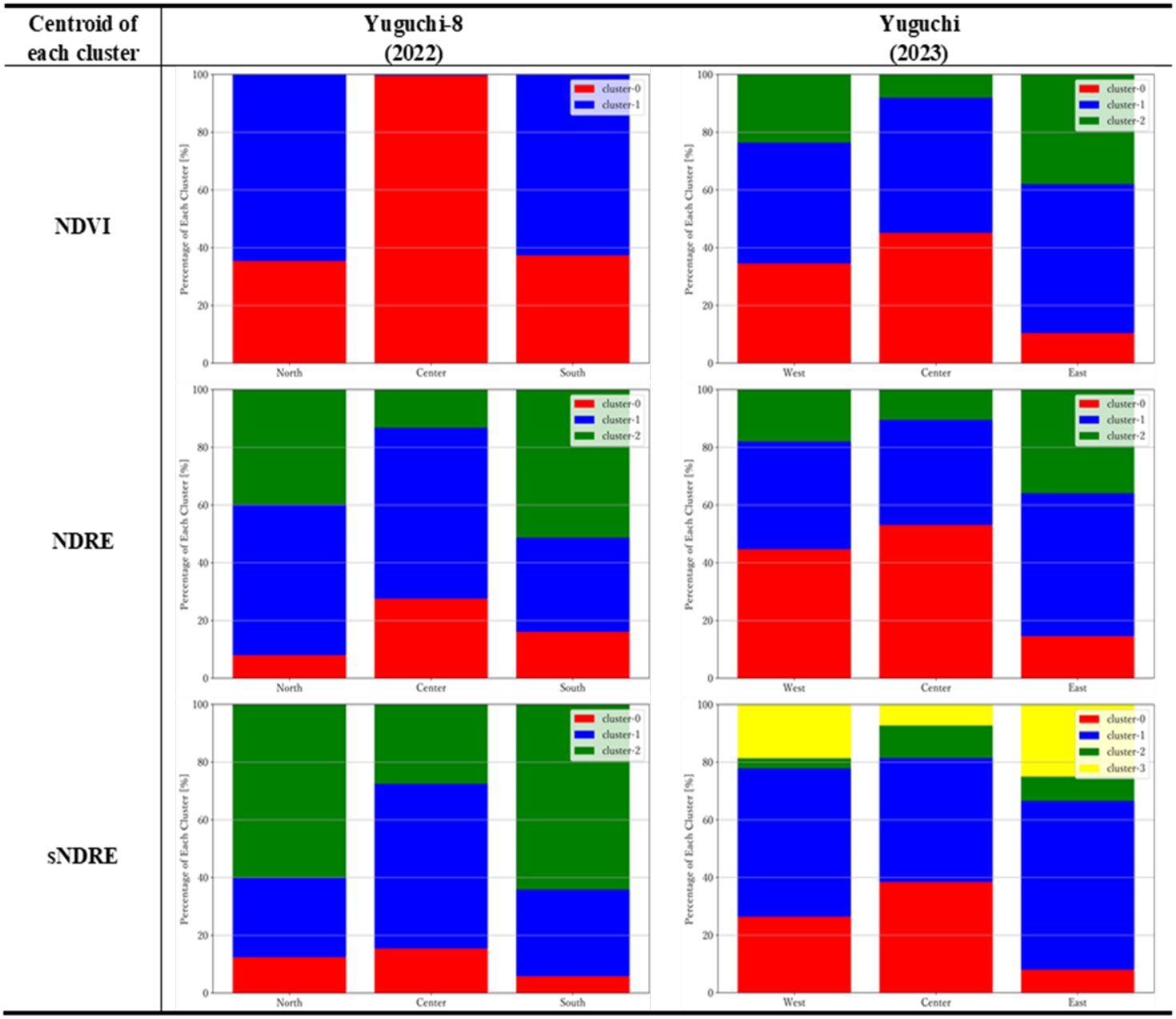
Centroid of each cluster	Yuguchi-8 (2022)	Yuguchi (2023)
NDVI		
NDRE		
sNDRE		

Table 8: Percentage of each cluster (entire period)



[24]. To optimize the optimization method of monitoring period, first, the initial period was extracted, for which the minimum period included all data with a Relative Importance of one or more. Next, the period was expanded or deduced to the sum of the Relative Importance of 49% to 51%, in contrast to the entire period shown in Figure 6.

The optimized periods of the EAT determined using this method are listed in Table 9. The rightmost column shows the EATs of the optimized period converted to dates using the EAT from May 7 in a normal year. The optimal period for all VIs was approximately one month. This period was from the young panicle forming to the milk ripe for NDVI and NDRE, and from heading to the full ripe for sNDRE. In Table 10, the results are applied to two sites that were excluded when calculating the optimized period (i.e., Yuguchi -8 in 2023 and Yuguchi in 2022). For the match rate of the cluster ID classification of the test data between the Entire RF and TSC (p_{ent}), and the Optimized RF and TSC (p_{opt}), p_{opt} was greater than p_{ent}. They are not necessarily 100% because the training data comprise 25% of all the meshes and the remaining 75% of the test data. The rightmost column shows the p-value of the test for the difference between two percentages. No VI was significant at the 5% level. In other words, the monitoring period can be optimized without significantly reducing the classification accuracy. In addition, although the 2023 data were not used to derive the optimized period, p_{opt} of Yuguchi had high accuracy (over 98%). This indicates that the optimized period can be derived using the EAT without depending on the year.

4. Discussion

4.1. Visualization of Paddy Rice Growth

In the TSC during the optimized period, most results showed that growth was different in the center of the fields and near the ridge. In the NDVI and NDRE, the area near the ridge was classified into low-value clusters, so it is thought that the vegetation growth and stress states were not better than those in the center of the field. In contrast, for the sNDRE at Yuguchi-8, the near-ridge meshes were classified into high-value clusters. It is thought that the amount of nitrogen was high because the fertilizer was washed away by heavy rain in August, (Figure 3, Table 7).

However, at Yuguchi-8, all monitoring occurred after heading, and at Yuguchi, the first time was before young panicle formation and the second time during heading. In short, no monitoring was conducted between the start date of the optimized period and the fertilization date at either site. From the results of the optimized TSC period, it was not possible to visualize the growth before fertilization. Thus, vegetation differences are discussed using the results for the entire period (Table 6). From the time-series trends of each cluster, it can be seen that NDVI and NDRE increased between the first and second monitoring days. This suggests that nutrient supplementation with fertilizers improves vegetation and stress conditions. In addition, sNDRE is not suitable for comparing these conditions with other days because of the standardization of each day.

Table 9: Optimized period of EAT of each VI

VI	Optimized period of EAT [℃]	Date for a Normal Year
NDVI	509 – 913	Jul. 10 – Aug. 10
NDRE	588 – 1063	Jul. 16 – Aug. 21
sNDRE	860 – 1332	Aug. 6 – Sep. 11

As described above, TSC of the VIs was able to visualize the health status of paddy rice and predict the appropriate fertilizer amount. On the other hand, in order to further improve predictive.

Table 10: Results of mesh classification match percentage between RF and TSC and McNemar's test

VI	Site	p_{ent}	p_{opt}	p-value
NDVI	Yuguchi-8	96.3%	100.0%	< 0.001
	Yuguchi	96.5%	99.4%	0.002
NDRE	Yuguchi-8	94.3%	99.0%	< 0.001
	Yuguchi	96.0%	98.1%	0.050
sNDRE	Yuguchi-8	91.0%	95.9%	0.003
	Yuguchi	93.4%	99.4%	< 0.001

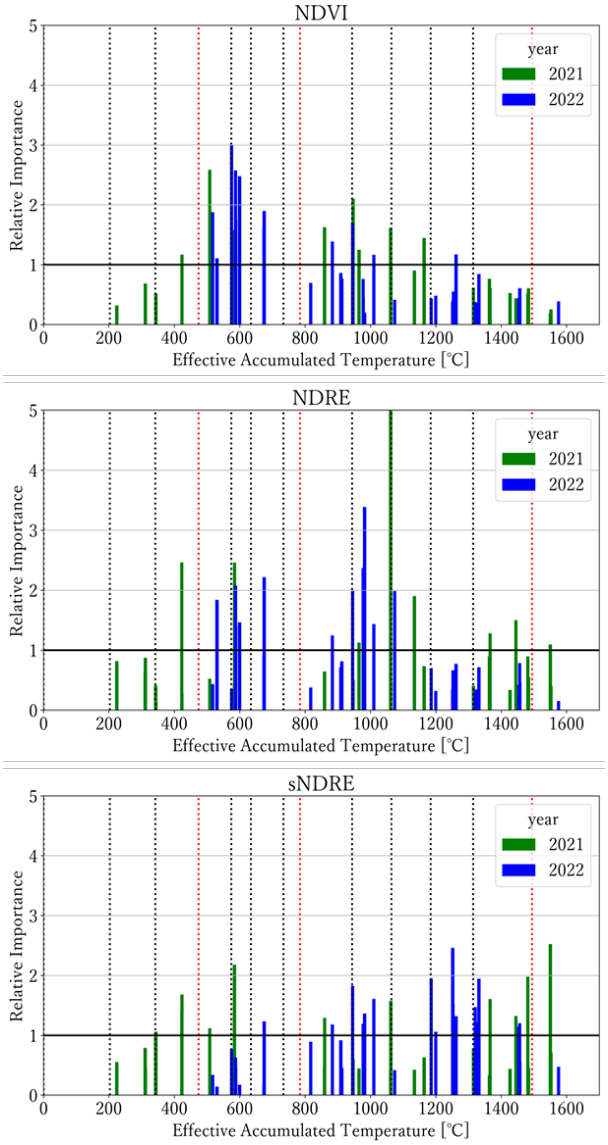


Figure 6: Relative Importance

For vegetation growth changes before and after fertilization, there were no monitoring data before fertilization included in the optimized period for either field (Table 6, Table 9).

Usually, additional fertilization is conducted on a day when EAT is around 500℃ to 650℃ (about 2 to 4 weeks before heading). Although sNDRE was not included in this period, NDVI and power, the following two things need to be considered. The first is the use of a broader range of wavelengths beyond visible light. Some of the VIs that have been developed so far use these bands (e.g., NDNI, CAI, etc.) [5], [25]. By using these VIs after confirming their effectiveness with ground truth data such as RGB images and site inspections, it becomes possible to visualize water stress, nutritional deficiencies, pest infestation, and so on. For ensuring the accuracy and reliability, it is necessary to verify them using detailed measurement data obtained from sensors such as air temperature and water temperature in mesh unit. Future research should utilize such ground truth data for TSC to observe various aspects of crop conditions. The second is to perform accurate spatial analysis. In this paper, for fast processing, the data used in the TSC was reduced in size by determining the mesh size

(approximately 3m x 3m) based on the resolution. It is desirable to perform TSC with various mesh sizes and verify the mesh size that increases the accuracy of the optimal fertilizer amount prediction and yield prediction. Also, the spatial variability of vegetation indices within the rice field and its implications for fertilizer management may be caused by the evenness of the rice field, location of water outlets and water flow within the rice field, and so forth. We hope to identify the causes of this spatial heterogeneity using spatial analysis techniques and the analytical methods of this study.

4.2. Evaluation of Optimization

In the previous section 4.1, it was analyzed in both fields that the appropriate fertilizer amount was the lowest amount field by the TSC using the optimized period. The bottom panel of Table 5 shows the evaluation of this analysis from the viewpoint of yield. As mentioned in the previous chapter, this is the p-value of the test for significant differences in yield between fields when the null hypothesis is set as "the average of sampled yields of all rice fields are equal." Based on this test, the null hypothesis was adopted, even at the 10% significance level. In other words, it cannot be said that there is a difference in yield between paddy fields, and it also cannot be said that changing the fertilizer amount will change the yield. Therefore, the field with the least amount of fertilizer is appropriate from the viewpoint of avoiding fertilizer costs, excessive fertilization, and environmental impacts. This result agrees with the TSC analysis results.

In local practice, the fertilizer applied to all fields is 4.2 gN/m^2. Assuming that this was changed to an appropriate 2.8 gN/m^2, the fertilizer was reduced to 87 and 91 kg, and the amount of leached nitrogen was reduced to 0.070mg/L for both Yuguchi-8 and Yuguchi. It was small compared to the water pollution standards. This is the case when fertilizer is applied during irrigation, and the leaching amount is only approximately 3% of that during surface drainage [26], [27][28]. Therefore, TSC method reduces the amount of excess fertilizer used, so it is expected to reduce the environmental impact. Furthermore, the development of such appropriate fertilization techniques will contribute to the sustainability of agriculture by creating soil that can produce food over the long-term without degrading the environment [29]. And continued analysis of this study will promote long-term soil health by reducing chemical fertilizers and increasing organic ones without reducing yields. This is a topic for future work.

As described above, the amount of excess fertilizer was estimated by using TSC for VI from UAV. In order to ensure accuracy and reliability, this research will need to address the following three things in the future. First is the use of data fusion technology. This is important to enhance the predictive power of the model and provide a more comprehensive understanding of factors influencing crop response to fertilization. Future research would like to try to analyze differences in soil and past crop management practices for each rice field.

5. Conclusions

In this research, the VIs time-series data were obtained using a UAV. Additionally, the monitoring dates were converted into growth stages (EAT) using a paddy rice growth model from related research [19], [20]. On this data, three analyzes were performed. In the first analysis, it was possible to visualize healthy vegetation growth and stress in paddy rice using the TSC. We also confirmed that healthy vegetation growth and stress of crops were improved by fertilizer. In the second analysis, the Kruskal–Wallis test confirmed the appropriate amount from the results of TSC. In the third analysis, the optimized period was determined based on the importance of each independent variable using RF. Based on these findings, our work have discovered the technologies can help reduce fertilizer costs, excessive fertilization and environmental impacts, and promote the use of UAV.

A future challenge is the development of a growth-prediction model. This study deals with ex-post evaluations, such as the visualization and classification of crop conditions. Also iIt is necessary to make future predictions for crop yields, early disease detection, and variable fertilization using VIs. Currently, there are no absolute standards for VIs, and no study has been found that generalizes the desirable values at each growth stage. However, a research has established a target NDVI for each growth stage using a portable device for NDVI measurement [30], [31]. This study collects and analyzes VI data via UAVs to improve productivity and reduce costs. For this purpose, we will develop a system that uses deep learning to provide decision support for fertilization in natural language based on the features of monitoring images [32] and methods such as farm income prediction based on yield prediction using regression, and so on. Then, it would be desirable to be able to connect UAVs with these systems and provide information to farmers in real time. As the IoT advances, security issues are a concern, so it will be important to develop autonomous decentralized systems that increase data tamper resistance using blockchain technology [33], [34], [35], [36].

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to thank the Research Center for Industrial Science and Technology, Iwate University, and the farmers in the experimental field for providing monitoring and yield survey data. We would also like to thank Editage (www.editage.jp) for English language editing. Finally, we are grateful to the referees for their helpful comments.

References

[1]. T. Ito, K. Minamino and S. Umeki, "Analysis of Vegetation Indices by Time Series Clustering of Drone Rice Monitoring Data," 3rd International Symposium on Electrical, Electronics and Information Engineering (ISEEIE 2023), 2023, DOI: 10.1049/icp.2023.1853.

[2]. Ministry of Agriculture, Forestry and Fisheries, "2020 nen nou-rin-gyou census kekka no gaiyou (kakuteichi)," URL: https://www.maff.go.jp/j/tokei/kekka_gaiyou/noucen/2020/index.html, [Accessed 5 February 2024], (article in Japanese).

[3]. K. Miyama, "Nou-sakumotsu · dozyou no bunkou-hansya-tokusei - nou-you-chi he no remote sensing gizyutsu no tekiyou-sei (I)," Journal of the Agricultural Engineering Society, Japan, **56**(12):1197-1202, 1988, (article in Japanese).

[4] J. L. Hatfield and J. H. Prueger, "Value of Using Different Vegetative Indices to Quantify Agricultural Crop Characteristics at Different Growth

Stages under Varying Management Practices," Remote Sensing, **2**(2):562-578, 2010, DOI: 10.3390/rs2020562.

[5]. J. Xue and B. Su, "Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications," Journal of Sensors, **2017**:1-17, 2017, DOI: 10.1155/2017/1353691

[6]. K. Tanaka and A. Kondoh, "Mapping of Rice Growth Using Low Altitude Remote Sensing by Multicopter," Journal of The Remote Sensing Society of Japan, **36**(4):373-387, 2016, DOI: 10.11440/rssj.39.S1.

[7]. K. Harashina, K. Yamamoto, M. Maki, Y. Muto and E. Kurashima, "Monitoring Growth of Water-seeded Rice in Tsunami-stricken Paddy Fields Using UAV-mounted Multispectral Sensor," Journal of the Japanese Society of Irrigation, Drainage and Rural Engineering, **87**(2):121-126, 2019, (article in Japanese with a title in English).

[8]. A. Matsuzaki, Inasaku dai hyakka dai 2 han dai 3 kan saibai no zissai/sehi-gizyutsu, Rural Culture Association, 2004, (article in Japanese).

[9]. Kubota Co., Ltd., "Mizu to hiryou ni yoru control | tanbo no kanri to higai-taisaku | okome ga dekirumade | Kubota no tanbo [manannde tanoshii! tanbo no sougou zyouhou site]," URL: https://www.kubota.co.jp/kubotatanbo/rice/management/manure.html, [Accessed 21 December 2023], (article in Japanese).

[10]. National Institute of Science and Technology Policy, "NISTEP PDF cover 200612," URL: https://nistep.repo.nii.ac.jp/record/5536/files/NISTEP-STT069.pdf, [Accessed 21 December 2023], (article in Japanese).

[11]. Ministry of the Environment, "Suishitsu-osen ni kakawaru kankyo-kizyun," URL: https://www.env.go.jp/kijun/mizu.html, [Accessed 7 February 2024], (article in Japanese).

[12]. T. Ito, K. Minamino and S. Umeki, "Analysis of Vegetation Indexes by Time Series Clustering of Drone Rice Monitoring Data," The 21st Forum on Information Technology (FIT2022), **21**(4):105-110, 2022, (article in Japanese with a title in English).

[13]. T. Ito, K. Minamino and S. Umeki, "Analysis of Fertilization Effects on Rice and Wheat by Time-Series Clustering of Vegetation Index Data," Proceedings of the 5th International Symposium on Advanced Technologies and Applications in the Internet of Things (ATAIT 2023), 2023.

[14]. Da-Jiang Innovations Science and Technology Co., Ltd., "AGRAS MG-1S Series - DJI," URL: https://www.dji.com/mg-1s, [Accessed 8 February 2024].

[15] .Da-Jiang Innovations Science and Technology Co., Ltd., "P4 MultispectralDJI," URL: https://www.dji.com/p4-multispectral, [Accessed 8 February 2024].

[16]. A. Karnieli, N. Agam, R. T. Pinker, M. Anderson, M. L. Imhoff, G. G. Gutman, N. Panov and A. Goldberg, "Use of NDVI and Land Surface Temperature for Drought Assessment: Merits and Limitations," Journal of Climate, **23**(3):618-633, 2010, DOI: 10.1175/2009JCLI2900.1

[17]. H. Zheng, W. Ji, W. Wang, J. Lu, D. Li, C. Guo, X. Yao, Y. Tian, W. Cao, Y. Zhu and T. Cheng, "Transferability of Models for Predicting Rice Grain Yield from Unmanned Aerial Vehicle (UAV) Multispectral Imagery across Years, Cultivars and Sensors," Drones, **6**(12:423):1-19, 2022, DOI: 10.3390/drones6120423.

[18]. L. Osborn, "NDVI vs. NDRE: What's the Difference? - Sentera," URL: https://sentera.com/resources/articles/ndvi-vs-ndre-whats-the-difference/, [Accessed 8 February 2024].

[19]. E. Kanda, Y. Torigoe and T. Kobayashi, "A Model to Estimate the Increase of Leaf Number on the Main Culm of the Rice Plant," Japanese Journal of Crop Science, **69**(4):540-546, 2000, DOI: 10.1626/jcs.69.540, (article in Japanese with an abstract in English).

[20]. E. Kanda, Y. Torigoe and T. Kobayashi, "A Simple Model to Predict the Developmental Stages of Rice Panicles Using the Effective Accumulative Temperature," Japanese Journal of Crop Science, **71**(3):394-402, 2002, DOI: 10.1626/jcs.71.394, (article in Japanese with an abstract in English).

[21]. A. K. Jain, M. N. Murty and P. J. Flynn, "Data Clustering: A Review," ACM Computing Surveys, **31**(3):264-323, 1999, DOI:10.1145/331499.331504.

[22]. R. Tavenard, "tslearn.clustering.TimeSeriesKMeans — tslearn 0.6.3 documentation,"URL:https://tslearn.readthedocs.io/en/stable/gen_modules/clus tering/tslearn.clu stering.TimeSeriesKMeans.html,

[23]. C. Shi, B. Wei, S. Wei, W. Wang, H. Liu and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," EURASIP Journal on Wireless Communications and Networking, **31**:1-16, 2021, DOI: 10.1186/s13638-021-01910-w.

[24]. S. Ronaghan, "The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark," URL: https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3, [Accessed 11 February 2024].

[25]. The IDB Project, "Index DataBase," URL: https://www.indexdatabase.de/, [Accessed 23 April 2024].

[26]. Toyama Prefecture, "Chika-sui kanyou no suishin ni mukete ," URL: https://www.pref.toyama.jp/documents/7739/00745883.pdf, [Accessed 17 February 2024], (article in Japanese).

[27]. C. Imagawa, "Nitrogen Dynamics Modeling to Support Land and Water Management in Rice Paddy Watershed," Soil/Water Research Group Materials, **30**, 2013, (article in Japanese with a title in English).

[28]. M. Tsuji and M. Takei, "Degrees of Water Purification in Cultivated Rice Fields and Fallow Fields," Research bulletin of the Aichi-ken Agricultural Research Center, **43**:1-6, 2011, (article in Japanese with an abstract in English).

[29]. M. M. Tahat, K. M. Alananbeh, Y. A. Othman and D. I. Leskovar, "Soil Health and Sustainable Agriculture," Sustainability, **12**(12), 2020, DOI: 10.3390/su12124859.

[30]. NIKON-TRIMBLE CO., LTD.,"GreenSeeker 2," URL:https://www.nikon-trimble.co.jp/products/product_detail.html?tid=428, [Accessed 15 February 2024].

[31]. Y. Kaneta, M. Nishida, F. Takakai and T. Sato, "New growth diagnosis standards of high-yielding rice and demonstration of high yielding using NDVI by GreenSeeker handheld crop sensor," Japanese Journal of Soil Science and Plant Nutrition, **91**(6):417-425, 2020, DOI: 10.20710/dojo.91.6_417, (article in Japanese with an abstract in English).

[32]. S. Ayoub, Y. Gulzar, F. A. Reegu and S. Turaev, "Generating Image Captions Using Bahdanau Attention Mechanism and Transfer Learning," Symmetry, **14**(12), 2022, DOI: 10.3390/sym14122681.

[33]. F. A. Reegu, H. Abas, Y. Gulzar, Q. Xin, A. A. Alwan, A. Jabbari, R. G. Sonkamble and R. A. Dziyauddin, "Blockchain-Based Framework for Interoperable Electronic Health Records for an Improved Healthcare System," Sustainability, **15**(8), 2023, DOI: 10.3390/su15086337.

[34] . A. A. Dar, M. Z. Alam, A. Ahmad, F. A. Reegu and S. A. Rahin, "Blockchain Framework for Secure COVID-19 Pandemic Data Handling and Protection," Computational Intelligence and Neuroscience, **2022**(7025485):1-11, 2022, DOI: 10.1155/2022/7025485.

[35]. M. Z. Alam, F. Reegu, A. A. Dar and W. A. Bhat, "Recent Privacy and Security Issues in Internet of Things Network Layer: A Systematic Review," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), 2022, DOI: 10.1109/ICSCDS53736.2022.9760927.

[36]. F. A. Reegu, W. A. Bhat, A. Ahmad and M. Z. Alam, "A review of importance of blockchain in IOT security," International Conference on Emerging Trends in Materials, Computing and Communication Technologies (ICETMCCT 2021), **2587**(1), 2023, DOI: 10.1063/5.0150432.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 4, 35-50 (2024)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Early Detection of SMPS Electromagnetic Interference Failures Using Fuzzy Multi-Task Functional Fusion Prediction

Declan Mallamo*, Michael Azarian, Michael Pecht

Center for Advanced Life Cycle Engineering, University of Maryland, College Park, MD 20742, USA

ARTICLE INFO

Article history:
Received: 01 May, 2024
Revised: 04 July, 2024
Accepted: 05 July, 2024
Online: 26 July, 2024

Keywords:
Prognostics Health Management
Functional Data Analysis
Electromagnetic Interference
Multitask Lasso Regression
Aluminum Electrolytic Capacitors

ABSTRACT

This study addresses the need for improved prognostics in switch-mode power supplies (SMPS) that incorporate electromagnetic interference (EMI) filters, with a focus on aluminum electrolytic capacitors, which are critical for the reliability of these systems. The primary aim is to develop a robust model-based approach that can accurately predict the degradation and operational lifetime of these capacitors under varying environmental conditions. To achieve this, the research employs a generalized state space averaging technique to simulate a population of impending degradation trajectories for the capacitors. Environmental and degradation effects are modeled comprehensively. Frequency-based test features are derived from the gain, control, and impedance transfer functions of the filter and SMPS. These features are fitted with b-spline functionals for resampling and subsequently analyzed using functional principal component analysis to project the data onto the principal modes of variation. The extracted features serve as inputs to a fuzzy multi-task functional fusion predictor, which estimates the state of health at critical frequencies. The effectiveness of this model-based approach is validated through extensive experimentation, demonstrating its potential to significantly enhance the predictive maintenance strategies for SMPS with EMI filters.

1. Introduction

Switch-mode power supplies (SMPS) are an integral part of modern electronic systems, known for their efficiency and providing stable power supply across a variety of output levels using high-speed switching of transistors to regulate voltage and current. This switching action introduces electromagnetic interference (EMI) that is becoming more significant with high frequency operation [1]. This has created the need for input EMI filters to protect both the power line and the switch mode power supply. This EMI arises conductive, radiated, and near-field coupling of the created by resonant peaks in the frequency response, as well as switching transients caused by the closed-loop duty cycle corrections. In Figure 1, an LC-type filter is shown in-series with a buck-boost SMPS.

EMI-filtered closed loop SMPS with constant power and controlled output voltage or current are susceptible to a condition called negative input resistance, where increases in input voltage lead to decreases in input current. This creates oscillations that can destabilize the system, dampened out by the input filter capacitors' equivalent series resistance (ESR). Aging or variability in electrical parameters can cause a cross-over of the SMPS input impedance and EMI-filter output impedance suddenly under certain operating conditions [2]. It has been found that many SMPS devices currently being utilized in industry, due to their nonlinear behavior, would fail to meet EMC standards without extensive up-stream filtering [3]. The Middlebrook criteria dictate that for a cascaded system, like that of an EMI-filtered SMPS, stability is maintained if the output of an upstream subsystem is less than the input impedance of the downstream subsystem, preventing oscillations and preserving electrical integrity. Typically, this cross-over occurs near the resonance frequency of the LC components of the filter, where the output impedance peaks, but at higher frequencies, with increased effects of the higher dv/dt and di/dt values, predicting the cross-over frequency range that leads to increased EMI using model-based becomes intractable.

The primary research objectives are to develop and validate parametric models for SMPS that integrate harmonic analysis and consider the degradation of both input and output filter capacitors. These models aim to precisely simulate the effects of aging and environmental variation on capacitor behavior, thereby enhancing

*Corresponding Author: Declan Mallamo, University of Maryland, +1(631)871.8725, dmallamo@umd.edu

https://dx.doi.org/10.25046/aj090405

our understanding of how such factors influence degradation. Utilizing this data, the study seeks to improve prognostic capabilities for SMPS by predicting potential failure points and frequency vulnerabilities. The effectiveness of these prognostic models will be assessed through a comparative analysis with existing predictive maintenance strategies. This analysis aims to demonstrate early detection and preemptive management of failures by identifying frequency ranges with the largest variation tied to the system degradation.

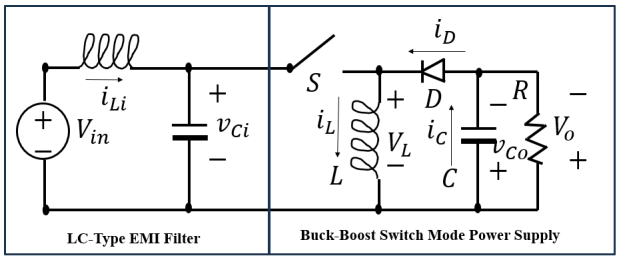


Figure 1: Buck-boost switch mode power supply with a general low pass LC-input filter.

1.1. Literature Review

1.1.1. Methods of Assessing and Modeling Degradation and Reliability of EMI filtered SMPS

There has been a large amount of research into the aging effects on SMPS components and their effects on the overall frequency response of the converter. It was found that the degradation trajectory of the rise and fall time of a MOSFET switch followed random variation as a result of induced thermal aging [1, 2]. The main sources of electromagnetic emissions were the power MOSFETs, and leakage inductance from the main transformer [3, 4]. The output diodes and output inductor can also be considered as emission sources; however, they provide a more secondary contribution. Sensitivities created from passives parasitics in the printed circuit board and sub-components that require targeted identification to systematically mitigate potential causes of noise.

Several studies have found the output filter capacitor was identified as a main source of failure in SMPS that causes the increased noise and critically impacts the performance of the converter, resulting in increased stress on the peripheral components [5], [6], [7], [8].

The studies have been conducted that provide in-depth insights into the degradation mechanisms of aluminum electrolytic capacitors under varying conditions, particularly highlighting their impact on the reliability of SMPS. A diagram of a typical Aluminum electrolytic capacitor can be found in Figure 2.

In [8], the author focused on the effects of thermal overstress noting that a reduction in electrolyte volume from evaporation directly decreases capacitance and increases ESR due to a shorter liquid path length. It showed that thermal overstress from storage beyond room temperature conditions significantly compromised their longevity and performance.

In [9], the author explained that models were created to address that over half of SMPS failures are attributed to output smoothing electrolytic capacitors and proposes new models to incorporate temperature variations to forecast degradation, influenced by time, core temperature, and operation frequency.

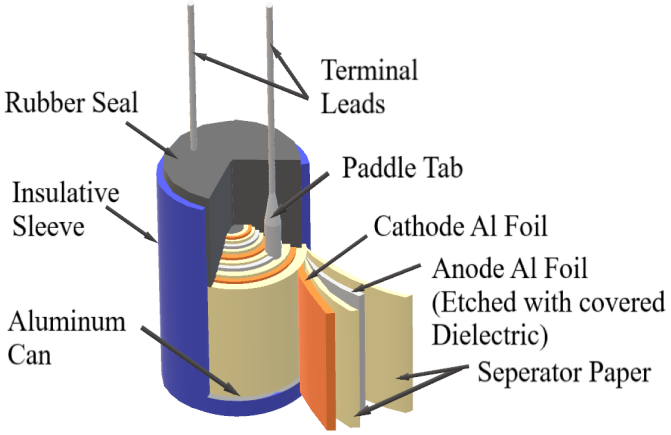


Figure 2: Exploded view of an Aluminum electrolytic capacitor, showing the different layers of the wound jellyroll which is immersed in the electrolytic solution.

SMPS with EMI filters are increasingly analyzed with advanced mathematical modeling techniques and system optimization methods are often employed to understand system behavior comprehensively, leading to more effective design decisions. These traditional approaches can be extended from a holistic system perspective to optimize the overall system architecture [10]. These methods often require extensive modeling of each system component without necessarily focusing on reducing the experimental or simulation effort to only the most influential factors affecting system stability and performance.

In [11], the author proposed a genetic algorithm-based method for designing front-end rectifier inductors, taking into account the effects of DC link capacitors. However, the computational intensity of genetic algorithms poses limitations. These algorithms require significant computational resources and time, particularly as system complexity increases, which can limit their practicality in iterative design processes.

In continuous conduction mode, the switch remains closed long enough for the inductor current to never fall to zero, making the system's behavior more predictable and easier to model using linear approximations [12]. In contrast, discontinuous conduction mode is characterized by the inductor current reaching zero within each switching cycle, which introduces non-linearities that are critical for understanding the SMPS performance under light load conditions [12], [13]. The integration of these modes into a unified model involves crafting state space analyses that capture the essential dynamics of the SMPS, including its susceptibility to noise and interference as influenced by the duty cycle variations.

These state space models elucidate the intricate output to input relationships within an EMI-filtered SMPS device, highlighting how variations in the duty cycle affect the overall stability and efficiency of the power supply. The transfer functions derived from these models are instrumental in predicting the system's behavior in response to external disturbances and internal parameter changes.

To further refine these models and achieve a more generalized representation, generalized state space averaging (GSSA) methods are employed. GSSA is a technique that averages the various state equations over one switching period to produce a smoothed model, effectively managing the rapid switching characteristics of SMPS. This method simplifies the complex dynamics of switching power supplies into manageable forms, allowing for the integration of non-linear and dynamic aspects of the system into a coherent framework. By averaging the states, GSSA reduces the computational complexity and enhances the model's ability to predict long-term behavior under a broad range of operating conditions.

1.1.2. Time and Frequency Stability Analysis Methods

Non-invasive online condition monitoring techniques for capacitors often utilize the spectral content of voltage and current, employing various approaches estimating ESR and capacitance with minimal error. While comprehensive, this analysis requires sweeping through a broad frequency range, which can be time-consuming and may not pinpoint the most critical frequencies causing instability in complex systems.

Additionally, AC analysis methods provide a robust framework for assessing the global stability of closed-loop systems by exploring the system's behavior across a broad frequency spectrum. This analysis is critical as it helps identify resonance peaks, phase shifts, and gain changes, offering valuable insights into how the system responds to different frequency inputs and highlighting potential instabilities. Understanding gain and phase margins is pivotal in ensuring global system stability. Gain margin refers to the amount by which the gain of a system can increase before the system becomes unstable, while phase margin is the amount of additional phase lag required to bring the system to the brink of instability.

In [14], the author used frequency and time-domain methods to assess conducted emissions, focusing on the signal characteristics of SMPS emissions, which include a mix of medium- and high-frequency components and significant spectrum leakage. The study found that there is a need to tune and optimize these analytical methods to achieve reliable results. The Prony method tracks changes in amplitude at specific frequencies using the least squares approach, but this method requires a complete understanding of the frequencies associated with degradation for a global population of units. It has been shown that by understanding the specific EMI profile to be mitigated, selective trade-offs can be made to reduce the filter footprint, weight, and cost while maintaining performance [3].

1.1.3. Prognostic and Health Management Techniques

Accurately anticipating and addressing these shifts requires analyzing the impact of component degradation on the system's performance. The prognostic and health management (PHM) system optimizes maintenance processes based on diagnostic and prognostic outcomes to prevent failures and enhance lifecycle management [8][15]. For SMPS, capacitance was identified as a suitable failure precursor for system failure [8].

Manufacturers often recommend a methodical approach for estimating the useful life of aluminum electrolytic capacitors using datasheet specifications. The rated ripple current, $I_{AC,\,R}$ at the capacitor's maximum specified temperature is identified and the actual operating ripple current I_{AC} is used to calculate their quotient. This ratio and the ambient temperature are used to estimate the capacitor's remaining life by interpolating a given life expectancy graph and accounting for frequency variations from the standard test condition frequency, usually 100 Hz. [16].

A data-driven fault detection algorithm was introduced specifically designed for identifying failures in multilayer ceramic capacitors [17]. The algorithm utilizes regression analysis, residual detection, and prediction analysis to enhance the accuracy and reliability of fault detection. A key component of their methodology is the use of Mahalanobis distance for anomaly detection in the test data.

In [18], the author used accelerated life testing for aluminum electrolytic capacitors to evaluate how conditions such as electrolyte leakage can affect capacitance, quality factor, and dissipation factor. The study utilized statistical time-domain feature extraction and correlation-based feature selection to accurately monitor capacitor health and predict failures.

In [19], the author proposed a method that leverages noninvasive condition monitoring via time-frequency analysis of conducted EMI to evaluate the health of DC-link capacitors in three-phase inverters. This method involves a combined EMI filter and measurement board placed on the DC bus, which not only filters conducted EMI to comply with MIL-STD-461 G but also facilitates EMI measurements for condition monitoring. A continuous wavelet transform is used to create characteristic switching images, which are then used to train support vector machine (SVM) models to classify the health of DC-link capacitors into one of five stages with high accuracy. This approach uses broad-spectrum analysis, which may include frequencies that are not always relevant to condition monitoring.

A PHM system was presented that is designed to preempt failures and enhance lifecycle management for insulated gate bipolar transistors [15]. This process is organized into three main stages: Observation, Analysis, and Action. The Observation stage involves monitoring and data processing, the Analysis stage includes health evaluation and future state forecasting, and the Action stage focuses on maintenance implementation based on assessments. To overcome challenges in detecting subtle degradation signals, Principal Component Analysis (PCA) is used for feature engineering to reveal hidden trends. These trends are then fed into a Deep Neural Network for classification, enhancing the system's ability to detect and predict failures accurately.

A PHM framework was proposed to combine traditional model-based and data-driven approaches, utilizing extensive sensor data for remaining useful life predictions [11]. This approach is designed to analyze subtle time-series patterns in large datasets by treating sensor data as continuous random processes. These functional relationships can encapsulate a significant amount of variation information across different equipment in a compact, resamplable form. This capability to adapt to time-

varying data makes the approach particularly useful for the EMI-filtered SMPS usage case.

Using multivariate functional relationships as predictors for state of health can alleviate big data concerns and can be further improved by sparsity-induced optimization methods, which learn multiple classification tasks while simultaneously performing feature selection. A method of multi-task feature learning, used for analyzing brain imaging data with varied functional data sets collectively, was developed to enhance predictability and accuracy [20]. This approach can be adapted to other domains, including the health monitoring of EMI-filtered SMPS, to improve the accuracy and efficiency of remaining useful life predictions.

1.1.4. Identifying and Addressing Research Gaps

Despite extensive research on degradation and failure mechanisms in SMPS, gaps remain in predicting system instabilities caused by frequency vulnerabilities. Traditional studies, which focus on component aging and electromagnetic emissions, often employ complex models that overlook critical frequency regions impacting system stability. Moreover, these models lack the interpretability necessary to identify how specific component degradations influence overall performance in frequency-sensitive scenarios.

To address these deficiencies, we propose an approach that utilizes multivariate functional analysis and multitask least absolute shrinkage and selection operator (LASSO) regression. This method transforms complex, high-dimensional datasets into manageable, infinite-dimensional functionals conducive to resampling. It significantly enhances model accuracy and robustness, allows for the learning of multiple classification problems simultaneously, and focuses specifically on identifying critical frequency vulnerabilities in SMPS.

By isolating specific EMI components and narrowing the analysis to essential frequencies, this approach not only reduces data complexity but also improves measurement precision and system stability. This targeted analysis streamlines research, boosts efficiency, and provides clearer insights into EMI behaviors, substantially enhancing system vulnerability assessments and predictive maintenance strategies.

*1.2. Fuzzy Multi-Task Functional Fusion Predictor*s

This study aims to develop and validate an innovative prognostic framework that leverages discrete event simulation (DES) of the EMI-filtered closed-loop SMPS and degradation models of the EMI-filter input filter capacitor and SMPS output filter capacitor. The goal is to accurately predict the degradation and remaining useful life of aluminum electrolytic capacitors in electromagnetic interference filters of switch-mode power supplies, thereby significantly enhancing system reliability and performance.

The developed SMPS system prognostic approach, termed Fuzzy Multi-Task Functional Fusion Predictor (FMT-FFP), is an advanced predictive model that combines fuzzy logic with multitask LASSO regression and B-spline convolutional-integral-based cross-correlations, integrated through functional PCA for robust and precise state-of-health forecasting in complex systems, while preserving interpretability of the input features. For the given case study, the developed approach focuses on identifying the impact of critical frequency regions associated with the degradation trajectories of aluminum electrolytic filter capacitors within an EMI-filtered SMPS system.

Generalized state-space averaging models of the LC-filtered buck-boost SMPS are developed early on in the prototyping design phase to express the k-th order harmonic content and derive the output voltage to input voltage gain frequency response, the frequency response for the output voltage to duty cycle control transfer function, and the frequency response for the EMI filter output and SMPS input impedance.

Once frequency responses for each test are collected, they are extended using convolutional integrals for each combination. Functional Principal Component Analysis (FPCA) is then used to perform dimensionality reduction and feature engineering to restrict the functional data to the principal variational modes. Multitask LASSO regression is employed to make probabilistic State of Health (SoH) estimations while identifying a sparse set of features associated with frequency-response criteria and ranges. An outline of the prognostic approach can be seen in Figure 3 below.

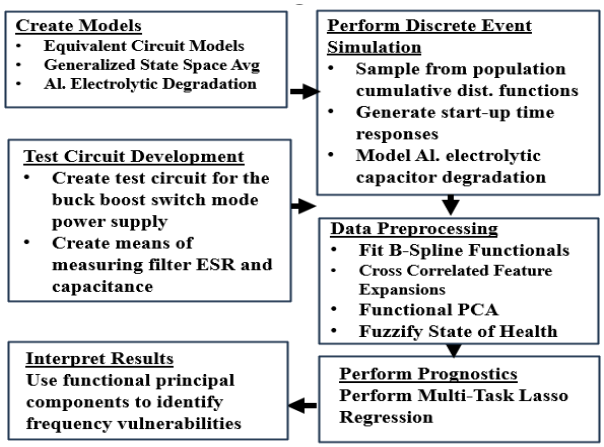


Figure 3: Methodology to create the interpretable failure prognostics for the GSSA model of the buck-boost SMPS.

This study introduces a refined modeling approach tailored to the unique operational characteristics of SMPS with EMI filters. The key contributions of our research are threefold. First, a generalized state-space averaging model was developed to effectively capture the dynamic interactions within the SMPS and the input EMI filter, facilitating a deeper understanding of system behavior across varying operational states and filter capacitor degradation. This model is particularly adept at identifying critical frequency regions that directly impact the stability and efficiency of SMPS. Finally, the application of these advanced modeling techniques has led to the establishment of an efficient testing scheme. This scheme enhances traditional maintenance strategies by providing precise diagnostic capabilities that can preemptively address potential failures, thereby significantly improving system reliability and operational lifespan. Through these contributions, our research offers substantial advancements in the predictive

maintenance and reliability assessment of EMI-filtered SMPS, ensuring more stable and efficient power supply systems.

The remainder of the report outlines the prognostic approach. Section 2 lays the theoretical groundwork, detailing the derivation of the generalized state-space model for the buck-boost SMPS, the discrete event simulation methodology used to create the training data, and how the probabilistic SoH estimations are converted into remaining useful life estimates. This section also elucidates the technique for identifying frequency vulnerabilities using the results of the Multitask LASSO Regression. Section 3 presents a buck-boost SMPS test fixture, featuring a microcontroller with proportional integral derivative (PID) control and voltage-based current sensors for the input filter and SMPS inductor currents. These sensors use a high-impedance, time-constant matching network for accurate current measurement and are used to collect impedance information. Section 4 presents the research analysis of the results and discusses any wider implications. Finally, in Section 5, a summary of the research findings and conclusions is provided.

2. Theory

In this research, the intricacies of EMI filter and SMPS In this research, the intricacies of EMI filter and SMPS design are addressed utilizing the generalized state-space averaging (GSSA) model. This modeling approach decomposes the state signals using Fourier analysis to account for the harmonic content evolution related to the effects of the switching frequency harmonics and the degradation effects in aluminum electrolytic capacitors within the SMPS and input filters. The analysis employs a buck-boost DC-DC SMPS topology with a general input EMI LC-type low-pass filter, as illustrated in Figure 1 in Section 1.

Parasitic resistances and losses, including the input filter inductor's parasitic resistance, the equivalent series resistance (ESR) of input and output capacitors, and the copper losses in the SMPS inductor, are assumed. Additionally, inputs associated with the input voltage, forward voltage drops across the switching transistor and diode, and load current perturbations are considered.

This comprehensive approach allows for a detailed understanding of how various parasitic elements and operational parameters influence the performance and stability of the SMPS, thereby providing a robust foundation for predictive maintenance and reliability assessment.

2.1. State Space Averaging

The development of a linearized model for a SMPS with an EMI filter employs Kirchhoff's Voltage Law across circuit loops, focusing on continuous conduction mode (CCM) operation. In CCM, the system oscillates between states of active switching transistor and conducting diode. By averaging these states, the model smooths out non-linear transitions due to the switch's high slew rate, simplifying the analysis. By performing the nodal analysis and solving for the state time derivatives, state space representation of the EMI filtered SMPS can be found below in (1).

$$\begin{aligned} \dot{x} &= A \cdot x + B \cdot u \\ y &= C \cdot x + D \cdot u \end{aligned} \tag{1}$$

$$x = \begin{bmatrix} i_{Li} \\ i_L \\ v_{Ci} \\ v_{Co} \end{bmatrix}, u = \begin{bmatrix} v_{in} \\ v_M \\ v_D \\ i_O \end{bmatrix}, y = \begin{bmatrix} v_o \\ i_{Li} \\ i_L \\ v_{Ci} \\ v_{Co} \end{bmatrix}$$

where *x, u, y* are the state, input, and output vectors, respectively, consisting of the input and SMPS inductor current, v_{Ci} and v_{Ci}, and input and output filter capacitor voltages, v_{Ci} and v_{Co}, and input voltage, v_{in}, switch and diode forward bias voltages, v_M and v_D, and load current perturbations, i_O.

2.2. Steady State

The steady state behavior is derived by setting the state derivative to zero and solving for the steady state response as shown in (2).

$$\dot{x} = 0 \rightarrow X = -A^{-1} \cdot B \cdot U \tag{2}$$

2.3. Linearization

Linearization is used to create a small-signal model that accounts for duty cycle perturbations and their interactions with state variables:

$$x = \tilde{x} + X \qquad u = \tilde{u} + U$$

2.4. Proportional integral derivative Control Laws

The proportional integral derivative (PID) controller logic for the output voltage tracking objective can be below in (3).

$$\tilde{\delta} = K_p\left(\tilde{v}_{ref} - \tilde{v}_o\right) + K_i \int \left(\tilde{v}_{ref} - \tilde{v}_o\right)dt + K_d \frac{d}{dt}\left(\tilde{v}_{ref} - \tilde{v}_o\right) \tag{3}$$

The PID controller is formulated to manage output voltage deviations as follows in (4).

$$\frac{d}{dt}\tilde{\delta} = K_p \frac{d}{dt}\tilde{v}_O + K_i\left(\tilde{v}_{ref} - \tilde{v}_O\right) - K_d \frac{d}{dt}\frac{d}{dt}\tilde{v}_O \tag{4}$$

where $K_P, K_I,$ and K_D are the respective proportional, integral, and derivative gains, the voltage reference input, $\tilde{v}_{ref}$, and the output voltage, $\tilde{v}_O$. This differential form leverages linear relationships of state and input vectors to adjust the duty cycle dynamically.

2.5. Generalized State Space Averaging

Generalized state space averaging is applied to express the state as a sum of sinusoidal functions over one switching period, enhancing the model's ability to capture the dynamic interactions within the SMPS and input EMI filter:

$$x(t) = \int_{t_0}^{t_0+T} A \cdot x(t) + B \cdot u(t)\, dt + x(t_0)$$

$$= \sum_{k=-n}^{n} < x >_k (t)\, e^{i\omega k t}$$

where $\omega = \frac{2\pi}{T}$ is the angular frequency in terms of switching period, T. From the nodal analysis of the LC filtered buck boost circuit depicted in Figure 1, the original state space had four-states associated with each inductor current and capacitor voltage.

$$x_i(t) = x_i + 2 * Cos(\omega t)x_{i+4} - 2 * Sin(\omega t)x_{i+5} + 2 * Cos(2\omega t)x_{i+12} - 2 * Sin(2\omega t)x_{i+13}$$

The x-coefficients are found from the real and imaginary component for each Fourier coefficient, $<x>_k$, represents the amplitude of the k-th harmonic frequency component.

$$< x >_k = \int_{t-T}^{t} x(t)e^{-i\omega kt}\, dt$$

Using product rule and the chain rule, the expression for the time derivative of the Fourier coefficient is seen below in (5).

$$\frac{d}{dt} < x >_k = -j\omega k\tau < x >_k (t) + < \frac{d}{dt}x >_k (t) \quad (5)$$

The k-th order convolution coefficient is found as the sum of the product of the Fourier coefficients of the two signals noting the ordering of the average subscript combinations.

$$< x \cdot y >_k = \sum_{k=-n}^{n} < x >_i (t) * < y >_{k-i} (t)$$

The negative k-th average harmonic relates the complex conjugate of the signal, seen below in (6).

$$< x >_{-k} = < x >_k^* \quad (6)$$

2.6. Extracting Transfer Function Information

The use of analyzing AC frequency response features has long been used in insuring global stability of closed loop dynamic systems [2]. From linear systems theory, the complete family of output-to-input transfer function relationships can be found for a linear time-invariant multi-input multi-output dynamic system in state space representation can be found below in (7).

$$G(s) = C \cdot (s * I - A)^{-1} \cdot B \quad (7)$$

The output voltage to input gain, $G_g = v_o/v_{in}$ and control, $G_c = v_o/dc$ are used to provide information concerning the different gain and phase margins which are measures of system's stability. The EMI filter output impedance, $Z_{f,out} = v_{f,out}/i_{f,out}$, and SMPS input impedance, $Z_{in} = v_{in}/i_{in} \approx v_{in}/i_L * \mathrm{DC}$, are also used to evaluate the system's dynamic stability relying on the Middlebrook criterion as a method to assess compatibility between the SMPS and the LC-input filter, taking into account the effects of negative incremental impedance on the constant power controller voltage buck-boost SMPS [2].

2.7. Aluminum Electrolytic Degradation Models

The longevity of aluminum electrolytic capacitors is chiefly compromised by electrolyte evaporation, a consequence of elevated operating temperatures and heat from ripple currents [8][9]. A thermal model, as shown in Figure 4, simplifies the system by considering the hotspot (T_{HS}) and case temperature (T_C) to be approximately equal. This model integrates ripple current, capacitor ESR, and case-to-ambient thermal resistance, offering a streamlined approach to evaluating capacitor thermal behavior.

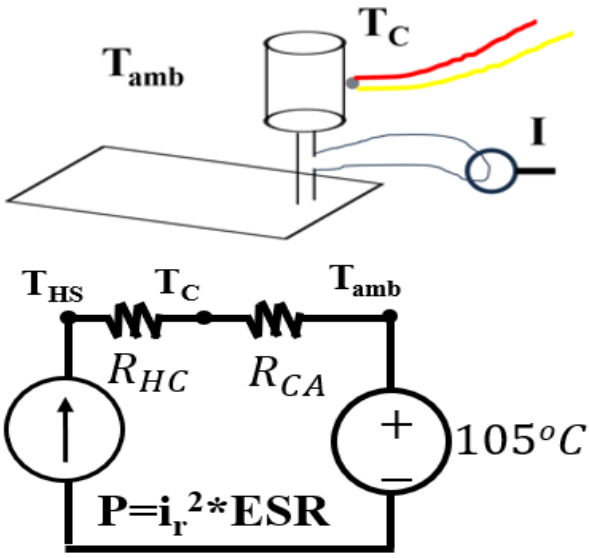


Figure 4: Testing configuration to measure case-to-ambient thermal resistance assuming a negligible difference between case temperature, T_C vs hotspot temperature T_{HS}.

$$R_{CA} = \frac{T_C - T_{amb}}{I^2 * ESR}$$

$$T_{max} = R_{CA} * I_{max}^2 * ESR + T_{amb}$$

Models have been developed to account for the environmental and temporal degradation-based effects on an aluminum liquid electrolytic capacitor's capacitance and ESR [9]. These models are contained in (8) and (9) below.

$$Cap(T,t) = Cap_0 * e^{\frac{B_0}{T(t)}} * \left(1 - \int_0^t L[T(t)]dt\right) \quad (8)$$

$$ESR(T,t) = \frac{ESR_0}{1 + D_0 * T(t)} * e^{\int_0^t M[T(t)]dt} \quad (9)$$

where $L[T(t)] = A_0 * e^{\frac{E_{a1}}{\kappa * T(t)}}$ and $M[T(t)] = C_0 * e^{\frac{E_{a2}}{\kappa * T(t)}}$, are the different Arrhenius models that relate the temporal degradation rate to temperature for the device in terms of dimensionless parameters A_0 and C_0 and activation energies Ea_1 and Ea_2, are found using experimentation [9].

2.8. Discrete Event Simulation

A discrete event simulation (DES) was performed to model the degradation trajectories of aluminum electrolytic capacitors' capacitance and ESR. The simulation initialized circuit attributes and parasitic elements with a 10% variation to mimic real-world deviations. Throughout the simulation, the duty cycle was dynamically adjusted based on PID control logic, targeting a 15V output from a 12V input overlaid with 10% white noise to represent input perturbations. An overview of the DES is depicted in the directed graph in Figure 5.

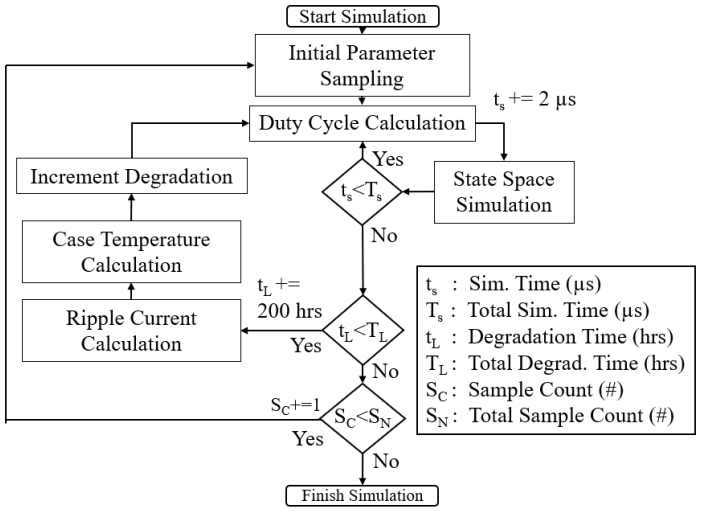


Figure 5: Discrete event simulation routine used to simulate the small time scale responses and the large time scale capacitor degradation responses.

The lifecycle of three sample SMPS was simulated by analyzing startup transient responses in 2-microsecond increments over a span of 600 increments, achieving a steady state. At this steady state, the ripple current through each capacitor was calculated for every time response. This data was then applied to incrementally update the temporal degradation. The degradation cycle was advanced every 200 hours, continuing up to a maximum of 3000 hours or until a failure condition occurred. Failure was defined as an event where the output experienced an overshoot or ripple voltage exceeding 2 volts.

2.9. State of Health Prediction Methods

The following sections cover the derivation of the different prognostic approaches. The three methods focus on dimensionality reduction routines and extracting tacit information from time-series data.

2.9.1. Method 1: (PCA-DNN)

The first method involves reducing the high dimensional data using principal component analysis for feature engineering and a deep neural network for classification of the state of health (SoH) into probabilistic fuzzy estimations.

2.9.1.1. Feature Engineering with Principal Component Analysis

The features from the AC analysis create a high dimensional data set that captures the underlying behavior but are too complex for efficient analysis. Principal component analysis (PCA) is a statistical technique that reduces the dimension of the data while preserving most of the variation by identifying correlations within the data using a covariance matrix, seen below.

$$\Sigma = \frac{1}{n-1} XX^T \tag{10}$$

Eigen analysis is used to create synthetic variables that are linear combinations of the original features.

$$\Sigma \mathrm{v} = \lambda \mathrm{v} \tag{11}$$

These variables can be truncated to only include the most significant details of the variation in the original data.

2.9.1.2. Neural Networks

The reduced-dimension PCA output data serves as an input to the first layer of a deep neural network (DNN). A neural network is composed of layers of neurons that take a linear combination of inputs, x, assigns a respective weight to each input, w, and a bias, b, and applies an activation function, $f(\cdot)$.

$$y = f(\Sigma_{i=1}^{n} w_i x_i + b) \tag{12}$$

A common activation function used for the hidden neurons is the ReLu function which adds non-linearities into the model to help learn complex patterns.

$$ReLU(x) = \max(0, x) \qquad (Hidden\ layer)$$

To create the binned state-of-health probabilities in the output, a SoftMax function is used which normalized exponentials to create a probability distribution.

$$Softmax(x_i) = \frac{e^{x_i}}{\Sigma_{j=1}^{n} e^{z_i}} \ (Output) \tag{13}$$

The PCA/DNN pipeline for the AC analysis feature data can be seen in Figure 6.

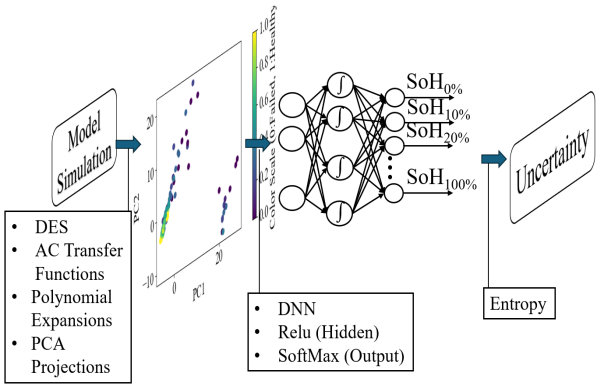


Figure 6: Diagram of the PCA-DNN framework, where high dimensional frequency response data is projected using the truncated discrete principal components and used as inputs to a deep neural network.

2.9.2. Method 2: CWT-SVM

Continuous Wavelet Transforms (CWTs) are useful in analyzing time signals that exhibit non-stationary behavior. Such as state of health classification [19]. The time and frequency analysis capability of the CWTs make them ideal at analyzing state-of health effects from harmonics associated with the interactions between the output of the LC-type input EMI-filter and the input of the closed-loop SMPS.

2.9.2.1. Continuous Wavelet Transform

The CWT formula used in this report can be found below in (14).

$$CWT(s,\tau) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{\infty} x(t)\psi\left(\frac{t-\tau}{s}\right)dt \tag{14}$$

where, $\psi(t)$ is the wavelet function, s adjusts the wavelet's width and τ translates the wavelet thereby affecting time resolution.

2.9.2.2. Support Vector Machine

The CWT is utilized to analyze the time-frequency characteristics of the system's output voltage. This analysis generates a high-dimensional image array, capturing intricate time-scale variations within the signal. These image arrays serve as the input for a Support Vector Machine (SVM), a powerful machine learning technique used for classification tasks. SVM operates by identifying optimal boundary support vectors among the data points that represent different classes. It constructs a hyperplane that maximizes the margin, which is the distance between the nearest data point of each class and the hyperplane itself. This maximization is crucial as it contributes to the robustness of the classification against new data. The objective function of SVM is formulated to minimize:

$$Objective = \frac{1}{2}\left|\left|w_d\right|\right|^2 \quad (15)$$

where w_d is the normal to the hyperplane. SVM uses constraints to ensure that all data points correctly classify by maintaining a distance from the hyperplane:

$$y_i(w^T x_i + b) \geq 1, \ \forall\, i \quad (16)$$

where x_i are the CWT feature vectors, y_i are the respective class labels ranging from +1 to -1 and $w^T x_i + b$ is a linear decision function.

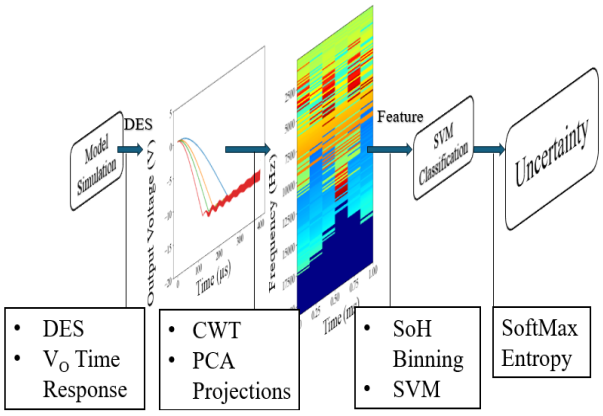


Figure 7: Diagram of the CWT-SVM framework, where high dimensional time-frequency response data is projected using PCA and then used as input to a support vector machine algorithm used for classification of the state of health.

To assist in creating separability in the data, a radial basis function is used, that measures the similarity between the data points.

$$K(x_i, x_j) = e^{-\gamma\left|\left|x_i - x_j\right|\right|^2} \ \textit{(Radial Basis Function)}$$

Using the kernel trick changes the constraints to the following form:

$$y_i\left(\sum_{j=1}^{n} \alpha_j y_j K(x_i, x_j) + b\right) \geq 1, \ \forall\, i \quad (17)$$

where α_j are Lagrange multipliers, which are optimized during training.

A diagram of the CWT-SVM methodology can be found in Figure 7, illustrating how time series data from the output voltage is processed into CWT image arrays. These arrays are then used as inputs to the kernel SVM, which classifies the system's state based on learned patterns from the training phase.

2.9.3. Method 3: Fuzzy Multi-Task Functional Fusion Predictors

The fuzzy multi-task fusion predictor uses b-spline functional curves to form versatile low dimensional representations of the AC frequency features created from the transfer function representations. The method creates fusion between the signals by extending the features using convolution integral-based cross-correlations and multi-task learning with LASSO regression to identify a sparse subset of test conditions conducive to increasing the performance of the prognostic.

2.9.3.1. B-Spline Resampling

B-spline interpolation resamples time series data to a standard temporal scale, utilizing piecewise polynomials defined over specific intervals as can be found below in (18).

$$f(t) = \sum_{i=0}^{n} P_i B_{i,k}(t)$$

$$B_{i,1}(t) = \begin{cases} 1, & if, t_i \leq t \leq t_{i+1} \\ 0, & \text{otherwise} \end{cases}$$

$$B_{i,k}(t) = \frac{t - t_i}{t_{i+k-1} - t_i} B_{i,k-1}(t) + \frac{t_{i+k} - t}{t_{i+k} - t_{i+1}} B_{i+1,k-1}(t) \quad (18)$$

2.9.3.2. Extending the Features using Derivatives

Derivatives of the gas path parameters are approximated using the forward finite difference method to enrich the feature set, enabling more detailed analyses.

$$f'(x) = \frac{f(x+h) - f(x)}{h} \quad (19)$$

2.9.3.3. Generating Cross-Correlated Features Expansions

The individual AC features can be extended using polynomial expansion of the discrete AC waveforms and their derivatives using the expression below in (20) used to calculate the convolution integrals between two signals:

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x) \cdot g(x - \delta)\, d\delta \quad (20)$$

where f and g represent different AC transfer functions, δ represents a shift in the frequency domain, and $x = 2\pi j f_s$, and f_s is the signal frequency.

2.9.3.4. Functional Principal Component Analysis

It is possible to extract functional principal components from the functional b-spline representations of the extended states, paralleling how PCA extracts the primary variational modes from discrete multivariate cross-sectional data. This is achieved by

using the inner product to generalize the multivariate principal component eigen analysis problem.

$$\int C(s,t)\phi_k(t)dt = \lambda_i\phi_k(s) \quad (21)$$

where $\phi_i(t)$ are principle components selected from a selected from infinite set, λ_i are the sorted eigenvalues and $C(s,t)$ serves as an empirical covariance function given in the following form below.

$$C(s,t) = \frac{1}{n}\sum_{i=1}^{n}\big(X_i(s)-\mu(s)\big)\big(X_i(t)-\mu(t)\big) \quad (22)$$

where $X_i(s)$ consists of a vector of observed functional data (i.e., curves and spectra), $\mu(s)$ is a mean function estimated from the data, and s and t are different points across the sample curves.

2.9.3.5. Functional Data Projections

The i[th]-observation within the data can be projected onto the most significant functional principal components, to summarize the principal variational aspects of the functional signal while preserving essential patterns and trends:

$$\hat{X}_i(s) = \mu(s) + \sum_{k=1}^{K}\int\big(X_i(s)-\mu(s)\big)\phi_k(s)ds\,\phi_k(s) \quad (23)$$

2.9.3.6. Creating probabilistic State of Health Estimations

State of health estimations are represented probabilistically, converting crisp inputs into a binned format to facilitate uncertainty analysis and enhance model explainability. The crisp inputs were converted to 10-bins that make up the sequential values between 0 and 1.

$$Bin(x) = \lfloor SoH * N_{bins} \rfloor \quad (24)$$

2.9.3.7. Predicting State of Health using Multi-Task LASSO Regression

Multi-task LASSO regression is utilized to predict state of health by exploiting commonalities across tasks and applying L1 regularization for effective feature selection. Shared information among different tasks is used to create SoH estimations that are robust to individual feature anomalies by minimizing the following cost function:

$$Objective = \min_{W,b}\frac{1}{N}\sum_{i=1}^{N}||Y-(W\Theta+b)||_F^2 + \lambda\sum_{j=1}^{D}||W||_1 \quad (25)$$

Multitask LASSO regression employs L1 regularization to promote sparsity across different tasks by zeroing out certain predictor weights, ideal for handling high-dimensional datasets. This method computes the state of health by aligning predictor weights to minimize the root mean square error (RMSE) between the predicted and actual health states. The resulting predictions are then processed through a SoftMax function, in (26) below, to generate probability density functions, enhancing the interpretability of the model's outputs.

$$SoftMax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n}e^{x_{ij}}} \quad (26)$$

2.9.3.8. Developing Measure of Uncertainty using Shannon Entropy

The Shannon entropy function is used to quantify the expected information contained in a random variable or distribution, with higher entropy values indicating more unpredictability or disorder within the distribution. Equation 27 below has the i-th samples entropy calculated using the state of health probabilities.

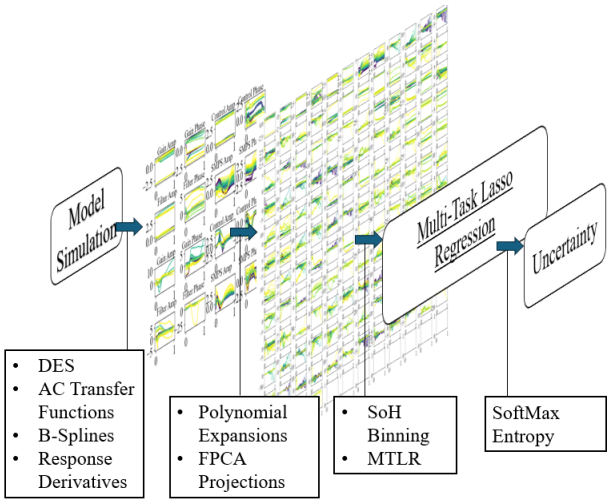


Figure 8: Diagram of the FMT-FFP framework, where high dimensional frequency response data is smoothed using functional b-splines and extended using functional derivatives and polynomial expansion using convolution integrals. Multi-Task LASSO regression is used for feature selection and produce a probabilistic assessment of state-of-health.

$$entropy(x_i) = \sum_{j=1}^{n}x_{ij}\log(x_{ij}) \quad (27)$$

A scaling is applied to the entropy function to create a more discernable measure of uncertainty,

$$uncertainty(x_i) = \frac{e^{1-\xi*entropy(x_i)}}{e^1} \quad (28)$$

2.9.3.9. Predicting Remaining Useful Life from State of Health and Usage Time.

A composite function in (29) is used to predict the remaining useful life of a system based on its current State of Health (SoH) and accumulated Usage Time (UT).

$$RUL(SoH,UT) = a*e^{-b*SoH} + c*UT^d + e \quad (29)$$

2.10. Comparative Analysis of Methods Using Error-Based Performance Metrics

In this study, we assess three distinct methods for predicting SoH and the remaining useful life (RUL). The primary metric for evaluating the accuracy of these predictions is the RMSE of the output voltage compared to the reference voltage, as defined in (30):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\big(V_{out,i}-V_{ref,i}\big)^2} \quad (30)$$

Additionally, uncertainty measurements play a crucial role in understanding the reliability of each method's predictions. Binned

probabilistic measurements of SoH have underlying uncertainty that can be used to compare the methods. By quantifying the probabilities of different SoH states, we can gauge the confidence level in the predictions made by each method.

3. Simulation

The simulation was run according to the diagram in Figure 5 using the variables for the circuit, and degradation models found below in Table 1.

Table 1: Simulation parameter values and ranges

Initial Electrical Parameter Selections					
Parameter	Ci, C_o	r_{Ci}, r_{Co}	L_i	L	r_M, r_D, r_{Li}, r_L
Value	0.1±0.02 mF	0.1±0.02 mF	10.0±2.0 µH	2.4±0.2 mH	0.1±0.04 Ω
Aluminum Capacitor Degradation Model Parameters [al elct testing report]					
Parameter	A_0	Ea1	Ea2	B_0	C_0
Value	3.462M	0.773	0.694	-111.8	3.999M
Parameter	D_0	Cap_0	ESR_0	R_{CA}	k_b
Value	0.407m	Nominal Ci, C_o	Nominal r_{Ci}, r_{Co}	100°C	8.6173 eV/K
Simulation Parameter Selection					
Parameter	t_s	t_L	T_s	T_c	T_M
Value	0.1±0.02	0.1±0.02	10.0±2.0	2.4±0.2	0.1±0.04

The data was collected and assigned a state of health label. The output voltage time response for a sample unit can be seen in Figure 9 below. As the degradation in the input and output filter progresses it causes parametric shifts that affect the ripple characteristics of the SMPS device.

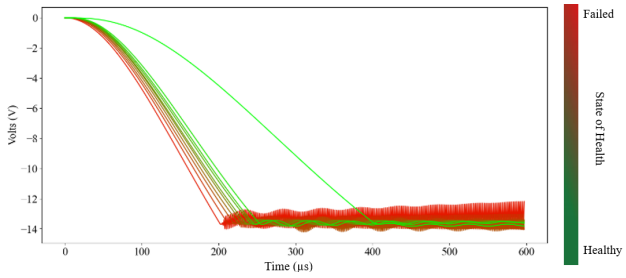


Figure 9: Output voltage start-up transient response for the buck boost SMPS from the generalized state space average model with input EMI filter exhibiting aluminum electrolytic capacitor degradation over time.

Using the state derivative information, the ripple current over the capacitors can be reconstructed, seen below in Figure 10, and used to calculate the estimated case temperature that is then used to dictate the degradation of the capacitors' capacitance and ESR.

The case temperature, ESR, and capacitance and ripple for the sample device can be seen below in Figure 11. The device is utilized in an ambient temperature of 105°C which creates environmental effects on the capacitance and ESR by causing the aluminum and film layers to be more closely packed with higher temperature. Over time degradation is associated with electrolytic liquid drying and temperature effects which causes capacitance to decrease and ESR to increase.

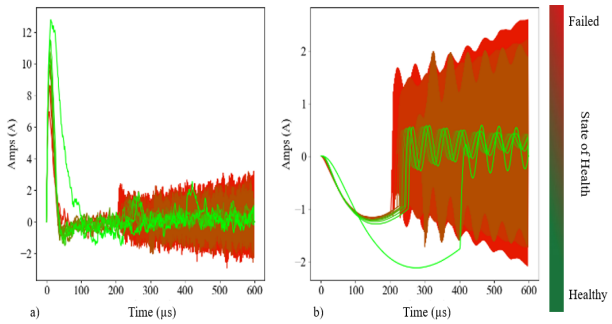


Figure 10: Current waveforms input EMI filter exhibiting aluminum electrolytic capacitor degradation over time.

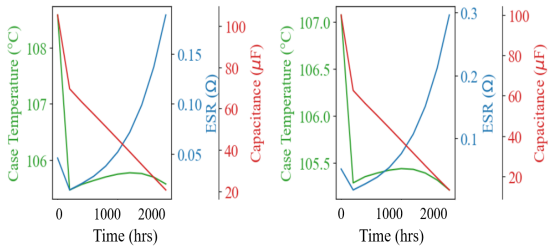


Figure 11: Output voltage start-up transient response for the buck boost SMPS from the generalized state space average model with input EMI filter exhibiting aluminum electrolytic capacitor degradation over time.

4. Results

The following section presents visual representations of input feature samples for each of the three methods, followed by graphs illustrating the SoH and RUL estimates with calculated uncertainty. It concludes with a comparison of each method's performance through tabulated error metrics, providing a clear evaluation of their predictive accuracy.

4.1 Visualization of Reduced Dimensional Features

The PCA-DNN method employs Principal Component Analysis (PCA) to reduce high-dimensional data to three principal components, focusing on capturing the most critical variances. This streamlined data set forms the basis for further deep neural network analysis, ensuring that essential features are retained while excluding less informative variables.

Figure 12 displays a 3D scatter plot of the reduced-dimensional features. This visual representation plots each observation according to the three principal components, helping to elucidate the data's underlying structure and highlighting potential patterns or anomalies within the reduced feature space.

The CWT features are visualized through a set of heatmaps, representing the CWT magnitudes, ranging from 0 to 1600 hours, seen below in Figure 13. Each panel represents a distinct time slice showing the frequency content of the data evolves, revealing changes in signal properties due to operational impacts like aging or wear. By examining these patterns, you can detect critical events or degradation, aiding in predictive maintenance and system monitoring.

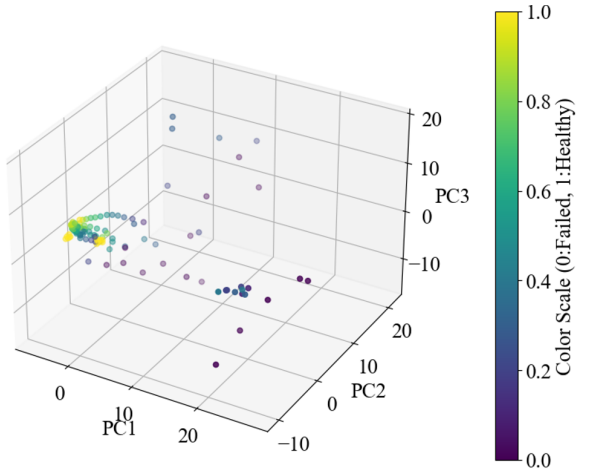


Figure 12: Current waveforms input EMI filter exhibiting aluminum electrolytic capacitor degradation over time.

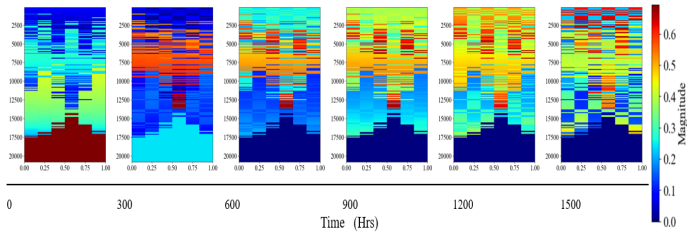


Figure 13: Sequential CWT heatmaps from 0 to 1600 hours, showing the evolution of signal frequencies over time.

The FMT-FFP method sample input features are shown in Figure 14 showing amplitude and phase responses for gain and control transfer functions and EMI Filter output, and SMPS input impedances. The magnitudes are standardized and domain normalized, ranging from 0 to 1. Each colored line represents variable conditions, highlighting the dynamic interactions within the system.

The state of health (SoH) classification methods outlined in Section 2 were applied to the data and the results can be seen below in Figure 15. For inputs into the respective method, extended AC features were used for the PCA-DNN and FMT-FFP routines and the raw output voltage time-response features were used for the CWT-SVM method. The predictions for the binned state of health outputs were turned into probability density functions (greyscale in the plot) using the SoftMax function and the uncertainty formula was used to create a set of empirical confidence intervals for each of the SoH estimates (in red).

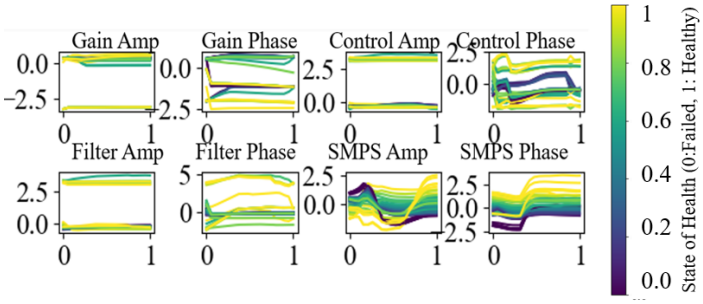


Figure 14: Current waveforms input EMI filter exhibiting aluminum electrolytic capacitor degradation over time.

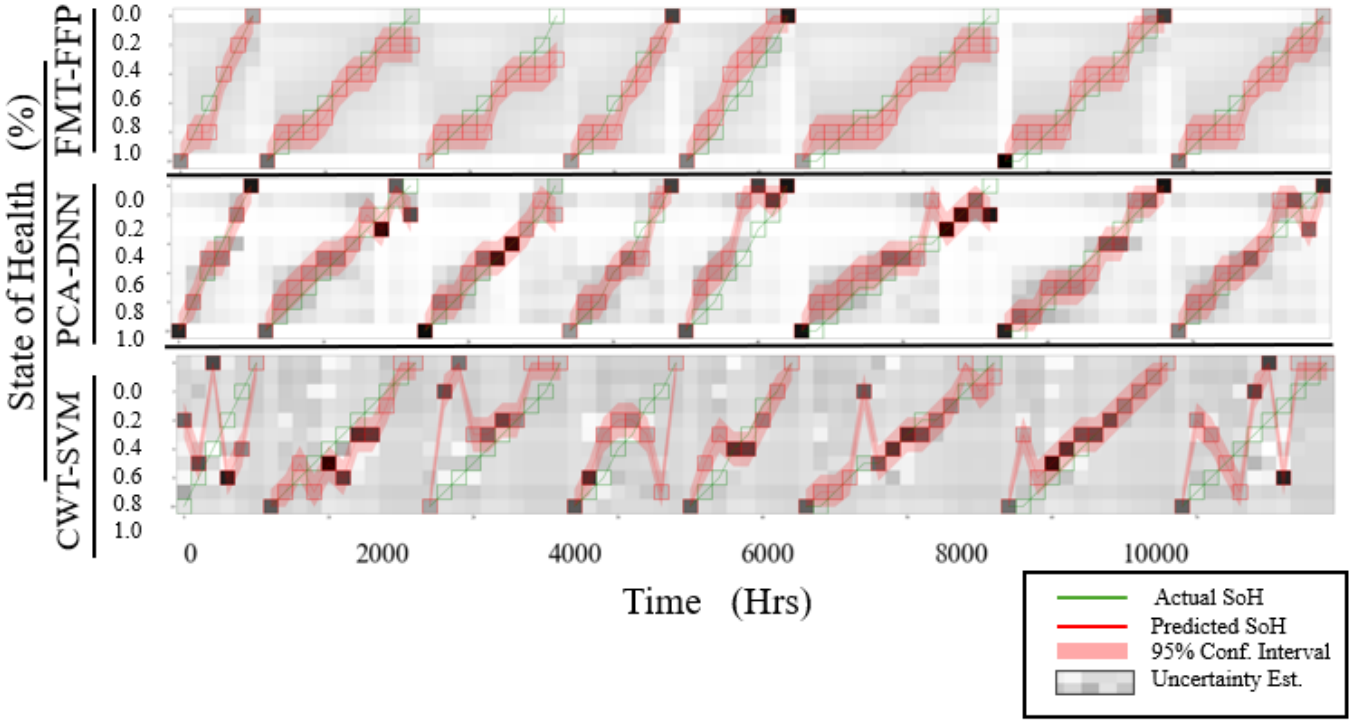


Figure 15: State-of-health prediction from the multi-task LASSO regression for the testing data.

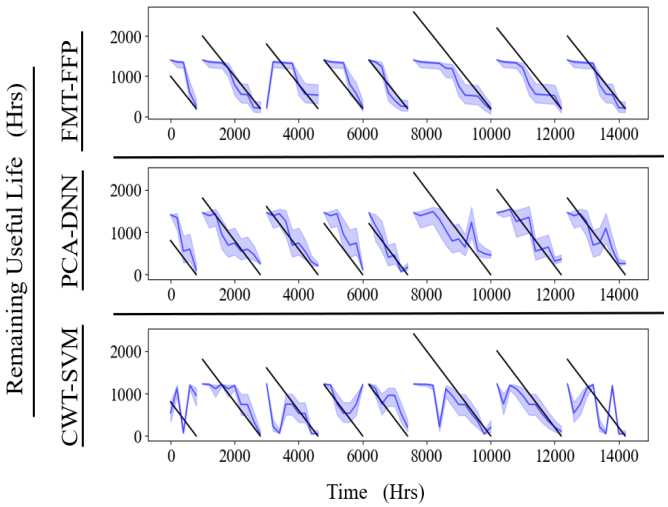


Figure 16: Remaining useful life prediction from the composite model contained in (30).

The SoH estimates were used as inputs into the RUL calculation given in Section 2, and the results can be seen for each method in Figure 16. The plot compares three methods for predicting RUL of components: FMT-FFP, PCA-DNN, and CWT-SVM. Each model's predicted RUL is plotted against the actual RUL over time.

The FMT-FFP method shows decent trend following with moderate prediction uncertainty that increases as RUL decreases. The PCA-DNN method exhibits tighter confidence intervals close to the end of life for each component but suffers from misalignments during mid-life usage predictions. The CWT-SVM has the widest confidence intervals indicating higher uncertainty and shows considerable fluctuation, suggesting sensitivity to noise.

All methods struggle with long-term predictions, reflected by widening confidence intervals over time, but the FMT-FFP has the most predictable temporal behavior which resulted in a monotonically decreasing SoH estimate and bounds. The variability in prediction accuracy at different life stages suggests that combining models or refining calibration might yield better results. No single method consistently outperforms the others across the entire lifecycle. The RMSE error for both the SoH and RUL estimate can be found in Table 2 below.

Table 2: RMSE values for the separate methods found in the literature compared to the proposed method.

	SoH (%) Training	**SoH (%) Testing**	**RUL (Hrs) Training**	**RUL (Hrs) Testing**	**Ref**
FMT-FFP	0.176±0.032	0.108±0.047	395.6±29.5	366.3 ±37.7	8
PCA-DNN	0.098 ±0.047	0.129±0.414	377.9±55.26	388.2±56.6	8*
CWT-SVM	0.181±0.029	0.250±0.014	4501.7 ±23.0	569.8 ±14.12	3*

Each method has its own benefits that are difficult to compare quantitatively. The PCA-DNN method can establish nonlinear trends in the data that provides superior predictive capability but would be difficult to interpret without additional approaches to interrogate the activated weight paths that lead to predictions with low uncertainty, and increased performance. The CWT-SVM method benefits from the fact that it uses an output voltage signal for feature generation and requires minimal testing compared to using the AC response features. This led to increased uncertainty in the SoH estimations that translated to misalignments in the RUL curves.

The performance of each method generally improves with more data, however the PCA-DNN is subjected to the vanishing gradient phenomena where over training causes initial weights to trend towards zero, and the CWT-SVM has hyperparameters associated with the chosen kernel that could lead to overfitting. The FMT-FFP method has a λ regularization parameter associated with the MTLR objective functions that was found via grid search and ξ-hyperparameter that scales the entropy in the uncertainty function that was found using k-fold cross validation, the results of which are in Figure 17.

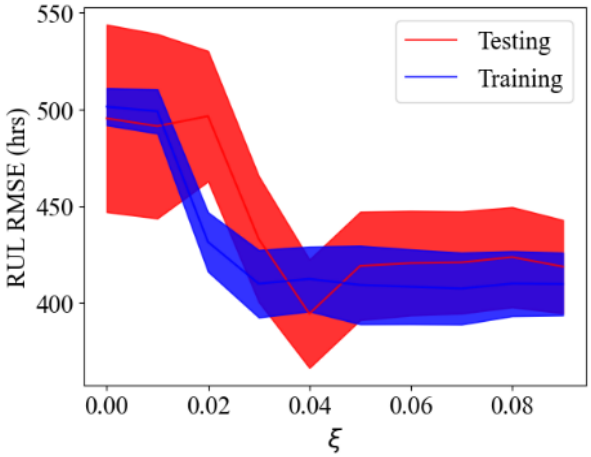


Figure 17: k-fold cross validation results used to find an optimal value of ξ=0.04 for the entropy scaling parameters in the SoH uncertainty calculation.

The FMT-FFP method provides the most explainable results in that the selected features are associated with specific frequency points in the AC-response information, found in Figure 18.

These points are found to be in line with the second harmonic of the switching frequency (10 kHz) and a value of 15.8 kHz which is consistent across several of the features. This allows for a targeted data sampling routine that that does not require data collection over the full frequency range.

4.2. Data Validation using the test circuit

To validate the developed prognostic, a test circuit apparatus was built shown in Figure 19 was used to simulate degradation. The results of the output filter capacitor ESR degradation trajectories show a mean response of 0.1-0.3 ohms over the lifetime of the component and for the input filter and mean ESR degradation trajectory going from 0.05 to 0.15. Using parallel circuit arrangements, a digital potentiometer to vary a resistance in-series with each capacitor. This effectively created a means of simulating the ESR degradation, ignoring the effects of the diminished capacitance seen in the simulation results.

4.3. Using a test circuit for validation

The FMT-FFP method was validated using a test circuit, designed with features selected via multi-task LASSO regression. The circuit components include a 2.7mH inductor, a 220µF capacitor with 7.6 ohms series resistance, an IRLZ44N NPN MOSFET, and a 1N5109 Schottky diode. An MCP602 operational amplifier is also employed in a differential topology for voltage scaling.

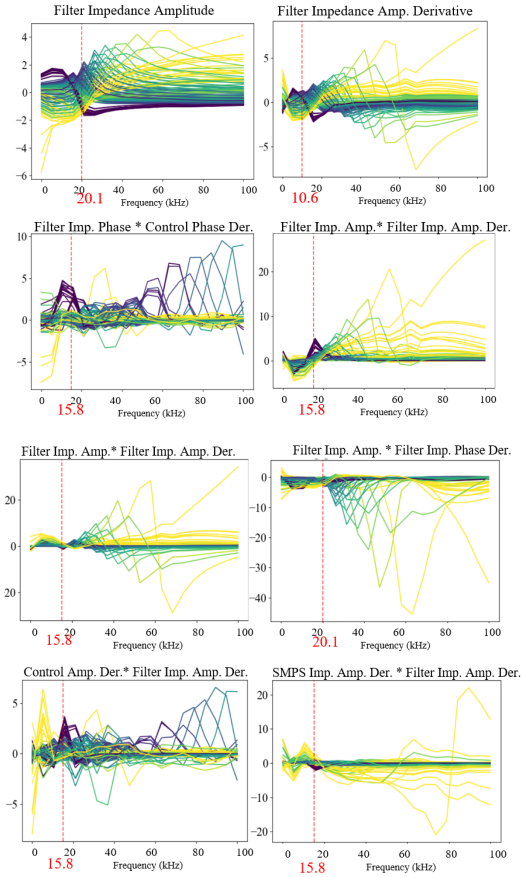


Figure 18: Feature importance visualizations for the down selected features identified in the FMT-FFP method.

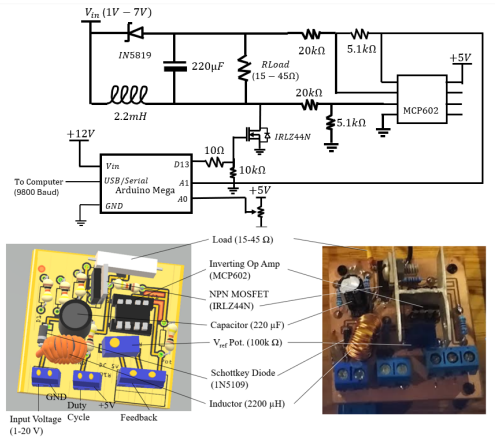


Figure 19: A schematic of the test circuit topology, a 3D model and actual image of the test circuit.

The frequency features for the validation were taken from a subset of AC features identified from simulation results and used to train and validate the prognostic method. ESR degradation trajectories where modelled using a 15Ω rheostat modulates ESR degradation from the nominal 7.6 Ω to 22.6 Ω using a similar exponential relationship to that identified in Figure 10. The procedure generated a comprehensive dataset for testing.

A power voltage divider circuit was used to generate the AC response features, by exciting the input signal at specific frequencies identified through simulation. The process was streamlined by limiting measurements to control amplitude gains, derived by dividing the output voltage gain by the duty cycle gain for a given input disturbance. This was chosen because only a reading of the output voltage, and a readily available duty cycle value are required. The control amplitude measurements were enhanced with numerical derivatives at ±2kHz and further extended via polynomial expansion. The overall approach to generating the features can be seen in the diagram in Figure 20.

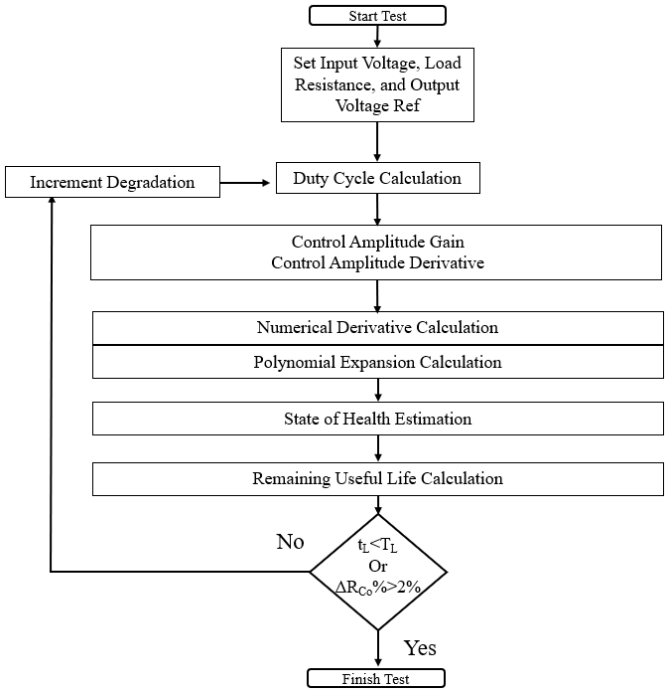


Figure 20:. Test circuit results using the down selected features identified in the FMT-FFP method.

The data were standardized and split into training and testing subsets and multi-task LASSO regression was used with grid search on the λ-value to derive state-of-health probabilities and uncertainty measures, seen in Figure 21.

The training state of health RMSE error was found to be 1.99 +/- -0.07 and the testing RMSE error was 1.61 +/- 0.05. The state of health and usage time where combined using equation 30 to form the following RUL predictions seen in Figure 22. The training remaining useful life RMSE error was found to be 68.55+/- 0.73 and the testing RMSE error was 224.96+/- 1.89.

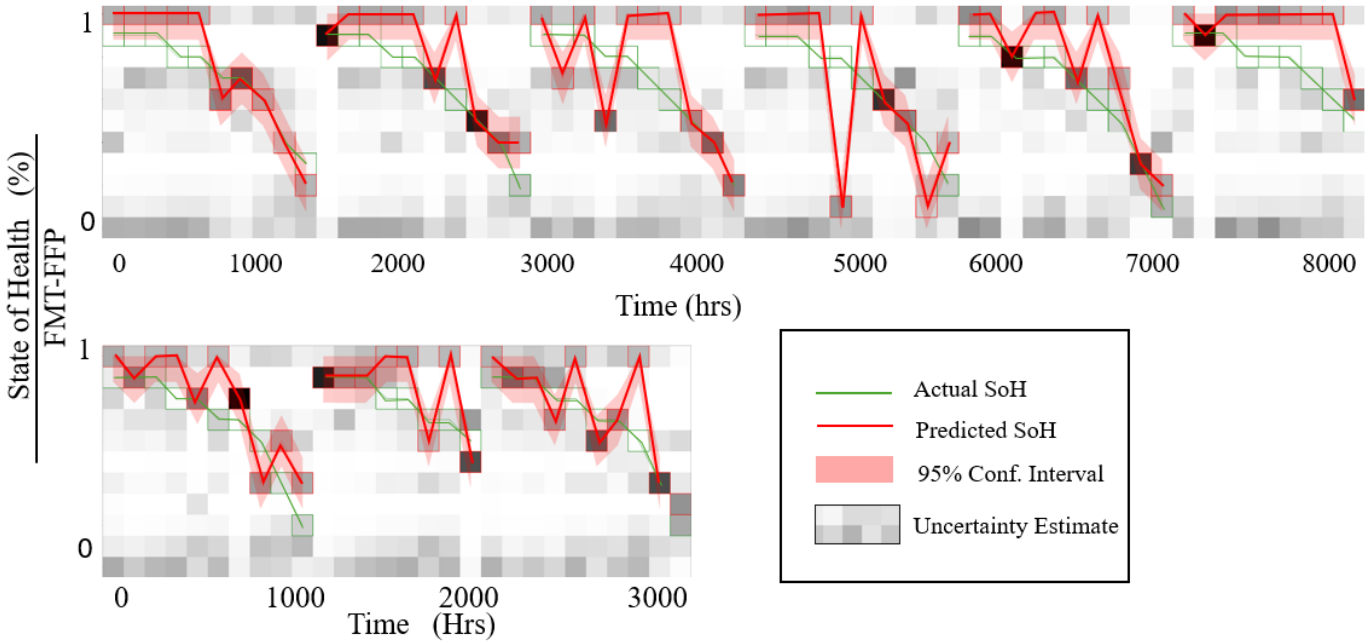


Figure 21: Test circuit results using the down selected features identified in the FMT-FFP method.

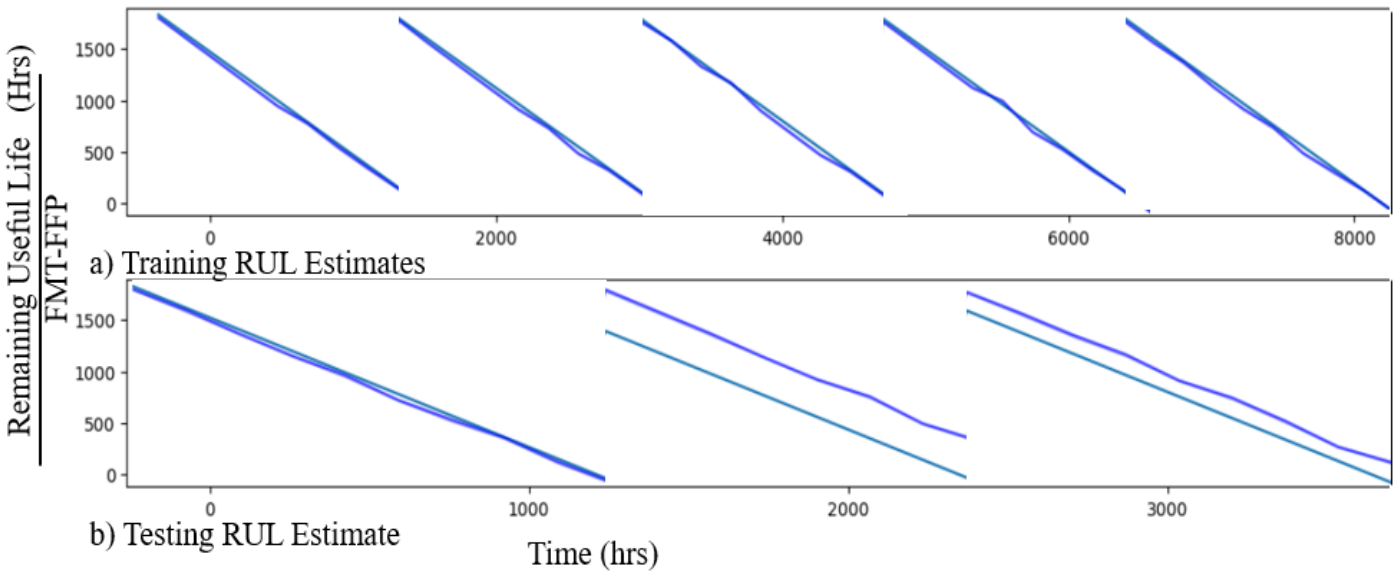


Figure 22: Test circuit results using the down selected features identified in the FMT-FFP method.

5. Discussion

This This research introduces a novel Fuzzy Multi-Task Functional Fusion Predictor (FMT-FFP) method for predicting Remaining Useful Life (RUL) that takes into account prediction uncertainty and incorporates multivariate functional fusion of the AC analysis raw data. The accuracy of the FMT-FFP method is improved by using functional principal component projections to handle noise that would otherwise reduce the model's effectiveness. The method can be generalized across different systems and conditions to account for specific failure scenarios.

The FMT-FFP method is not as sensitive to the choice of hyperparameters compared to more black-box methods, making it conducive to practical implementation across varied contexts to provide domain knowledge of impending failure modes. Experimental validation has shown that the FMT-FFP framework can be used to monitor systems in real time, providing timely insights into system health and potentially averting failures before they occur.

Comparison analysis shows that the FMT-FFP method performs well against more black-box methods of prediction, such as the PCA-DNN and CWT-SVM approaches. This comparative

analysis highlights the robustness and practical applicability of the FMT-FFP method in diverse operational scenarios.

6. Conclusion

This research aims to advance model-based design strategies for power electronics by developing a framework that reduces EMI in SMPS, with a particular focus on mitigating conducted emissions and their subsequent effects on radiated emissions. A key achievement of this study is the utilization of multitask LASSO regression to define data-driven stability regions, which enables the sparse mapping of design parameters directly linked to EMI sources. Employing a strategy that integrates low-dimensional feature mapping with sparsity-induced optimization, the method extracts and utilizes extended frequency-based features from the transfer functions and impedance characteristics of input filters and SMPS to create a sparse linear model. This method establishes AC frequency test criteria and addresses complex degradation behaviors associated with higher-order harmonics and specified stability criteria under high-EMI conditions.

The models have been validated through LTSPICE simulations and a test circuit, demonstrating their adaptability to diverse operational conditions and confirming their real-world feasibility and effectiveness.

Future research should focus on further validating these methods across a broader range of SMPS device topologies and operational conditions. Additionally, incorporating more comprehensive signal observations could further enhance the predictive accuracy and robustness of the model. Comparative analyses with existing methods have underscored the unique interpretability and practical effectiveness of our approach, highlighting its significant potential for industry adoption. This study not only demonstrates a significant step forward in the field of power electronics but also emphasizes the critical role of advanced analytics and modeling in boosting the reliability and efficiency of power systems amidst rapidly evolving technology demands.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] S. Douzi, M. Tlig, J.B.H. Slama, 2015. “Experimental investigation on the evolution of a conducted-EMI buck converter after thermal aging tests of the MOSFET,” Microelectronics Reliability, **55**(9-10), 1391-1394, 2015, doi:10.1016/j.microrel.2015.07.009.

[2] J. I. Corcau, L. Dinca, "Stability Studies of Power Systems for More Electric Aircraft," in 2022 International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM), Sorrento, Italy, 861-864, 2022, 10.1109/SPEEDAM53979.2022.9842040.

[3] D. Nemashkalo, N. Moonen, F. Leferink, "Practical consideration on power line filter design and implementation," 2020 International Symposium on Electromagnetic Compatibility - EMC EUROPE, Rome, Italy, 1-6, 2020, doi: 10.1109/EMCEUROPE48519.2020.9245777.

[4] R. Trois, G. Viscillo, G. F. Volpi, S.A. Pignari, "Accurate prediction of conducted emissions in switch-mode power supplies for space applications," in 2020 International Symposium on Electromagnetic Compatibility - EMC EUROPE, Rome, Italy, 1-6, 2020, doi: 10.1109/EMCEUROPE48519.2020.9245814.

[5] L. Eliasson, “Aluminium electrolytic capacitor’s performance in Very High Ripple Current and Temperature Applications,” in CARTS Europe, 2007.

[6] D. Goodman, J. Hofmeister, J. Judkins, "Electronic prognostics for switched mode power supplies," Microelectronics Reliability, **47**(12), 1902-1906, 2007.

[7] R. F. Orsagh, D. W. Brown, P. W. Kalgren, C. S. Byington, A. J. Hess and T. Dabney, "Prognostic health management for avionic systems," in 2006 IEEE Aerospace Conference, Big Sky, MT, USA, 7, 2006, doi: 10.1109/AERO.2006.1656086.

[8] C.S. Kulkarni, J.R. Celaya, K. Goebel, G. Biswas, "Physics based electrolytic capacitor degradation models for prognostic studies under thermal overstress." in PHM Society European Conference,**1**, No. 1, 2012, doi:10.36001/phme.2012.v1i1.1423

[9] B. Sun, X. Fan, C.A. Yuan, C. Qian, G. Zhang, "A degradation model of aluminum electrolytic capacitors for LED drivers," in 2015 16th International Conference on Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems, Budapest, Hungary, 1-4, 2015, doi:10.1109/EuroSimE.2015.7103124.

[10] M.M. Jovanović, "Power supply technology–past, present, and future," in Power Conversion and Intelligent Motion China Conf. for Power Electronics (PCIM China) Shanghai, China, 3-15, 2007.

[11] Z. Wang, Conducted EMI Noise Prediction and Filter Design Optimization, Ph.D Dissertation, Virginia Tech, 2016.

[12] Z. Li, K.W. E. Cheng, J. Hu, "Modeling of basic DC-DC converters," in 2017 7th International Conference on Power Electronics Systems and Applications - Smart Mobility, Power Transfer & Security (PESA), Hong Kong, China, 1-8, 2017, doi: 10.1109/PESA.2017.8277782.

[13] A. Ghosh, K. Mayank, State-space average modeling of dc-dc converters with parasitic in discontinuous conduction mode (dcm), Bachelor’s Thesis, National Institute of Technology, Rourkela, 2010.

[14] L. Sandrolini, A. Mariscotti, "Waveform and spectral characteristics of supraharmonic unsymmetrical conducted EMI of switched-mode power supplies," Electronics, **11**(4), 591, 2022, doi:10.3390/electronics11040591.

[15] A. Ismail, L. Saidi, M. Sayadi, M. Benbouzid, “A new data-driven approach for power IGBT remaining useful life estimation based on feature reduction technique and neural network,” Electronics, 9(10), 1571, 2020, doi:10.3390/electronics9101571.

[16] A.G. EPCOS, "Aluminum electrolytic capacitors – general technical information’. White Paper, 2014.

[17] J. Gu, M. Pecht, "Prognostics and health management using physics-of-failure," in 2008 Annual Reliability and Maintainability Symposium, Las Vegas, NV, USA, 481-487, 2008, doi: 10.1109/RAMS.2008.4925843.

[18] A.B. Kareem**,** J.W. Hur**, “**A feature engineering-assisted CM technology for SMPS output aluminium electrolytic capacitors (AEC) considering D-ESR-QZ parameters**,”** Processes, **10**(6), p. 1091**,** 2022, doi: 10.3390/pr10061091.

[19] T. McGrew, V. Sysoeva, C.H. Cheng, C. Miller, J. Scofield, M. J. Scott, "Condition Monitoring of DC-Link Capacitors Using Time–Frequency Analysis and Machine Learning Classification of Conducted EMI," IEEE Transactions on Power Electronics**,** **37**(10), 12606-12618, 2022, doi: 10.1109/TPEL.2021.3135873.

[20] A. Altmann, B. Ng, "Joint Feature Extraction from Functional Connectivity Graphs with Multi-task Feature Learning," in 2015 International Workshop on Pattern Recognition in NeuroImaging, Stanford, CA, USA, 29-32, 2015, doi: 10.1109/PRNI.2015.17

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

Appendices

Appendix A. Linearized Closed Loop State Space Matrices for an LC-Filtered Switch Mode Power Supply

State Space Result

The state matrices associated with the plant, A, input, B, Output, C, and feed-through response, D, are found below.

$$A=\begin{bmatrix} A_{11} & A_{12} & A_{13} & 0 & A_{15} \\ A_{21} & A_{22} & A_{23} & A_{24} & A_{25} \\ A_{31} & A_{32} & 0 & 0 & A_{35} \\ 0 & A_{42} & 0 & A_{44} & A_{45} \\ A_{51} & A_{52} & A_{53} & A_{54} & 0 \end{bmatrix}$$

$A_{11}=-\frac{r_{C_i}+r_{L_i}}{L_i}$; $A_{12}=\frac{D\,r_{C_i}}{L_i}$; $A_{13}=-\frac{1}{L_i}$; $A_{15}=\frac{I_L r_{C_i}}{L_i}$;

$A_{21}=-\frac{D\,r_{C_i}}{L}$; $A_{22}=K_1$; $A_{23}=\frac{D}{L}$; $A_{24}=\frac{R(D-1)}{L(R+r_{C_o})}$;

$A_{25}=\frac{1}{L(R+r_{C_o})}(I_L(R\,r_{C_o}+R\,r_D+R\,r_L+r_{C_o}r_D+r_{C_o}r_L-(R+r_{C_o})(r_{C_i}+r_L+r_M))+R(I_O r_{C_o}+V_{C_o})+(R+r_{C_o})(I_{L_i}r_{C_i}+V_{C_i}-V_D+V_M)$;

$A_{31}=\frac{1}{C_i}$; $A_{32}-\frac{D}{C_i}$; $A_{35}=-\frac{I_L}{C_i}$; $A_{42}=\frac{R(1-D)}{C_o(R+r_{C_o})}$; $A_{44}=-\frac{1}{C_o(R+r_{C_o})}$; $A_{45}=\frac{R(-I_L+2\,I_O)}{C_o(R+r_{C_o})}$;

$A_{51}=\frac{DK_dR\,r_{C_i}r_{C_o}(1-D)}{L\,L_i(R+r_{C_o})}+\frac{DK_pR\,r_{C_i}r_{C_o}(1-D)}{L(R+r_{C_o})}+\frac{Dr_{C_i}}{L}\left(\frac{K_dK_1R\,r_{C_o}(1-D)}{(R+r_{C_o})}+\frac{K_dR^2(1-D)}{C_o(R+r_{C_o})^2}\right)+\frac{DK_dR\,r_{C_o}(1-D)}{C_i\,L_i(R+r_{C_o})}$;

$A_{52}=\frac{D^2K_dR\,r_{C_i}^2r_{C_o}(1-D)}{L\,L_i(R+r_{C_o})}+\frac{K_iR\,r_{C_o}(1-D)}{(R+r_{C_o})}+\frac{K_pK_1Rr_{C_o}(1-D)}{(R+r_{C_o})}+K_1\left(\frac{K_dK_1R\,r_{C_o}(1-D)}{(R+r_{C_o})}+\frac{K_dR^2(1-D)}{C_o(R+r_{C_o})^2}\right)+\frac{K_pR^2(1-D)}{C_o(R+r_{C_o})^2}+\frac{R(1-D)\left(\frac{-K_dR^2\,r_{C_o}(1-D)^2}{L(R+r_{C_o})^2}-\frac{K_dR}{C_o(R+r_{C_o})}\right)}{C_o(R+r_{C_o})}-\frac{K_dD^2R\,r_{C_o}(1-D)}{C_iL(R+r_{C_o})^2}$;

$A_{53}=\frac{DK_dR\,r_{C_i}r_{C_o}(1-D)}{L\,L_i(R+r_{C_o})}+\frac{DK_pR\,r_{C_i}r_{C_o}(1-D)}{L(R+r_{C_o})}+\frac{D\left(\frac{K_dK_1R\,r_{C_o}(1-D)}{(R+r_{C_o})}+\frac{K_dR^2(1-D)}{C_o(R+r_{C_o})^2}\right)}{L}$;

$A_{54}=\frac{K_iR}{R+r_{C_o}}-\frac{K_pR}{C_o(R+r_{C_o})}-\frac{K_pR^2r_{C_o}(1-D)^2}{L(R+r_{C_o})^2}-\frac{-\frac{K_dR^2r_{C_o}(1-D)^2}{L(R+r_{C_o})^2}+\frac{K_pR}{C_o(R+r_{C_o})^2}}{C_o(R+r_{C_o})}+\frac{R(D-1)\left(\frac{K_dK_1R\,r_{C_o}(1-D)}{(R+r_{C_o})}+\frac{K_dR^2(1-D)}{C_o(R+r_{C_o})^2}\right)}{L(R+r_{C_o})}$,

$$B=\begin{bmatrix} B_{11} & 0 & 0 & 0 & 0 \\ 0 & B_{22} & B_{23} & B_{24} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & B_{42} & 0 & 0 & 0 \\ B_{51} & B_{52} & B_{53} & B_{54} & B_{55} \end{bmatrix}$$

$B_{11}=\frac{1}{L_i}$;$B_{22}=-\frac{R\,r_{C_o}(1-D)}{L(R+r_{C_o})}$;$B_{23}=-\frac{D}{L}$; $B_{24}=-\frac{1-D}{L}$;

$B_{42}=\frac{R\,(1-2D)}{C_o(R+r_{C_o})}$; $B_{51}=\frac{DK_dR\,r_{C_i}r_{C_o}(1-D)}{L\,L_i(R+r_{C_o})}$;

$B_{52}=\frac{K_iRr_{C_o}(1-2D)}{R+r_{C_o}}+K_pR^2r_{C_o}^2\cdot\frac{(1-D)(D-1)}{L(R+r_{C_o})2}+R\frac{r_{C_o}(D-1)\left(K_dR\frac{r_{C_o}(1-D)\left(-D(R+r_{C_o})(r_{C_i}+r_L+rM)+(D-1)(RrCo+RrD+RrL+rCorD+rCorL)\right)}{L(R+r_{C_o})^2}+\frac{K_d^2R(1-D)}{C_o(R+r_{C_o})^2}\right)}{L(R+r_{C_o})}+\frac{K_pR^2(1-2D)}{C_o(R+r_{C_o})^2}-\frac{R(1-2D)\left(\frac{K_dR^2r_{C_o}(1-D)^2}{L(R+r_{C_o})^2}-\frac{K_dR}{C_o(R+r_{C_o})^2}\right)}{C_o(R+r_C_o)}$;

$B_{53}=-\frac{D\,K_pR\,r_{C_o}(1-D)}{L(R+r_{C_o})}-\frac{D}{L}\left(\frac{K_dRr_{C_o}(1-D)\left(-D(R\,r_{C_o})(r_{C_i}+r_L+r_M)+(D-1)(Rr_{C_o}+Rr_D+R\,r_L+r_{C_o}r_D+r_{C_o}r_L)\right)}{L(R+r_{C_o})^2}+\frac{K_dR^2(1-D)}{C_o(R+r_{C_o})^2}\right)$.

$B_{54}=-\frac{K_pR\,r_{C_o}(1-D)^2}{L(R+r_{C_o})}-\frac{(1-D)}{L}\left(\frac{K_dRr_{C_o}(1-D)\left(-D(R+r_{C_o})(r_{C_i}+r_L+r_M)+(D-1)(Rr_{C_o}+Rr_D+R\,r_L+r_{C_o}r_D+r_{C_o}r_L)\right)}{L(R+r_{C_o})^2}+\frac{K_dR^2(1-D)}{C_o(R+r_{C_o})^2}\right)$;

$B_{55}=-K_i$;

$$C=\begin{bmatrix} 0 & \frac{Rr_{C_o}(1-D)}{R+r_{C_o}} & 0 & \frac{R}{R+r_{C_o}} & \frac{Rr_{C_o}(-I_L-2\,I_O)}{R+r_{C_o}} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$D=\begin{bmatrix} 0 & \frac{Rr_{C_o}(1-2D)}{R+r_{C_o}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 2, 45-54 (2024)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Exploring Current Challenges on Security and Privacy in an Operational eHealth Information System

Viktor Denkovski[*, 1], Irena Stojmenovska[1], Goce Gavrilov[1], Vladimir Radevski[1], Vladimir Trajkovik[2]

[1]*University American College Skopje, School of Computer Science and Information Technology, Skopje, 1000, North Macedonia*

[2]*University "Ss. Cyril and Methodius", Faculty of Computer Science and Engineering, Skopje, 1000, North Macedonia*

ARTICLE INFO

Article history:
Received:27 February, 2024
Revised: 08 March, 2024
Accepted: 05 April, 2024
Online: 16 April, 2024

Keywords:
eHealth Information System
Security and Privacy Issues
Medical Providers
eHealth Challenges

ABSTRACT

Bearing in mind that patient data is extremely sensitive, it is crucial to establish strong protection when the security and privacy of healthcare data are concerned. Prioritizing data security and privacy is essential for the overall healthcare industry in order to maintain the reliability of electronic healthcare (eHealth) information systems. This study explores the gathered data and information from the surveys and interviews by looking at the security and privacy concerns in using eHealth information technologies. The surveys and interviews were performed on the medical practitioners in N. Macedonia. The main goal is to find out how well-informed are the medical practitioners on the already in-place privacy measures that have been implemented by the medical authorities and to assess their attitudes regarding the need for additional improvements of the system. From the executed interviews, eight healthcare professionals participated in a thorough email interview in order to discover security and privacy issues associated with eHealth systems usage. This information served as the groundwork for administrating an online survey, to which 370 medical practitioners responded from primary and secondary healthcare. The findings emphasize how essential it is to promptly address the system usability concerns on the security and privacy procedures that are implemented when using eHealth technologies.

1. Introduction

This paper is an extension of work originally presented at the 2023 IEEE International Mediterranean Conference on Communications and Networking (MeditCom) [1] to highlight how critical it is to promptly tackle system usability issues related to privacy and security protocols while utilizing eHealth technologies. In recent years there have been revolutionary developments concerning the integration of eHealth information technology, changing the delivery of healthcare services, and providing several benefits, such as raised efficiency and improved accessibility to medical data [2, 3]. Technology has shown increased development that has sped up the integration of eHealth systems way above initial expectations. In 2004, the European Union (EU) established the eHealth systems as a national strategy and policy after realizing their crucial role play in today's modernized healthcare [4]. The research and development strategies from the EU are actively supporting the international progress of the eHealth systems promoting non-EU and EU country cooperation [4].

Even before progressing to national projects, the deployment of eHealth systems needs to go under extensive research and testing in order to address the adaptation and ethical concerns [5]. The mere fact that N. Macedonia's eHealth system was initially implemented without following any clear national policies or strategies indicates the need for additional financial support and support from the government to ensure a smooth national integration later on [6, 7].

The main purposes of electronic health systems are making the medical staff's paperwork easier, lowering costs, and raising satisfaction [8]. To enhance the quality of health care services, well-defined workflow procedures and online data exchange between primary and secondary care facilities must be implemented and adopted [9]. Furthermore, using eHealth technologies makes it much simpler to contact medical specialists, especially those who are working in big cities and larger hospitals. This promotes cooperation in overall health care and improves patient accessibility [8].

To increase efficiency in managing patients and saving paperwork, eHealth systems are an essential tool for patient management and data sharing. This involves digitizing,

[*]Corresponding Author: Viktor Denkovski, viktor.denkovski@uacs.edu.mk

https://dx.doi.org/10.25046/aj090206

exchanging, and archiving all relevant information [7]. However, worries over the security and privacy of the patient's private medical information are raised by the growing dependencies on digital platforms. Trust maintenance in the health care systems and also guaranteeing the patient's data privacy and security requires safeguarding the patient data from illegal breaches, access, and misuse [10]. In alignment with the regulatory protocols, people must be guaranteed that their data is being secured and that the security of digital data is being considered a top priority in eHealth systems [7].

Per a systematic study in [11], the usage of big data analysis and its applications in the healthcare industry has emerged recently in several studies and research projects. Three modules comprise the health care system's structure, of which, medical practitioners, medical consumers, and health services related to medical treatment, like research and health insurance. Furthermore, although it is still in its early stages, the pharmacy business and shareholders started their analysis of big data even more regularly in order to gain knowledge and an overview of numerous activities related to their sector [12].

By the end of 2021, 22 nations that were already a part of the digital service infrastructure for eHealth systems, began exchanging patient information and electronic prescriptions, indicating that healthcare data digitalization in some EU countries was already at an advanced stadium [13]. According to the World Health Organization (WHO) Regional Office for Europe in [14], research of healthcare big data in the healthcare sector presented vital importance, however, they were not sufficiently explored, and there were only a few methods and arrangements that offered help in this field. Having in mind this assessment, just 13% (six nations) of the EU had integrated national policies and rules in motion for big data in this area, while the private sector controlled just a small portion of this, or just 9% (four nations). The commercial sector would take over and seize the chance to capitalize on the potential of big data analysis in the healthcare industry if the public medical authorities fail to investigate this issue.

As was already pointed out above, big data in healthcare information systems presents several security and privacy issues when it comes to patient medical records. At the top, data centers that house sensitive data possess various levels of security. Because of this, it was unable to directly apply conventional security options and solutions to sizable and diverse data sets [15]. The complexity of security in many software systems that can handle different sources and data formats is compounded by the growing demand for cloud solutions in the healthcare sector [16]. Big data had to be properly maintained for this reason to be made available for data analysis.

Utilizing the accumulated big data for any kind of progress across all industries may present various difficulties. The lifecycle of the data was divided into data, management, and process classification when it comes to those difficulties [17]. The representation of data characteristics (variety, visualization, volume, variability, etc.) was one type of such difficulty. Security and privacy issues presented one of the challenges to be managed, also, issues that arise during data processing (analysis and modeling, data mining and implementation, etc.), and the absence of understanding the data. According to [18], management of big data must be improved by using operations of cleaning, processing, analyzing, securing, and granting access to the data due to its lack of flexibility, scalability, and performance. Furthermore, security and privacy protection were essential for the widespread usage of big data in all areas especially in the healthcare sector. Thus, the privacy and security issues related to the topic at hand were breakdown into categories that address infrastructure security (cloud security DoS attacks, Hadoop), data management (auditing and monitoring, key management, and data provenance), and data privacy (including data anonymization, encryption, and access control) [19]. For instance, considering big data in healthcare, it is crucial to protect sensitive patient data especially when dealing with privacy and security issues. Additionally, in [16] it was illustrated the distinctions between privacy and security issues that arise when dealing with big data, in which the primary security objective is to safeguard the information and not adequate to address privacy issues. In their research, they explain that, while privacy was more dependent on individual management of data (implementing policies to ensure that customer data was being gathered, used, and shared appropriately), security was more concerned with the protection of the data from malicious attacks and acquiring profit from the stolen data.

Data interoperability across primary and secondary healthcare systems was essential for the improvement of the end-user experience when discussing the national eHealth information system. To make it easier to integrate and share data between several platforms, it was crucial to develop common technological standards for data interchange between the primary and secondary healthcare systems. Adoption of common standards could be considered for this purpose, such as DICOM [20] (Digital Imaging and Communications in Medicine), SNOMED CT [21] (designed as a multilingual international core set for electronic clinical data exchange that can be used in Electronic Health Records (HER) systems), and HL7 FHIR [22] (next-generation standard for interoperability developed by Fast Healthcare Interoperability Resources and Health Level 7 to provide efficient and quick health data exchange) [23].

The sensitive patient data that must be legally protected and secured raised concerns when dealing with eHealth systems that monitor all the patient information and data as well as the medical staff who provide healthcare. According to WHO in [14], over 80% of the EU member states have created laws to protect the privacy of sensitive health information, which almost 30% from 2009 to 2016 was raised exponentially. Furthermore, the medical personnel must address additional challenges related to financial difficulties, computer expertise, information technology support, etc. when implementing eHealth systems [24]. In [25] it was pointed out that eHealth systems must have strong privacy and security policies in place to be able to handle these issues. The medical practitioners as the major users of these health systems must be aware of and abide by the established security procedures to reduce the threats associated with the digital transfer and storage of medical data [26].

The dynamic nature of security and privacy in healthcare was highlighted in [27], who additionally point out the significance of current security practice reviews and observation in order to respond to emerging vulnerabilities and threats. The medical staff members must take part in frequent education and awareness

programs in order to be informed about the latest security requirements and best practices. In [28] the focus was put on privacy and security concerns in the era of digital healthcare, emphasizing the need for robust security protocols to manage the issues and challenges posed by the rapid advancements of technology and the expanding amount of data in the healthcare sector. A question was posed on privacy protection and sensitive information, addressing the national data protection laws and the General Data Protection Regulation (GDPR). The topic of discussion was "Is it decently implemented in the medical field?". As a member of the Council of Europe since 2006, N. Macedonia has benefited from the creation of the Convention for Protection of Individuals related to the Automatic Processing of Personal Data, known as Convention 108, and on December 5th, 2019, was made the 37th nation to sign the updated Convention 108+ [29, 30]. The two main objectives of this revised Convention 108+ were to effectively reinforce the Convention's execution and address different issues that may arise from employing the new Information and communication technology. According to the convention, member nations must ensure that the individuals are informed about which and what type of information was being gathered about them, have access to the data, and can make corrections to their data. To keep up with the EU's GDPR, the new law on personal data protection was passed by Macedonian Parliament that went into effect in February 2020 [31, 32]. This law establishes responsibilities for the data processors and controllers to respect and safeguard the individuals' right to privacy. Also, this law regulates the processing of sensitive health information as personal data.

The opinion of the medical practitioners on eHealth technology security and privacy safeguards was the focus of this study. Their familiarity with the security mechanisms in place was evaluated, and possible areas of improvement – specifically, with the "Moj Termin" (My Appointment) system currently operational in N. Macedonia were highlighted. Understanding these problems would enable recommendations to be made to strengthen privacy of patient data and enhance the system's performance. To investigate these privacy and security concerns, the medical practitioners were subjected to a comprehensive email interview and a structured survey to evaluate their awareness of the security protocols and opinions regarding the protection of private medical information. By focusing on the system "Moj Termin", the results from the survey will be used to provide guided suggestions for enhancing the overall security and privacy of health care information systems.

The remaining sections of this research are arranged as follows: The study's tools and methods were detailed in the Materials and Methods section. The Results section includes the study's findings. The Discussion section interprets the survey and interview data and highlights the important findings. The conclusion section summarizes the findings of this research and offers recommendations for the medical authorities, organizations, and various IT bodies, emphasizing the value of providing ongoing education and training for medical practitioners to familiarize them with various privacy and security measures.

2. Materials and Methods

To gain a better grasp of the effect that using eHealth information systems has on the medical practitioners' awareness of the security and privacy of sensitive medical data, indebted e-mail interviews were the initial method used to conduct this research. The e-mail interviews presented an excellent alternative to phone and in-person interviews. Furthermore, the participants who were difficult to get in touch with lived in distant locations, or just didn't have the time because of their jobs may benefit more from this kind of qualitative research [33, 34]. The primary problem with this approach was that, in many cases, the interview participants weren't responsive to their requests to take part in the interview. To avoid this challenge, timed reminders were necessary [33].

Before conducting the e-mail interviews, a set of materials was put together, which included interview questions, a completed and scanned interview permission form, interview instructions, and an invitation to the interview by e-mail. Finally, there were no financial incentives offered as rewards to the study's participants.

The focus of this research was the addressed questions about security and privacy measures of sensitive medical data and the viewpoints of the participants in the context of challenges and difficulties. These questions included:

- What security measures were available to you while handling sensitive eHealth information belonging to the patients?
- What kinds of issues did you encounter when utilizing the eHealth information system "Moj Termin"?
- What features, data, services, etc. do you believe the eHealth information system "Moj Termin" requires?
- How did the administrators of the medical department help you when you first started using this information system "Moj Termin"?
- Which security measures were disposal to you when handling the sensitive data belonging to the patients?

The Interview transcripts were examined, and the questions were reviewed. The data was evaluated and examined to uncover similar statements on issues and concerns related to security and privacy.

An online survey was used in the second technique of this study to include a considerably bigger number of participants for the research relevance and further investigation of the topic at hand. Only the first section of the questionnaire, which included the demographic characteristics of physicians of primary and secondary level, and the third section, which included the issues and challenges were taken into consideration for this study out of the total six sections that the questionnaire was composed of.

Survey of primary and secondary physicians was carried out in N. Macedonia in two-months, between February 7th and April 7th, 2022. Considering the physicians' busy schedules, closed hospitals except in cases of emergency, and the reason that many of them were ill and isolated at home, the medical personnel were difficult to contact during the COVID-19 outbreak.

2.1. Study population

The standardized formula for minimal sample size was used to estimate the minimal sample size for this study to be relevant and achieve its primary goal [35].

$$n = \frac{\frac{z^2\pi(1-\pi)}{e^2}}{1 + \frac{z^2\pi(1-\pi)}{e^2 N}} \quad (1)$$

The formula considers the following factors:

- Sample size (n),
- Population proportion (π),
- Desired confidence level (z),
- Acceptable sampling error (e), and
- Population size (N).

At the time of access, in N. Macedonia were registered 4636 specialists in outpatient clinics, 1793 pharmacists, 1554 general physicians, and 160 gynecologists [36].

Based on the 9386 total population, 370 participants were necessary according to (1) to complete the questionnaire for the study to be relevant with a 95% confidence level and a 5% margin of error. All medical practitioners who worked with or had knowledge of the eHealth system "Moj Termin" usage were eligible to participate in this study. The technical workers at the healthcare facilities, those who had never utilized the eHealth system, and those who had not given informed consent were excluded from the study.

2.2. Statistical analysis

IBM SPSS Statistics version 23.0, a statistical program for the Windows operating system was utilized to examine the collected data [37]. The data analysis's categorical variables were provided as frequencies and percentages. To find any significant link between the variables, a cross-tabulation was performed. Phi and Cramer's V were utilized for calculation of the effect size between the variables, and to test the correlation, the Chi-Square test was used. A two-tailed p-value < 0,05 was considered significant and for the null hypothesis, it was assumed that there would be no relationship between the variables.

3. Results

3.1. Participants' demographic characteristics

3.1.1. Results from in-depth interviews

Between April and September 2020, the study's participants began to receive invitations for interviews, out of which, eight medical practitioners took part in the study (five being primary and three secondary care physicians). The demographic details of those who took part in the interview procedure are displayed in Table 1. They all finished the email interview, and two of them additionally had a follow-up email exchange. The interviews lasted between 20 to 40 minutes (on average) and it took the physicians one to two weeks (8,25 days on average) to respond after they received the invitation.

Table 1: Demographic characteristics of medical practitioners from the in-depth interviews

Characteristics	
Gender – No. (%)	
Male	1 (12,5%)
Female	7 (87,5%)
Age, years – Mean (±SD)	33,8 (±10,5)
Level of Education – No. (%)	
University	6 (75%)
PhD	1 (12,5%)
Other	1 (12,5%)
Profession – No. (%)	
Primary healthcare physician	5 (62,5%)
Specialist	3 (37,5%)
Work experience of primary healthcare physician, years – Mean (±SD)	7,8 (±10,3)
Work experience of specialist, years – Mean (±SD)	5,7 (±4,1)
Experience in patient management before starting to use the eHealth system – No. (%)	3 (37,5%)
No experience in patient management before starting to use the eHealth system – No. (%)	5 (62,5%)

3.1.2. Results from the online survey

As shown in Figure 1, most of the participants who took part in the online survey were from primary healthcare.

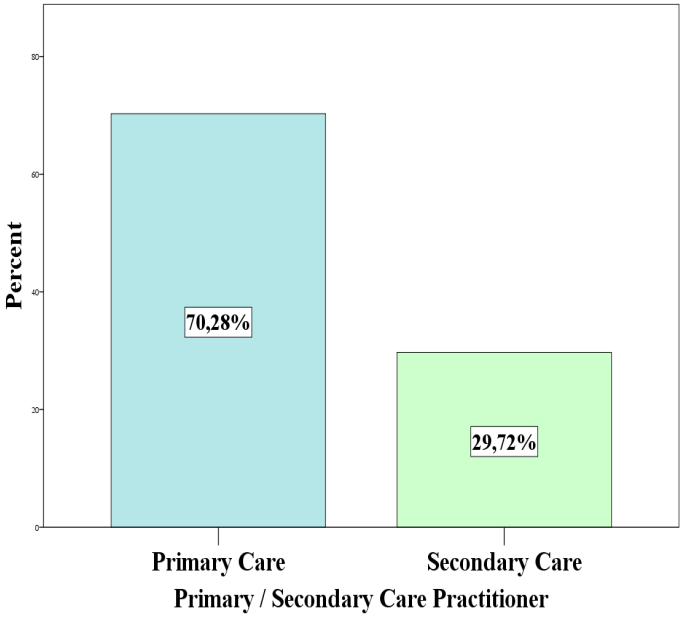


Figure 1: Medical practitioners' distribution within the healthcare sectors

The majority of the medical practitioners who took part in this survey were female, married, and had finished their specialty training as presented in Table 2. Furthermore, it was estimated that the typical medical practitioner had 19 (+/-10,5) years of overall work experience, however, the average years of experience with the eHealth system were estimated to be 7,5 (+/-3,8) years. Finally, as shown in Table 2, most of the participants worked in an urbanized environment.

Table 2: Results from the online survey - Demographic characteristics of medical practitioners

Characteristics	
Age, in years – Mean (±SD)	46,9 (±10,9)
Gender – N (%)	
Female	244 (66,1)
Male	125 (33,9)
Marital status – No. (%)	
Married	314 (84,9)
Unmarried	37 (10,0)
Widowed	17 (4,6)
Other	2 (0,5)
Education – N (%)	
Specialized studies	206 (56,4)
Graduate	123 (33,7)
PhD	19 (5,2)
Postgraduate	16 (4,4)
Other	1 (0,3)
Practitioners' Healthcare Sector – N (%)	
Primary Care	253 (70,3)
Secondary Care	107 (29,7)
Working experience of the medical practitioner, in years – Mean (±SD)	19,1 (±10,5)
Location of the medical facility – N (%)	
Urban Area	321 (87,2)
Rural Area	47 (12,8)
eHealth system patient management experience, in years – Mean (±SD)	7,5 (±3,8)

The distribution of the participants was approximately evenly distributed, with the majority falling within the age range of 28 and 62, as seen in Figure 2.

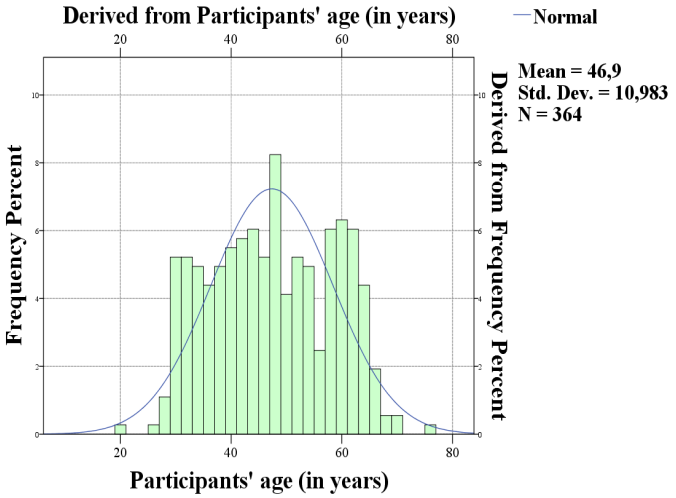


Figure 2: Results from the online survey - Age distribution of the medical practitioners

The specializations of the medical practitioners included in this analysis are shown in Table 3. Of these, the majority were registered as other specialized physicians, followed by general practitioners, and dentists.

Table 3: Medical practitioners' area of expertise

Physician's role		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Gynecologist	21	5,7	6,1	6,1
	Dermatologist	2	,5	,6	6,7
	Dermatologist and venereologist	2	,5	,6	7,3
	Internist	14	3,8	4,1	11,4
	Effectologist	4	1,1	1,2	12,5
	Clinical pharmacologist	1	,3	,3	12,8
	Maxillofacial surgeon	1	,3	,3	13,1
	Neurologist	3	,8	,9	14,0
	Neuropsychiatrist	1	,3	,3	14,3
	Neurosurgeon	1	,3	,3	14,6
	Nuclear medicine	2	,5	,6	15,2
	General Medicine	66	17,8	19,2	34,4
	Orthopedist	3	,8	,9	35,3
	Otorhinolaryngologist	8	2,2	2,3	37,6
	Ophthalmologist	3	,8	,9	38,5
	Pediatrician	21	5,7	6,1	44,6
	Pneumophthisiologist	1	,3	,3	44,9
	Psychiatrist	2	,5	,6	45,5
	Radiologist	2	,5	,6	46,1
	X-ray specialist	1	,3	,3	46,4
	Social medicine	4	1,1	1,2	47,5
	Specialist	13	3,5	3,8	51,3
	Sports medicine	1	,3	,3	51,6
	Dentistry	58	15,7	16,9	68,5
	Transfusiologist	2	,5	,6	69,1
	Urologist	3	,8	,9	70,0
	Physical medicine and rehabilitation	7	1,9	2,0	72,0
	Surgeon - Specialist	5	1,4	1,5	73,5
	Other specialist doctors	91	24,6	26,5	100,0
	Total	343	92,7	100,0	
Missing		27	7,3		
Total		370	100,0		

3.2. Challenges and Issues (Contextual Aspects)

The opinions of the medical practitioners on the challenges and issues they encountered when utilizing the eHealth system are shown in Table 4. The majority of the medical practitioners in this survey reported that throughout their working experience in the eHealth system, they occasionally ran across different persistent technical challenges, nevertheless, this group was closely followed by those who often faced consistent issues. When faced with various technical issues, the majority of the medical practitioners stated they were rarely given assistance or some technical support by the medical authorities. However, 16,1% stated they never received any assistance or support from medical authorities. When the medical practitioners first began using the eHealth information system, the majority of them did not received introduction or explanation of the security and privacy measures that were implemented to protect the sensitive patient medical data and information that was being stored in the eHealth information system. Lastly, the majority of the medical practitioners in this study strongly agreed that the information system required a variety of significant and critical enhancements to the system functionalities. These system functionalities are to be integrated and upgraded to the current implementation based on their working experience with the system and its effectiveness in managing patients.

Table 4: Medical practitioners' opinions on challenges and issues in utilizing the eHealth information system

Context Factors	N (%)
Frequently run into recurring technical issues when using the eHealth system	
Always	54 (14,6)
Very often	123 (33,3)
Sometimes	133 (36,0)
Rarely	48 (13,0)
Never	11 (3,0)
Frequently got assistance or technical support from the medical authorities when faced with technical difficulties	
Always	46 (12,5)
Very often	74 (20,2)
Sometimes	91 (24,8)
Rarely	97 (26,4)
Never	59 (16,1)
Presented with different privacy and security measures to safeguard the patients' sensitive medical data stored in the eHealth system	
To a great extent	53 (14,4)
Somewhat	93 (25,3)
Very little	89 (24,3)
Not at all	132 (36,0)
Significant functional improvements are required for the eHealth system	
Strongly agree	210 (56,9)
Agree	82 (22,2)
Undecided	60 (16,3)
Disagree	14 (3,8)
Strongly disagree	3 (0,8)

Figure 3 presents an introduction to the various security measures implemented concerning the security and privacy protection of the patient's sensitive medical data by the practitioners' gender, distribution within the healthcare sectors, and the location of their medical facility.

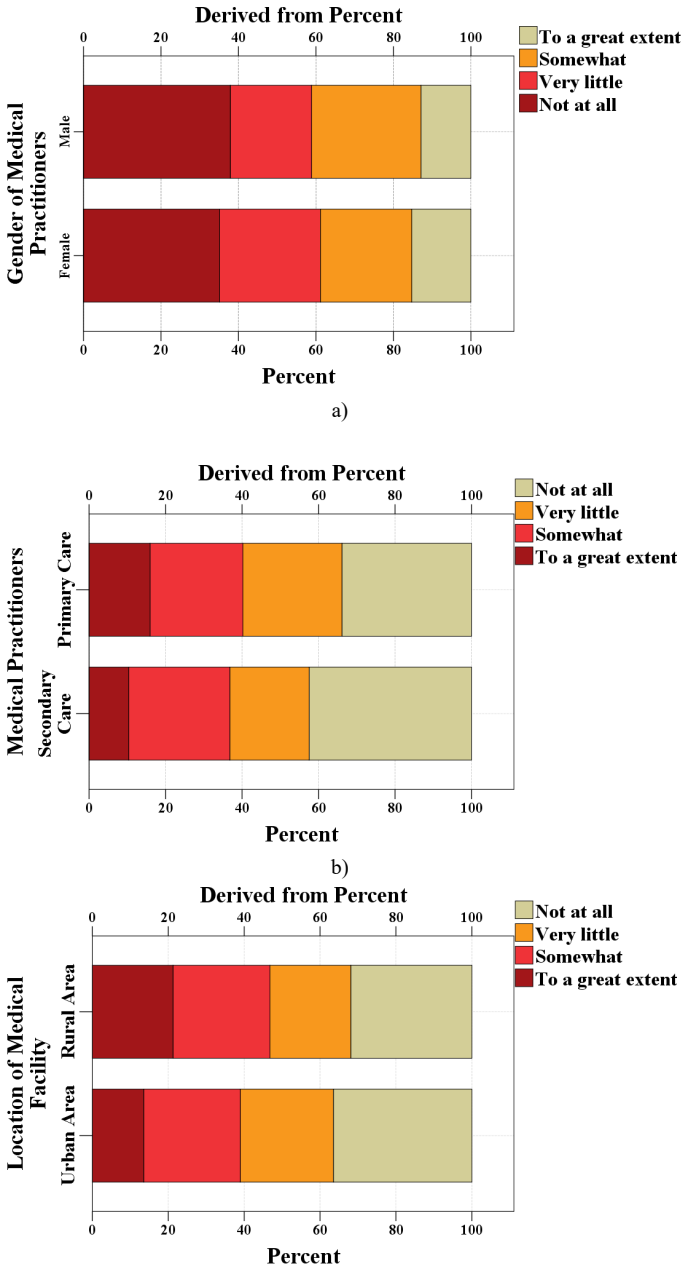


Figure 3: Various privacy and security measures in relation to protecting patients' sensitive medical data presented by a) Practitioners' gender; b) Distribution within the healthcare sectors; and c) Location of their medical facility

A crosstabulation between the variables presented various security measures to safeguard the patients' sensitive medical data stored in the system and often receives technical support and assistance when encountering technical difficulties while working in the system is presented in Table 5. A greater number of the medical practitioners than anticipated did not receive any assistance or support from the medical authorities, nor were they informed of the numerous privacy and security procedures

implemented to safeguard patients' private medical information. The number of medical practitioners who occasionally, frequently, and consistently received assistance and support from the medical authorities was lower than anticipated. Additionally, there was a complete lack of introduction of security protocols regarding the safeguarding of patients' sensitive medical data.

Table 5: Crosstabulation of the variables Often receive assistance and technical support from the medical authorities and Presented with various privacy and security measures

Frequently got assistance or technical support from the medical authorities when faced with technical difficulties		Presented with different privacy and security measures to safeguard the patients' sensitive medical data stored in the eHealth system				
		Not at all	Very little	Somewhat	To a great extent	Total
Never	Count	32	12	6	9	59
	Expected Count	21,2	14,4	15,1	8,3	59,0
Rarely	Count	39	28	23	6	96
	Expected Count	34,5	23,5	24,5	13,5	96,0
Sometimes	Count	28	25	30	7	90
	Expected Count	32,4	22,0	23,0	12,6	90,0
Very often	Count	18	20	24	11	73
	Expected Count	26,3	17,8	18,7	10,2	73,0
Always	Count	14	4	10	18	46
	Expected Count	16,6	11,2	11,8	6,4	46,0
Total	Count	131	89	93	51	364
	Expected Count	131,0	89,0	93,0	51,0	364,0

Certain medical practitioners were somewhat introduced with different security and privacy preferences taken in regard to protecting the patients' sensitive data, however, they never, rarely, or always received any aid and support from the medical authorities when was required. There were more medical practitioners who sometimes and frequently got assistance and support from the medical authorities than expected. They were somewhat introduced to different privacy and security measures that were taken in relation to protecting the sensitive patients' data.

There were certain medical practitioners who rarely and sometimes received assistance and support from the medical authorities than to be expected, that to a great extent were introduced to different privacy and security measures taken in relation to protecting the sensitive patients' medical data. There were more medical practitioners who never, frequently, and always received assistance and support from the medical authorities than to be expected, that to a great extent were introduced with different privacy and security measures taken in relation to protecting the sensitive patients' medical data.

There was a statistically significant difference (X^2=53,194, p=0,000) between the opinions that introduction to different privacy and security measures to safeguard the sensitive medical data of the patients and often received assistance and support from the medical authorities. Since the p-value in this instance was less than the alpha standard value of 0,05, the null hypothesis, which stated that the two variables were independent of each other was rejected. The data indicates that the variables' introduction to different privacy and security measures for the protection of the sensitive data of the patients and frequently receiving assistance and support from the medical authorities were related to one another. The effect of this association was almost moderate (0,221), meaning that the assistance provided by the medical authorities may have had some role in the way in which the medical practitioners answered the given questions. The remaining variables, such as age group, marital status, location of the medical facility, education, gender, employment history, etc., did not reveal any significance of impact on the opinions of the medical practitioners.

4. Discussion

The in-depth interviews and online surveys were the two main research methodologies used in this study to examine the difficulties and concerns related to privacy and security when handling the sensitive information of the patients. The fact that a wide range of medical practitioners with different specialties (Table 3) offered their opinions on the use of the eHealth information system enhanced the level of quality and importance of this research. The government and medical authorities have the responsibility of authorizing the adoption of these eHealth systems without giving any thought to the security integrity of the private medical data of the patients [38]. For comparable eHealth information systems in other developing nations, this study would likewise present relevant data and support the government leaders and medical authorities.

This study was carried out among primary and secondary care practitioners (Table 3) to find out what they thought of the privacy and security features of the system, what was lacking, and what features and services should be enhanced to provide patients with better treatment. The answers to the survey revealed that the experiences of the medical practitioners with the security measures implemented by the medical authorities varied greatly. A significant portion of the survey participants, who believed that the different privacy and security measures were not sufficiently introduced and presented to them, emphasized the needs for improving the awareness and training programs to be initiated more frequently (Table 4). By filling in these knowledge gaps with different training and seminars, the healthcare authorities should ensure that the medical practitioners have the skills required to safeguard and protect the patient's sensitive data and respect the privacy and protection legislation [39].

Data interoperability requirements and protecting the privacy and security of sensitive patient medical data are potential challenges when working with external documents associated with

medical data exchange between primary and secondary care. It may be challenging to integrate the patients' medical data records or other healthcare-related documents into the broader eHealth system since they may be kept in several external systems at various locations. To address this challenge and issue with interoperability, appropriate funding is required for collaboration with external systems and data integration technologies to ensure operational sharing of medical sensitive data. As was already indicated, an additional difficulty was guaranteeing the privacy and security of the patient's medical data because, the system holds the personal data of the patients, such as their insurance information and medical history. Considering this, these various healthcare systems need to implement the necessary safeguards to protect sensitive data from unauthorized access and cyber threats [23].

As an example, across the user experience of patients, medical practitioners, and other end-users may have greatly impacted to the integration of privacy and security measures in the national eHealth information systems. From one perspective, robust security and privacy measures may enhance the trust in the healthcare system and the end-user confidence by providing reassurance that their private medical information is being shielded from unwanted access, disclosure, and use. This would improve the overall user experience as they would be more likely to feel content and at ease in using the eHealth system and its features. On the other hand, inadequate privacy and security measures may lead to unfavorable user experiences because the end-users could be concerned that their private medical information which must remain confidential with implemented safeguards may be compromised or that too strict safety measures would make it impossible for them to access or share their data. This could cause distrust, annoyance, and a reluctance to utilize the eHealth information system [23, 39].

The Law on Health Protection includes regulations for the security of patient data specifically for the healthcare industry. Because of this, the health data is regarded as sensitive information and is given extra protection under Macedonian law. However, there may be several obstacles to effectively implementing and enforcing these regulations and laws in day-to-day life, not just in N. Macedonia, but in any other country. The Macedonian healthcare practitioners must understand their responsibilities under the data privacy laws and take the appropriate safety measures to safeguard the patient's sensitive medical data. The EU has regulations in place to protect individuals' rights when it comes to the gathering and usage of personal sensitive data. These regulations include the General Data Protection Regulation (GDPR) and the Directive on the Protection of Individuals concerning the processing of personal data by adequate authorities, both of which have as their main objectives the prevention, investigation, detection, prosecution, or the implementation of penalties for criminal offenses. For these reasons, all nations, especially N. Macedonia, must ensure that the right policies are implemented to safeguard the rights of the population and that they always remember the ethical principles when handling medical data It meant guarantee that the data was being collected and used openly, that sufficient organizational and technical security measures were being put in place, and that people's consent was being appropriately requested. To ensure that people are not treated unfairly in light of the findings of big data analysis, it was also crucial to address concerns of prejudice and discrimination when analyzing the medical data [29, 30, 31, 32].

This study also assessed the perceptions of medical practitioners on the functionality of the eHealth information system. This was crucial since system updates could enhance both the user experience and the overall security of the system (Table 4). By concentrating on the precise areas where the medical practitioners believe that improvements are necessary, the IT staff can focus their efforts on resolving these issues and with that offering better security measures.

This research adds to the amount of knowledge already available on privacy and security issues in healthcare systems especially in medical information systems. It aligns with the increased focus on patient privacy, data security, and compliance with legal frameworks such as the GDPR and the Health Insurance Portability and Accountability Act (HIPPA). Enhancing the security and privacy of eHealth systems is imperative, so it's crucial to identify useful insights the parties involved may apply to improve eHealth systems [40].

This study has its limitations because it concentrated on the use of the "Moj Termin" eHealth information system that is currently in use in N. Macedonia. Second, there was a lack of a straightforward open discussion between the interviewer and the participants because the in-depth interviews were conducted remotely via e-mail. Thirdly, the online survey sample utilized in this study was restricted to only Macedonian healthcare professionals working in primary and secondary healthcare level, as such, it may not accurately reflect the majority of medical practitioners both inside and outside the nation. Thus, more research is required to examine the boundaries of the various eHealth systems in use in other more developed countries, as well as to find out what many relevant medical and technical professionals think about the issue at hand, which directions should be pursued, and how to provide further upgrades for the eHealth systems to develop properly.

5. Conclusion

The study's findings demonstrate the urgent need for better security and privacy safeguards when medical practitioners utilize the eHealth systems. The healthcare authorities and IT companies must act quickly in response to the recognized difficulties with the eHealth systems, the lack of expertise and knowledge of the current security and privacy standards. By resolving these issues, medical practitioners' confidence in the security of protecting private patient information could be increased. This will eventually aid in the successful integration of the eHealth system technology into the delivery of overall healthcare.

Integrating privacy and security safeguards into the national healthcare and medical information systems may have a substantial impact on how patients, medical practitioners, and other end-users interact with the system. Users would therefore likely feel more satisfied and comfortable utilizing the eHealth information system and its features, which would enhance the user experience. Inadequate privacy and security measures, however, may have the opposite negative effect. End-users may worry that their personal medical data won't be secure, that the security measures in place will be compromised, or that they won't be able

to access or share their data easily because of overly strict regulations and security measures. This could cause mistrust, reluctance, and annoyance to utilize the eHealth system. In consideration of this, integrating sufficient security and privacy features into the system would be essential to maintaining the confidence of the patients. It's also essential to implement this under the relevant regulations. Privacy and security measures should be included in the design and implementation of eHealth systems in a way that creates a balance between the need for data protection, usability and accessibility. The following are some suggestions for a user-centered integration of security and privacy safeguards into national medical information systems, as well as for presenting these ideas as a kind of implementation roadmap: Start a user research project and utilize the results to assist the medical organization in better understanding the user requirements, expectations, and preferences regarding privacy and security safeguards that are in line with those demands; Healthcare organization should involve the users in the design process of privacy and security measures. Through user testing and other feedback sessions, they can ensure that the implemented measures are clear, in line with expectations and requirements, and easy to use; Provide users with accurate and transparent information about the privacy and security safeguards implemented through data usage agreements and privacy policies. This would increase users' trust in the system and increase their level of satisfaction in general; User-friendly authentication techniques, like single log-in or biometric authentication, to provide users with easier, secure system access; and finally, monitor and assess the effectiveness of the privacy and security safeguards that have been put in place to make any necessary adjustments along the road to ensure that they continue to fulfill the requirements and expectations of the end-users.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] V. Denkovski, I. Stojmenovska, G. Gavrilov, V. Radevski, V. Trajkovik, "Investigating Privacy and Security Concerns in a Running eHealth Information System," in 2023 IEEE International Mediterranean Conference on Communications and Networking (MeditCom), 39-44, 2023, doi:10.1109/MeditCom58224.2023.10266398.

[2] T. O. Abolade, "The benefits and challenges of e-Health applications in developing nations: A review," in 14th iSTEAMS Conference, 14, 37-44, 2018.

[3] B. Atanasovski, M. Bogdanovic, G. Velinov, L. Stoimenov, D. Sahpaski, I. Skrceska, M. Kon-Popovska, D. Jankovic, B. Jakimovski, "Transforming an Enterprise E-Health Syste,m from Process Oriented to Model Driven Architecture," in 7th International Conference on Information Society and Technology ICIST, 159-162, 2021.

[4] S. Olsson, A. Lymberis, D. Whitehouse, "European Commission activities in eHealth", International Journal of Circumpolar Health, **63**(4), 310–316, 2004, doi:10.3402/ijch.v63i4.17747.

[5] P. G. Svensson, "eHealth Applications in Health Care Management," Ehealth international, **1**(1), 5, 2002, doi:10.1186/1476-3591-1-5.

[6] S. Faber, M. van Geenhuizen, M. de Reuver, "eHealth adoption factors in medical hospitals: A focus on the Netherlands," International journal of medical informatics, 100, 77–89, 2017, doi:10.1016/j.ijmedinf.2017.01.009.

[7] R. Mitrovska, V. Pachovski, A. Bozinovski, "Challenges, benefits and expectations from implementing an e-health system in Macedonia," in 2015 Proceedings of the 12th International Conference on Informatics and Information Technologies (CiiT 2015), 268-272, 2015.

[8] G. Dussault, L. Lapão, "The contribution of ehealth and mhealth to improving the performance of the health workforce: a review," Public Health Panorama, 3(3), 463-471, 2017.

[9] A. La Rocca, T. Hoholm, "Coordination between primary and secondary care: the role of electronic messages and economic incentives," BMC Health Serv Res 17, 149, 2017, doi: 10.1186/s12913-017-2096-4.

[10] R. A. Tariq, P.B. Hackert, "Patient Confidentiality", in StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing, 2023, Available from: https://www.ncbi.nlm.nih.gov/books/NBK519540/, [Last Accessed: January 3, 2024].

[11] S. Khanra, A. Dhir, A.K.M.N. Islam, M. Mäntymäki, "Big data analytics in healthcare: a systematic literature review", Enterprise Information Systems, **14**(7), 878–912, 2020, doi: 10.1080/17517575.2020.1812005.

[12] S. Kumar, M. Singh, "Big data analytics for healthcare industry: impact, applications, and tools", Big Data Mining and Analytics, **2**(1), 48–57, 2019, doi: 10.26599/BDMA.2018.9020031.

[13] European Commission, "Exchange of Electronic Health Records across the EU", Shaping Europe's digital future, Policy, 2021, Available from: https://ec.europa.eu/digital-single-market/en/exchange-electronic-health-records-across-eu#:~:text=The%20eHDSI%20connects%20eHealth%20national,such%20health%20information%20by%202021, [Last Accessed: January 3, 2024].

[14] World Health Organization Regional Office for Europe, "From innovation to implementation eHealth in the WHO European Region", WHO Regional Office for Europe, 2016, Available from: https://intranet.euro.who.int/__data/assets/pdf_file/0012/302331/From-Innovation-to-Implementation-eHealth-Report-EU.pdf, [Last Accessed: January 3, 2024].

[15] T. Yuan, "Towards the development of best data security for big data", Communications and Network, **9**(4), 291-301, 2017, doi: 10.4236/cn.2017.94020.

[16] P. Jain, M. Gyanchandani, N. Khare, "Big data privacy: a technological perspective and review", Journal of Big Data, **3**(1), 25, 2016, doi: 10.1186/s40537-016-0059-y.

[17] U. Sivarajah, M.M. Kamal, Z. Irani, V. Weerakkody, "Critical analysis of big data challenges and analytical methods", Journal of Business Research, **70**, 263–286, 2017, doi: 10.1016/j.jbusres.2016.08.001.

[18] A. Oussous, F.Z. Benjelloun, A. Ait Lahcen, S. Belfkih, "Big data technologies: A survey", Journal of King Saud University - Computer and Information Sciences, **30**(4), 431-448, 2017, doi: 10.1016/j.jksuci.2017.06.001.

[19] H. Ye, X. Cheng, M. Yuan, L. Xu, J. Gao, C. Cheng, "A survey of security and privacy in big data", in 2016 16th International Symposium on Communications and Information Technologies (ISCIT), 268-272, 2016.

[20] The Medical Imaging Technology Association (MITA), "DICOM® — Digital Imaging and Communications in Medicine", 2024, Available from: https://www.dicomstandard.org/, [Last Accessed: January 4, 2024].

[21] The National Library of Medicine, "SNOMED CT a standard for electronic exchange of clinical health information", 2024, Available from: https://www.nlm.nih.gov/healthit/snomedct/index.html, [Last Accessed: January 4, 2024]

[22] The FHIR Core Work Group, "FHIR ®© HL7.org 2011+. FHIR R5 hl7.fhir.core#5.0.0", 2024, Available from: https://hl7.org/fhir/overview.html, [Last Accessed: January 4, 2024].

[23] A. Torab-Miandoab, T. Samad-Soltani, A. Jodati, P. Rezaei-Hachesu, "Interoperability of heterogeneous health information systems: a systematic literature review", BMC medical informatics and decision making, 23(1), 18, doi:10.1186/s12911-023-02115-5.

[24] I. Khan, G. Xitong, Z. Ahmad, F. Shahzad, "Investigating factors impelling the adoption of e-Health: A perspective of African expats in China," SAGE Open, **9**(3), 2019, doi: 10.1177/2158244019865.

[25] A. H. Seh, M. Zarour, M. Alenezi, A. K. Sarkar, A. Agrawal, R. Kumar, R. A. Khan, "Healthcare Data Breaches: Insights and Implications," Healthcare (Basel, Switzerland), **8**(2), 133, 2020, doi:10.3390/healthcare8020133.

[26] D. El Majdoubi, H. El Bakkali, S. Sadki, Z. Maqour, A. Leghmid, "The systematic literature review of privacy-preserving solutions in smart healthcare environment," Security and Communication Networks, 1–26, 2022, doi:10.1155/2022/5642026.

[27] M. Javaid, A. Haleem, R. P. Singh, R. Suman, "Towards insighting cybersecurity for Healthcare Domains: A comprehensive review of recent practices and Trends," Cyber Security and Applications, 1, 2023, doi:10.1016/j.csa.2023.100016.

[28] M. Pyrrho, L. Cambraia, V. F. de Vasconcelos, "Privacy and Health Practices in the Digital Age", The American journal of bioethics : AJOB, **22**(7), 50–59, 2022, doi:10.1080/15265161.2022.2040648.

[29] Council of Europe, "Convention 108+ Convention for the protection of individuals with regard to the processing of personal data", 2018, Available from: https://coe.int/en/web/data-protection/convention108-and-protocol, [Last Accessed: January 4, 2024].

[30] Council of Europe, "Modernization of convention 108 - data protection -", 2024, Availabe from: https://www.coe.int/en/web/data-protection/convention108/modernised, [Last Accessed: January 4, 2024].

[31] Official Gazette of the Republic of North Macedonia no. 42/20 and 294/21, "DP Law", 2020, [Last Accessed: January 4, 2024].

[32] DLA PIPER., "Data Protection Laws of the World North Macedonia", 2022, Available from: https://www.dlapiperdataprotection.com/index.html?t=law&c=MK#:~:text=The%20DP%20Law%20defines%20personal,his%20or%20her%20personal%20identification, [Last Accessed: January 4, 2024].

[33] J. E. Hawkins, "The practical utility and suitability of email interviews in qualitative research," The Qualitative Report, 23(2), 493-501, 2018, doi: 10.46743/2160-3715/2018.3266.

[34] R. L. Fritz, R. Vandermause, "Data collection via in-depth email interviewing: Lessons from the field," Qualitative Health Research, **28**(10), 1640–1649, 2018, doi: 10.1177/1049732316689067.

[35] D. M. Levine, D. F. Stephan, A. S. Kathryn, Statistics for managers: Using Microsoft excel, 9th edition, Pearson, 2018.

[36] Health Insurance Fund of Republic of North Macedonia, "Code of physicians", 2022, Available from: https://fzo.org.mk/en/node/1126, [Last Accessed: January 4, 2024].

[37] SPSS Inc., "IBM SPSS Statistics", 2023, Available from https://www.ibm.com/products/spss-statistics, [Last Accessed: January 4, 2024].

[38] H. C. Ossebaard, L. Van Gemert-Pijnen, "eHealth and quality in health care: implementation time," International journal for quality in health care: journal of the International Society for Quality in Health Care, **28**(3), 415–419, 2016, doi:10.1093/intqhc/mzw032.

[39] A.T. Alanazi, "Clinicians' Perspectives on Healthcare Cybersecurity and Cyber Threats", Cureus, 14, **15**(10), e47026, 2023, doi: 10.7759/cureus.47026.

[40] P. Yamcharoen, O. S. Folorunsho, A. Bayewu, O. E. Fatoye, "Advancing Healthcare Security: Developing a Composite Set of Cybersecurity Requirements for the Healthcare Industry", Computing, Information Systems, Development Informatics & Allied Research Journal, **14**(1), 9-20, 2023, doi: 10.22624/AIMS/CISDI/V14N1P2.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 10, No. 1, 01-06 (2024)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences 2024

ASTESJ

ISSN: 2415-6698

Lightning Detection System for Wind Turbines Using a Large-Diameter Rogowski Coil

Sarawuth Pramualsingha*, Kazuo Yamamoto, Rikuto Tanaka

Graduate School of Electrical and Electronic Engineering, Chubu University, Kasugai, 487-8501, Japan

ARTICLE INFO

Article history:
Received: 15 March, 2024
Revised: 31 December, 2024
Accepted: 01 January, 2025
Online: 16 January, 2025

Keywords:
Lightning detection
Wind turbines
Large-diameter Rogowski coil
Wideband integrator
High-current measurement

ABSTRACT

A lightning detection system based on a large-diameter Rogowski coil and an analog integrator was developed for wind turbine applications and is presented in this paper. To accurately detect lightning current, the Rogowski coil was designed with a lower cutoff frequency of 0.1 Hz. The analog integrator, comprising an inverting active integrator, and an amplifier, was used to restore the original waveform of the lightning current. Tests were conducted to verify the bandwidth of the system using alternating current, damped sinusoidal current, and a rectangular pulse. The system demonstrated a bandwidth of 0.1 Hz to 100 kHz, which is within the standard acceptable range. This lightning detection system offers a significantly lower cost compared to commercial alternatives, making it highly suitable for practical applications.

1. Introduction

For sustainable development and achieving carbon neutrality, wind turbines are increasingly being used for power generation [1]. Over the past decade, Japan has been the second-largest wind power installer in the world, with a cumulative capacity of 5,213 MW [2]. Wind turbines operate effectively in areas with strong winds, such as coastal or offshore locations, but this also places them at a high risk of being struck by lightning [3]-[5]. It is important to stop wind turbines and perform manual inspections after a lightning event is detected to prevent further damage and ensure safety. To achieve this, a lightning detection system (LDS) is deployed to identify such events.

As stated by the JEM and IEC standards [6],[7], the LDS consisting of a large-diameter Rogowski coil and an analog integrator is one of the alternative methods for the detection of lightning events. In practice, this coil is securely installed at the wind turbine tower near the foundation, wrapped around its exterior. This configuration allows the detection of impulse current from the tower.

Rogowski coil induces voltage is proportional to the time derivative of the lightning current (i_l), and its transfer function exhibits a positive slope gain. To retrieve the original shape of the lightning current i_l and maintain its gain constantly, an analog integrator is used. It has a negative slope gain in its transfer function, which cancels out the slope gain of the Rogowski coil.

In this paper, a cost-effective lightning detection system for wind turbine applications is presented. The system has a bandwidth of 0.1 Hz to 100 kHz and comprises a large-diameter Rogowski coil and an analog integrator. The design details are presented in Section 2. Frequency characteristic tests were performed to ensure sufficient bandwidth and are presented in Sections 3 and 4. Finally, the conclusion is presented in Section 5.

2. Lightning Detection System

The possibility of a lightning strike at the blade tip of wind turbines is high, as it typically becomes the highest point during operation. A permanently mounted air-termination system, known as a receptor, serves as the contact point between wind turbines and lightning [7]. Once lightning strikes, lightning current i_l flows from the receptor, a down-conducting system, wind turbine tower, and eventually reaches an earth-termination system at a foundation, as illustrated in Figure 1.

As lightning current i_l flows to the earth-termination system, it generates a time-varying magnetic field (B) around its conductor. This enables current sensors, such as the Rogowski coil, to detect lightning events if installed anywhere along this current path.

Commercially available Rogowski coil-based LDS units can cost a million yen or more each. In contrast, the developed LDS has a total cost of approximately 33,000 yen, comprising 23,000

*Corresponding Author: Sarawuth Pramualsingha, Chubu University, sarawuth.pr@kmitl.ac.th

https://dx.doi.org/10.25046/aj100101

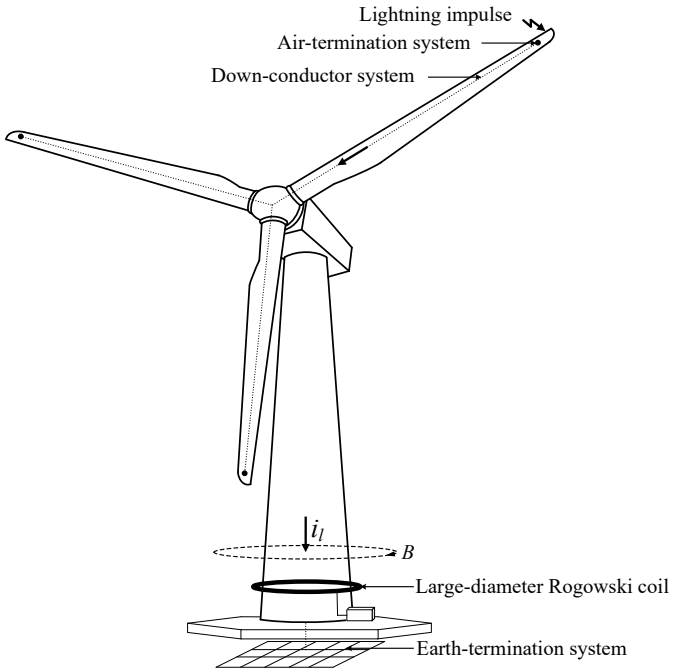


Figure 1: Wind turbine lightning protection system components and the location of the Rogowski coil-based lightning detection system [7].

yen for the large-diameter Rogowski coil and 10,000 yen for the analog integrator making them significantly cost-effective.

Since the output voltage from the Rogowski coil feeds into the integrator, the frequency characteristics of the LDS can be analyzed by multiplying the Rogowski coil transfer function $H_r(s)$ and the integrator transfer function $H_i(s)$ as in (1)

$$H_l(s) = H_r(s) \cdot H_i(s) \tag{1}$$

Details for obtaining the Rogowski coil and integrator transfer function are as follows.

2.1. Large-diameter Rogowski coil

Rogowski coils were introduced in 1912 by German physicist [8]. For simplicity in investigation, it is modeled using lumped parameters, as illustrated in Figure 2, where M represents mutual inductance, L_s represents self-inductance, R_s represents self-resistance, C_s represents stray capacitance, R_d represents damping resistance, i_2 represents coil loop current, and v_r represents the coil output voltage [9]. Applying Kirchhoff's voltage law, loop voltage is

$$M\, di_l(t)/dt = L_s\, di_2(t)/dt + R_s i_2(t) + v_r(t) \tag{2}$$

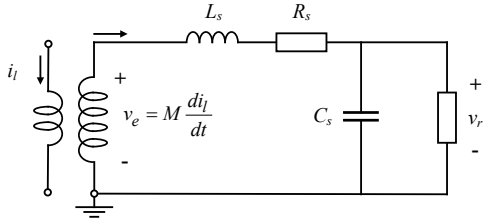


Figure 2: Lumped parameter equivalent circuit of the large-diameter Rogowski coil [9].

Assume that $L_s \cdot di_2(t)/dt \gg R_s \cdot i_2(t) + v_r(t)$, and stray capacitance C_s are neglected, so voltage drops across self-resistance R_s and damping resistance R_d can be neglected. Hence, loop current $i_l(t) \approx L_s \cdot i_2(t)/M$. This type of coil is known as a self-integrating Rogowski coil, suitable for detecting very short-duration current pulses [10].

To obtain the coil output voltage in the time domain $v_r(t)$, apply Kirchhoff's current law, and the current loop is

$$i_2(t) = C_s\, dv_r(t)/dt + v_r(t)/R_d \tag{3}$$

Substitute (3) into (2) and then apply integration with respect to time (t). In practice, stray capacitance C_s and L_s / R_d are assumed to be approximately zero. The voltage drops across self-inductor L_s and the current consumed by stray capacitor C_s can, therefore, be neglected. Thus,

$$i_l(t) \approx (R_s + R_d/R_d\, M) \int_{t=0}^{t=\infty} v_r(t)dt \tag{4}$$

From (4), lightning current i_l can be reconstructed using an external integrator. This type of Rogowski coil is known as an external-integrating Rogowski coil, which is suitable for relatively long-duration current pulses and was utilized as part of LDS in this study [11].

The frequency characteristics of Rogowski coils can be determined from their transfer function by substituting (3) into (2) and applying Laplace transform. The transfer function of the Rogowski coil, $H_r(s)$, is

$$H_r(s) = sMR_d \Big/ \left(s^2 L_s C_s + sL_s + sR_s C_s R_d + R_s + R_d\right) \tag{5}$$

Rogowski coil is a helical toroidal coil of wire. Lead from one end returns to another end through the center of the helical coil. Hence, both terminals are positioned at the same end. Mutual inductance M, self-inductance L_s, stray capacitance C_s, and self-resistance R_s can be calculated from its dimensions as in (6) to (9), respectively [12]-[15].

$$M = (\mu_0 N/2)\left(\sqrt{r_o} - \sqrt{r_i}\right)^2 \tag{6}$$

$$L_s = MN \tag{7}$$

$$R_s = \frac{\rho}{A}\left[\pi(r_i + r_o) + N\pi\sqrt{\left((r_o + r_i)/N\right)^2 + (r_o - r_i)^2}\right] \tag{8}$$

$$C_s = 4\pi^2 \varepsilon_0 (r_o + r_i) \Big/ \log_{10}\left((r_o + r_i)/(r_o - r_i)\right) \tag{9}$$

Where r_i is the coil inner radius, r_o is the coil outer radius, N is the number of turns, u_0 is the permeability of free space ($4\pi \times 10^{-7}$ H/m), ε_0 is the permittivity of free space (8.85×10^{-12} F/m), and ρ is the resistivity of the coil wire.

The developed large-diameter Rogowski coil has an inner radius r_i of 2 m, allowing it to wrap around the tower of a wind

turbine. To achieve a bandwidth of 0.1 Hz to 100 kHz, Rogowski coils are designed with an outer radius r_o of 2.04 m and 1400 turns. The coil, made from commercially available wire-reinforced PVC ducting with a diameter of 40 mm, costs approximately 23,000 yen. Aluminum wire is used for the coil, with a cross-sectional area of 1.78×10^{-8} m² and resistivity ρ of 2.7×10^{-8} Ω/m. The damping resistance R_d is derived from the impedance of an RG-58 50 Ω coaxial cable, resulting in an infinite damping resistance R_d. From (6) to (9), mutual inductance M, self-inductance L_s, stray capacitance C_s, and self-resistance R_s are calculated as 15.65 µH, 0.22 mH, 766 pF, and 291.3 Ω, respectively. The gain and phase characteristics of the coil, analyzed by (5), are shown in Figure 4 and 5 respectively. It exhibits a linear gain slope of approximately 20 dB/decade and a phase shift of roughly -90 degrees for a frequency range from 0.1 Hz to 100 kHz, which satisfies the target range. The external integrator used with the Rogowski coil has been designed to cancel out the coil gain slope, and the detailed design is presented in the following section.

2.2. Analog Integrator

Analog integrators for the Rogowski coil serve two purposes: to regain the original shape of the lightning current i_l and to maintain a constant gain. In previous studies, integrators were usually designed with two poles and one or no zeros [10],[16],[17]. The interaction between poles can lead to more complex dynamic behavior in the integrator [18]. The analog integrator used in this study is shown in Figure 3. It is designed to have a single pole with no zero, resulting in more predictable and stable behavior. The analog integrator consists of an inverting active integrator and an amplifier. Bias resistors R_2 and R_3 minimize the effect of offset voltage, while R_5 and R_8 limit the bias current. The amplifier circuit is utilized to make it more effective for the digital recorder to record the output voltage v_o.

Let V_n and V_p denote inverting and non-inverting operational amplifier (Op-amp) terminal voltages, respectively, with respect to ground. I_n and I_p denote bias currents entering inverting and non-inverting Op-amp terminals, respectively. For simplicity in analysis, let $V_p \approx V_n$, $I_p \approx 0$, and $I_n \approx 0$ [19]. The transfer function of the developed analog integrator, $H_i(s)$, is

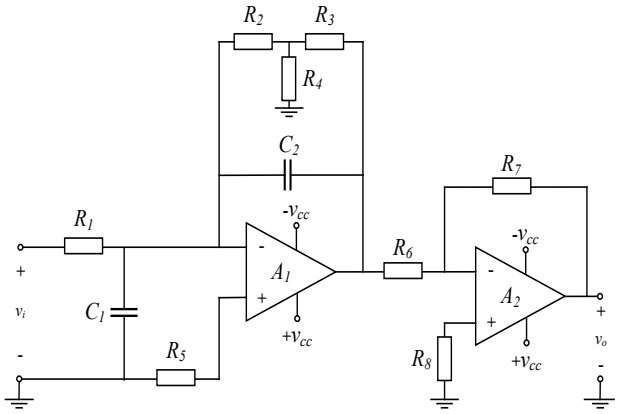


Figure 3: Developed analogue integrator.

$$H_i(s) = \frac{R_3/R_1 + R_2/R_1 + R_2R_3/R_1R_4}{1 + s\left(C_2R_3 + C_2R_2 + C_2R_2R_3/R_4\right)} \cdot \frac{R_7}{R_6} \tag{10}$$

To achieve a gain of about -20 dB/decade and phase shift of -90 degrees from a frequency of 0.1 Hz to 100 kHz, circuit parameters for R_1, C_1, R_2, R_3, R_4, R_5, C_2, R_6, R_7, and R_8 were set to 10 kΩ, 2.25 µF, 200 kΩ, 200 kΩ, 100 kΩ, 6.8 kΩ, 2.25 µF, 10 kΩ, 100 kΩ, and 10 kΩ, respectively. Resulting in a pole at approximately -0.556, ensuring the stability of the system. LT112 operational amplifiers were used for both Op-amps A_1 and A_2 due to their high slew rate of 60 V/µs. Power supplies were used. An output voltage was set to 15 V_{DC} and -15 V_{DC} for V_{cc} and V_{ee}, respectively. The overall cost of assembling the analog integrator was approximately 10,000 yen. The frequency characteristics of the developed integrator, analyzed by (10), show a linear gain slope of -20 dB/decade and a phase shift of approximately -90 degrees as illustrated in Figure 6 and 7 respectively for the frequency range of 0.1 Hz to 100 kHz, which satisfies the target range.

3. LDS Frequency Characteristic Test

Frequency characteristic tests were performed to verify that LDS is capable of measuring lightning current i_l without attenuation or phase error. To ensure that each component in the LDS is functioning as designed, three separate frequency characteristic tests were conducted: a bandwidth test on the large-diameter Rogowski coil, the analog integrator, and the LDS.

To demonstrate the alignment between the test results and the analysis, the assessment includes the calculation of the maximum relative error (MRE), defined as

$$MRE = \max_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{\hat{y}_i} \right| \tag{11}$$

where $\hat{y}_i$ and y_i represent the ith element of analysis and test result, and n is the total number of considered data points [20].

3.1. Frequency Characteristic Test of the Rogowski Coil

The test was conducted using alternating current (AC) voltage as recommended by the JEM standard [6]. A sinusoidal waveform from a function generator and a bipolar voltage amplifier were used to generate a variable-frequency voltage. A sufficient bandwidth shunt resistor (R_{sh}) was employed to draw current from the source and serve as a reference current measuring point. Schematic for the test is shown in Figure 8, with the analog integrator removed. The test results for gain and phase shift are presented in Figures 4 and 5 respectively.

At frequencies lower than 100 kHz, the MRE of gain and phase is 0.31 and 0.11, occurring at frequencies of 50 kHz and 1 Hz, respectively, demonstrating good alignment between the analysis and test results. At frequencies higher than 100 kHz, the stray capacitance between each turn of the coil wire becomes significant, impacting the coil characteristics and resulting in large errors between the analysis and test results. However, these frequencies are beyond the intended use and do not affect the overall performance of the LDS.

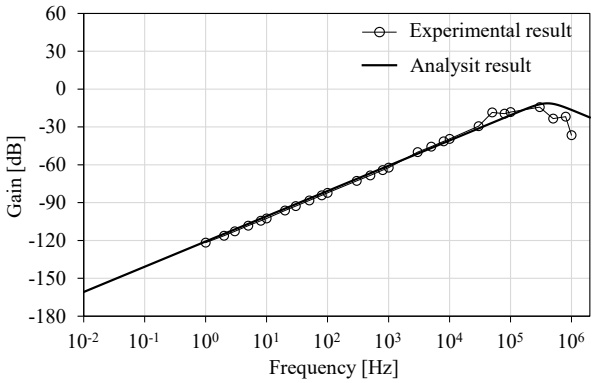


Figure 4: Analysis and test results of gain characteristics of the developed large-diameter Rogowski coil.

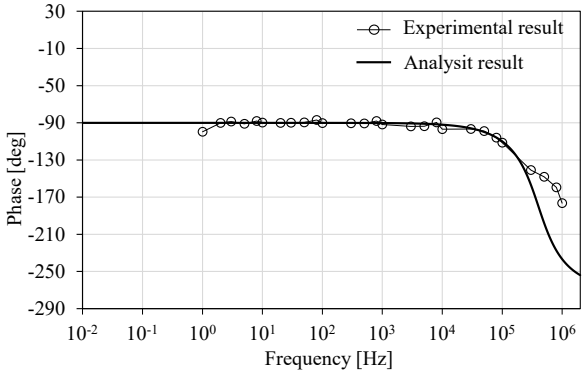


Figure 5: Analysis and test results of phase shift characteristics of the developed large-diameter Rogowski coil.

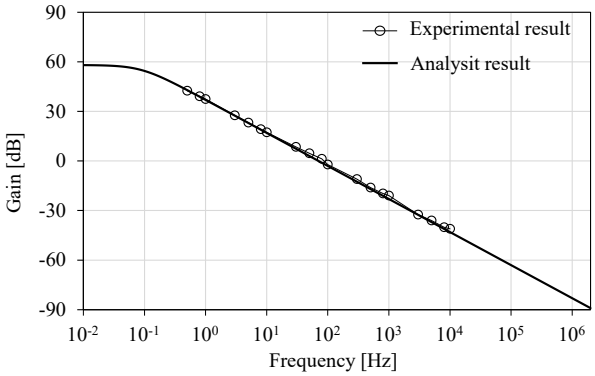


Figure 6: Analysis and test results of gain characteristics of the developed analog integrator.

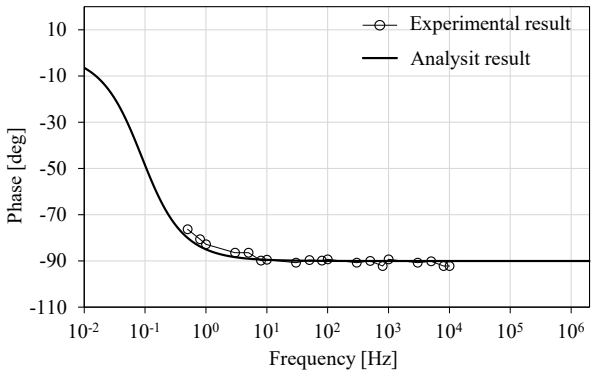
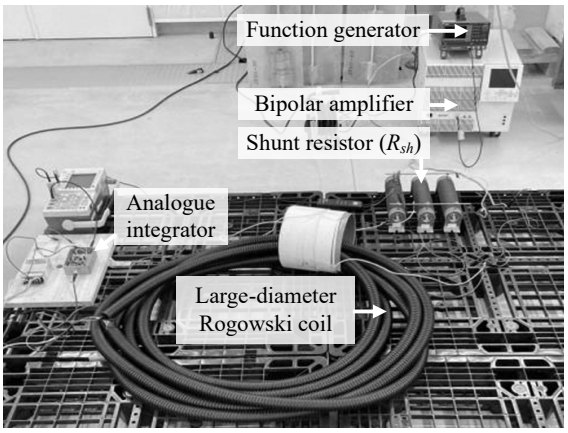
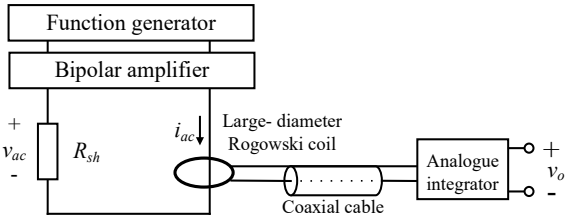


Figure 7: Analysis and test results of phase shift characteristics of the developed analog integrator.



(a)



(b)

Figure 8: Alternating current (AC) power circuit and LDS under test: (a) Actual test setup, (b) Schematic diagram.

3.2. Frequency Characteristic Test of Analogue Integrator

To assess the frequency characteristics of the developed analog integrator, a sinusoidal waveform from a function generator was directly applied to the integrator. Comparing the amplitude of the input and output voltages, the test results for gain and phase shift are presented in Figures 6 and 7 respectively. MSE is 2.25 and 0.04 for gain and phase at frequencies of 80 Hz and 0.8 Hz respectively ensuring well alignment of analysis and the test results.

3.3. Frequency Characteristic Test of LDS

Frequency characteristic testing of LDS was divided into two ranges: below 30 kHz and above 30 kHz. The test schematic and actual test are depicted in Figure 8. Test results for gain and phase shift are shown in Figures 9 and 10, respectively. Analysis using (1) and test results are well aligned. For frequencies below 100 kHz, the MSE of gain and phase is 0.04 and 0.08, occurring at approximately 30 kHz and 100 kHz, respectively. At higher frequencies, the analog integrator components exhibit bandwidth limitations. However, these frequencies are beyond the range of interest and will not affect the LDS performance in practical use. The test details are as follows.

- LDS test at a frequency below 30 kHz, the test was conducted with its schematic and actual test configuration depicted in Figure 8, using AC voltage as a source. Input current and output voltage were measured to determine gain and phase shift.
- LDS test at a frequency above 30 kHz, a relatively high current is required to assess frequency characteristics. To overcome this challenge, tests were conducted using damped alternating current (DAC), the waveform suggested as an alternative by the JEM standard [6].

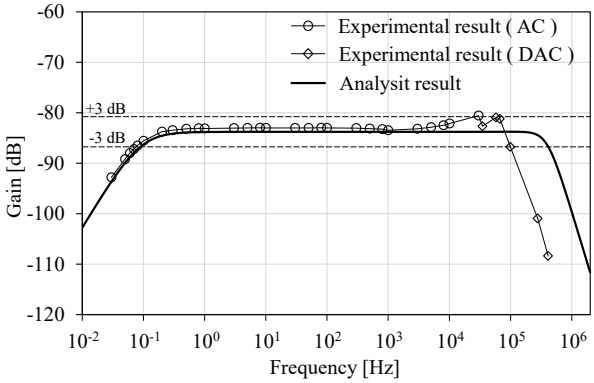


Figure 9: Analysis and test results of gain characteristics of the LDS.

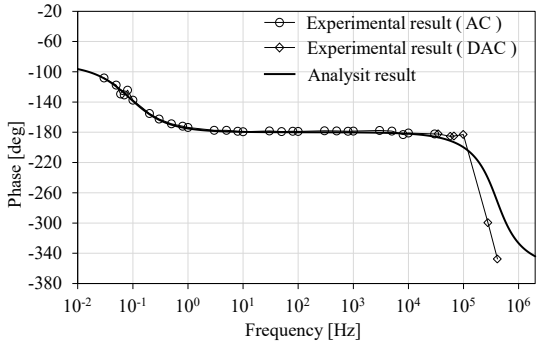


Figure 10: Analysis and test results of phase shift characteristics of the LDS.

4. LDS Low Cutoff Frequency Verification

A low cutoff frequency of LDS is a crucial parameter as it directly impacts measurements of total charge. To verify the cutoff frequency, a rectangular pulse response of LDS is compared to that of a known low-pass filter with a specific cutoff frequency.

Due to its sharp transitions, a rectangular pulse contains a broad spectrum of sinusoidal frequencies. When applied to a low-pass filter, the flat top of the pulse gradually decays toward zero. The higher the cutoff frequency, the faster this decay occurs. To confirm the cutoff frequency of the LDC at 0.1 Hz, a first-order low-pass filter with a cutoff frequency of about 0.08 Hz was selected. The filter was composed of a 15 kΩ resistor and a 130 µF capacitor [21].

The test was conducted as schematic shown in Figure 11. The pulse had an amplitude of approximately 1000 A and a duration of 450 ms. The reference pulse response was obtained by replacing the LDS with the filter. Input current and output voltage were measured, and the results, shown in Figure 12, demonstrate good alignment for the time range of approximately 0 to 450 ms. This confirms that the LDS has a lower cutoff frequency of at least about 0.08 Hz, meeting the intended design specification of 0.1 Hz.

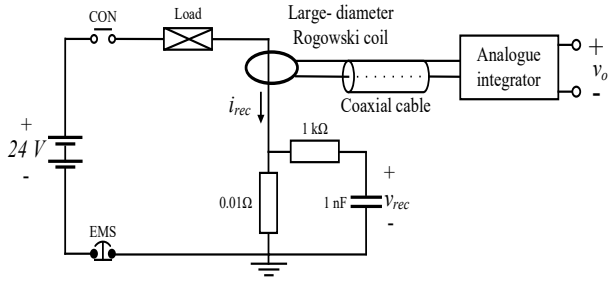


Figure 11: Schematic of rectangular pulse power circuit and LDS under test.

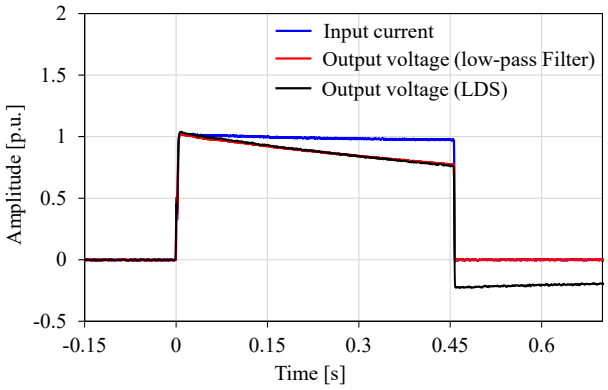


Figure 12: Rectangular pulse response of the LDS and the reference low-pass filter.

5. Conclusion

A lightning detection system for wind turbine applications is presented in this paper. The system consists of the large-diameter Rogowski coil and the analog integrator. It operates with a bandwidth range of 0.1 Hz to 100 kHz and a linear gain of approximately -83 dB. The developed large-dimeter Rogowski coil is designed with a diameter of 4 m, 1400 turns in total, and a slope of 20 dB/decade. The developed analog integrator has a slope gain of -20 dB/decade and consists of the inverting active integrator and an amplifier.

A series of tests were conducted to assess the frequency characteristics of the LDS, particularly the lower and upper cutoff frequencies. LDS frequency characteristics were measured using alternating current (AC) for frequencies below 30 kHz and damped alternating current (DAC) for higher frequencies. The results align well with the analysis, with a maximum gain error of less than 3 dB, confirming the upper cutoff frequency to be 100 kHz. Additionally, the lower cutoff frequency of the LDS was verified using a rectangular pulse response, confirming that it is lower than 0.1 Hz.

The developed LDS complies with the standard required bandwidth while offering significant cost-effectiveness compared to commercial alternatives. This makes the LDS suitable for field applications in lightning detection for wind turbines.

Acknowledgment

Authors would like to give special acknowledgment to High Voltage Laboratory, Graduate School of Electrical and Electronic Engineering, Chubu University for the support and facilities.

References

[1] G. Council, "GWEC global wind report 2019," *Global Wind Energy Council*, **(2017)**: [page numbers missing], 2017

[2] G. Council, "GWEC global wind report 2024," *Global Wind Energy Council*, **(2024)**: [page numbers missing], 2024

[3] K.L. Cummins, "Overview of the Kansas Windfarm 2013 Field Program," *Lightning Detection Conference, Vaisala*, **(2014)**: [page numbers missing], 2014

[4] D. Wang, N. Takagi, T. Watanabe, H. Sakurano, M. Hashimoto, "Observed characteristics of upward leaders that are initiated from a windmill and its lightning protection tower," *Geophysical Research Letters*, **35(2)**: [page numbers missing], 2008

[5] M. Ishii, M. Saito, M. Chihara, D. Natsuno, "Transferred charge and specific energy associated with lightning hitting wind turbines in Japan," *IEEJ Transactions on Power and Energy*, **132(3)**: 294–295, 2012

[6] N.P.G.S. Lightning current detection-Part, "1515: Test methods of Lightning current detection system for wind turbines," *The Japan Electrical Manufacturers' Association, JEM*, **(2023)**: 1515, 2023

[7] W. Turbines-Part, "24: Lightning protection," *International Electrotechnical Commission, IEC*, **(2010)**: 61400-24, 2010

[8] W. Rogowski, W. Steinhaus, "Die messung der magnetischen spannung," *Electrical Engineering (Archiv für Elektrotechnik)*, **1(4)**: 141–150, 1912

[9] I.T. Transformers-Part, "10: Lightning protection," *International Electrotechnical Commission, IEC*, **(2017)**: 61869-10, 2017

[10] Y. Liu, F. Lin, Q. Zhang, H. Zhong, "Design and construction of a Rogowski coil for measuring wide pulsed current," *IEEE Sensors Journal*, **11(1)**: 123–130, 2010

[11] W. Li, C. Mao, J. Lu, "Study of the virtual instrumentation applied to measure pulsed heavy currents," *IEEE Transactions on Instrumentation and Measurement*, **54(1)**: 284–288, 2005

[12] G. Robles, M. Argueso, J. Sanz, R. Giannetti, B. Tellini, "Identification of parameters in a Rogowski coil used for the measurement of partial discharges," *IEEE Instrumentation & Measurement Technology Conference*, **(2007)**: 1–4, 2007

[13] J. Hlavacek, R. Prochazka, K. Draxler, V. Kvasnicka, "The Rogowski coil design software," *IMEKO TC4 International Symposium*, **(2008)**: 295–300, 2008

[14] M. Xiang, H. Gao, B. Zhao, C. Wang, C. Tian, "Analysis on transfer characteristics of Rogowski coil transducer to travelling wave," *International Conference on Advanced Power System Automation and Protection*, **2**: 1056–1059, 2011

[15] M. Rezaee, H. Heydari, "Mutual inductances comparison in Rogowski coil with circular and rectangular cross-sections and its improvement," *IEEE Conference on Industrial Electronics and Applications*, **(2008)**: 1507–1511, 2008

[16] W. Limcharoen, P. Yutthagowith, "Rogowski coil with an active integrator for measurement of switching impulse current," *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, **(2013)**: 1–4, 2013

[17] B. Wang, D. Wang, W. Wu, "A Rogowski coil current transducer designed for wide bandwidth current pulse measurement," *International Power Electronics and Motion Control Conference*, **(2009)**: 1246–1249, 2009

[18] K. Ogata, *Modern Control Engineering*, United Kingdom: Prentice Hall, 2010

[19] C.K. Alexander, M.N.O. Sadiku, M. Sadiku, *Fundamentals of Electric Circuits*, Boston, MA, USA: McGraw-Hill Higher Education, 2007

[20] J.D. Hoffman, S. Frankel, *Numerical Methods for Engineers and Scientists*, CRC Press, 2018

[21] D. Chattopadhyay, P.C. Rakshit, *Fundamentals of Electric Circuit Theory*, S. Chand Publishing, 2000

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 9, No. 4, 51-59 (2024)
www.astesj.com
Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTES Journal
ISSN: 2415-6698

Integrating Speech and Gesture for Generating Reliable Robotic Task Configuration

Shuvo Kumar Paul*, Mircea Nicolescu, Monica Nicolescu

Department of Computer Science and Engineering, University of Nevada, Reno, 89557, USA

ARTICLE INFO

Article history:
Received: 24 April, 2024
Revised: 17 July, 2024
Accepted: 25 July, 2024
Online: 02 August, 2024

Keywords:
Task Configuration
Robotic Task
Gesture Recognition

ABSTRACT

This paper presents a system that combines speech and pointing gestures along with four distinct hand gestures to precisely identify both the object of interest and parameters for robotic tasks. We utilized skeleton landmarks to detect pointing gestures and determine their direction, while a pre-trained model, trained on 21 hand landmarks from 2D images, was employed to interpret hand gestures. Furthermore, a dedicated model was trained to extract task information from verbal instructions. The framework integrates task parameters derived from verbal instructions with inferred gestures to detect and identify objects of interest (OOI) in the scene, essential for creating accurate final task configurations.

1. Introduction

The rapid advancement of robotics, automation, and artificial intelligence has ignited a revolution in robotics. While industrial robots have proliferated over the past few decades, there's been a recent surge in the integration of robots into our daily lives. This shift has led to a significant change in robotics research focus, moving from industrial applications to service robots. These robots now serve as assistants in various tasks such as cooking, cleaning, and education, among others. Consequently, this transformation has redefined the role of human users, evolving them from primary controllers to collaborative teammates, fostering increased interaction between humans and robots.

While robots can autonomously handle tasks in certain scenarios, human interaction is often necessary. Unlike industrial robots that perform repetitive tasks, service robots are designed to engage with humans while carrying out their functions. In such contexts, it's crucial for interactions to feel natural and intuitive.

To achieve this, interaction components should mirror those commonly observed in human-to-human interactions. Human interactions typically involve gestures, gaze, speech, and facial expressions. While speech effectively conveys complex information, gestures can indicate direction, location within a scene, and common task-specific actions. Combining speech and gestures enhances the interaction experience by enabling intuitive communication and conveying meaningful commands.

In this work, our focus was on integrating pointing and four distinct hand gestures with verbal interactions. We developed a neural network model to extract task parameters from verbal instructions, utilizing a dataset of 60,769 annotated samples. For recognizing pointing gestures, we employed AlphaPose to capture skeletal joint positions and calculated the forearm's angle and length ratio to determine the pointing direction. Additionally, we predicted the Object of Interest (OOI) based on the shortest distance from the pointing direction vector. Finally, we identified four common hand gestures—bring, hold, stop, and point—using hand landmarks from Google's Mediapipe and trained a 3-layer fully connected neural network for gesture recognition. This integration not only enhances natural interaction but also gathers crucial additional information and context, thereby aiding in disambiguating and inferring missing task parameters. By combining speech with gestures, our system enhances the richness and clarity of interactions, which is essential for service robots designed to assist in everyday tasks. To this effect, the following are our contributions:

1. Multimodal Integration: Unlike existing approaches that often rely on a single mode of interaction, our research integrates pointing gestures, four distinct hand gestures, and verbal instructions. This multimodal integration is crucial for creating interactions that are more natural and intuitive, closely mirroring how humans communicate with each other.

2. Enhanced Task Parameter Estimation: By combining verbal commands with gestures, our system is able to disambiguate

*Corresponding Author: Shuvo Kumar Paul, 1664 N Virginia St, Reno, NV 89557 & shuvokumarp@unr.edu

and infer missing task parameters more effectively. This leads to more accurate and reliable task configurations, which is a significant advancement in the field of human-robot interaction.

3. Real-time Processing: Our framework operates in real-time, managing multiple inputs concurrently. This capability is vital for practical applications where timely and responsive interactions are required.

4. Experimental Validation: We conducted experiments to validate our approach, demonstrating its efficacy in generating reliable task configurations. Our results show that the integration of gestures and verbal instructions significantly improves the system's performance in real-world applications.

Our work introduces a framework that seamlessly integrates multiple forms of communication. The ability to interpret and combine verbal commands with pointing and hand gestures represents a significant step forward in creating more intuitive and effective human-robot interactions. This multimodal approach not only enhances the naturalness of interactions but also provides the robot with richer contextual information, enabling it to perform tasks more accurately and efficiently.

The paper follows this structure: the subsequent section provides a concise overview of prior research on gesture recognition techniques and Natural Language Understanding in Human-Robot Interaction (HRI) design. We then proceed to elaborate on the methodology of our work. Subsequent chapters incorporate our evaluation, including experimental results and observations. Finally, we summarize our findings in the concluding section of this paper.

2. Related Works

2.1. Natural Language Understanding in HRI

In [1], the author presented a hierarchical recurrent network coupled with a sampling-based planner to enable the comprehension of sequences of natural language commands within a continuous configuration space. Similarly, in [2], the author devised a system that interprets natural language directions for robots by extracting spatial description clauses, using a probabilistic graphical model that grounds landmark phrases, evaluates spatial relations, and models verb phrases. In [3], the author explored the application of statistical machine translation techniques to enable mobile robots to interpret unconstrained natural language directions, effectively mapping them onto environment maps, leveraging physical constraints to manage translation complexity and handle uncertainty. Additionally, in [4], the author demonstrated the robot's capability to learn action sequences' conditions from natural language, promptly updating its environment state knowledge and world model to generate consistent new plans, highlighting both specific operational success and the dialogue module's scalability and responsiveness to untrained user commands. In [5], the author explored spatial relationships to create a natural communication channel between humans and robots, showcasing in their study how a multimodal robotic interface integrating linguistic spatial descriptions and data from an evidence grid map enhances natural human-robot interaction. In addition, in [6], the author introduced Generalized Grounding Graphs, a dynamic probabilistic graphical model that interprets natural language commands for autonomous systems navigating and manipulating objects in semi-structured environments.

While prior research predominantly addressed navigational tasks, our approach extends this by employing deep learning techniques to extract specific parameters from single instructions pertinent to collaborative tasks.

2.2. Gesture Recognition In HRI

In [7], the author proposed a two-stage Hidden Markov Model (HMM) approach aimed at enhancing Human-Robot Interaction (HRI) by enabling intuitive robot control via hand gestures. The first stage identifies primary command-like gestures, while the second stage focuses on task recognition, leveraging Mixed Gaussian distributions within HMM to improve recognition accuracy. In [8], the author introduced a robust HRI system that continuously performs gesture recognition to facilitate natural human-robot interaction by employing online-trained ad-hoc Hidden Markov Models to accommodate intra-user variability, evaluated through studies on hand-formed letters and natural gesture recognition scenarios. In [9], the author introduced an HRI system using gesture recognition that incorporates multiple feature fusion, failure verification, and validation through real-world testing with a mobile manipulator. In [10], the author presented a gesture-based human-robot interaction framework, utilizing wearable sensors and an artificial neural network for gesture classification, and introducing a parameterization robotic task manager for intuitive robot task selection and validation in collaborative assembly operations. In [11], the author introduced a parallel convolutional neural network (CNN) method optimized for recognizing static hand gestures in complex environments, particularly suited for space human-robot interaction tasks, demonstrating superior accuracy over single-channel CNN approaches and other existing methods.

We have implemented two gesture recognition methods: a heuristic-based pointing gesture recognition and pointing direction estimation, and neural network based hand gesture recognition. In both methods, we leveraged the extracted landmarks from body skeleton and hands respectively.

3. Methodology

3.1. Extracting information from verbal commands

In a typical human collaboration, shared instructions often encompass specific details like the required action, the target object, navigation directions, and the particular location of interest within the scene. Additionally, we frequently use descriptive attributes such as size, relative position, shape, pattern, and color to specify objects, as seen in phrases like "bring the green box," "the book on the right," or "hold the blue box" [12]. These details outline various aspects of a task, as depicted in Figure 1, which showcases various task parameters linked to particular instructions.

In our research, we developed a dataset tailored for collaborative robotic commands, comprising verbal instructions that specify actions and include details on at least one of the following attributes:

object name, object color, object location, or object size. This dataset contains 60,769 samples, each annotated with five labels. We thoroughly assessed eight different model architectures for training, ultimately determining that the single-layer Bi-directional Long Short-Term Memory (Bi-LSTM) model delivered the best performance.

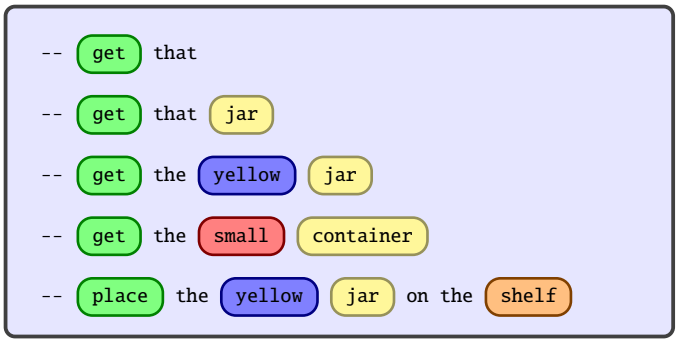


Figure 1: The task action is denoted by the green box, while the object's location in the scene is highlighted by the orange box. The red box indicates the size of the object, while the yellow and blue boxes respectively highlight the object of interest and its corresponding attributes.

Figure 2 depicts the model architecture, which comprises three neural layers. The model starts with an embedding layer, followed by a Bi-LSTM layer which is connected to a fully connected layer (FCN). The dataset vocabulary size, denoted as V, is used to one-hot encode each word, resulting in a vector size $W \in \mathbb{R}^{1\times \mathtt{V}}$. The input sequences, consisting of n words, are processed by an embedding layer represented as $\mathcal{E}$. The output from Bi-LSTM cells is concatenated and then passed through four layers. The resulting outputs from the FCN layer are subjected to softmax activation for the classification of five task parameters. Each classifier is evaluated using Cross Entropy loss $\mathcal{L}_c$. To update the model, we compute the mean of these losses as $\mathcal{L}_m = \frac{1}{5}\sum_{c=1}^{4}\mathcal{L}_c$.

3.2. Recognition of pointing gestures

We employed AlphaPose [13] to capture the skeletal joint positions for predicting pointing gestures and their overall direction. For simplicity, we assumed that the user uses one hand at a time for pointing. Following the categorization by [14], the authors distinguished pointing gestures into two types: extended (large) and bent arm (small) gestures. Furthermore, we generalized the forearm's orientation concerning the body into three categories: across, outward, and straight, as depicted in Figure 3(b).

We analyze the forearm angle θ_a (Figure 3(a)), comparing it against a predefined threshold θ_t to distinguish between across and outward pointing gestures. When the user isn't pointing (Figure 3(b)), the forearm angle is smaller compared to when they are pointing. When pointing directly towards the camera (robot's vision) (Figure 3(c)), the angle approaches 0. To refine this analysis, we introduce the forearm length ratio ρ_a. If the user isn't pointing, both forearms show similar lengths (Figure 3(b)). Conversely, a noticeable difference suggests the user is pointing directly (or very close) towards the camera with that arm (Figure 3(b)). Additionally, we determine the pointing direction d by analyzing the relative positions of the wrist and elbow of the pointing arm, enhancing navigational command interpretation.

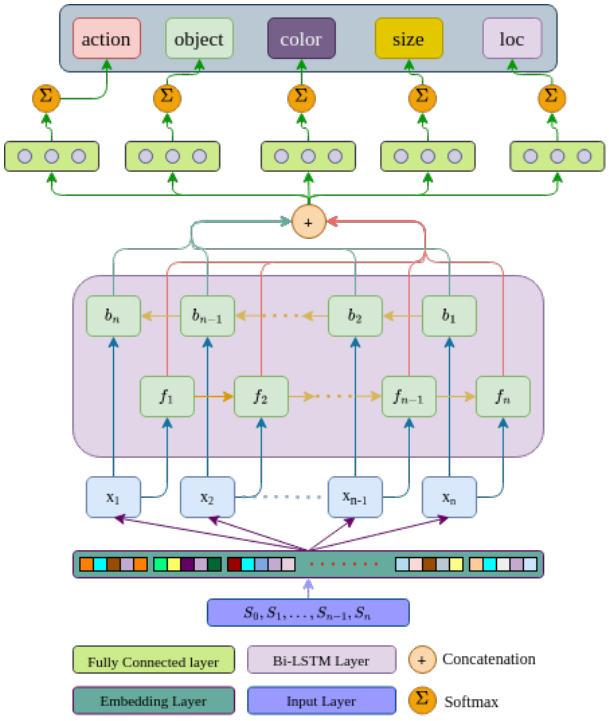


Figure 2: Neural Network (NN) model for parameter extraction from verbal commands.

3.2.1. Deriving θ_a from the positions of the wrist and elbow

We specifically need the locations of certain skeletal joints. These are locations of the left elbow, left wrist, right elbow, and the right wrist. This ensures our method remains effective even if some body parts are obscured, as long as the pointing hand's joints are detected. Let (x_1, y_1) denote the coordinates of the elbow, and (x_2, y_2) denote those of the wrist. By defining the 2D vector from the elbow to the wrist as $\vec{d} = (x_2 - x_1, y_2 - y_1)$ and using $\vec{v} = (0, 1)$ as the reference vertical vector, we can calculate the pointing angle θ_a using Equation 1:

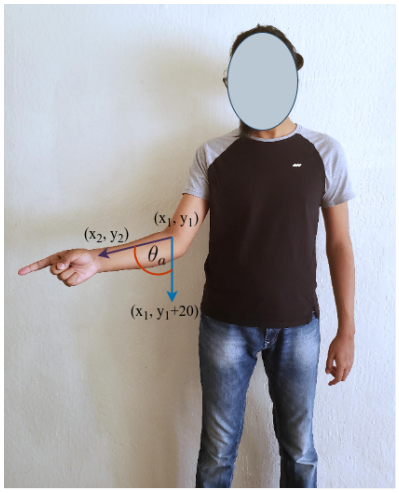
$$\theta_a = \cos^{-1}\frac{\vec{d}\cdot\vec{v}}{|\vec{d}||\vec{v}|} \tag{1}$$

If θ_a exceeds θ_t, the corresponding forearm is identified as performing the pointing gesture. Next, we assess the x coordinates of the wrist and elbow to determine the overall pointing direction within the scene—either left or right relative to the body. Additionally, we evaluate the forearm length ratio $\rho_a = \frac{\text{Length of the arm of interest}}{\text{Length of the other arm}}$ against a predefined ratio ρ_t to determine if the user is pointing directly ahead. Specifically, ρ_t is set to 0.8 and θ_t to 15°.

3.3. OOI Prediction

For every object identified, we establish its central point as a reference. Next, the perpendicular distance from the center of each object to the direction vector is computed.. The object found to have

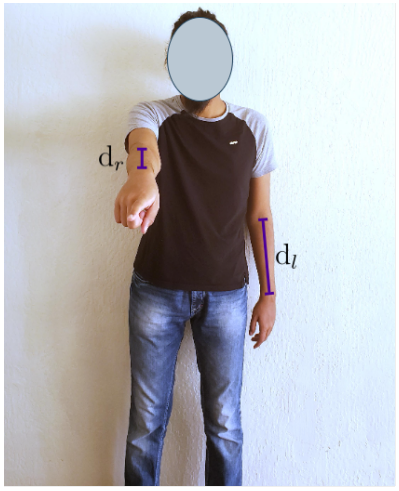
the shortest distance is considered the Object of Interest (OOI) is shown in figure 4.



(a)



(b)



(c)

Figure 3: (a) Generated angle θ_a , (b) length of forearms d_l, d_r when not pointing, and (c) pointing straight

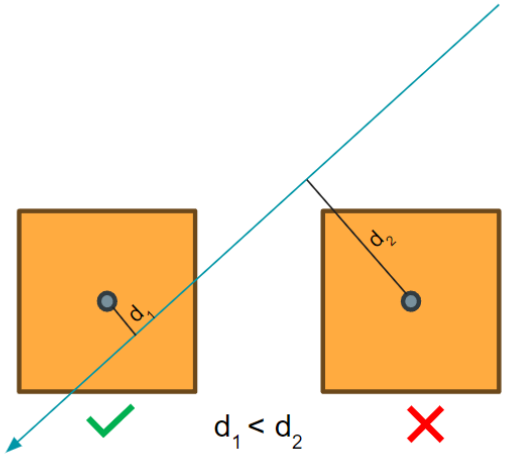


Figure 4: Determining Object of Interest (OOI) from pointing direction.

3.4. Gesture recognition

We have identified four common gestures for instructing robots: bring, hold, stop, and point gestures (see Figure 5). This capability enables the robot to navigate toward either an object or a designated location within the scene. Utilizing Google's Mediapipe library [15], we extracted hand landmarks, providing 21 landmark points for each hand (see Figure 6). These landmarks were captured for both hands during the aforementioned gestures to compile a dataset. Subsequently, the dataset underwent training using a 3-layer fully connected neural network model. Each fully connected layer's outputs were subjected to a dropout layer and then activated by ReLU (Rectified Linear Unit). The model's architecture is illustrated in Figure 7.



(a) Bring



(b) Hold



(c) Stop



(d) Point

Figure 5: Gesture categories

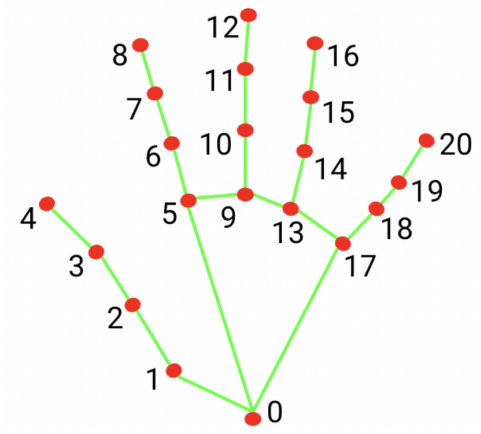


Figure 6: Extracted hand landmark [16]

4. Experimental Result

We have combined pointing gestures and hand gesture recognition systems with the task parameter extraction module and evaluated them separately.

4.1. Pointing Gesture With Verbal Command

During our experiments, participants were assigned to perform precise pointing gestures in predefined scenarios. Each scenario depicted a scene with three distinct objects: two books and a Cheez-It box. Participants were directed to point to one object at a time. For instance, in a specific scenario, they were instructed to extend their right hand and point to the leftmost object. Thus, from this particular data sample, we could analyze that the participant executed a pointing gesture with their right hand, directing it towards their left, aiming at the object positioned farthest to the right (from their perspective). This dataset served as the foundation for our quantitative evaluations.

The experiments involved positioning the user at distances of 1.22, 2.44, 3.66, and 4.88 meters from the camera. Each system component underwent separate evaluation, encompassing tasks such as extracting parameters from verbal commands, detecting the active hand, estimating pointing direction and predicting the object of interest. The extracted task parameters were then presented in tabular format to illustrate the results.

We evaluated each frame's prediction against its label, assessing accuracy, precision, and recall. For instance, if a frame's label specifies "Right hand: pointing; Left hand: not pointing," a correct prediction of "Right hand: pointing" counts as a True Positive; otherwise, it registers as a False Negative. Conversely, predicting "Left hand: pointing" when the label indicates otherwise is a False Positive, while accurately predicting "Left hand: not pointing" is a True Negative. Table 1 details the accuracy, precision, and recall metrics across various distances.

Table 2 illustrates various sample scenarios and their corresponding task parameters extracted from data sources. The column labeled "Structured Information" showcases data derived from both the Pointing State and Verbal Command. Each row pertains to a specific scenario, beginning with an indication of whether pointing was involved, followed by the experiment number. Verbal commands, typically involving the fixed task action "get," are listed alongside extracted information from both verbal commands and simultaneous pointing states. Predicted Objects of Interest (OOI) that necessitate action are noted, along with the system's corresponding response. Instances of ambiguity are highlighted in bold within the cells.

Table 1: Pointing Gesture Recognition

Distance (m)	Accuracy	Precision	Recall
4.88	1	1	1
3.66	0.995	1	0.99
2.44	0.995	1	0.99
1.22	0.995	1	0.99

Ambiguity occurs when the object of interest (OOI) cannot be identified solely from the verbal command and pointing gesture provided. In these situations, the system informs the user with the message "Additional information needed to identify the object," and it waits for the user to provide more input, either by repeating the pointing gesture or by adjusting the command given.

In Table 2, observations indicate that ambiguity arises in different scenarios. When the system is in the "Not Pointing" state, ambiguity occurs due to insufficient object attributes (e.g., Exp# 1, 3), which hinder the unique identification of the OOI, leading the system to request more information. Conversely, in the "Pointing" state, ambiguity arises when the pointing direction does not intersect with any object boundaries. Verbal commands play a crucial role in reducing this ambiguity by providing additional information.

4.2. Hand Gesture Recognition With Verbal Command

The system processes verbal commands by identifying and extracting up to five distinct task parameters, which are subsequently stored for sequential task execution. The transcription of the verbal commands and their corresponding extracted parameters is presented in Table 3.

If no matches are identified, the respective parameters are denoted as ***None***. Each command initiates a task, recorded in the order of execution. Furthermore, Figure 8 illustrates the performance comparison among different models. Subfigure a illustrates the overall accuracy, while subfigure b highlights the accuracy of Object of Interest (OOI) prediction tasks. Across both evaluations, the Bi-LSTM based model consistently outperforms all other models.

Table 4 showcases the accuracy, recall, and f1-score achieved in recognizing four specific gestures. The model consistently demonstrates high accuracy in interpreting user gestures. Figure 9 illustrates the confusion matrix for these gestures, highlighting occasional misclassifications where the 'Bring' gesture is mistakenly identified as 'Stop.' However, considering users receive feedback until the correct action is chosen, these rare errors hold minimal consequence.

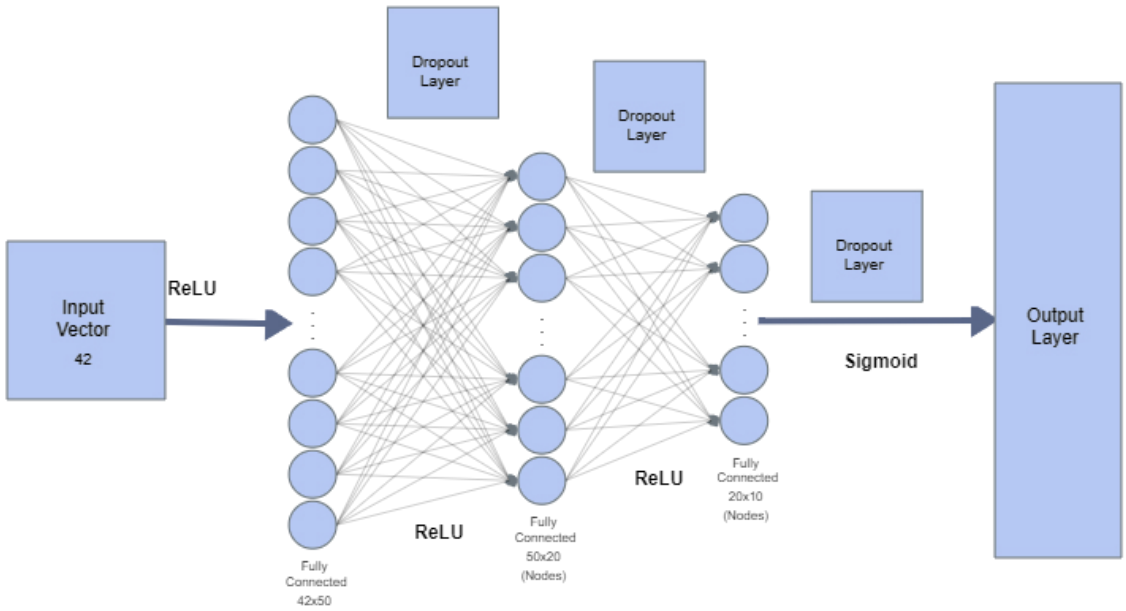


Figure 7: Gesture recognition model architecture.

Table 2: Generated Task Parameters With Pointing State

Pointing State	Exp #	Verbal Command	Structured information	Identified Object	Feedback
Pointing	1	get that, get me that	{*action: "get", pointing_identifier: True, object: "book", object_identifiers:* {*attributes: null, position: null*}}	"book-1"	None
	2	get the red book	{*action: "get", pointing_identifier: True, object: "book", object_identifiers:* {*attributes: "red", position:* }}	"book-2"	None
	3	get that red thing	{*action: "get", pointing_identifier: True, object: null, object_identifiers:* {*attributes: "red", position:* }}	"cheez-it"	None
Not Pointing	1	get that, get me that	{*action: "get", pointing_identifier: False, object: null, object_identifiers:* {*attributes: null, position: null*}}	**None (ambiguous)**	**"Additional information is needed to identify object"**
	2	get the red book	{*action: "get", pointing_identifier: False, object: "book", object_identifiers:* {*attributes: "red", position: null*}}	"book-2"	None
	3	get that red thing	{*action: "get", pointing_identifier: False, object: null, object_identifiers:* {*attributes: "red", position: "right"*}}	**None (ambiguous)**	**"Additional information is needed to identify object"**

Table 3: Extracted task parameters from various verbal commands

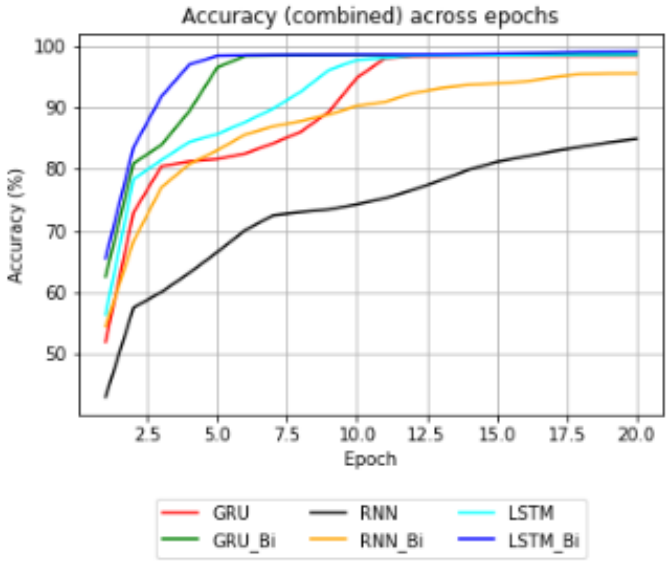
Verbal command: "bring me the jar"
Object: jar — Action: bring — Attributes: None — Location: None

Verbal command: "give me that black box"
Object: box — Action: give — Attributes: [black] — Location: None

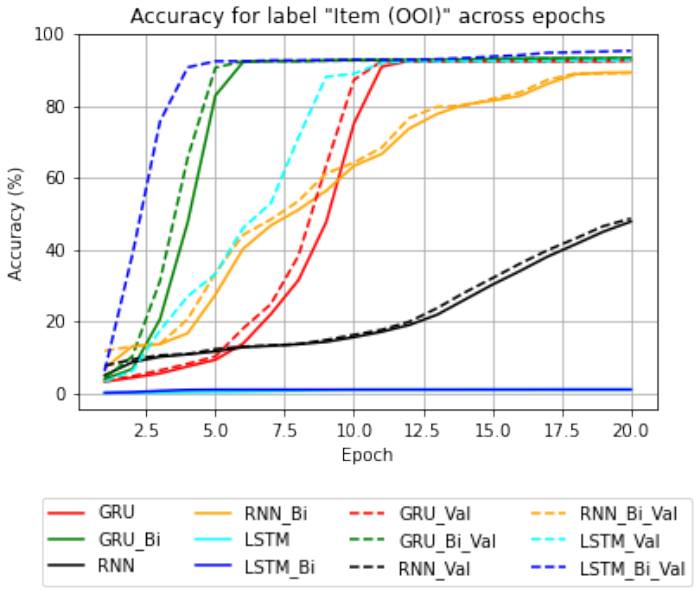
Verbal command: "turn right"
Object: None — Action: turn — Attributes: None — Location: right

Verbal command: "hold the small white jar on your left"
Object: jar — Action: hold — Attributes: [white, small] — Location: left

Verbal command: "place the box on the shelf"
Object: box — Action: place — Attributes: None — Location: shelf



(a)



(b)

Figure 8: The comparative performance of various models: (a) the aggregate accuracy of all five extracted task parameters over epochs, and (b) the accuracy for the parameter "Item (OOI)" over epochs.

Table 4: Performance metrics

Gesture	precision	recall	f1-score
Bring	0.93	1.00	0.97
Hold	1.00	0.99	0.99
Point	1.00	0.99	0.99
Stop	1.00	0.95	0.98

Subsequently, we explored scenarios where users performed gestures alongside predefined natural language instructions. Extracted information was utilized to establish task parameters, with follow-up responses issued in case of ambiguity. Table 5 delineates the sequential steps of a sample interaction, wherein gestures assist in identifying crucial task elements such as 'action' and 'object.' Notably, in step 2, the system requests additional information to identify the object of interest (OOI). Conversely, in step 4, although no verbal instructions are provided, the system maintains the previous OOI and executes the 'hold' action accordingly. This table underscores the utility of combining gesture and verbal cues for robust task configurations.

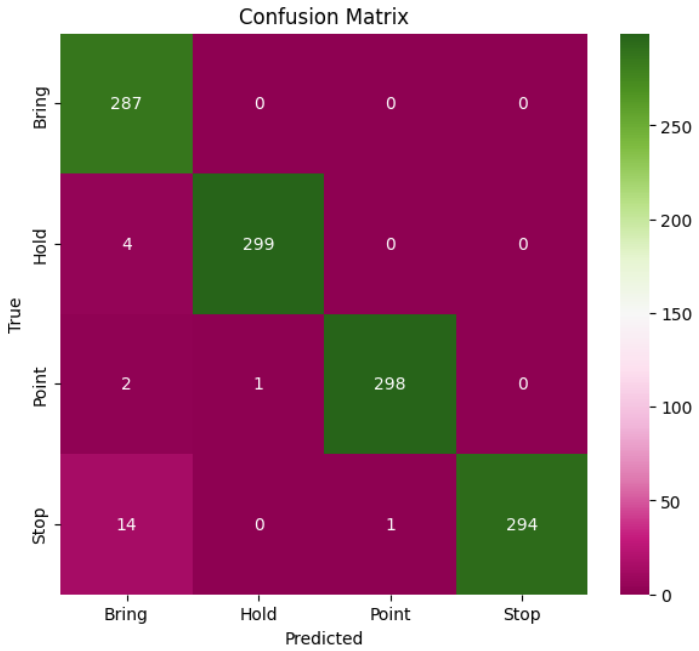


Figure 9: Confusion matrix for the four gestures

4.3. Qualitative Insights and Analysis

The results of our experiments provide significant insights into the effectiveness of integrating pointing gestures, hand gestures, and verbal commands for enhancing robotic task configurations. By evaluating each component separately, we were able to assess the accuracy, precision, and recall of the system in identifying the operating hand, estimating pointing direction, and predicting the object of interest (OOI).

Our findings indicate high accuracy in pointing gesture recognition across various distances, as demonstrated in Table 1. This suggests that the system can reliably interpret pointing gestures even from a distance of up to 4.88 meters, which is crucial for practical applications in diverse environments.

The integration of verbal commands with gestures significantly improves the system's ability to disambiguate and infer task parameters. As shown in Table 2, the combined use of pointing gestures and verbal instructions enhances the system's capability to identify objects and actions accurately. However, instances of ambiguity still arise, particularly when the OOI cannot be determined solely from the given inputs. In such cases, the system effectively prompts the user for additional information, demonstrating a robust error-handling mechanism.

Additionally, the hand gesture recognition component, when paired with verbal commands, consistently achieved high accuracy, recall, and f1-scores, as illustrated in Table 4 and Figure 9. This

Table 5: Extracted Task Parameters With Gesture Recognition

Step#	Gesture Performed	Verbal Instruction	Structured Information	Feedback
1	Stop	-	action: stop, object: None, identifier: None, location: None	-
2	Bring	give me that	action: give, object: None, identifier: None, location: None	”Additional information is needed to identify object”
3	Point	bring me that book	action: bring, object: book, identifier: pointed direction, location: None	-
4	Hold	-	action: hold, object: jar, identifier: None, location: None	-
5	Point	bring it here	action: bring, object: jar, identifier: None, location: pointed location	-
6	Bring	that red jar on the shelf	action: bring, object: jar, identifier: red, location: shelf	-
7	Point	put it here	action: put, object: None or , identifier: None, location: None	-
8	Point	go there	action: go, object: None, identifier: None, location: pointed location	-

indicates the system's reliability in interpreting user gestures and extracting relevant task parameters, as further evidenced by the sequential task execution detailed in Table 5.

Overall, the implications of these results contribute to the overarching goals of the paper by showcasing the potential of multimodal interaction systems to facilitate natural and efficient human-robot interactions. The high accuracy and robustness of the system in various scenarios underline its practicality for real-world applications, where reliable task configurations are paramount for effective robotic assistance.

5. Conclusion

This paper presents a Human-Robot Interaction (HRI) framework tailored for extracting parameters essential for collaborative tasks between humans and robots. Operating in real-time, the framework concurrently manages multiple inputs. Verbal communication is leveraged to capture detailed task information, encompassing action commands and object attributes, complemented by gesture recognition. The amalgamation of these inputs yields named parameters, facilitating subsequent analysis for constructing well-structured commands. These commands seamlessly communicate task instructions to robotic entities and streamline the task execution processes.

To detect pointing gestures and infer their directions, we utilized a third-party library for skeleton landmark extraction. Additionally, we introduced a hand gesture recognition system capable of identifying four distinct hand gestures. This involved extracting hand landmarks and training a model to interpret these gestures. Furthermore, verbal commands captured by sensors are transcribed into text and processed through a pre-trained model to extract task-specific parameters. The amalgamation of this information culminates in the creation of the final task configuration. In instances where required parameters are lacking or ambiguities arise, the system offers appropriate feedback.

We evaluated the system's performance by subjecting it to various natural language instructions and gestures to generate task configurations. The extracted task parameters, corresponding to different verbal commands and gesture states, were arranged in a table to illustrate the effectiveness of our methodology.

It is important to highlight that our Human-Robot Interaction (HRI) framework showcases a robust capability to integrate verbal communication and gesture recognition in real-time, significantly enhancing the accuracy and efficiency of task parameter extraction. This integration is crucial for developing more intuitive and natural human-robot collaborative environments.

Furthermore, our experimental results validate the system's reliability in interpreting complex task instructions, which underscores its potential for practical applications in diverse settings. The high accuracy achieved in recognizing gestures and extracting task-specific parameters indicates that our approach can greatly improve the seamless execution of tasks by robotic entities.

Looking ahead, we see promising future research directions in exploring more intricate interaction scenarios. Investigating interactions involving multiple users, dynamic and continuous gestures, and complex dialogues will not only enhance the robustness of our system but also contribute to the broader evolution of HRI systems. By addressing these challenges, we aim to develop even more sophisticated and meaningful interaction frameworks that can further bridge the communication gap between humans and robots.

References

[1] Y.-L. Kuo, B. Katz, A. Barbu, “Deep Compositional Robotic Planners That Follow Natural Language Commands,” in 2020 IEEE International Conference on Robotics and Automation (ICRA), 4906–4912, IEEE, 2020, doi:10.1109/ICRA40945.2020.9197464.

[2] T. Kollar, S. Tellex, D. Roy, N. Roy, “Toward Understanding Natural Language Directions,” in 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 259–266, IEEE, 2010, doi:10.1109/HRI.2010.5453186.

[3] C. Matuszek, D. Fox, K. Koscher, “Following Directions Using Statistical Machine Translation,” in 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 251–258, IEEE, 2010, doi:10.1109/HRI.2010.5453189.

[4] R. Cantrell, K. Talamadupula, P. Schermerhorn, J. Benton, S. Kambhampati, M. Scheutz, “Tell Me When and Why to Do It! Run-Time Planner Model Updates via Natural Language Instruction,” in Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, 471–478, 2012, doi:10.1145/2157689.2157840.

[5] M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, D. Brock, “Spatial Language for Human-Robot Dialogs,” IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), **34**(2), 154–167, 2004, doi:10.1109/TSMCC.2004.826273.

[6] S. Tellex, T. Kollar, S. Dickerson, M. Walter, A. Banerjee, S. Teller, N. Roy, “Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation,” in Proceedings of the AAAI Conference on Artificial Intelligence, volume 25, 2011, doi:10.1609/aaai.v25i1.7979.

[7] N. Nguyen-Duc-Thanh, S. Lee, D. Kim, “Two-stage hidden markov model in gesture recognition for human robot interaction,” International Journal of Advanced Robotic Systems, **9**(2), 39, 2012, doi:10.5772/50204.

[8] S. Iengo, S. Rossi, M. Staffa, A. Finzi, "Continuous gesture recognition for flexible human-robot interaction," in 2014 IEEE International Conference on Robotics and Automation (ICRA), 4863–4868, IEEE, 2014, doi:10.1109/ICRA.2014.6907571.

[9] G. H. Lim, E. Pedrosa, F. Amaral, N. Lau, A. Pereira, P. Dias, J. L. Azevedo, B. Cunha, L. P. Reis, "Rich and robust human-robot interaction on gesture recognition for assembly tasks," in 2017 IEEE International conference on autonomous robot systems and competitions (ICARSC), 159–164, IEEE, 2017, doi:10.1109/ICARSC.2017.7964069.

[10] P. Neto, M. Simão, N. Mendes, M. Safeea, "Gesture-based human-robot interaction for human assistance in manufacturing," The International Journal of Advanced Manufacturing Technology, **101**, 119–135, 2019, doi:10.1007/s00170-018-2788-x.

[11] Q. Gao, J. Liu, Z. Ju, Y. Li, T. Zhang, L. Zhang, "Static hand gesture recognition with parallel CNNs for space human-robot interaction," in Intelligent Robotics and Applications: 10th International Conference, ICIRA 2017, Wuhan, China, August 16–18, 2017, Proceedings, Part I 10, 462–473, Springer, 2017, doi:10.1007/978-3-319-65289-4_44.

[12] F. H. Previc, "The Neuropsychology of 3-D Space." Psychological Bulletin, **124**(2), 123, 1998.

[13] H.-S. Fang, S. Xie, Y.-W. Tai, C. Lu, "RMPE: Regional Multi-person Pose Estimation," in ICCV, 2017.

[14] C.-B. Park, S.-W. Lee, "Real-Time 3D Pointing Gesture Recognition for Mobile Robots With Cascade HMM and Particle Filter," Image and Vision Computing, **29**(1), 51–63, 2011, doi:10.1016/j.imavis.2010.08.006.

[15] Google, "Google/mediapipe: Cross-platform, customizable ML solutions for live and streaming media." https://github.com/google/mediapipe, accessed: 2022-03-13.

[16] "Hand landmarks," https://developers.google.com/static/mediapipe/images/solutions/hand-landmarks.png, accessed: 2023-12-12.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 9, No. 3, 62-71 (2024)
www.astesj.com
Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTES Journal
ISSN: 2415-6698

Leveraging Machine Learning for a Comprehensive Assessment of PFAS Nephrotoxicity

Anirudh Mazumder, Kapil Panda*

University of North Texas, Texas Academy of Mathematics and Science, Denton, 76203, United States of America

ARTICLE INFO

Article history:
Received: 04 March, 2024
Revised: 18 May, 2024
Accepted: 19 May, 2024
Online: 12 June, 2024

Keywords:
Machine Learning
Kidneys
Polyfluoro-Alkyl Substances
Toxicokinetics

ABSTRACT

Polyfluoroalkyl substances (PFAS) are persistent chemicals that accumulate in the body and environment. Although recent studies have indicated that PFAS may disrupt kidney function, the underlying mechanisms and overall effects on the organ remain unclear. Therefore, this study aims to elucidate the impact of PFAS on kidney health using machine learning techniques. Utilizing a dataset containing PFAS chemical features and kidney parameters, dimensionality reduction and clustering were performed to identify patterns. Machine learning models, including XGBoost classifier, regressor, and Random Forest regressor, were then developed to predict kidney type from PFAS descriptors, estimate PFAS accumulation in the body, and predict the ratio of glomerular surface area to proximal tubule volume, which indicates kidney filtration efficiency. The kidney type classifier achieved 100% accuracy, confirming that PFAS exposure alters kidney morphology. The PFAS accumulation model attained an R^2 of 1.00, providing a tool to identify at-risk individuals. The ratio prediction model reached an R^2 of 1.00, offering insights into PFAS effects on kidney function. Furthermore, PFAS descriptors and anatomical variables were identified through analyses using feature importance, demonstrating discernible links between PFAS and kidney health, offering further biological significance. Overall, this study can significantly contribute to the current findings on the effect of PFAS while offering machine learning as a contributive tool for similar studies.

1. Introduction

Polyfluoro-alkyl substances (PFAS) have garnered significant attention in recent years due to their presence in a wide range of consumer and industrial products and their existence in the environment [1]. However, the characteristics that make PFAS so prevalent in our society also underscore their challenges to human and environmental health [2].

PFAS are a group of synthetic organic compounds characterized by their perfluoroalkyl chains, which consist of carbon atoms fully saturated with fluorine atoms [3]. This unique chemical structure results in one of the most robust and most stable bonds in organic chemistry, the carbon-fluorine (C-F) bond, which is responsible for PFAS's exceptional resistance to heat, chemical degradation, and biological breakdown processes, earning it the name of the "forever chemical" [4]. The fluorine atoms in PFAS molecules form a protective shield around the carbon backbone, rendering these compounds highly hydrophobic, which contributes to their utility in various industrial applications, such as the production of non-stick coatings and water-resistant textiles [5]. However, this is also the reason behind their persistence in the environment and their ability to bioaccumulate in organisms [6].

PFAS's hydrophobic properties make it highly insoluble in water, which prevents the chemical from dissolving into aqueous environments, allowing it to persist in the soil, water, and sediment for extended periods [7]. Furthermore, this hydrophobicity disrupts normal metabolic pathways, as it partitions into fatty tissues rather than remaining in aqueous solutions [8]. This phenomenon leads to bioaccumulation in organisms, as PFAS are absorbed through ingestion or absorption and accumulate in fatty tissues over time, resulting in elevated concentrations within organisms throughout the food chain [9].

Over the past few years, various research has revealed the growing dangers associated with PFAS on the body, encompassing concerns such as cancer, thyroid disorders, developmental anomalies in children, and immune system dysfunction [10, 11]. It is well known that the carcinogenic potential of PFAS has been associated with various cancers due to the PFAS-induced oxidative stress that plays a role in cellular damage and DNA mutations, contributing to cancer development [12, 13]. Furthermore, studies have shown that by interfering with hormonal regulation systems and suppressing immune system function, PFAS can lead to disorders like hypothyroidism,

*Corresponding Author: Kapil Panda, 1155 Union Cir, kapilpanda@my.unt.edu

developmental anomalies, such as stunted growth and delayed cognitive development in children, and neurotoxic effects [14, 15].

However, recent studies have uncovered the notion that PFAS, in fact, also has the potential to infiltrate further and disrupt fundamental physiological processes, namely the kidneys [16]. The kidneys, often called the body's natural filtration and waste management system, play a pivotal role in maintaining homeostasis. Their intricate network of nephrons and tubules ensures the efficient removal of waste products, excess fluids, and electrolytes from the bloodstream [17]. Consequently, the exploration of the relationship between PFAS exposure and kidney function has assumed a position of paramount importance in the realm of health research [18]. Extensive research in recent years has shed light on the profound impact of PFAS on kidney health. When introduced into the body, PFAS compounds can infiltrate renal tissues, interacting with various cellular components and initiating molecular responses [19]. These interactions can lead to structural changes in the kidneys, potentially altering the distribution of kidney types, which has been identified in previous studies. Such changes in kidney morphology have significant implications for kidney function and overall health [20]. Moreover, studies have indicated that PFAS exposure can disrupt the finely tuned balance of hormonal regulation systems, potentially leading to disorders such as hypothyroidism, which can further affect kidney health [21].

Therefore, this research aims to gain further insight into the relationship between PFAS and kidney function [22]. Utilizing machine learning techniques, we aim to unravel the complexities of this relationship and shed more light on how PFAS accumulation may impact the intricate structure and function of the kidneys [23]. Utilizing a dataset containing PFAS chemical features and kidney parameters, exploratory data analysis and dimensionality reduction were performed using PCA to identify patterns and correlations within the data. To ensure and verify the correlation between PFAS and kidneys, an XGBoost Classifier was used to predict kidney type from PFAS descriptors. Next, an XGBoost Regressor was used to estimate PFAS accumulation in the organ, assessing the impact of PFAS on the kidneys. Finally, a Random Forest Regressor was developed to determine the ratio of Glomerular Total Surface Area to Proximal Tubule Volume to offer insights into kidney function. The models were trained on 70% of the dataset and evaluated using metrics such as R-squared, confusion matrices, Mean Absolute Error, and residual analysis. Hyperparameter tuning through methods like Grid Search and Cross-Validation was also conducted.

2. Materials and Methods

2.1. Dataset

A dataset found at [24] was used. The dataset is critical to providing a few key pieces of information in depth. The first key piece of information that it provides is the chemical features of PFAS. For example, the dataset has information about the PFAS inside of the body's lipophilicity, the vapor pressure of PFAS, the water solubility of PFAS, and more critical descriptors of PFAS chemically. Additionally, the dataset contains information about the actual animal being looked at; for example, it has information about the species type and the gender of the animal. Furthermore, some critical physiological features of the animal were kept within the dataset, such as information about the animal's body mass, while also keeping physiological characteristics of the kidney inside of the body, which has vital information about things like the diameter of the proximal tubules. Utilizing this entire feature set, we conducted some methods to see if the data could be used to create important machine learning algorithms proving correlations between different parts of PFAS and kidney functions.

2.2. Data Analysis

Before the machine learning algorithms were created, exploratory data analysis was undertaken to see the key patterns and statistical backing behind the data before any machine learning was run, ensuring that the algorithms depend on there being learnable patterns. Additionally, conducting the analysis would allow for essential pieces of data that would mess with the accuracy of the machine learning models in the future, like outliers or potential data errors that may be found.

The following data analysis technique was a dimensionality reduction utilizing a Principal Component Analysis (PCA). PCA is perfect for finding the key correlations and patterns in the data, which directly leads to variation. Additionally, Figure 1 shows how the higher dimensionality is reduced to lower dimensionality spaces, making it easier to visualize data patterns while also allowing us to understand the explained variance ratio for the principal components. This allows us to see the primary source of variation in the data and how the first few components are vital in depicting the first few data points. Additionally, we can find clusters and patterns in the data, allowing us to find relationships that are not very noticeable or subgroups of data.

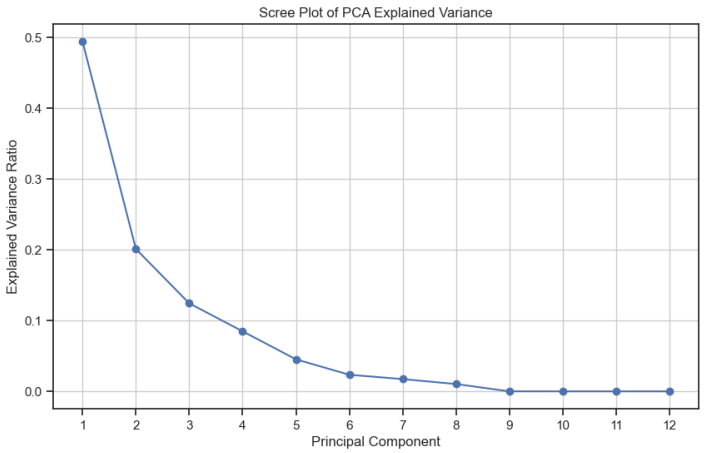


Figure 1: Principal Component Analysis and Explained Variance Ratio Showing a key change at The Principal Component 2

Furthermore, using methods like clustering on top of PCAs, we can better understand the patterns inside of the dataset [25]. This is a crucial method because clustering while using PCA allows us to find the natural groups of the patterns within the data, as PCA simplifies the data while keeping the essential characteristics, so the clusters of this allow us to understand the groupings of these characteristics allowing us to find these specific groups within the

data [26]. Additionally, K-Means can cluster the points based on their similarity, allowing us to find the similarity between the different characteristics delineated as essential due to the PCA algorithm. As seen in Figure 2, there seems to be an essential characteristic within the dataset that can show how some parts of the principal component analysis are very close. However, there are sections in the middle where multiple clusters are near one another.

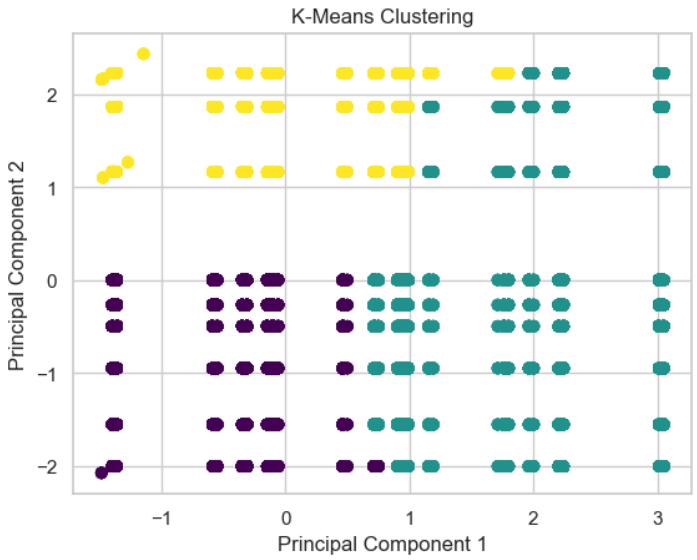


Figure 2: Clustering Analysis Using PCA and K-Means Algorithm to Identify Natural Data Patterns and Similarity Showing the Clusters of the First and Second Principal Components

2.3. Kidney Type Prediction

Utilizing our data, we attempted first to establish and confirm a correlation between PFAS and kidneys, which we would later conduct the rest of our analysis on. This correlation can describe if there is a relationship between the information about the PFAS and precisely how much it affects the kidney while providing the actual characteristics of the organ. To do this, we created a classifier that would classify what type of kidney was being affected by the PFAS inside of the species, which could allow for early measures and precautions to be taken based on the predicted effects of PFAS on the organ.

2.3.1. Feature Selection

The data was split up into all of the pieces of data which actively described any property or action of the PFAS chemical was used as a X variable, while just the kidney type was left as the Y variable.

2.3.2. Model Selection & Hyperparameter Tuning

The machine learning algorithm utilized an XGBoost Classifier(XGBC) to classify the PFAS data into the two different types of kidneys in the dataset. XGBC is critical to creating multiple weaker decision trees, which are then compiled up and added into a singular, more robust predictor while correcting previous tree errors by selectively creating new trees, allowing the algorithm to make the predictions more accurate [27]. As a result, the ensemble-based algorithm is critical to creating a robust classifier, as it uses the concept that if each tree has a vital understanding of the data, then when they all are combined, the final decision that the model is going to be outputting will understand all of the dimensionalities of the dataset. Additionally, the algorithm has inbuilt regularization techniques, which is critical because the dataset that we were working with had a ton of data points, meaning that the algorithm was very prone to just learning the pattern in training data without having a resemblance to the output data because it would be overfitting.

Furthermore, XGBC allows us to look into the specific features with the highest correlations, especially within the context of the machine learning algorithm's trained understanding of the patterns within the data [28]. Utilizing the XGBC feature importances, we can see which columns are the most important for the training of the algorithm and which columns may potentially be negatively affecting the understanding of the data, which would allow us to go back to the feature selection step and ensure that they are removed from the training dataset. Additionally, we can see which features have the highest correlation with the target output that we are looking for using this feature selection to understand better the patterns within the data, which otherwise would not be understandable.

Then, after the model had been trained and the feature selector had been used to see which features were key, hyperparameter tuning was used to create the most optimal machine learning model, especially in the context of XGBC where there are multiple parts to the actual algorithm including important pieces like loss functions which could change. Thus, essential concepts like Grid Search and Cross-Validation were used to conduct hyperparameter tuning so the classifier could achieve the best accuracy possible.

2.3.3. Model Training & Evaluation

The model was trained using a majority of the data, but a train test split of 70% - 30% was used to evaluate the model and prevent overfitting of the data. Then, to understand the relationship between the different pieces of data, a few metrics were used, including a confusion matrix to understand how the model was doing when predicting the testing data, and an accuracy score was used to understand the difference between the predicted and the actual outputs.

2.4. PFAS Accumulation Model

After establishing if there is a correlation between the kidney type and the information about PFAS, it is vital to see the amount of PFAS that is going to accumulate in the kidney based on the PFAS descriptors. We must see this information because it is critical to assessing the impact of PFAS on the kidney. Additionally, it is essential to understand if there is a correlation between the chemical descriptors of PFAS and the actual effect of PFAS when inside the human body. As a result, a machine learning model was created to make these assessments and understand if we can predict the accumulation of PFAS inside the human body.

2.4.1. Feature Selection

A fundamental difference between this algorithm and the kidney-type prediction algorithm is that this one needs to output a numerical

rather than a categorical response. This delineation is critical to understanding how the data will be used, as many features exist in the datasets. However, some of them are categorical, so the data must be encoded into numerical data. After the data was imputed, the data was split into all the PFAS descriptors: the X data. At the same time, the singular Y data was the amount of PFAS accumulating inside the body.

2.4.2. Model Selection & Hyperparameter Tuning

Similarly to the Kidney Type Prediction model, an XGBoost Regression algorithm was used to predict the amount of PFAS that would accumulate in the body. XGBoost regression was used for this algorithm due to its innate ability to understand nonlinear data through its use of gradient-boosted decision trees [29]. Using these trees, the algorithm can perform an ensemble method, reducing overfitting inside the algorithm. Additionally, since it is a gradient-boosting algorithm, it can optimize its performance through an iterative process where it continuously improves itself [30]. Also, it can understand larger datasets with many features, delineate which ones are the most essential parts for it to learn, and then create the most optimal outputs after it has achieved the response it is attempting to get [31].

2.4.3. Model Training & Evaluation

The model was trained on a 70%-30% train test split. Using the train test split, the model could be evaluated on its ability to truly learn the training data and still apply it to the testing data without just outputting it based on the answers it was already given. On this, an R^2 and MAE were calculated to output the model's accuracy to see how closely the model truly understood and predicted the data.

2.5. Glomerular Total Surface Area vs Proximal Tubule Predictor

Following the prediction of PFAS accumulation in kidneys, we aimed to discern the effects of PFAS on the kidneys' function. To do this, we sought to estimate the ratio of Glomerular Total Surface Area (GlomTotSA) to the Volume of the Proximal Tubule (ProxTubTotVol) within the kidneys, which provides insight into the structural dynamics of the organ. A higher ratio may indicate efficient filtration and reabsorption processes, suggesting healthier kidney function. In comparison, a lower ratio might suggest potential kidney morphology and function alterations, providing early indicators of kidney health issues.

2.5.1. Feature Selection

The feature selection process involved carefully examining the relevance and significance of each feature in the dataset. Using Recursive Feature Elimination (RFE) and correlation analysis, we identified the most informative attributes for our algorithm. Furthermore, domain expertise was crucial in selecting pertinent features, ensuring that the model captured the essence of PFAS accumulation within the kidneys.

Feature engineering, however, was not confined to feature selection alone. It extended to creating engineered features, such as interaction terms between PFAS accumulation descriptors and anatomical variables, enabling the model to capture nuanced relationships within the data. These engineered features served as the basis for the algorithm to make precise predictions regarding the GlomTotSA/ProxTubTotVol ratio.

2.5.2. Model Selection & Hyperparameter Tuning

The choice of a Random Forest Regressor was deliberate, given its aptitude for handling complex datasets and capturing linear and nonlinear relationships within the data. Our dataset encompassed a multitude of variables, including PFAS exposure descriptors, anatomical features, and the target variable of GlomTotSA/ProxTubTotVol ratio. These variables interact in intricate ways that may need to be more linear and straightforward. The ensemble nature of Random Forest, comprising multiple decision trees, allows the algorithm to capture these complex relationships effectively. Each decision tree contributes its unique perspective, and the ensemble aggregates these insights to yield robust and accurate predictions. This robustness is particularly advantageous when dealing with biological data, which often exhibits intricate and nonlinear interactions. The ensemble nature of the Random Forest algorithm, comprising multiple decision trees, enhances its robustness and predictive power [32]. Another compelling aspect of the Random Forest Regressor is its innate ability to provide insights into feature importance [33]. Understanding which features are most influential in predicting the GlomTotSA/ProxTubTotVol ratio is essential for gaining mechanistic insights into the relationship between PFAS exposure and kidney morphology [34]. Random Forest calculates feature importance scores, enabling us to identify which PFAS descriptors and anatomical variables are pivotal in predicting the target variable [35].

On the other hand, hyperparameter tuning, a pivotal aspect of model development, involved an exhaustive search for optimal parameters. We employed techniques such as grid search and cross-validation to find the parameter configuration that yielded the best predictive accuracy and generalization performance.

2.5.3. Model Training & Evaluation

The Random Forest Regressor was trained on the training data, utilizing a 70/30 train test split, allowing the selected features and engineered descriptors to learn the intricate relationships between PFAS accumulation and the GlomTotSA/ProxTubTotVol ratio.

To gauge the model's performance, we employed several evaluation metrics, including Mean Absolute Error (MAE) and R^2, which quantified the model's predictive accuracy. Additionally, we assessed the model's ability to generalize to unseen data through k-fold cross-validation, ensuring robustness and mitigating overfitting concerns.

3. Results

3.1. Accuracy

3.1.1. Kidney Type Prediction

To assess the performance of the model, the accuracy of the model had to be calculated by comparing the model's prediction of the training dataset against the actual values of the training dataset. Using this metric to calculate the model's accuracy, the model achieved an accuracy of 100%. The accuracy of this model can be depicted through the confusion matrix as seen in Figure 3. The figure shows that the predictions give the same output as is expected from the actual data, allowing for the machine learning model to be associated with perfect accuracy. Additionally, other performance metrics were used to measure the machine learning algorithm's performance, which can be seen in Table 1. These metrics can depict the fact that the accuracy can perfectly classify what type of kidney is affected by PFAS and the spread of PFAS throughout the body.

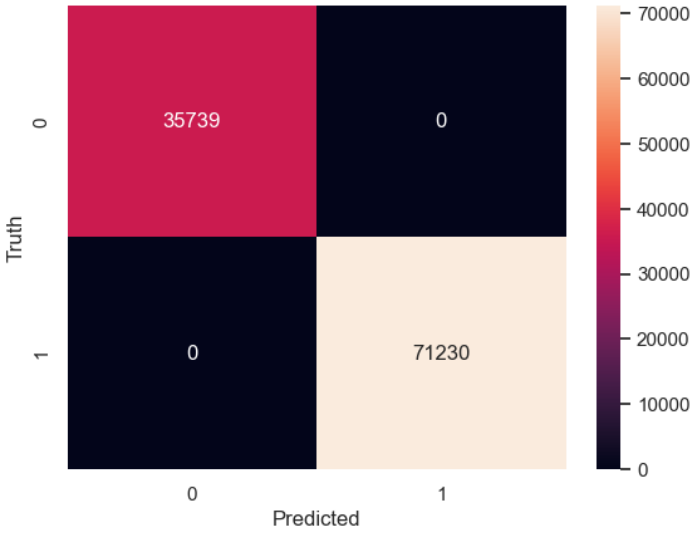


Figure 3: Confusion Matrix Showing the Machine Learning Model Predicting with Perfect Accuracy

Table 1: Machine Learning Model Performance Metrics

Labels	Precision	Recall	F1 Score	Support
Multirenculated	1.00	1.00	1.00	35739
Unipapillary	1.00	1.00	1.00	71230
Accuracy			1.00	106969
Macro Average	1.00	1.00	1.00	106969
Weighted Average	1.00	1.00	1.00	106969

Cross-validation scores were also calculated to ensure that the accuracy of the model was generealizable to data that wasn't specific to training the machine learning model itself. When done on a 5-fold cross validation model, all five scores contained 100% accuracy and and the standard deviation from these scores was 0%.

3.1.2. PFAS Accumulation Model

This machine learning model's accuracy was calculated using a few different methods since the model is a regression algorithm rather than a classifier. The first method used to calculate the accuracy of the model was to statistically evaluate the created regression line, which is done utilizing a R^2 value and the Mean Absolute Error(MAE). The R^2 value was 1.00, and the MAE was 0.00. These values indicate that the model was able to model the data perfectly. These values can be seen in Figure 4, where you can see how the graph depicts the alignment between the predicted and actual predictions. Additionally, these values were further justified through a cross-validation algorithm, which can prove the robustness of the model by verifying that the data split is not done by chance; it shows if the machine learning model is genuinely learning the data. The cross-validation statistics also validated that the machine learning model perfectly predicted the accumulation of PFAS, as it gave the same 1.00 R^2 value and 0.00 MAE. Additionally, these values can be further depicted in Figure 5, which shows the distribution of the residuals, which are calculated by subtracting the actual value from the predicted value, and the figure shows how the values are very close to 0, indicating that the model is predicting to almost perfection.

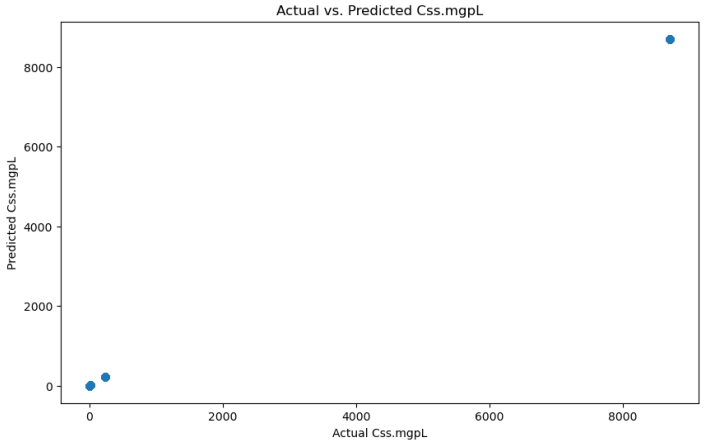


Figure 4: Comparison of the Predicted and Actual PFAS Accumulation Values

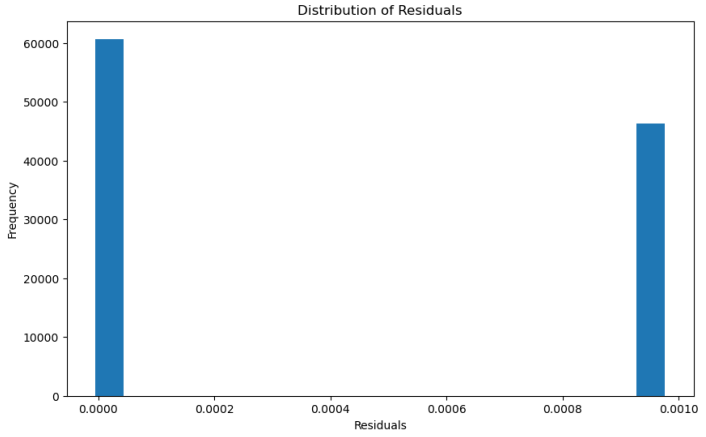


Figure 5: Residual Distribution Plot based on the Real PFAS Accumulation Values and the Predicted PFAS Accumulation Values

The cross validation results that were achieved for this model was a 0.00 MAE and a 1.00 R^2 on a 5-fold cross validation model.

3.1.3. Glomerular Total Surface Area vs Proximal Tubule Predictor

Similarly to the PFAS Accumulation Model, the accuracy of the machine learning model was calculated by finding the statistical values, which were verified through cross-validation, and the residuals were calculated. The initial statistical values calculated for the machine learning model were a 0.00 MAE and a 1.00 R^2 statistic. Furthermore, using cross-validation, the model still achieved an accuracy of 1.00. These values can be depicted through Figure 6, where it can be seen how the model is predicting the values with an incredibly high accuracy. Additionally, in Figure 7, you can see the plot of the distribution of the residuals, which shows an apparent discrepancy that spans a lot of data points. However, most of the data points are relatively close to the actual value.

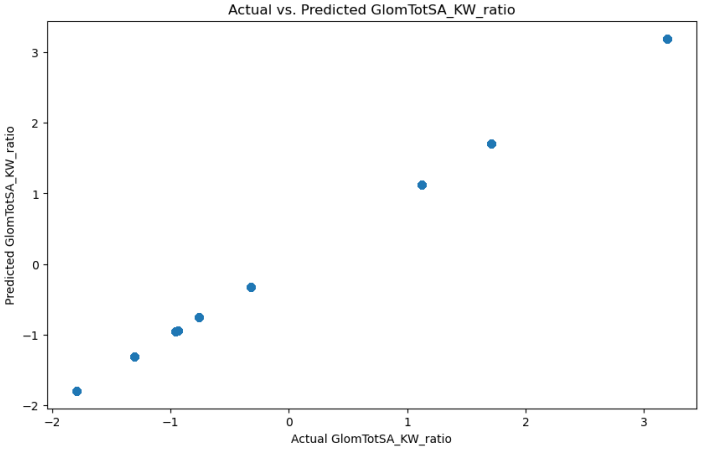


Figure 6: Comparisons of the Predicted and Actual Glomerular Total Surface Area and Proximal Tubule Ratio Values

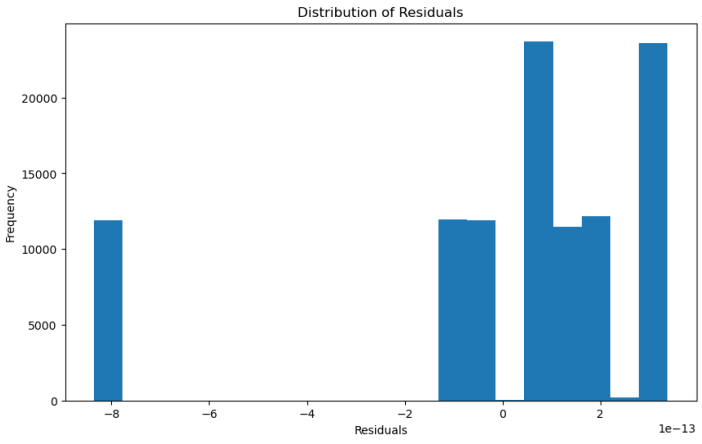


Figure 7: Residual Distribution Plot based on the Real Glomerular Total Surface Area and Proximal Tubule Ratio Values and the Predicted Glomerular Total Surface Area and Proximal Tubule Ratio Values

On a 5-fold cross validation the model is achieves a R^2 value of 1.00 and a MAE value of 0.00, showing the model is not just bound to the training data that was fed into it and it has based its information off of.

3.2. Feature Importance

After understanding the performance of the models, it is also essential to understand the key features indicating these specific outputs. Additionally, due to the specific models chosen as discussed in the methodology, it can be seen what features were critical to the specific models training rather than just looking at correlations within the dataset.

3.2.1. Kidney Type Prediction

The feature importances for the first model can be seen in Figure 8. The first model had some features with very high indications, but most did not correlate much with the type of kidney. The key features that can be seen in Figure X are the species of the anime and the proximal tubule diameter, which both seem to be understandable indicators. They both are reasonable indicators because the species directly affects the type and function of the kidney, which is critical to understanding which type it will be. Additionally, the proximal tubule diameter is a physical characteristic of kidneys, which indicates the type of kidney because it is a very clear descriptor of the kidney being affected.

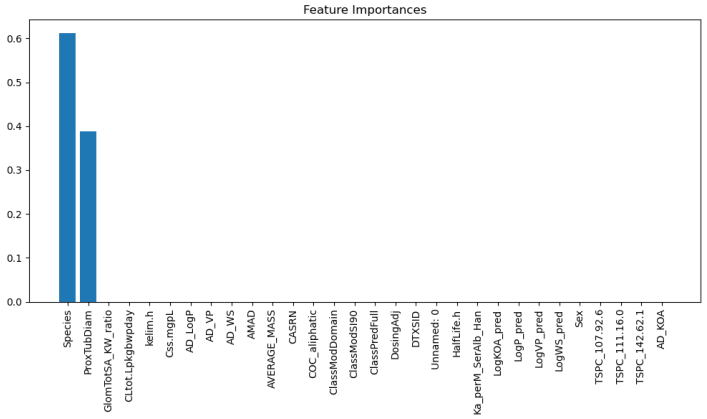


Figure 8: Feature Importance Plot for the Kidney Type Prediction Model

3.2.2. PFAS Accumulation Model

For this machine learning model, one feature that the model indicates has a direct correlation with the output of the model. The feature that the model is showing is the ClassPredFull, which is a metric inside of the dataset that tells us the level of exposure the body is to PFAS, which indicates that using this level, the machine learning algorithm is learning how to accurately predict how much PFAS is going to accumulate inside of the body and damage the kidney. This can be seen in Figure 9

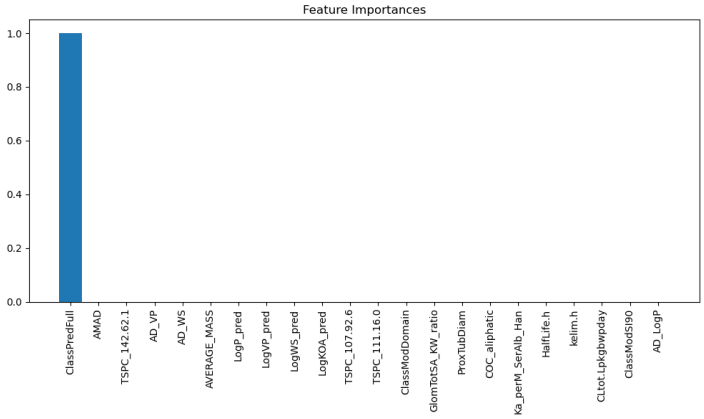


Figure 9: Feature Importance Plot for the PFAS Accumulation Model

3.2.3. Glomerular Total Surface Area vs Proximal Tubule Predictor

As seen in Figure 10, the third machine learning model indicates that two key features allow it to achieve a high accuracy. The two features that can indicate this are GlomTotSA_KW_ratio, which shows the ratio between the glomerular total surface area to the kidney weight, and the other feature, which can show a correlation to the output of this model is the diameter of the proximal tubule. Both of these features are understandable because they are descriptors of the physiological characteristics of the kidney or the specific proximal tubule, allowing for them to be correlated to the measurements that the machine learning model is concerned with predicting.

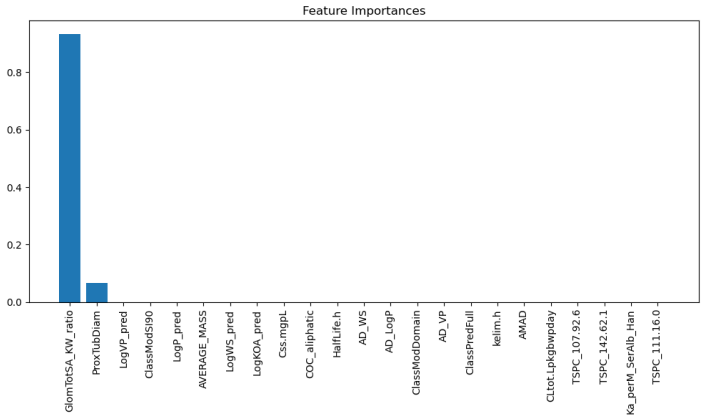


Figure 10: Feature Importance Plot for the Glomerular Total Surface Area vs Proximal Tubule Model

4. Biological Significance

As seen by the accuracy of the Kidney Type Prediction Model, we can safely verify previous studies and conclude that there is indeed a correlation between PFAS descriptors and kidneys to base the rest of our analysis. The ability of PFAS descriptors to predict kidney type with perfect accuracy suggests that PFAS compounds are not passive bystanders in the body but actively influence kidney morphology. As mentioned, PFAS, once absorbed into the bloodstream, can infiltrate renal tissues, where they interact with cellular components and trigger molecular responses [36]. This interaction may lead to structural changes in the kidneys, potentially altering the distribution of kidney types [37]. The fact that PFAS descriptors alone can reliably distinguish between different kidney types underscores the pronounced and biologically meaningful impact of PFAS exposure on renal structures.

Furthermore, the model provides significance in early diagnoses and organ assessment by predicting the amount of PFAS accumulation in the kidneys utilizing an XGBoost Regressor. Understanding how PFAS chemicals accumulate in vital organs like the kidneys is crucial for assessing the long-term health impacts of PFAS exposure. By employing XGBoost, researchers gain a powerful tool for modeling and predicting these accumulations, allowing for more accurate risk assessments and informed policy decisions. This predictive capability can guide regulatory bodies and healthcare providers in designing effective strategies to mitigate PFAS exposure and protect public health. By forecasting PFAS accumulation, we move beyond reacting to environmental contamination and instead proactively manage and mitigate its impacts.

The estimation of the ratio of Glomerular Total Surface Area (GlomTotSA) to the Volume of the Proximal Tubule (ProxTubTotVol) also offers significant scientific insight into the health and performance of our kidneys. The GlomTotSA-to-ProxTubTotVol ratio holds the key to understanding the structural efficiency of the kidneys. The glomerular surface area represents the vast expanse of tiny filtering units within the kidney, where blood is meticulously sieved to remove waste and excess substances [38]. The volume of the proximal tubule reflects the space available for the reabsorption of essential substances into the bloodstream, a crucial process in maintaining bodily homeostasis [39]. When this ratio is higher, it suggests that the kidneys are adept at filtration and reabsorption, indicative of a well-functioning organ [40]. This signifies that the kidneys efficiently clear waste and retain vital compounds, reflecting good renal health. Conversely, a lower GlomTotSA-to-ProxTubTotVol ratio can signify potential kidney morphology and function issues. This may point to structural alterations within the kidneys that affect their ability to filter and reabsorb substances optimally [41]. On a chemical level, this ratio can also be linked to the renal clearance of various substances, including drugs and toxins [42]. Understanding these chemical processes is pivotal for predicting how the kidneys will handle different compounds, informing medication dosages and toxicity assessments.

This ratio further provides insight into the complex interplay of filtration, reabsorption, and structural integrity within the kidneys [43]. It allows researchers to investigate the effects of PFAS exposure on these processes, potentially uncovering early indicators of kidney health issues. Moreover, it has applications beyond PFAS research, as it can be used as a valuable biomarker for assessing kidney function and diagnosing renal diseases. The ability to estimate this ratio with precision, informed by predictive PFAS accumulation models, represents a powerful tool for advancing our understanding of kidney health and clinical interventions and treatments.

However, the GlomTotSA-to-ProxTubTotVol ratio also offers a fruitful perspective on the intricate relationship between kidney structure and its functional role, particularly in the context of estimating glomerular filtration rate (eGFR) [44]. The eGFR is a critical

indicator of kidney function, primarily based on factors like creatinine levels, age, sex, and race, ultimately reflecting how efficiently the kidneys filter waste products from the bloodstream [45]. The GlomTotSA-to-ProxTubTotVol ratio and eGFR (estimated glomerular filtration rate) in kidneys are interconnected in a complex and critical manner, biologically and chemically. This relationship is pivotal for understanding renal function, particularly how well the kidneys filter waste and maintain overall homeostasis in the body. The GlomTotSA-to-ProxTubTotVol ratio is essentially a representation of the glomerular surface area relative to the proximal tubular volume, and eGFR is an estimate of the rate at which the glomeruli in the kidneys filter blood [46].

Biologically, this relationship hinges on the intricate anatomy and physiology of the renal system. The glomeruli, small tuft-like structures within the kidney, are the primary filtration units. They filter blood and allow water and solutes to enter the tubular system while retaining larger molecules like proteins. The proximal tubules, on the other hand, are involved in reabsorbing valuable substances such as glucose and electrolytes [42]. The ratio of the glomerular surface area to the proximal tubular volume reflects the balance between filtration and reabsorption in the kidney. A high GlomTotSA-to-ProxTubTotVol ratio suggests efficient filtration relative to reabsorption, typically associated with better renal function [47]. When the kidneys function optimally, the eGFR is higher because the glomeruli are efficiently filtering a greater blood volume, and waste products are being excreted. Conversely, if the ratio is skewed or compromised due to kidney damage or disease, the eGFR decreases, indicating impaired filtration capacity and potential renal dysfunction [48].

Chemically, the interaction is also governed by the intricate molecular exchange processes within the kidney. The glomeruli filter blood by using a combination of pressure-driven physical forces and the selective permeability of their basement membranes and podocyte cells [49]. These structures allow small molecules and water to pass into the renal tubules while preventing larger molecules like proteins from crossing over. The proximal tubules then selectively reabsorb essential molecules and regulate the concentration of electrolytes and waste products in the urine [50]. The ratio signifies the efficiency of these filtration and reabsorption processes. When the balance between filtration and reabsorption is disrupted, such as in kidney damage, inflammation, or glomerular dysfunction, the GlomTotSA-to-ProxTubTotVol ratio can change [51]. This imbalance decreases eGFR, as less blood is effectively filtered and more waste products may accumulate in the bloodstream [52]. In this chemical interplay, disruptions in the GlomTotSA-to-ProxTubTotVol ratio can serve as a valuable indicator of kidney function and provide insights into renal health or dysfunction. This ratio may also signify how effectively the kidneys handle the excretion of various substances. Efficient filtration and reabsorption are vital for maintaining electrolyte balance and eliminating waste, pharmaceuticals, and toxins from the body [53]. Therefore, if the ratio tilts towards a more significant glomerular surface area relative to the proximal tubule volume, it may suggest that the kidneys can process substances more effectively, potentially contributing to higher eGFR values.

5. Discussion

5.1. Limitations

The models were based off of a single dataset, so a dataset with more diversity and different populations is necessary to truly assess the causality and relationships that the models are indicating are true from the data provided. Additionally, investigations into the biological mechanisms which are highlighted through the research study and experimental validation are necessary for understanding the true biological processes and pathways that are associated with PFAS and its effect on the body. Also, there are some confounding variables which could have potential effects on the results of the finding, such as environmental exposures which have the ability to effects the observed relationships.

5.2. Conclusion

This study demonstrates machine learning's potential to elucidate the intricate impacts of PFAS on kidney health. The models presented confirm discernible links between PFAS and renal function. The kidney type classifier verifies that PFAS alters morphology. The PFAS accumulation regressor enables clinical monitoring to protect at-risk groups. Estimating the GlomTotSA/ProxTubTotVol ratio provides significant insights into PFAS's effects on filtration and reabsorption efficiency.

The models showcase PFAS's multifaceted effects on kidney structure and function. The techniques pave the way for enhanced risk assessment, improved clinical surveillance, and targeted therapeutics. This study underscores the power of machine learning to unravel PFAS toxicity mechanisms. It makes significant contributions to the current understanding of PFAS nephrotoxicity.

The high predictive accuracy, feature importance analyses, and model interpretability reveal concrete biological impacts of PFAS on kidneys. PFAS descriptors actively influence morphology and accumulation. The ratio estimation sheds light on filtration and reabsorption dynamics. These findings confirm PFAS's pronounced effects on renal physiology.

This research demonstrates machine learning's immense potential for elucidating PFAS nephrotoxicity. The models provide actionable clinical insights and expand mechanistic knowledge of PFAS-kidney interplay. This paves the way for more informed risk analysis, earlier diagnosis of PFAS-associated kidney damage, and targeted therapies. The study sets a strong foundation for future machine-learning investigations into PFAS toxicity.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment We would like to thank the University of North Texas for providing us with the resources and support to conduct this research. The invaluable guidance and encouragement from our professors and mentors have been instrumental in shaping the direction and scope of this study. We would also like to acknowledge the National Kidney Foundation for their support and inspiration in conducting this project. Finally, we would also like to thank our families for supporting us throughout our research.

References

[1] J. Glüge, M. Scheringer, I. T. Cousins, J. C. DeWitt, G. Goldenman, D. Herzke, R. Lohmann, C. A. Ng, X. Trier, Z. Wang, "An overview of the uses of per- and polyfluoroalkyl substances (PFAS)," Environmental Science: Processes & Impacts, **22**, 2345–2373, 2020, doi:10.1039/D0EM00291G.

[2] P. Dewapriya, L. Chadwick, S. G. Gorji, B. Schulze, S. Valsecchi, S. Samanipour, K. V. Thomas, S. L. Kaserzon, "Per- and polyfluoroalkyl substances (PFAS) in consumer products: Current knowledge and research gaps," Journal of Hazardous Materials Letters, **4**, 100086, 2023, doi:10.1016/j.hazl.2023.100086.

[3] B. E. Blake, S. E. Fenton, "Early life exposure to per- and polyfluoroalkyl substances (PFAS) and latent health outcomes: A review including the placenta as a target tissue and possible driver of peri- and postnatal effects," Toxicology, **443**, 152565, 2020, doi:10.1016/j.tox.2020.152565.

[4] D. Renfrew, T. W. Pearson, "The Social Life of the "Forever Chemical"," Environment and Society, **12**, 146–163, 2021, doi:10.3167/ares.2021.120109.

[5] M. Kotthoff, J. Müller, H. Jürling, M. Schlummer, D. Fiedler, "Perfluoroalkyl and polyfluoroalkyl substances in consumer products," Environmental Science and Pollution Research, **22**, 14546–14559, 2015, doi:10.1007/s11356-015-4202-7.

[6] S. A. Bălan, V. C. Mathrani, D. F. Guo, A. M. Algazi, "Regulating PFAS as a Chemical Class under the California Safer Consumer Products Program," Environmental Health Perspectives, **129**, 2021, doi:10.1289/EHP7431.

[7] M. N. Ehsan, M. Riza, M. N. Pervez, M. M. O. Khyum, Y. Liang, V. Naddeo, "Environmental and health impacts of PFAS: Sources, distribution and sustainable management in North Carolina (USA)," Science of The Total Environment, **878**, 163123, 2023, doi:10.1016/j.scitotenv.2023.163123.

[8] J. Fabregat-Palau, M. Vidal, A. Rigol, "Examining sorption of perfluoroalkyl substances (PFAS) in biochars and other carbon-rich materials," Chemosphere, **302**, 134733, 2022, doi:10.1016/j.chemosphere.2022.134733.

[9] G. Jha, V. Kankarla, E. McLennon, S. Pal, D. Sihi, B. Dari, D. Diaz, M. Nocco, "Per- and Polyfluoroalkyl Substances (PFAS) in Integrated Crop–Livestock Systems: Environmental Exposure and Human Health Risks," International Journal of Environmental Research and Public Health, **18**, 12550, 2021, doi:10.3390/ijerph182312550.

[10] B. E. Blake, S. M. Pinney, E. P. Hines, S. E. Fenton, K. K. Ferguson, "Associations between longitudinal serum perfluoroalkyl substance (PFAS) levels and measures of thyroid hormone, kidney function, and body mass index in the Fernald Community Cohort," Environmental Pollution, **242**, 894–904, 2018, doi:10.1016/j.envpol.2018.07.042.

[11] M. L. Ljubicic, A. Madsen, A. Juul, K. Almstrup, T. H. Johannsen, "The Application of Principal Component Analysis on Clinical and Biochemical Parameters Exemplified in Children With Congenital Adrenal Hyperplasia," Frontiers in Endocrinology, **12**, 2021, doi:10.3389/fendo.2021.652888.

[12] M. Bonato, F. Corrà, M. Bellio, L. Guidolin, L. Tallandini, P. Irato, G. Santovito, "PFAS Environmental Pollution and Antioxidant Responses: An Overview of the Impact on Human Field," International Journal of Environmental Research and Public Health, **17**, 8020, 2020, doi:10.3390/ijerph17218020.

[13] K. M. Fraley, H. N. Fraley, D. Arthur, E. J. Walther, "Per- and Polyfluoroalkyl Substances (PFAS): Anglers May Be Exposed to Harmful Chemicals in Their Catch," Fisheries, **45**, 138–144, 2020, doi:10.1002/fsh.10389.

[14] L. Anderko, E. Pennea, "Exposures to per-and polyfluoroalkyl substances (PFAS): Potential risks to reproductive and children's health," Current Problems in Pediatric and Adolescent Health Care, **50**, 100760, 2020, doi:10.1016/j.cppeds.2020.100760.

[15] S. H. Baker, A. Kinde, "The Pathway to a Green New Deal: Synthesizing Transdisciplinary Literatures and Activist Frameworks to Achieve a Just Energy Transition," Environs: Environmental Law and Policy Journal, **44**, 1–40.

[16] P.-I. D. Lin, A. Cardenas, R. Hauser, D. R. Gold, K. P. Kleinman, M.-F. Hivert, A. M. Calafat, T. F. Webster, E. S. Horton, E. Oken, "Per- and polyfluoroalkyl substances and kidney function: Follow-up results from the Diabetes Prevention Program trial," Environment International, **148**, 106375, 2021, doi:10.1016/j.envint.2020.106375.

[17] W. PFALLER, M. RITTINGER, QUANTITATIVE MORPHOLOGY OF THE RAT KIDNEY, 17–22, Elsevier, 1980, doi:10.1016/B978-0-08-025517-0.50011-4.

[18] J. Yun, E.-C. Jang, S.-C. Kwon, Y.-S. Min, Y.-J. Lee, "The association of perfluoroalkyl substances (PFAS) exposure and kidney function in Korean adolescents using data from Korean National Environmental Health Survey (KoNEHS) cycle 4 (2018–2020): a cross-sectional study," Annals of Occupational and Environmental Medicine, **35**, 2023, doi:10.35371/aoem.2023.35.e5.

[19] J. W. Stanifer, H. M. Stapleton, T. Souma, A. Wittmer, X. Zhao, L. E. Boulware, "Perfluorinated Chemicals as Emerging Environmental Threats to Kidney Health," Clinical Journal of the American Society of Nephrology, **13**, 1479–1492, 2018, doi:10.2215/CJN.04670418.

[20] B. Conway, A. Badders, T. Costacou, J. Arthur, K. Innes, "Perfluoroalkyl substances and kidney function in chronic kidney disease, anemia, and diabetes," Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy, **Volume 11**, 707–716, 2018, doi:10.2147/DMSO.S173809.

[21] F. Coperchini, L. Croce, G. Ricci, F. Magri, M. Rotondi, M. Imbriani, L. Chiovato, "Thyroid Disrupting Effects of Old and New Generation PFAS," Frontiers in Endocrinology, **11**, 2021, doi:10.3389/fendo.2020.612320.

[22] A. Raza, S. Bardhan, L. Xu, S. S. R. K. C. Yamijala, C. Lian, H. Kwon, B. M. Wong, "A Machine Learning Approach for Predicting Defluorination of Per- and Polyfluoroalkyl Substances (PFAS) for Their Efficient Treatment and Removal," Environmental Science & Technology Letters, **6**, 624–629, 2019, doi:10.1021/acs.estlett.9b00476.

[23] C.-W. Huang, R. Lu, U. Iqbal, S.-H. Lin, P. A. Nguyen, H.-C. Yang, C.-F. Wang, J. Li, K.-L. Ma, Y.-C. Li, W.-S. Jian, "A richly interactive exploratory data analysis and visualization tool using electronic medical records," BMC Medical Informatics and Decision Making, **15**, 92, 2015, doi:10.1186/s12911-015-0218-7.

[24] D. E. Dawson, C. Lau, P. Pradeep, R. R. Sayre, R. S. Judson, R. Tornero-Velez, J. F. Wambaugh, "A Machine Learning Model to Estimate Toxicokinetic Half-Lives of Per- and Polyfluoro-Alkyl Substances (PFAS) in Multiple Species," Toxics, **11**, 2023, doi:10.3390/toxics11020098.

[25] R. Paul, A. S. M. L. Hoque, "Clustering medical data to predict the likelihood of diseases," in 2010 Fifth International Conference on Digital Information Management (ICDIM), 44–49, 2010, doi:10.1109/ICDIM.2010.5664638.

[26] X. Gong, L. Liu, S. Fong, Q. Xu, T. Wen, Z. Liu, "Comparative Research of Swarm Intelligence Clustering Algorithms for Analyzing Medical Data," IEEE Access, **7**, 137560–137569, 2019, doi:10.1109/ACCESS.2018.2881020.

[27] J. Ma, Z. Yu, Y. Qu, J. Xu, Y. Cao, "Application of the XGBoost Machine Learning Method in PM2.5 Prediction: A Case Study of Shanghai," Aerosol and Air Quality Research, **20**, 128–138, 2020, doi:10.4209/aaqr.2019.08.0408.

[28] H. Liang, K. Jiang, T.-A. Yan, G.-H. Chen, "XGBoost: An Optimal Machine Learning Model with Just Structural Features to Discover MOF Adsorbents of Xe/Kr," ACS Omega, **6**, 9066–9076, 2021, doi:10.1021/acsomega.1c00100.

[29] J. Li, X. An, Q. Li, C. Wang, H. Yu, X. Zhou, Y. ao Geng, "Application of XGBoost algorithm in the optimization of pollutant concentration," Atmospheric Research, **276**, 106238, 2022, doi:10.1016/j.atmosres.2022.106238.

[30] A. Ogunleye, Q.-G. Wang, "XGBoost Model for Chronic Kidney Disease Diagnosis," IEEE/ACM Transactions on Computational Biology and Bioinformatics, **17**(6), 2131–2140, 2020, doi:10.1109/TCBB.2019.2911071.

[31] X. Zhang, C. Yan, C. Gao, B. A. Malin, Y. Chen, "Predicting Missing Values in Medical Data Via XGBoost Regression," Journal of Healthcare Informatics Research, **4**, 383–394, 2020, doi:10.1007/s41666-020-00077-1.

[32] R. Sapir-Pichhadze, B. Kaplan, "Seeing the Forest for the Trees: Random Forest Models for Predicting Survival in Kidney Transplant Recipients," Transplantation, **104**, 905–906, 2020, doi:10.1097/TP.0000000000002923.

[33] A. Subasi, E. Alickovic, J. Kevric, Diagnosis of Chronic Kidney Disease by Using Random Forest, 589–594, 2017, doi:10.1007/978-981-10-4166-2_89.

[34] M. Z. Alam, M. S. Rahman, M. S. Rahman, "A Random Forest based predictor for medical data classification using feature ranking," Informatics in Medicine Unlocked, **15**, 100180, 2019, doi:10.1016/j.imu.2019.100180.

[35] P. G. Polishchuk, E. N. Muratov, A. G. Artemenko, O. G. Kolumbin, N. N. Muratov, V. E. Kuz'min, "Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity," Journal of Chemical Information and Modeling, **49**, 2481–2488, 2009, doi:10.1021/ci900203n.

[36] E. Gorrochategui, S. Lacorte, R. Tauler, F. L. Martin, "Perfluoroalkylated Substance Effects in Xenopus laevis A6 Kidney Epithelial Cells Determined by ATR-FTIR Spectroscopy and Chemometric Analysis," Chemical Research in Toxicology, **29**, 924–932, 2016, doi:10.1021/acs.chemrestox.6b00076.

[37] J. Zhao, P. Hinton, J. Chen, J. Jiang, "Causal inference for the effect of environmental chemicals on chronic kidney disease," Computational and Structural Biotechnology Journal, **18**, 93–99, 2020, doi:10.1016/j.csbj.2019.12.001.

[38] R. B. Jain, A. Ducatman, "Associations between the concentrations of α-klotho and selected perfluoroalkyl substances in the presence of eGFR based kidney function and albuminuria: Data for US adults aged 40–79 years," Science of The Total Environment, **838**, 155994, 2022, doi:10.1016/j.scitotenv.2022.155994.

[39] B. M. Brenner, J. L. Troy, T. M. Daugharty, I. F. Ueki, D. P. Nicholas, C. F. Wong, "On the Mechanism of Inhibition in Fluid Reabsorption by the Renal Proximal Tubule of the Volume-Expanded Rat," Journal of Clinical Investigation, **50**, 1596–1602, 1971, doi:10.1172/JCI106647.

[40] R. Quigley, "Androgens stimulate proximal tubule transport," Gender Medicine, **5**, S114–S120, 2008, doi:10.1016/j.genm.2008.03.011.

[41] A. Quan, S. Chakravarty, J.-K. Chen, J.-C. Chen, S. Loleh, N. Saini, R. C. Harris, J. Capdevila, R. Quigley, "Androgens augment proximal tubule transport," American Journal of Physiology-Renal Physiology, **287**, F452–F459, 2004, doi:10.1152/ajprenal.00188.2003.

[42] A. A. McDonough, "Mechanisms of proximal tubule sodium transport regulation that link extracellular fluid volume and blood pressure," American Journal of Physiology-Regulatory, Integrative and Comparative Physiology, **298**, R851–R861, 2010, doi:10.1152/ajpregu.00002.2010.

[43] B. Kaissling, I. Hegyi, J. Loffing, M. Hir, "Morphology of interstitial cells in the healthy kidney," Anatomy and Embryology, **193**, 1996, doi:10.1007/BF00186688.

[44] P. Delanaye, R. P. Radermecker, M. Rorive, G. Depas, J. M. Krzesinski, "Indexing glomerular filtration rate for body surface area in obese patients is misleading: concept and example," Nephrology Dialysis Transplantation, **20**, 2024–2028, 2005, doi:10.1093/ndt/gfh983.

[45] J. R. Nyengaard, T. F. Bendtsen, "Glomerular number and size in relation to age, kidney weight, and body surface in normal man," The Anatomical Record, **232**, 194–201, 1992, doi:10.1002/ar.1092320205.

[46] C. C. Geddes, Y. M. Woo, S. Brady, "Glomerular filtration rate what is the rationale and justification of normalizing GFR for body surface area?" Nephrology Dialysis Transplantation, **23**, 4–6, 2007, doi:10.1093/ndt/gfm662.

[47] G. Vervoort, B. Veldman, J. H. M. Berden, P. Smits, J. F. M. Wetzels, "Glomerular hyperfiltration in type 1 diabetes mellitus results from primary changes in proximal tubular sodium handling without changes in volume expansion," European Journal of Clinical Investigation, **35**, 330–336, 2005, doi:10.1111/j.1365-2362.2005.01497.x.

[48] W.-A. S. Mula-Abed, K. A. Rasadi, D. A. Riyami, "Estimated Glomerular Filtration Rate (eGFR): A Serum Creatinine-Based Test for the Detection of Chronic Kidney Disease and its Impact on Clinical Practice," Oman Medical Journal, **27**, 108–113, 2012, doi:10.5001/omj.2012.23.

[49] S. Rayego-Mateos, R. Rodrigues-Diez, J. L. Morgado-Pascual, F. Valentijn, J. M. Valdivielso, R. Goldschmeding, M. Ruiz-Ortega, "Role of Epidermal Growth Factor Receptor (EGFR) and Its Ligands in Kidney Inflammation and Damage," Mediators of Inflammation, **2018**, 8739473, 2018, doi:10.1155/2018/8739473.

[50] J. Tang, N. Liu, E. Tolbert, M. Ponnusamy, L. Ma, R. Gong, G. Bayliss, H. Yan, S. Zhuang, "Sustained Activation of EGFR Triggers Renal Fibrogenesis after Acute Kidney Injury," The American Journal of Pathology, **183**, 160–172, 2013, doi:https://doi.org/10.1016/j.ajpath.2013.04.005.

[51] C. A. O'Callaghan, B. Shine, D. S. Lasserson, "Chronic kidney disease: a large-scale population-based study of the effects of introducing the ¡em¿CKD-EPI¡/em¿ formula for eGFR reporting," BMJ Open, **1**, e000308, 2011, doi:10.1136/bmjopen-2011-000308.

[52] N. Richards, K. Harris, M. Whitfield, D. O'Donoghue, R. Lewis, M. Mansell, S. Thomas, J. Townend, M. Eames, D. Marcelli, "The impact of population-based identification of chronic kidney disease using estimated glomerular filtration rate (eGFR) reporting," Nephrology Dialysis Transplantation, **23**, 556–561, 2008, doi:10.1093/ndt/gfm839.

[53] J. A. Hirst, M. D. L. A. V. Montes, C. J. Taylor, J. M. Ordóñez-Mena, E. Ogburn, V. Sharma, B. Shine, T. James, F. D. R. Hobbs, "Impact of a single eGFR and eGFR-estimating equation on chronic kidney disease reclassification: a cohort study in primary care," British Journal of General Practice, **68**, e524, 2018, doi:10.3399/bjgp18X697937.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 2, 67-73 (2024)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Malaysia's Renewable Energy Policy and its Impact on ASEAN Countries

WaiYie Leong*[1], LeeSun Heng[2], YuanZhi Leong[2]

[1] *INTI International University, Persiaran Perdana BBN Putra Nilai, 71800 Nilai, Negeri Sembilan*

[2] *Schneider Electric, 50 Kallang Ave, Singapore 339505*

ARTICLE INFO

Article history:
Received: 08 February, 2024
Revised: 17 April, 2024
Accepted: 18 April, 2024
Online: 30 April, 2024

Keywords:
Renewable Energy Policy
Green Technology
Sustainable Development
Energy Efficiency

ABSTRACT

As the global community increasingly shifts its focus towards sustainable development and combating climate change, renewable energy policies have become pivotal in shaping national and regional energy landscapes. Malaysia, as a developing nation with a rapidly growing economy, has recognized the importance of renewable energy sources in achieving its socio-economic goals while addressing environmental concerns. This paper explores Malaysia's renewable energy policy framework and assesses its impact on neighboring countries in the Southeast Asian region. Through a comprehensive analysis of policy measures, incentives, challenges, and achievements, this paper elucidates the significance of Malaysia's renewable energy endeavors in fostering regional energy security, sustainability, and cooperation.

1. Introduction

In the face of global climate change and the need for sustainable development, countries worldwide are increasingly turning towards renewable energy sources. Malaysia, a rapidly developing nation in Southeast Asia, has recognized the importance of renewable energy in its energy transition. The Malaysian government has implemented various policies and initiatives to promote renewable energy adoption, aiming not only to meet domestic energy needs but also to have a positive impact on the region's energy landscape [1-2]. This paper examines Malaysia's renewable energy policy framework and analyzes its impact on neighboring countries in the region.

1.1. Malaysia's Renewable Energy Policy Framework

Malaysia's journey towards renewable energy adoption began with the enactment of the Renewable Energy Act in 2011 [1]. This legislation laid the groundwork for various initiatives aimed at promoting renewable energy development. The key components of Malaysia's renewable energy policy framework include:

A) National Renewable Energy Policy and Action Plan: The Malaysian government formulated a comprehensive National Renewable Energy Policy and Action Plan, outlining targets and strategies for renewable energy deployment. This policy framework provides a roadmap for transitioning towards a more sustainable energy mix [2].

B) Feed-in Tariff (FiT) Mechanism: The FiT mechanism, introduced in 2011, offers attractive tariffs to renewable energy producers, thereby incentivizing investment in solar, biomass, and biogas projects. This policy has stimulated significant growth in renewable energy capacity across Malaysia.

C) Net Energy Metering (NEM) Scheme: The NEM scheme allows renewable energy system owners to offset their electricity bills by exporting excess energy to the grid [3]. This initiative encourages the adoption of rooftop solar systems among residential, commercial, and industrial consumers.

D) Renewable Energy Fund: The government has established a Renewable Energy Fund to support research, development, and deployment of renewable energy technologies [4]. This fund facilitates innovation and capacity building within the renewable energy sector.

The implementation of Malaysia's renewable energy policies has not only transformed its domestic energy landscape but also had significant implications for neighboring countries in the region.

2. Literature Review of Malaysia's Renewable Energy Policy

Malaysia, a country in Southeast Asia, has been actively pursuing renewable energy initiatives to address energy security concerns, reduce dependency on fossil fuels, and mitigate climate

*Corresponding Author: WaiYie Leong, INTI International University, Negeri Sembilan, waiyie@gmail.com

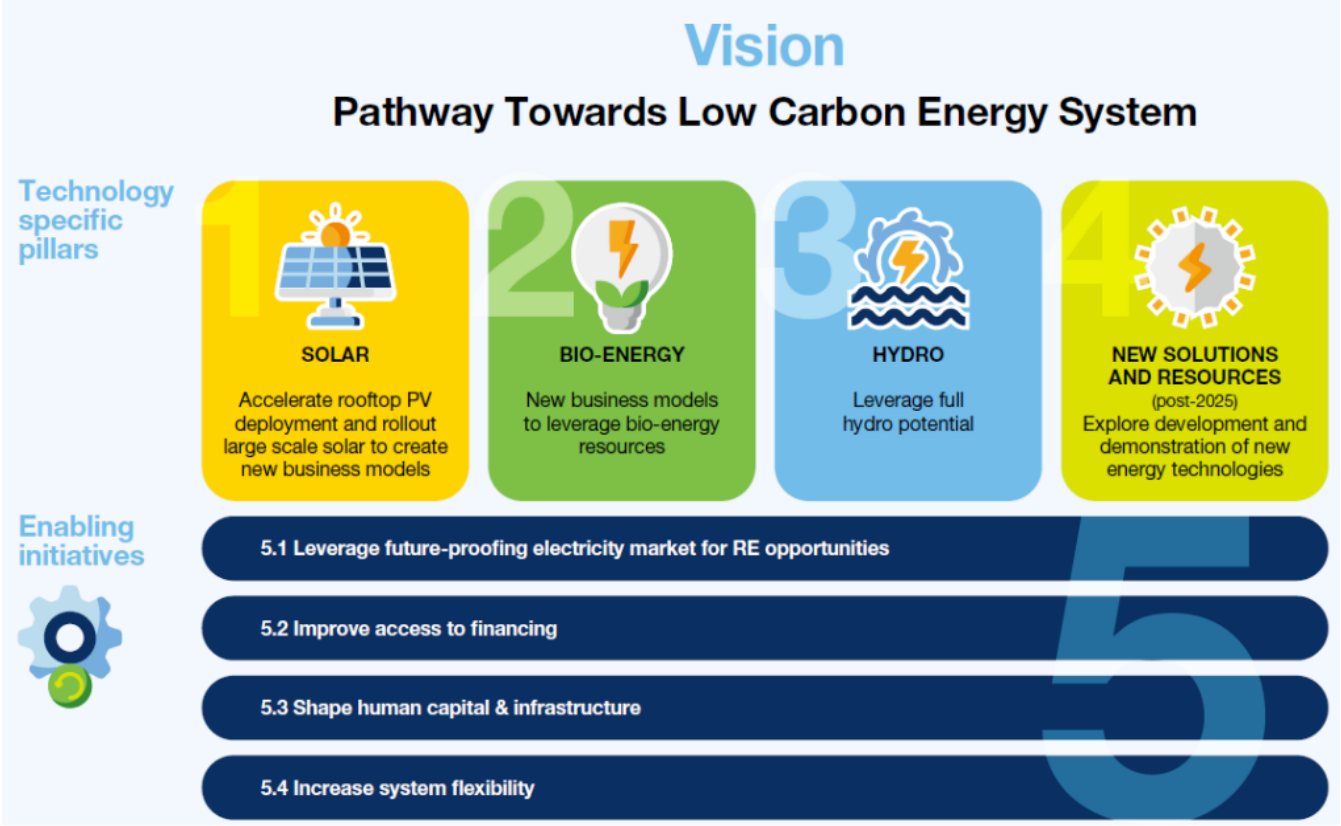


Figure 1: Strategic Plan and Framework for MyRER [1]

change. This section provides a historical overview and literature review of Malaysia's renewable energy policy development and its impact on neighboring countries in the region.

Malaysia's interest in renewable energy dates back to the early 2000s when the government recognized the need for sustainable energy development [1,5]. Key milestones include the establishment of the National Renewable Energy Policy and Action Plan in 2009, which set targets for renewable energy adoption and introduced policy mechanisms such as Feed-in Tariffs (FiT) and Net Energy Metering (NEM). Subsequent revisions and updates to the policy framework have aimed to accelerate renewable energy deployment and attract investment in the sector. Malaysia's renewable energy targets have been incorporated into national development plans, demonstrating the government's commitment to transitioning towards a low-carbon economy.

Numerous studies have examined various aspects of Malaysia's renewable energy policy and its impact on the region. Research has focused on the effectiveness of policy instruments such as FiT and NEM in incentivizing renewable energy investment and deployment. Scholars have also explored the potential of different renewable energy sources, including solar, wind, biomass, and hydro, in Malaysia's energy transition. Assessments of Malaysia's renewable energy market dynamics, investment climate, and regulatory framework have provided insights into the challenges and opportunities facing the sector.

The impact of Malaysia's renewable energy policy extends beyond its borders, influencing regional energy dynamics and cooperation. Enhanced regional energy security is achieved through diversified energy sources and reduced reliance on imported fossil fuels. Economic opportunities arise from cross-border investment and collaboration in renewable energy projects, fostering regional integration and development. Knowledge sharing and technology transfer initiatives contribute to capacity building and sustainable energy development in neighboring countries. Malaysia's leadership in renewable energy policy serves as a model for regional peers, influencing the formulation of renewable energy strategies and policies across Southeast Asia.

Malaysia's renewable energy policy has undergone significant development over the years, shaping the country's energy landscape and influencing regional energy dynamics. Through policy initiatives, incentives, and collaborations, Malaysia has made strides towards a more sustainable and resilient energy future. Continued research and cooperation among regional countries are essential for advancing renewable energy adoption and achieving shared climate and energy goals.

3. MyRER: Malaysia Renewable Energy Roadmap

The Malaysia Renewable Energy Roadmap (MyRER) is a strategic plan developed by the Malaysian government to promote the growth and adoption of renewable energy sources in the country. It aims to diversify Malaysia's energy mix, reduce dependence on fossil fuels, and mitigate the environmental impacts of energy generation.

Key objectives of MyRER typically include:

A) Increasing Renewable Energy Capacity: Setting targets for the installation of renewable energy capacity, such as solar, wind, hydroelectric, biomass, and biogas power plants.

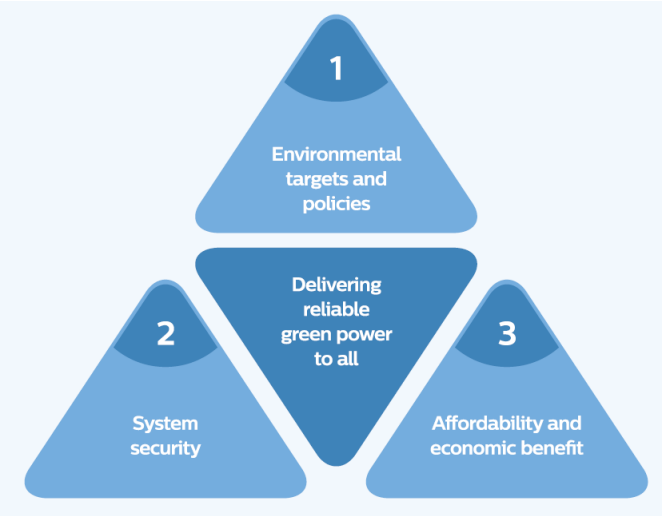


Figure 2: Principles to follow in the pursuit of dependable green power [1]

B) Policy and Regulatory Framework: Developing supportive policies, regulations, and incentives to attract investment in renewable energy projects and streamline approval processes.

C) Infrastructure Development: Investing in the necessary infrastructure, such as grid upgrades and energy storage facilities, to support the integration of renewable energy into the existing power system.

D) Research and Development: Encouraging research and development efforts to improve renewable energy technologies, reduce costs, and enhance efficiency.

E) Capacity Building and Awareness: Providing training programs, workshops, and public awareness campaigns to build capacity and promote the benefits of renewable energy adoption.

F) Collaboration and Partnerships: Fostering collaboration between government agencies, private sector entities, academia, and international organizations to achieve the goals of the roadmap.

MyRER serves as a guiding framework for the Malaysian government and stakeholders in the renewable energy sector to work towards sustainable energy development and achieve national energy security goals. It typically outlines specific targets, timelines, and action plans to monitor progress and ensure the successful implementation of renewable energy initiatives.

4. MyRER: The Milestones and Aspirations

The Malaysia Renewable Energy Roadmap (MyRER) typically outlines a series of milestones and aspirations to guide the country's transition towards a more sustainable and renewable energy future. These milestones and aspirations may vary depending on the specific version or iteration of MyRER, as well as evolving national priorities and goals. However, here are some common milestones and aspirations that could be included:

A) Setting Renewable Energy Targets: Establishing ambitious targets for the deployment of renewable energy capacity in Malaysia, including targets for solar, wind, hydroelectric, biomass, and biogas power generation.

B) Policy and Regulatory Reforms: Implementing policy and regulatory reforms to create a conducive environment for renewable energy investment, such as feed-in tariffs, net metering schemes, renewable energy auctions, and streamlined permitting processes.

C) Infrastructure Development: Investing in grid infrastructure upgrades, energy storage facilities, and smart grid technologies to support the integration of variable renewable energy sources into the electricity grid.

D) Promotion of Solar Energy: Encouraging the widespread adoption of solar photovoltaic (PV) systems, including rooftop solar installations on residential, commercial, and industrial buildings, as well as utility-scale solar farms.

E) Expansion of Wind Power: Identifying and developing suitable sites for wind energy projects, particularly in coastal and high-wind areas, to harness Malaysia's wind energy potential.

F) Hydropower Development: Continuing to develop hydropower projects in a sustainable manner, balancing energy generation with environmental and social considerations.

G) Biomass and Biogas Utilization: Promoting the use of biomass and biogas for electricity generation, industrial processes, and transportation fuels, leveraging Malaysia's abundant agricultural and organic waste resources.

H) Research and Innovation: Investing in research and innovation to drive down the costs of renewable energy technologies, improve efficiency, and address technical challenges.

I) Capacity Building and Education: Providing training programs, workshops, and educational initiatives to build local expertise and raise awareness about the benefits of renewable energy.

J) International Collaboration: Collaborating with international partners, organizations, and experts to share best practices, technology transfer, and financial assistance for renewable energy projects.

K) These milestones and aspirations are typically designed to accelerate the transition towards a more sustainable, low-carbon energy system in Malaysia, while also promoting economic growth, energy security, and environmental stewardship. Regular monitoring and evaluation of progress are essential to ensure that MyRER goals are being met and to make adjustments as needed to stay on track.

5. MyRER: The Strategic Framework

The strategic framework of the Malaysia Renewable Energy Roadmap (MyRER) typically encompasses a comprehensive approach to advancing renewable energy deployment and achieving sustainability goals [1,5]. While the specific details may vary, depending on the version and iteration of MyRER, the

strategic framework generally includes the following key components:

A) Policy and Regulatory Environment: Establishing a supportive policy and regulatory framework that incentivizes investment in renewable energy and facilitates its integration into the energy market [6]. This may include feed-in tariffs, renewable energy quotas, net metering schemes, tax incentives, and streamlined permitting processes.

B) Market Development and Financing: Stimulating market demand for renewable energy through mechanisms such as renewable energy auctions, power purchase agreements (PPAs), green procurement policies, and financial incentives for renewable energy projects [7]. Access to financing, including loans, grants, and venture capital, is also crucial for accelerating renewable energy deployment.

C) Infrastructure and Grid Integration: Investing in grid infrastructure upgrades, energy storage systems, and smart grid technologies to accommodate the variability and intermittency of renewable energy sources [8]. This includes grid expansion to remote areas with high renewable energy potential and implementing grid codes and standards to ensure the reliable and efficient operation of the grid with increased renewable energy penetration.

D) Resource Assessment and Planning: Conducting comprehensive resource assessments to identify and prioritize renewable energy sources, such as solar, wind, hydroelectric, biomass, and biogas, with the greatest potential for development. Long-term energy planning, including scenario analysis and forecasting, is essential for guiding investment decisions and optimizing the mix of renewable energy technologies [9].

E) Technology Innovation and Research: Supporting research, development, and innovation in renewable energy technologies to improve efficiency, reduce costs, and address technical challenges. This includes funding for research institutions, technology demonstration projects, and collaborative R&D initiatives with industry partners.

F) Capacity Building and Human Capital Development: Building local expertise and technical skills in renewable energy development, operation, and maintenance through training programs, education initiatives, and partnerships with academic institutions and industry associations. This includes vocational training, professional certification programs, and knowledge sharing platforms to enhance workforce readiness and competency.

G) Public Awareness and Stakeholder Engagement: Raising awareness among policymakers, businesses, communities, and the general public about the benefits of renewable energy, including environmental sustainability, energy security, job creation, and economic development. Engaging stakeholders through consultations, workshops, and outreach activities fosters support for renewable energy policies and projects.

H) Monitoring, Evaluation, and Review Mechanisms: Establishing robust monitoring, evaluation, and review mechanisms to track progress towards renewable energy targets, assess the effectiveness of policy measures, and

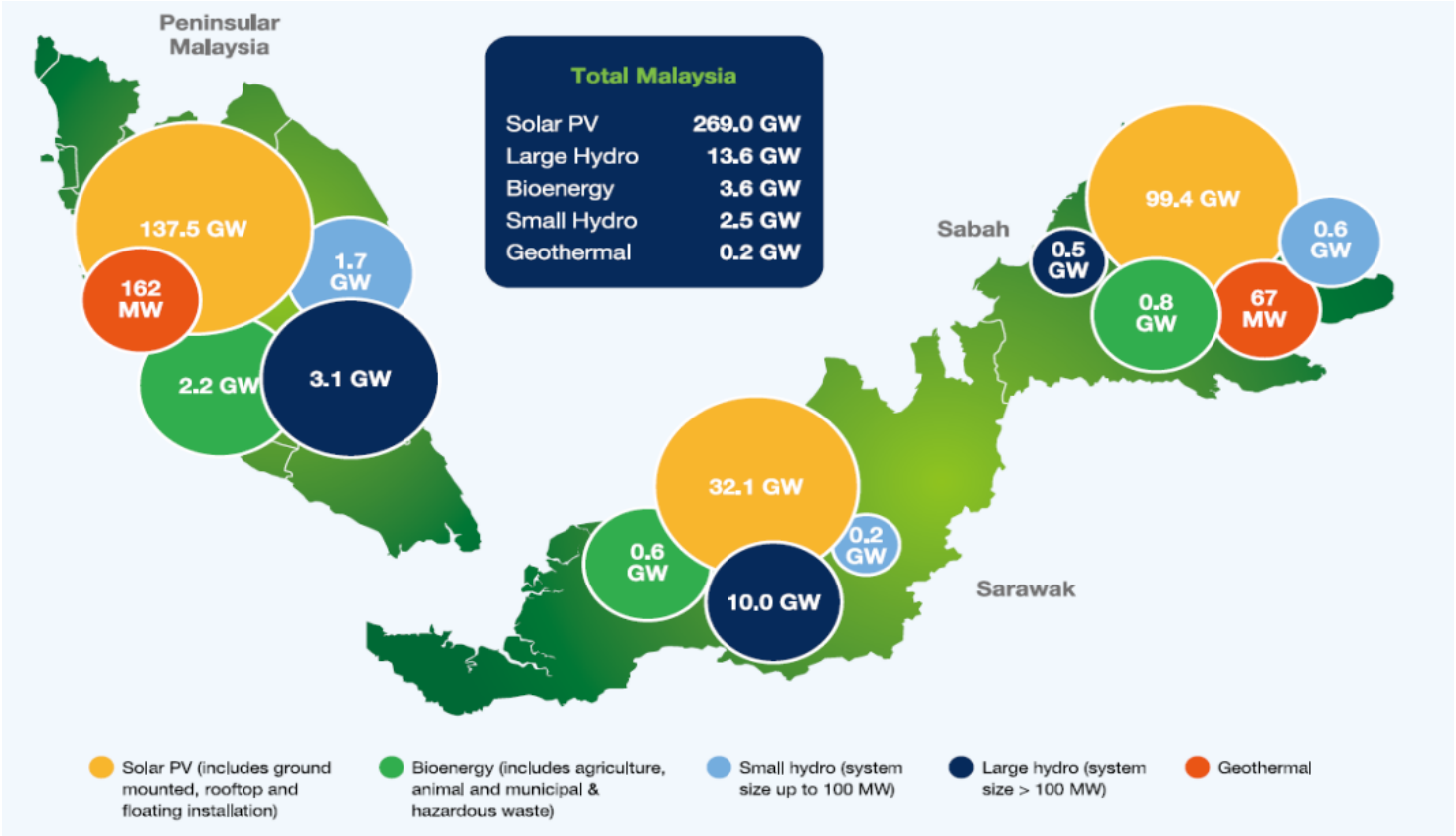


Figure 3: An overview of Malaysia's potential for RE resources [1]

identify areas for improvement. Regular performance monitoring and reporting ensure accountability and enable evidence-based decision-making to optimize the implementation of MyRER initiatives.

By addressing these strategic elements, MyRER aims to create an enabling environment for scaling up renewable energy deployment, driving sustainable economic growth, and mitigating climate change impacts in Malaysia. The strategic framework provides a roadmap for coordinated action by government agencies, private sector stakeholders, civil society organizations, and international partners to achieve shared renewable energy goals and objectives.

6. Growth Of Primary Energy Demand Is Expected To Decline

A decline in the growth of primary energy demand suggests a significant shift in energy consumption trends, typically influenced by various factors:

A) Energy Efficiency Measures: Increased adoption of energy-efficient technologies and practices across sectors like industry, transportation, and buildings can lead to reduced energy consumption per unit of economic output. Investments in energy-efficient appliances, vehicles, and infrastructure contribute to this decline in demand growth.

B) Renewable Energy Integration: The expanding deployment of renewable energy sources such as solar, wind, hydroelectric, and biomass can displace traditional fossil fuels for electricity generation and other energy needs. As renewable energy becomes more competitive and accessible, it can curb the growth of primary energy demand.

C) Economic Changes: Structural shifts in economies, such as a transition towards service-based industries or improvements in industrial processes, can affect energy demand. Economic diversification, coupled with technological advancements, can lead to more energy-efficient production methods and reduced energy intensity.

D) Policy and Regulatory Measures: Government policies and regulations aimed at promoting energy conservation, reducing greenhouse gas emissions, and increasing renewable energy deployment play a crucial role. Measures such as energy efficiency standards, carbon pricing mechanisms, renewable energy targets, and subsidies for clean energy technologies incentivize energy savings and renewable energy adoption.

E) Technological Innovations: Ongoing advancements in energy technologies, including energy storage, smart grids, and electrification of transportation, contribute to the decline in energy demand growth. Innovations in energy management systems, automation, and digitization optimize energy use and improve overall efficiency.

F) Changing Consumer Behavior: Increasing awareness of environmental issues and sustainability, coupled with changing consumer preferences, can drive demand for energy-efficient products and services. Lifestyle changes, such as the adoption of public transportation, cycling, and telecommuting, also contribute to reducing energy demand growth.

G) Global Trends: Global initiatives to address climate change, such as the Paris Agreement, and the growing recognition of the importance of sustainable development influence energy policies and investment decisions. International cooperation and commitments towards decarbonization further accelerate the transition towards cleaner and more efficient energy systems.

H) A decline in the growth of primary energy demand reflects efforts to decouple economic growth from energy consumption, promote sustainable development, and mitigate climate change. However, achieving sustained reductions in energy demand growth requires continued investment, innovation, and policy support across multiple sectors and stakeholders.

7. ASEAN: Increasingly Aggressive Renewable Energy Targets

The Association of Southeast Asian Nations (ASEAN) has indeed been increasingly aggressive in setting renewable energy targets in recent years. This trend reflects the region's commitment to promoting sustainable development, enhancing energy security, and mitigating climate change. Several factors contribute to ASEAN's growing focus on renewable energy [1,5-6]:

A) Energy Security Concerns: ASEAN member states recognize the importance of diversifying their energy sources to reduce dependence on fossil fuels and enhance energy security. Investing in renewable energy helps mitigate risks associated with energy supply disruptions and price volatility [7-8].

B) Climate Change Mitigation: ASEAN countries are vulnerable to the impacts of climate change, including extreme weather events, sea-level rise, and disruption of agricultural systems. By transitioning to renewable energy sources, ASEAN aims to reduce greenhouse gas emissions and contribute to global efforts to limit temperature rise and mitigate climate change impacts.

C) Economic Opportunities: The rapid growth of renewable energy markets presents significant economic opportunities for ASEAN countries. By investing in renewable energy infrastructure and technologies, ASEAN can stimulate job creation, attract investment, and promote industrial development, particularly in the manufacturing and services sectors.

D) Technological Advancements: Advances in renewable energy technologies, coupled with declining costs, have made renewable energy increasingly competitive with conventional fossil fuels. ASEAN member states can leverage these technological advancements to accelerate the deployment of renewable energy projects and achieve their targets cost-effectively[9-10].

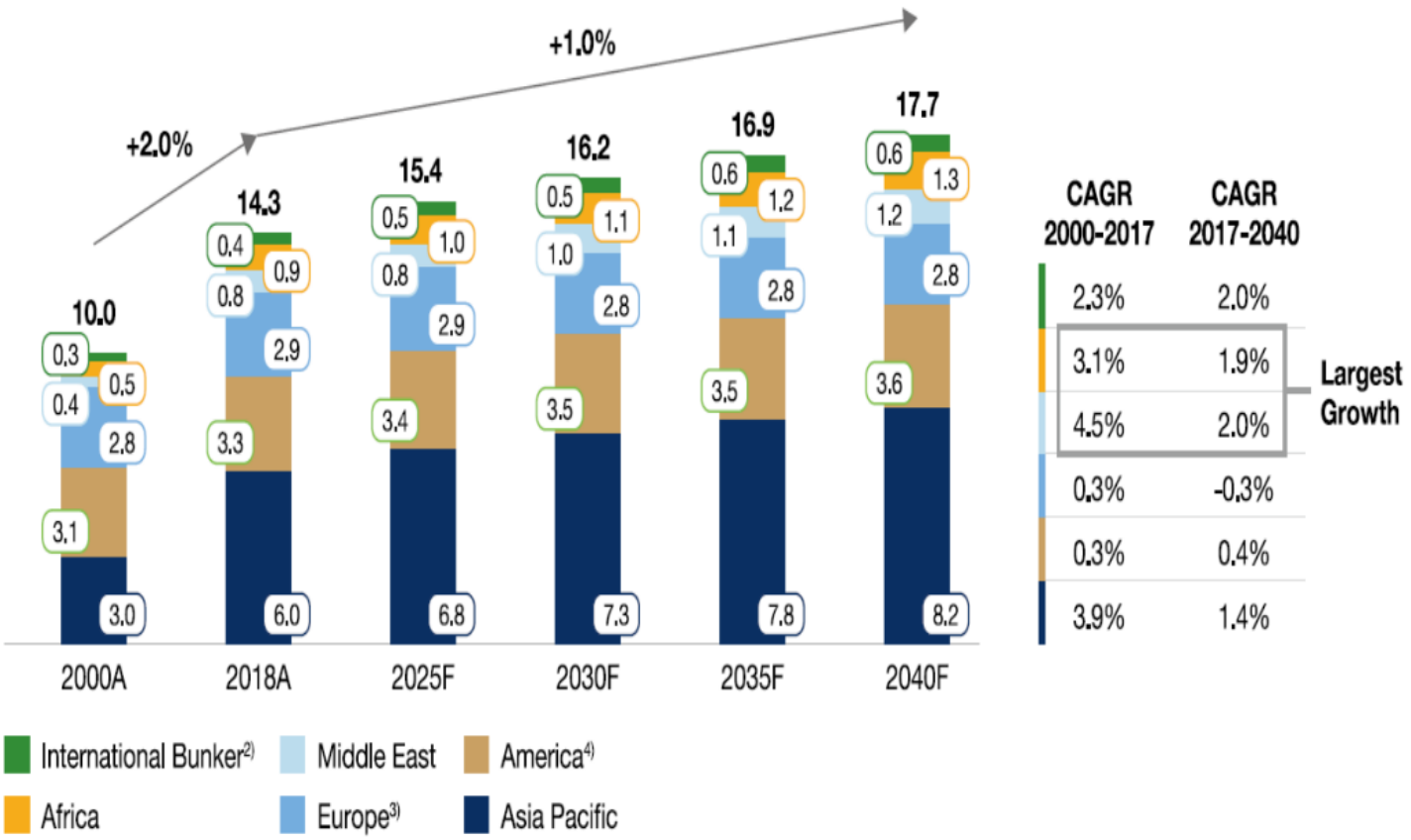


Figure 4: Regional variations in global primary energy demand (IEA New Policies Scenario), 2000–2040F [1]

E) Policy Support and Regional Cooperation: ASEAN has established frameworks and mechanisms to support renewable energy development, including the ASEAN Plan of Action for Energy Cooperation (APAEC) and the ASEAN Centre for Energy (ACE). These initiatives facilitate policy dialogue, capacity building, and knowledge sharing among member states to promote renewable energy deployment.

F) International Commitments: ASEAN member states are signatories to international agreements and initiatives aimed at promoting renewable energy and sustainable development, such as the Paris Agreement and the United Nations Sustainable Development Goals (SDGs). Meeting renewable energy targets aligns with ASEAN's broader commitments to these global initiatives [11-12].

	Malaysia	Indonesia	Philippines	Thailand	Vietnam	Singapore
Latest RE policy	MyRER 2035	National Energy Roadmap	Sectoral Energy Plan & Roadmap	Power Development Plan	Power Development Plan	Singapore's Energy Story
Year of latest RE policy	2020	2017	2017	2019	2019	2019
Overall RE targets	31% RE installed capacity by 2025, 40% by 2035	RE installed capacity by 45 GW by 2025, 168 GW by 2050, 31% of national primary energy supply in 2050	RE installed capacity of 20 GW by 2040	33% RE installed capacity by 2037 with RE mix as following • Solar 6 GW • Biomass 5.57 GW • Wind 3 GW • Hydropower 3.3 GW • Biogas 0.6 GW • MSW 0.5 GW	32% RE installed capacity by 2030, 45% by 2050	At least 2 GW of solar by 2030, and energy storage deployment target of 200 MW post 2025

Figure 5: Overview of key ASEAN countries' renewable energy share targets [1]

G) Public Awareness and Support: There is growing public awareness and support for renewable energy across ASEAN countries. Concerns about air pollution, environmental degradation, and the health impacts of fossil fuel combustion have fueled demand for cleaner and more sustainable energy solutions.

ASEAN's increasingly aggressive renewable energy targets underscore the region's commitment to transitioning towards a low-carbon and sustainable energy future. By harnessing the abundant renewable energy resources available in the region and leveraging technological innovations and policy support, ASEAN aims to achieve its renewable energy goals while promoting economic growth, energy security, and environmental sustainability.

8. Conclusions

Malaysia's Renewable Energy Policy has been a significant driver in promoting the adoption of renewable energy sources within the country, with ripple effects impacting neighboring regional countries. The policy framework, encapsulated in initiatives such as the Malaysia Renewable Energy Act and the National Renewable Energy Policy and Action Plan, sets ambitious targets for renewable energy capacity expansion and outlines various mechanisms to incentivize investment and deployment.

One key impact of Malaysia's Renewable Energy Policy on regional countries is the demonstration effect. As Malaysia implements and showcases successful renewable energy projects, it serves as a model for other countries in the region, inspiring them to pursue similar initiatives. This influence is particularly significant for ASEAN member states, given Malaysia's active participation and leadership within the ASEAN community.

Furthermore, Malaysia's efforts in renewable energy development contribute to regional energy security by diversifying the energy mix and reducing reliance on imported fossil fuels. This resilience is especially crucial for countries in Southeast Asia, which face geopolitical and supply chain risks associated with traditional energy sources.

Additionally, Malaysia's growing expertise and capabilities in renewable energy technology and project development offer opportunities for regional collaboration and knowledge exchange. Through partnerships, capacity-building initiatives, and joint ventures, neighboring countries can leverage Malaysia's experience to accelerate their own renewable energy transitions.

Moreover, Malaysia's Renewable Energy Policy contributes to the region's collective efforts in combating climate change. By reducing greenhouse gas emissions and promoting sustainable development practices, Malaysia sets a positive example for regional cooperation on climate action, aligning with international commitments such as the Paris Agreement.

Malaysia's Renewable Energy Policy has a multifaceted impact on regional countries, serving as a catalyst for renewable energy adoption, enhancing energy security, fostering collaboration, and advancing climate mitigation efforts across Southeast Asia. As Malaysia continues to progress in its renewable energy journey, its influence on the region's energy landscape is likely to grow, driving a more sustainable and resilient future for all.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Malaysia Renewable Energy Roadmap: Pathway Towards Low Carbon Energy System , SEDA Malaysia, 2021. https://www.seda.gov.my/reportal/myrer/

[2] R. Kumar, V. Jain, W. Y. Leong, S. Teyarachakul, Convergence of IoT, Blockchain, and Computational Intelligence in Smart Cities. CRC Press, 2023, DOI: 10.1201/9781003353034

[3] W. Y. Leong, Y. Z. Leong, W. S. Leong, "Smart Manufacturing Technology for Environmental, Social, and Governance (ESG) Sustainability". In Proc. 2023 IEEE 5th Eurasia Conference on IOT, Communication and Engineering (ECICE), 1-6, 2023, DOI: 10.1109/ECICE59523.2023.10383150

[4] W. Y. Leong, "Investigating and Enhancing Energy Management Policy and Strategies in ASEAN," 2023 Asia Meeting on Environment and Electrical Engineering (EEE-AM), Hanoi, Vietnam, 1-6, 2023, doi: 10.1109/EEE-AM58328.2023.10395060

[5] W. Y. Leong, L. S. Heng, Y. Z. Leong, "Malaysia renewable energy policy and its impact on regional countries". In Proc. 12th International Conference on Renewable Power Generation (RPG 2023), 7-13, 2023, DOI: 10.1049/icp.2023.2114

[6] W. Y. Leong, Y. Z. Leong, W. S. Leong, "Green Building Initiatives in ASEAN Countries". In Proc. 2023 Asia Meeting on Environment and Electrical Engineering (EEE-AM), 1-6, 2023, DOI: 10.1109/EEE-AM58328.2023.10395261

[7] W. Y. Leong, "Digital Technology for Asean Energy," 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India, 1480-1486, 2023, DOI: 10.1109/ICCPCT58313.2023.10244806

[8] W. Y. Leong, Y. Z. Leong, W. S. Leong, "Accelerating the Energy Transition via 3D Printing," 2023 Asia Meeting on Environment and Electrical Engineering (EEE-AM), Hanoi, Vietnam, 01-05, 2023, doi: 10.1109/EEE-AM58328.2023.10395773

[9] F. M. Han, W. Y. Leong, "Investigating Target Detection and Localization in Wireless Sensor Network", Procedia Engineering, **41**: 75-81, 2012, DOI:10.1016/j.proeng.2012.07.145

[10] W. Y. Leong, R. Kumar, "5G Intelligent Transportation Systems for Smart Cities", Convergence of IoT, Blockchain, and Computational Intelligence in Smart Cities, CRC Press, 1-25, 2023, DOI: 10.1201/9781003353034-1

[11] W. Y. Leong, L. S. Heng, Y. Z. Leong, "Smart city initiatives in Malaysia and Southeast Asia", IET Conference Proceedings, 2023, 1143-1149, DOI: 10.1049/icp.2023.2440

[12] W. Y. Leong, "ASEAN Federation of Engineering Organisations (AFEO) midterm meeting in Hanoi", Jurutera Bulletin, The Institution of Engineers, Malaysia (IEM), 35-36, 2016.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 3, 72-83 (2024)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Deploying Trusted and Immutable Predictive Models on a Public Blockchain Network

Brandon Wetzel, Haiping Xu

Computer and Information Science Department, University of Massachusetts Dartmouth, Dartmouth, MA 02747, USA

ARTICLE INFO

Article history:
Received: 25 March, 2024
Revised: 12 May, 2024
Accepted: 29 May, 2024
Online: 16 June, 2024

Keywords:
Machine learning
Predictive model
Public blockchain
Smart contract
On-chain storage

ABSTRACT

Machine learning-based predictive models often face challenges, particularly biases and a lack of trust in their predictions when deployed by individual agents. Establishing a robust deployment methodology that supports validating the accuracy and fairness of these models is a critical endeavor. In this paper, we introduce a novel approach to deploying predictive models, such as pre-trained neural network models, in a public blockchain network using smart contracts. Smart contracts are encoded in our approach as self-executing protocols for storing various parameters of the predictive models. We develop efficient algorithms for uploading and retrieving model parameters from smart contracts on a public blockchain, thereby ensuring the trustworthiness and immutability of the stored models, making them available for testing and validation by all peers within the network. In addition, users can rate and comment on the models, which are permanently recorded in the blockchain. To demonstrate the effectiveness of our approach, we present a case study focusing on storing vehicle price prediction models and review comments. Our experimental results show that deploying predictive models on a public blockchain network provides a proficient and reliable way to ensure model security, immutability, and transparency.

1. Introduction

Tasks that once required intensive manual labor can now be automated with the aid of artificial intelligence and machine learning (AI/ML) models, guaranteeing efficient completion in just minutes [1]. However, as AI/ML technologies revolutionize every aspect of our lives, ensuring the validity and fairness of these models becomes crucial. For example, in the field of recruitment, many organizations are incorporating machine learning methods into their hiring processes when dealing with high volumes of job applications. To address bias concerns, New York City implemented a groundbreaking law in July 2023 that requires all automated employment decision tools (AEDTs) to undergo bias audits prior to deployment, and the results of these audits must be made available to the public [2]. While the new regulations can help to ensure the trustworthiness of AI/ML models, it is widely recognized that such an approach shifts the responsibility for trust to a centralized regulatory body, which itself cannot guarantee its trustworthiness. An improved solution should allow users to validate the machine learning-based tools on their own and expose every facet of the tools to the public. This transparency would enable users to validate the functionality and efficiency of AI/ML tools and the absence of bias, thus fostering trust in these tools. This need for trust is not limited to AI/ML tools but pervades various domains where intricate algorithms may pose a challenge to human comprehension. Hence, there is a vital necessity to develop a practical and trustworthy mechanism for deploying immutable and publicly accessible AI/ML-based tools, such as predictive models, that can be easily tested and validated by users.

A blockchain network is a peer-to-peer, decentralized ledger that eliminates the need to trust a centralized audit mechanism. Blockchain was initially recognized for its key role in cryptocurrency systems such as Bitcoin and Ethereum, securing and decentralizing transaction ledgers [3], [4]. However, the use of blockchain technology extends far beyond cryptocurrency applications. Blockchain technology can be employed to ensure the integrity and permanence of data in diverse industries, signifying data that is verifiable and untamperable [5]-[7]. A blockchain securely organizes data into linked blocks using cryptographic techniques, ensuring a tamper-proof and immutable record of transactions and smart contracts. The decentralized design of the blockchain network, with its features of immutability, security, and transparency, naturally addresses the issue of public accountability for data stored in the blockchain network. Therefore, if AI/ML-based predictive tools are deployed on a public blockchain, users of the AI/ML systems no longer need to rely on internal auditing for assurance. This paper uses ML-based neural network predictive models as an example to demonstrate how to deploy such models on a public blockchain

*Corresponding Author: Haiping Xu, University of Massachusetts Dartmouth, Dartmouth, MA 02747, Email: hxu@umassd.edu

for predicting vehicle prices, an area that is susceptible to biases driven by company or dealer interests. Machine learning models might inadvertently set prices too high or too low for specific vehicle types in order to optimize profit. The main goal of this paper is to provide a secure way of deploying machine learning tools that support user validation of predictive models to ensure that the models operate in the best interest of the user, rather than serving corporate goals. To this end, the model should receive vehicle parameters and generate price predictions, allowing users to validate the predicted prices even if they do not have specialized knowledge. In addition, users should be able to obtain validation from other users, which could be supported by the establishment of a comprehensive rating system.

In this paper, we present a novel approach to deploying predictive models using smart contracts by storing the parameters of the predictive models on a public blockchain. This method allows a user to recreate the predictive models on a local computer using the stored model parameters. The user can then validate a version of the model by running certain test cases to ensure the required accuracy and fairness of the model predictions. In addition, we define a mutable meta-block (*MB*) attached to the beginning of the blockchain to record indexing information of all models deployed on the blockchain network. The *MB* can be used to efficiently retrieve the predictive models and their review comments, thereby significantly improving the performance of the blockchain network system.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 presents the public blockchain framework that supports the use of smart contracts to store predictive models. It also describes the structural design of the regular blocks for storing predictive models and review comments, and the meta-block for efficiently retrieving models and the review comments. Section 4 describes in detail the procedures for deploying and retrieving predictive models and reviews. Section 5 presents a case study and the analysis results. Section 6 concludes the paper and mentions future work.

2. Related Work

Blockchain is a secure means of storing immutable and protected transactional data through a peer-to-peer decentralized ledger system. Recent advances have extended the traditional application of blockchain in cryptocurrencies to the storage of medical records, including the use of InterPlanetary File System (IPFS) in blockchain based healthcare secure storage solutions [8]. In [9], the authors proposed a blockchain model for creating a secure system for storing and sharing Electronic Health Records (EHRs). Their strategy is to store large amounts of medical data in the cloud, while the blockchain is dedicated to storing metadata related to EHRs. In [10], the authors introduced a storage framework integrating blockchain and IPFS for efficient transaction storage within the blockchain. In their architecture, the primary patient reports are stored in a decentralized off-chain storage system using IPFS, and the blockchain exclusively stores hash values of these reports. This approach effectively reduces the overall block size within the blockchain. In [11], the authors overcame the security concerns of pure cloud storage and provided a robust and scalable solution that utilizes blockchain technology to handle big data storage of healthcare multimedia files. The approach introduces a hierarchical cloud-based blockchain framework for storing large amounts of healthcare data, resulting in significant efficiencies. Similarly, in [12] and [13], blockchain is again used as a step towards digital healthcare by storing medical images and EHRs directly in the blockchain instead of storing their metadata. Their goal in storing healthcare data via blockchain is to apply the security of blockchain to the healthcare system, mitigating the threat of tampering and increasing security. Blockchain technology has also been used in other areas to store data that may be public but must be trusted to be accurate, which is more similar to the intended application of this paper. For example, in [14], the authors proposed a public auditing mechanism for cloud storage systems using blockchain. They demonstrated that the proposed scheme is secure against type I/II/III/IV adversaries. In [15], the authors described a system that utilizes blockchain technology for video surveillance storage and sharing of videos that are encrypted but publicly accessible with a decryption key. Their approach consists of encrypting and storing camera-acquired video off-chain using distributed IPFS and storing the associated metadata in the blockchain. In addition, some other researchers proposed a secure method for sharing sensitive financial data using blockchain [16]. Their methodology involves recording access control rules, hash values, and storage addresses of financial data in the blockchain, while storing the actual financial data in a distributed database external to the blockchain. While blockchain technology has been effective in ensuring the immutability of stored data across various domains, our approach differs from existing methods in that, in addition to storing regular transactional data, we deploy ML-based predictive models in the blockchain network that are publicly available for validation and adoption by all users.

The challenge of bias and fairness in ML models has been a prevalent issue that currently lacks a universally accepted solution. A comprehensive survey on bias and fairness across various ML domains identified many areas of possible bias that can exist in AI/ML models [17]. ML models investigated in a number of real-world commercial use cases have shown that data integrity, learning parameters, and the lack of safeguards against bias can have a significant impact on the fairness of a model. The survey provides a taxonomy for determining the fairness of models and suggests that this is a pervasive problem for which researchers must find practical solutions. Furthermore, as noted in [18], mitigating bias in datasets for supervised ML is a pressing need for the emerging field of ML. Expanding on this idea, some researchers used statistical methods to assess the significant amount of bias in ML algorithms. For example, in [19], gender bias in facial recognition algorithms was measured and found to be prevalent, whereas in [20], the bias in the training data and its effect on the predictive results were detected and analyzed. The above approaches introduce methods for analyzing and assessing the amount of bias in ML models, and generally describe an approximation of the amount of bias in certain areas of ML. However, each of these proposed solutions for fairness and bias assessment requires trust in a centralized organization. In the absence of trust, e.g., when the model organizer has a strong incentive to manipulate the price of an object, a satisfactory solution remains elusive. In contrast, our approach utilizes the inherent security offered by blockchain technology for the storage of ML models. By employing decentralized hosting of ML models

within a blockchain network, each user can independently verify the accuracy of the models deployed by different organizations. This approach eliminates the need for users to worry about the fairness and potential bias of the models, as they can perform validation checks on each model individually.

The prevalence of substantial bias in ML models necessitates the need for dedicated research to mitigate bias. While bias mitigation relies on suitable bias assessment methods tailored to each unique problem, distinct solutions are needed. In [21], the authors introduced a novel approach to mitigating bias by visually displaying the attributes of an ML model to indicate bias in an intuitive manner. Subsequently, manual adjustments are required to rectify these attributes by modifying the dataset. In [22], the authors proposed a methodology for assessing fairness specific to dynamic pricing methods, where the fairness constraints on price prediction are assessed by a centralized entity. This provides a valuable mechanism for measuring and enforcing the fairness of model-based price prediction, even when the demand for the item is unpredictable. In [23], the authors described how to mitigate bias in neural network-based ML models trained on image data. The authors showed three main categories of bias mitigation methods, namely data-level techniques to balance the training data, model-level approaches to modify the learning algorithm, and adversarial approaches to identify biases in the data. Unlike the above methods, our approach reduces bias by employing a review system to ensure that users can easily find better-reviewed ML models, which tend to be less biased. Our bias mitigation strategy is centered on minimizing the bias directly encountered by users. Furthermore, our approach is versatile and applicable to a wide range of ML models, bypassing the reliance on statistical rigor and instead emphasizing user satisfaction.

The fusion of blockchain and machine learning is an emerging frontier, with ongoing research underscoring its promising potential. In [24], the authors presented various techniques concerning the preservation and processing of big data necessary for ML models. While this decentralized machine learning approach has advantages in terms of mitigating bias and improving fairness, the proposed solution falls short in describing the potentials for evaluating and validating the models, focusing mainly on the data stored in the blockchain. In [25], the authors described an ML-based approach where training gradients are validated by a decentralized blockchain network. They proposed a decentralized learning framework called LearningChain, where organizations can collaborate on training models by computing gradients together to update specific models. In [26], a similar approach to [25] was used to compute the parameters of an on-chain ML in a decentralized multi-threaded environment. Using association rule mining as an example, the authors showed how blockchain can be used to enable trusted machine learning. In [27], the authors further advanced the idea of using blockchain in the learning process by introducing explainable and traceable algorithms that make the learning process meaningful and trackable. Their study showed that smart contract-based training and prediction techniques produced mean square errors similar to scikit-learn-based prediction models. In [28], the authors analyzed the design space regarding the integration of machine learning into blockchain applications, a concept termed ML on chain. They presented a taxonomy for ML on chain, categorizing existing and prospective approaches based on design attributes and their characteristics. While the above methods inherently offer the fundamental security advantages associated with blockchain technology, our approach is different because we only store the model parameters on the blockchain, excluding the data used in training, as all training is done off-chain. Note that storing training data on a large public blockchain would be costly and impractical. Thus, by focusing on storing only the essential components needed to reconstruct ML models, our approach optimizes deployment efficiency on a blockchain platform. Furthermore, we allow users to review a model to verify its fairness. This enables even non-technical users to make informed decisions about the degree of bias present in the model. Table 1 summarizes the main contributions and novelties of our approach by comparing it with existing approaches in terms of five key features.

Table 1: Comparison with Existing Approaches

Approach	Trustless	Fairness	Scalability	Security	Performance
Bias Mitigation Methods [21]-[23]	No	Yes	N/A	No	N/A
Conventional Blockchain [25]-[28]	Yes	No	No	Yes	No
Our Approach	Yes	Yes	Yes	Yes	Yes

As shown in Table 1, a feature can be supported (Yes) or not supported (No) by an approach, or not applicable (N/A) to an approach. The features we consider include whether the method is trustless, whether the fairness of the model can be verified, whether the method is scalable and secure, and whether the method supports improved performance in terms of efficiently searching for stored predictive models. Methods that focus on bias mitigation are typically centralized [21]-[23], so they do not support trustless computing. Since the fairness constraints of stored models are assessed by centralized entities, no security mechanisms are provided to users. On the other hand, while conventional blockchain approaches support trustless and secure computing [25]-[28], they typically do not support bias mitigation for stored models and are also not scalable due to on-chain training of predictive models. Furthermore, since conventional blockchains do not contain a meta-block, they do not support efficient search for stored predictive models.

3. Deploying Predictive Models on a Public Blockchain

3.1. A Framework for a Public Blockchain with Smart Contracts

Each peer in a blockchain network may have a complete copy of the blockchain with a number of blocks. Figure 1 shows a framework for a public blockchain with smart contracts, as well as the list of transactions and the list of smart contracts stored in block B_2. As shown in the figure, each block in the blockchain consists of three main parts: a block header, a list of transactions, and a list of smart contracts. The block header is defined as a 5-tuple (*BID, HPB, BTS, NTR, NSC*), where *BID* is the block ID, *HPB* is the hash value of the previous block, *BTS* is the timestamp when the block was created, and *NTR* and *NSC* are the number of transactions and the number of smart contracts contained within the block, respectively. Transactions are used to record new transactional information added to the blockchain, such as associated data related to a predictive model and review data of the model provided by users. Each transaction is defined as a 5-tuple (*TID, UID, TTS, DES, DAT*), where *TID* is the ID of the transaction, *UID* is the ID of the user who initiated the transaction,

TTS is the timestamp when the transaction was created, *DES* is the detailed description of the transaction, and *DAT* is any associated data stored with the transaction.

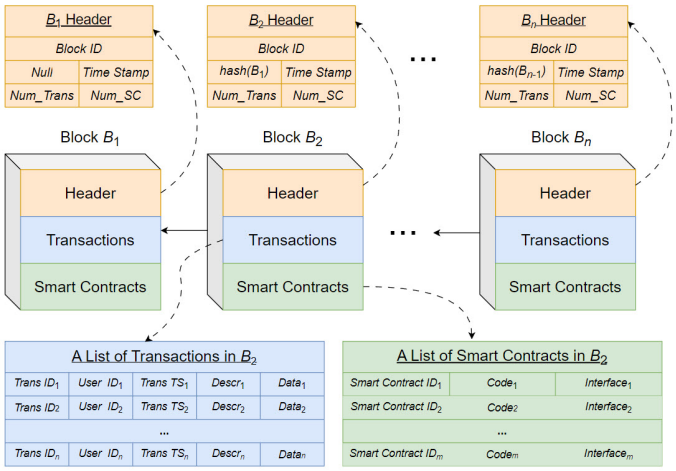


Figure 1: A Framework for a Public Blockchain with Smart Contracts

Smart contracts are initially defined as self-executing code stored in the blockchain and are often seen as the terms of an agreement between two parties. These terms are encoded directly into the smart contract, which is publicly accessible, allowing both parties to fully agree to these terms. In this paper, we broaden the concept of smart contracts to include methods for storing and accessing immutable data on the blockchain. We define a smart contract as a 3-tuple (*SID*, *COD*, *INF*), where *SID* is the identifier of the smart contract, *COD* is the code of the smart contract, which is usually stored as binary code, and *INF* is the interface of the smart contract, which defines the methods that can be invoked. The process of deploying a smart contract involves creating a transaction *T*, where *T.DES* describes the smart contract to be deployed, including the smart contract's location in the block, and *T.DAT* may include any associated data, e.g., a test dataset of a predictive model that is encoded in the smart contract.

Running unverified code, such as smart contracts, on a user's local machine can pose potential security risks. To address this issue, we can assign the responsibility of verifying the trustworthiness of smart contracts on a blockchain to full-node peers. Additionally, executing smart contracts on a virtual machine, such as the Ethereum Virtual Machine (EVM), can help minimize security vulnerabilities by restricting their interactions within the virtual environment. This approach ensures a deterministic execution of smart contracts, as the virtual machine maintains full control over the execution environment.

3.2. Storing Predictive Models Using Smart Contracts

Smart contracts being used as storage for predictive models provide a layer of abstraction that increases the efficiency for users to access and verify predictive models stored on a blockchain network. For example, a predictive model stored with a smart contract can be efficiently extracted by invoking a corresponding method defined in the smart contract. We call a smart contract that stores a predictive model a Model Smart Contract (*MSC*), which contains the serialized parameters of the model and a set of methods that can be used to access and verify the model. Every version of a model must have both a model ID (*MID*) and a version ID (*VID*), where *MID* and *VID* refer to the type of model and an instance of that model, respectively. For example, when *MID* refers to a set of predictive models used to forecast a company's stock price, each model version refers to a specific prediction model. Models sharing the same *MID* can be published by the same publisher at various times or by different publishers, but it is essential for each model to have a unique *VID*.

In our approach, predictive models are trained off-chain and only trained models are stored on-chain. Each peer can extract and recreate an exact copy of the trained model on a local machine by invoking a smart contract method defined in the corresponding *MSC*, which then allows the user to validate and execute the model without having to trust a centralized party. A smart contract containing a predictive model requires an associated transaction to be stored in the same block, documenting its deployment, location, and related data. The review comments of a model are also recorded in transactions but do not require an associated smart contract to access them. Instead, they are stored as plain text in a review transaction, directly accessible to peers. If transaction *T* tracks the deployment of an *MSC*, *T.DES* and *T.DAT* describe the *MSC* and its associated data, respectively. In contrast, if transaction *T* records the review data for a model version, *T.DES* and *T.DAT* contain the *VID* of the model and the corresponding review of the model, respectively. Figure 2 shows *m* versions of a predictive model that are stored on a blockchain network.

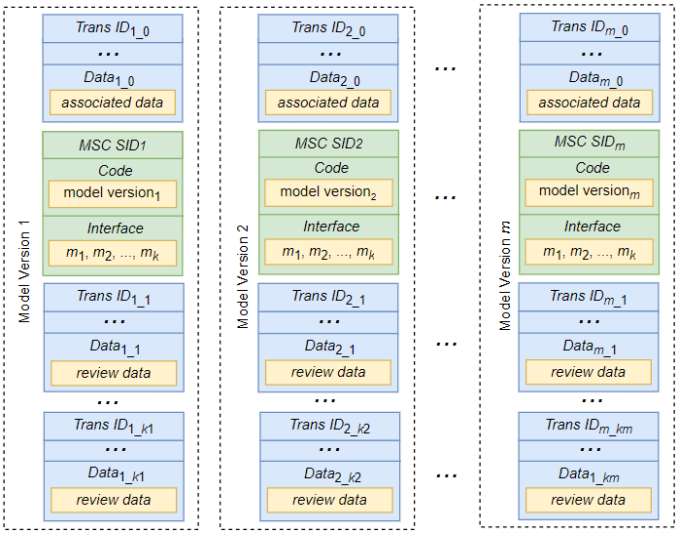


Figure 2: A Predictive Model with Multiple Versions and Review Data

As shown in Figure 2, each model version *VID* is stored in an *MSC* that defines its smart contract ID, the code of the model version, and an interface listing the methods that can be invoked by users. Each *MSC* model version *i*, where $1 \le i \le m$, stored in block *B* is accompanied by a transaction with ID *i*_0 that describes the model version and is stored in the same block *B*. Suppose model version *i* has k_i reviews. These reviews are stored in transactions with IDs from *i*_1 to *i*_k_i. Note that since the reviews for a model version are uploaded at different times, the review transactions can be stored in different blocks other than block *B*.

3.3. The Structure of a Meta-Block

In order to store indexing information for efficiently finding specific versions of a predictive model and all reviews about the

model versions, we introduce the *MB*, which is a mutable meta-block attached to the blockchain. When a new block is approved and added to the blockchain, each peer needs to update its *MB* so that it contains the latest indexing information about the models, versions, and review comments contained in the new block. Figure 3 shows the structure of the *MB* that consists of three parts, namely header, model index, and review index.

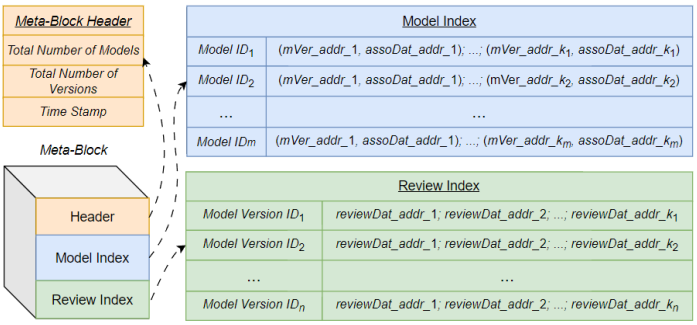


Figure 3: The Structure of a Meta-Block

The *MB* header contains the general information of the *MB*, which is defined as a 3-tuple (*NM*, *NV*, *TS*), where *NM* and *NV* are the total number of models and the total number of versions, respectively, and *TS* is the timestamp when *MB* was last updated. The model index and the review index modules contain indexing information about model versions, associated data and review data in terms of their addresses in the blockchain. The address of a model version is defined as a pair (*BID*, *ISC*), where *BID* is the ID of the block that stores the smart contract containing the model version, and *ISC* is the index of the smart contract within the list of smart contracts in the block. Similarly, the address of the model associated data or review data is defined as a pair (*BID*, *ITR*), where *BID* is the ID of the block that stores the associated data of the model or its review data, respectively, and *ITR* is the index of the transaction that records the information about the model version and review data within the list of transactions in the block. As shown in the figure, the model index module can be implemented in the form of a HashMap, where the key is an *MID* and the output value is a list of pairs (*mVer_addr*, *assoDat_addr*), containing the model version address and the model associated data address, i.e., the address of the smart contract with the predictive model and the address of the transaction that records information about the smart contract, respectively. Likewise, the review index module can also be implemented in the form of a HashMap, where the key is a *VID* of a model version, and the output value is a list of addresses that can be used to locate the transactions that record the reviews about the model version.

4. Deployment and Retrieval of Predictive Models

4.1. Process Overview

To support efficient peer usage of a blockchain network, we require a distinction between full-node peers and regular peers, where a full-node peer contains a full copy of the entire blockchain, while a regular peer is not required to maintain a full copy but can make a request to a full-node peer for information retrieval and uploading. In this paper, we allow regular peers to retrieve a predictive model from a full-node peer and also to send reviews to a full-node peer for publication. Full-node peers have the privilege of deploying models by adding an *MSC* and its accompanying transaction to a new block. When a new block is added, it is broadcast to all full-node peers for updating. Figure 4 shows the process of model deployment and extraction on a blockchain with the *MB* and *m* blocks. As shown in the figure, the blockchain contains models deployed by *n* full-node peers. There are also *r* regular peers, each of which is connected to a full-node peer and can utilize this connection to retrieve a model or post a review for the model. A full-node peer connected to the regular peer can add these retrieval transactions and reviews to a new block. Once the new block contains enough transactions, the full-node peer can broadcast the new block to all full-node peers on the network for approval and adoption.

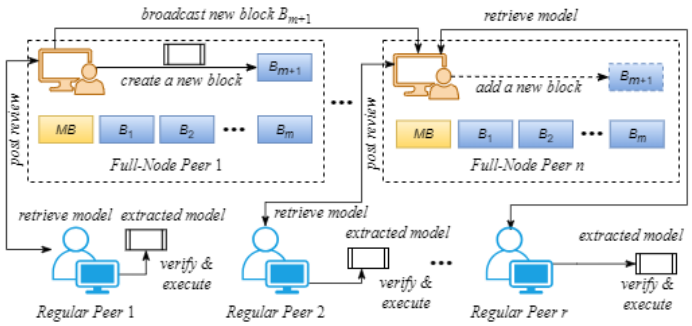


Figure 4: An Overview of the Model Deployment and Extraction Process

4.2. Generating a Meta-Block

The process of generating an *MB* involves searching each block in a blockchain. While searching a block, the addresses of models and reviews are found and added to the *MB*. Each model transaction *tr*, where *tr.DES* describes an *MSC* with an *MID*, must have the accompanying smart contract found in the same block. The address of *tr* and the address of the smart contract described in *tr.DES* are paired and added to the *MB*'s *Model Index* HashMap with *MID* as a key. If the pair represents the first instance of a model, an empty list of addresses is created before the pair of addresses can be added. On the other hand, if *tr* is a review transaction that contains a review on a particular model *VID*, its address is added to the *Review Index* HashMap with *VID* as a key. The complete process to generate an *MB* for a public blockchain is described in Algorithm 1. According to the algorithm, we search for transactions in each block of the blockchain. For each transaction, if it describes the deployment of an *MSC* with *MID* and the model has not been recorded in *MB*, the pair <*MID*, [(*mVer_addr, assoDat_addr*)]> is added to the *Model Index*, where *mVer_addr, assoDat_addr* are the address of the model version and the address of the associated data transaction, respectively. On the other hand, if the transaction is a review transaction for a model version *VID*, the pair <*VID*, [*reviewDat_addr*]> is added to the *Review Index*, where *reviewDat_addr* is the address of review transaction for the model version *VID*. This process is repeated for all transactions in each block of the blockchain until all blocks have been processed. While generating an *MB* on an existing blockchain can be a time-consuming process, this process only needs to be performed once to be usable, thereby significantly reducing the time it takes to search and extract data from the blockchain. Additionally, a peer can download an already existing *MB* from another peer that has the latest *MB*, thus eliminating the need to search through the blockchain to create a new *MB*. Note that when a new block is

added to the blockchain, the *MB* can be updated in a similar manner as defined in Algorithm 1.

Algorithm 1: Generating a Meta-Block for a Public Blockchain

Input: A public blockchain *PB* with *n* blocks
Output: A generated *MB*

1. Create an empty *MB* with *Model Index* and *Review Index* HashMap Π_M *and* Π_R, respectively
2. Initialize the total number of models *NM* to 0
3. Initialize the total number of model versions *NV* to 0
4. Set the time stamp *TS* to the current time
5. **for** each block *B* in *PB*
6. **for** each transaction *tr* in block *B*
7. **if** *tr.DES* describes a model with *MID* stored in *MSC*
8. **if** Π_M does not contain any versions with key *MID*
9. Add <*MID*, [(*mVer_addr, assoDat_addr*)]> to Π_M
10. Increase *NM* by 1
11. **else** Update the list of model versions in Π_M with key *MID*
12. **else if** *tr.DES* describes a review of a model version *VID*
13. **if** Π_R does not contain any reviews of *VID*
14. Add <*VID*, [*reviewDat_addr*]> to Π_R
15. Increase *NV* by 1
16. **else** Update the list of reviews in Π_R with key *VID*
17. **return** *MB*

4.3. *Deploying Models and Posting Reviews*

The process of deploying a classifier can only be done by a full-node peer. This process must first check whether the corresponding model version has been deployed. The *MB* is used to ensure that a particular version of a model has not yet been deployed to the blockchain by checking whether the version being deployed has been indexed by the *MB*. Formally, we define a model as a 4-tuple (*MDA, MID, VID, ASD*), where *MDA* is the data required to recreate the model, stored in a standard way of storing model data such as HDF5 [29]; *MID* and *VID* are the model ID and version ID, respectively; and *ASD* is the associated data of model version *VID*. Deploying a model α requires the creation of a transaction T_α and a model smart contract MSC_α, where MSC_α contains a method for extracting model data for the model version. The transaction T_α contains a description of MSC_α including the use cases of the model and the address of MSC_α. T_α must also contain any associated data, ASD_α. T_α and MSC_α are then added to a new block *B* that is being deployed. Deploying review β requires the creation of a review transaction, T_β, where T_β contains the review being posted. A new block *B* may contain multiple models or reviews, as defined by the full-node peer running the algorithm. Once block *B* is ready, it is broadcast to all full-node peers for approval before it can be published and added to the blockchain. If block *B* is successfully published, the *MB* must be updated to index the new data in block *B*. Algorithm 2 describes the process of generating and deploying a new block with multiple new models on a public blockchain. The efficiency of this process is predominantly influenced by the duration needed to broadcast block *B* and the quantity of transactions and smart contracts contained within the new block. Note that the process of publishing a review is similar to the process of deploying a model, but without the need to create an *MSC*. Therefore, in Algorithm 2, we do not show the process of adding a review transaction to the new block *B*. Once the new block *B* is approved and deployed, the *MB* needs to be updated to add the model information and the review data addresses to the HashMaps *Model Index* and *Review Index*, respectively.

Algorithm 2: Generate and Deploy a New Block

Input: A public blockchain *PB* with an *MB;* a list of new models L_{NM}
Output: Boolean value indicating successful or failed deployment

1. Let *B* be a new block to be generated and deployed
2. **for** each predictive model α in L_{NM}
3. **if** *MB* contains key α*.MID* in the *Model Index* HashMap
4. Let Γ_M be a list of deployed models with key α*.MID*
5. **if** Γ_M contains a model with version ID α*.VID*
6. **continue** // the model version has already been deployed
7. Create a new model smart contract MSC_α for model α
8. Add MSC_α to the list of smart contracts in *B*
9. Create a new transaction T_α with model data α*.ASD* and a description of MSC_α
10. Add T_α to the list of transactions in *B*
11. Broadcast *B* and await consensus approval determination
12. **if** new block *B* is not approved
13. **return** *false* // failed deployment
14. **else** Update *MB* to include new information from *B*
15. **return** *true* // successful deployment

4.4. *Extracting Models and Reviews*

For a regular peer, the process of extracting a model of *MID* from the blockchain requires a connection to a full-node peer in order to make the request. When a full-node peer receives such a request, it retrieves all its model versions of *MID* and their reviews from the blockchain, and then returns the result as a list of model instances to the regular peer. Algorithm 3 describes the process of extracting all predictive models of *MID* and their reviews from the public blockchain *PB* with an *MB*.

Algorithm 3: Extracting Predictive Models and their Reviews

Input: A public blockchain *PB* with an *MB*; a model *MID* to retrieve all its model versions and their reviews
Output: A list of model instances of *MID* along with their reviews

1. Initialize L_M to an empty list of model instances of *MID*
2. **if** *MB* contains the key *MID* in the *Model Index* HashMap
3. Let Γ_M be a list of (*mVer_addr, assoDat_addr*) with key *MID*
4. **else return** L_M // L_M is an empty list, i.e., no model is found
5. **for** each pair *(mVer_addr, assoDat_addr)* in Γ_M
6. Let the model to which the pair refers be model α with VID_α
7. Create a new empty model instance Λ
8. Extract model α stored at *mVer_addr*
9. Extract associated data *assoDat* stored at *assoDat_addr*
10. Add model α and associated data *assoDat* to Λ
11. **if** *MB* contains the key VID_α in the *Review Index* HashMap
12. Let Γ_V be a list of *reviewDat_addr* with key VID_α
13. **for** each *reviewDat_addr* in Γ_V
14. Extract the review *reviewDat* stored at *reviewDat_addr*
15. Attach *reviewDat* to Λ
16. Add Λ to the list of model instances L_M
17. **return** L_M

As shown in Algorithm 3, the full-node peer first searches for *MID* in its *Model Index* HashMap. If the model with *MID* does not exist, an empty list of model instances is returned. Otherwise, each model version must be extracted from the blockchain and added to a list of model instances L_M. For each pair (*mVer_addr*, *assoDat_addr*), both the model α stored at *MSC_addr* and the

model data *assoDat* stored at *assoDat_addr* are extracted from the corresponding blocks in the blockchain and added to a model instance Λ. Then, the full-node peer searches for *VID_α* in its *Review Index* HashMap. For each review stored at address *reviewDat_addr*, it is extracted from the corresponding block in the blockchain and attached to the model instance Λ. Once all review comments have been retrieved, Λ is added to the list of model instances L_M. Finally, when all model instances have been retrieved, L_M is returned to the regular peer with the retrieval request. Note that without the *MB*, the extraction process requires traversing the entire blockchain. With the *MB*, it is possible to quickly find the addresses of predictive models and their review comments, thus making the extraction process very efficient.

5. Case Study and Simulation Results

The case study presented in this section demonstrates the efficacy of storing predictive models on a public blockchain. The stored models are multilayer perceptron (MLP) neural network classifiers that predict vehicle prices based on given parameters such as age and mileage. MLP classifiers were chosen due to their relatively large storage requirements compared to other ML methods such as linear regression. By showcasing the successful storage of MLP classifiers on a public blockchain network, we enhance our confidence in the adaptability of our approach to other types of predictive models and domains. This provides a springboard for more ambitious future research efforts. The MLP classifiers used in this case study are deployed by vehicle dealerships with an *MID* for each model, and a *VID* for each version of these models. For simplicity, in this case study, the *MID* and *VID* contain the vehicle model and vehicle year as substrings, respectively. For example, the classifier that predicts the price of a 2011 Toyota Camry, deployed by car dealer *D* on March 15, 2024, may have an *MID* of "Toyota Camry" and a *VID* of "2011_*D*_03152024".

5.1. Environment Settings

We have developed two different modules: one using the Ethereum blockchain [4] and the other providing users with a web interface to interact with the blockchain. To experiment on the Ethereum blockchain, we employed Ganache, a tool commonly used for testing and development [30]. As with many smart contracts on Ethereum, the smart contracts developed in this case study were also coded in Solidity. While this approach is experimentally tested on the Ethereum network, it also involves simulating processes such as connecting to a full-node peer to retrieve data from the blockchain. In addition, since the approval of new blocks takes place through a consensus mechanism [31], we simulated the approval times of new blocks, assuming that the times are normally distributed with appropriate means and standard deviations (*STDs*). In this sense, we collect data points partly from measurements of real-world implementations and partly from simulated values that best fit real-world scenarios.

The web interface was created to allay the concerns of regular users about the complexity of blockchain usage and to establish a secure environment for conducting the case study. Regular users can interact with the user-friendly interface to search for classifiers and their review comments, select suitable classifiers, extract classifiers and execute them on their local machines. It also allows regular users to create, publish, and read reviews comments through the interface. The website interacts with Ethereum through the *Web3.js* tool, facilitating user engagement with the Ethereum nodes. A regular peer can use the website by connecting to a full-node peer and assembling a classifier locally with the support of the *TensorFlow.js* tool. Review comments on a classifier are uploaded and viewed through the website by regular users. Various types of users such as companies or dealers can create and train classifiers with TensorFlow before deploying them to the blockchain through the interface.

5.2. Space Efficiency for Storing Classifiers

Space efficiency is crucial in a public blockchain because the blockchain needs to be stored on every full-node peer. This experiment provides a basis for the expected spatial storage of our proposed solution. In an MLP classifier, parameters are numerical values that quantify the complexity of the classifier. TensorFlow contains efficient methods for storing a classifier that allow us to isolate these parameters. The number of parameters *nP* in an MLP classifier can be calcaulted as in (1).

$$nP = \sum_{i=1}^{n-1}\left(\left(Layer_i + 1\right) * Layer_{i+1}\right) \quad (1)$$

where $Layer_i$ refers to the number of neurons in layer i of the classifier. The rationale for this equation is that for each neuron in a layer, the number of connections to that node must be the same as the number of neurons in the previous layer plus a connection from a bias neuron. For example, a model with 5 input neurons, 1 output neuron, and 3 hidden layers, each with 128 neurons, would have a total of 33,536 parameters. This quantifies the complexity of the classifier, which allows for comparisons between the different scenarios in this case study.

In our approach, when determining the storage size of a classifier, we consider the size of the classifier itself as well as the overhead of storage, including the storage size of the methods defined in a smart contract and the size of the transactions that track the deployment of the classifier. The storage overhead is usually stable, while the size of the stored classifier can be efficiently measured by the number of bytes stored for its parameters. In addition, we must also account for the number of reviews for each classifier. We define a random number of reviews for each classifier by skewing the normal distribution [32], [33], and define its *pdf* (probability density function) and *cdf* (cumulative density function) using the *pdf* and *cdf* of the normal distribution as shown in (2-5).

$$pdf_{norm}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (2)$$

$$cdf_{norm}(x) = \int_{-\infty}^{\frac{x-\mu}{\sigma}} pdf_{norm}(t)dt \quad (3)$$

$$pdf_{skew}(x) = 2 * pdf_{norm}\left(\frac{x-\xi}{\omega}\right) * cdf_{norm}\left(\alpha\frac{x-\xi}{\omega}\right) \quad (4)$$

$$cdf_{skew}(x) = \int_{-\infty}^{\frac{x-\xi}{\omega}} pdf_{skew}(t)dt \quad (5)$$

In (2-5), the parameters μ and σ represent the mean and *STD* of a normal distribution and are set to 0 and 1, respectively. Parameters α, ξ, and ω represent the *shape*, *scale* and *location* of a skew normal distribution and their values are chosen to be 10, 0, and 200, respectively, to better match real-world scenarios. The rational for using this distribution and the chosen parameter values is based on the observations on websites such as Amazon

and others with reviews of items, where only a few items are highly popular and reviewed but most items are unpopular and have much fewer reviews. With this understanding, we can adjust the parameter values to create a right-skewed distribution we would find in the real world. The chosen values can produce a highly skewed distribution, similar to the distribution of product reviews on Amazon. This distribution is visualized as in Figure 5, where the distribution has a high variance and a significant skew to the right. The mode can be visually shown to be approximately 70 reviews, although the mean is expected to be larger due to the significant skew.

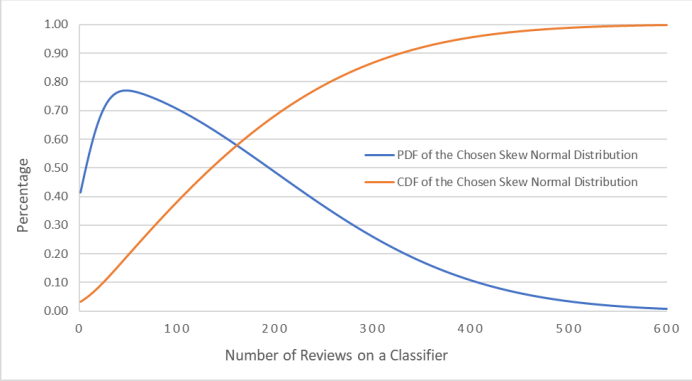


Figure 5: An Example of Skew Normal Distribution of Reviews

The mean, *STD* and skewness of a skew normal distribution can be calculated as in (6-8):

$$mean = \xi + \omega\delta\sqrt{\frac{2}{\pi}}, where\ \delta = \frac{\alpha}{\sqrt{1+\alpha^2}} \quad (6)$$

$$STD = \omega\sqrt{1-\frac{2\delta^2}{\pi}} \quad (7)$$

$$skewness = \frac{4-\pi}{2}\frac{(\delta\sqrt{2/\pi})^3}{\left(1-2\delta^2/\pi\right)^{3/2}} \quad (8)$$

Table 2 shows the resulting mean, *STD*, and skewness of the number of reviews, as well as the *MAX* and *MIN* chosen for the skew normal distribution.

Table 2: Chosen Parameters for the Size of Review Storage

	Number of Reviews[1]	Review Transaction Size (KB)[2]
Mean	159	0.35
STD	121	0.03
Max	600	0.6
Min	0	0.25
Skewness	0.96	0

[1] Number of Reviews follows a skew-normal distribution with skewness of 0.96
[2] Review Transaction Size follows a normal distribution (skewness = 0)

As shown in Table 2, we assume the review transaction size follows a normal distribution with a mean of 0.35KB, a standard deviation of 0.03, and a skewness of 0. Note that review transactions stored on Ethereum require a minimum transaction size, while the maximum size is set to prevent malicious attacks on the blockchain network, such as posting unreasonably large reviews. With a generated random number of reviews for a classifier, we can calculate the total space S_{cla} to store the classifier and its reviews on Ethereum as in (9).

$$S_{cla} = MSCSize(claComp) + RevSizes(numRev) \quad (9)$$

In (9), *claComp* is the classifier complexity (i.e., the number of model parameters), *MSCSize* is a function that computes the amount of storage space required to store the model's *MSC* in a block, *numRev* is a random variable representing the number of reviews, and *RevSizes*(*numRev*) is the size of *numRev* randomized review transactions. The function *MSCSize* was determined from experimental test data on the Ethereum network, where a linear relationship was found between the complexity of a classifier and the size of the associated *MSC* containing the method of extracting the model parameters of a classifier.

Figure 6 shows the 500 data points generated based on the chosen parameters listed in Table 2 and the storage space required on Ethereum. Each data point is generated by randomizing the model complexity (i.e., a random number of parameters in the model) and yields the storage space required to store the model and the associated transaction describing the model. We further generate the random number of reviews and their review transaction sizes based on the distributions provided in Table 2. The storage sizes of the classifier and all its reviews are summed to yield the total size required to store the model on the Ethereum blockchain network.

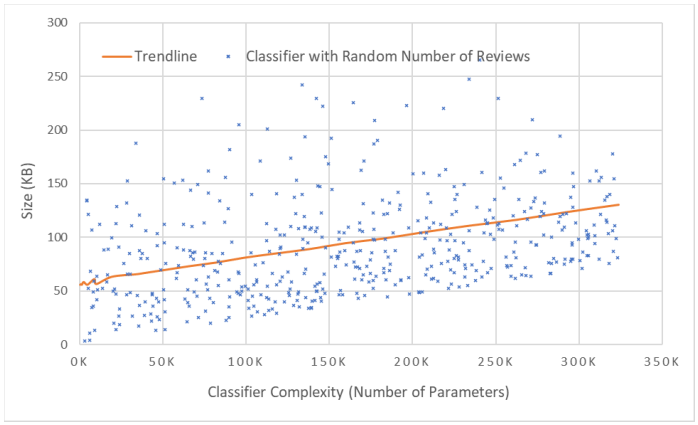


Figure 6: Storage Size of Uploaded Classifier and Their Reviews

As shown in Figure 6, the trendline clearly demonstrates that the average size of storing a model on Ethereum is reasonably small and increases with classifier complexity. While the presence of highly reviewed models puts the trendline slightly above the large clustering of models, the size required to store such classifiers is typically less than 250KB, even at the far end of the range where high-complexity classifiers are accompanied by many reviews. This result is promising and demonstrates both the spatial efficiency and scalability of this approach. For example, if there are 100 car dealerships, each deploying 50 models per year, with approximately 200K parameters in each model, for 10 years, the blockchain storage requirement would be merely 10GB. It is worth noting that while we used Ethereum for our experiments in this paper, our vision for the future is to develop dedicated public blockchain networks for storing specific types of predictive models. Thus, such dedicated public blockchain networks could be scalable and available for a long period of time (e.g., 10+ years). In addition, we could also consider mitigating the scalability issue further by using historical blockchains, such as those discussed in [34].

5.3. Time Efficiency for Deploying a New Block

The amount of time to deploy a new block containing multiple classifiers includes the time to broadcast a new block to other full-node peers and the time to approve the new block using a consensus mechanism. The broadcast time can be estimated based on the broadcast rate and sizes of the classifiers included in the new block. As shown in Table 3, we assume that the size of the classifier is normally distributed with a mean value of 200KB for storing the classifier parameters with different model complexity and the overhead required to store the model on Ethereum. Note that the classifier size distribution requires a maximum value to avoid unrestricted file upload attacks on the blockchain network and a minimum classifier size to avoid deployment of trivial models on the blockchain. Considering the uncertainty of network traffic and the possible bandwidth constraints that may be enforced by the full-node peers, we assume that the broadcast rate and the approval time (i.e., the consensus time) are also normally distributed with the parameters listed in Table 3.

Table 3: Parameters for Deploying a New Block with Multiple Classifiers

	Broadcast Rate (KB/s)	Classifier Size (KB)	Consensus Time (s)
Mean	500	200	12
STD	300	20	2
Max	1,000	1,000	100
Min	200	10	8

Our estimate of the approval time is based on the average time it takes to add a new block to the Ethereum blockchain, which is about 12 seconds [35]. Similarly, the consensus time must be able to time out (e.g., 100 seconds) when there are not enough full-node peers to approve a new block. Since classifiers must be deployed through blocks, the number of transactions or classifiers in each block directly affects the time required for deployment. The time required to broadcast a new block to other full-node peers increases with the number of transactions and classifiers, as well as the sizes of the transactions and classifiers. The time to deploy a new block T_{deploy} can be generated using (10), where *ClaSizes*(*numCla*) returns the size of *numCla* classifiers in the new block, *BCR* is a random broadcast rate, and *ConsTime* is a random consensus time.

$$T_{deploy} = \frac{ClaSizes(numCla)}{BCR} + ConsTime \quad (10)$$

Figure 7 shows the time required to deploy a new block with 1 to 10 classifiers. There are 30 data points for each number of classifiers, each representing a new block, and the variables are chosen randomly according to the distributions defined in Table 3. The trendline in the figure shows how much deployment time is expected to be needed as the number of classifiers increases. With 10 classifiers, the expected deployment time is about 18 seconds, which is a reasonable waiting time for the users. We note that among the above data points, the time used for the approval process accounts for the largest percentage of the total deployment time, while the percentage of the time used for broadcasting new blocks increases slightly with the number of classifiers. Based on the experimental results, it is desirable to include more classifiers in the same block when there is a high demand for uploading classifiers in a short period of time.

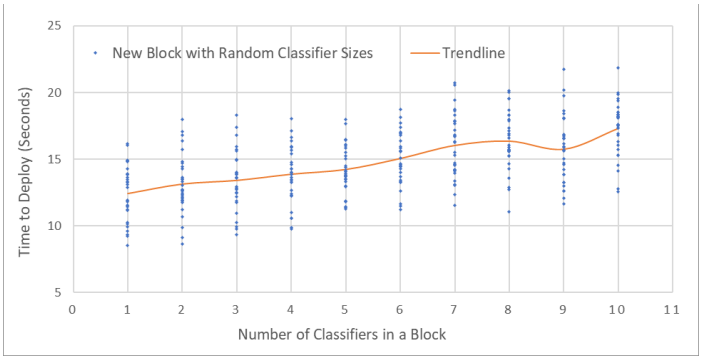


Figure 7: Time to Deploy a New Block

5.4. Generation of a Meta-Block

Another key use case for this approach is the generation of an *MB* by a full-node peer. This process is largely affected by the number of transactions in each block and the total number of blocks in the blockchain. Since there is usually an upper limit to the number of transactions that can be reasonably placed in a block, the time to generate an *MB* would be approximately linear with the number of blocks in the blockchain. In a real-world setting, a full-node peer needs to connect to other full-node peers to verify the consistency of its blockchain before generating an *MB*. The time to connect to other full-node peers for such verification is assumed to be normally distributed with a mean time of 10 seconds and a *STD* of 2 seconds. The time T_{genMB} to generate a meta-block can be calculated using (11), where *GenMB* is the time to generate an *MB* based on experimental tests, and *ConnPeer* is the randomly simulated time to connect to a full-node peer node.

$$T_{genMB} = GenMB + ConnPeer \quad (11)$$

Figure 8 shows the simulation time to generate an *MB* based on the number of blocks in the blockchain.

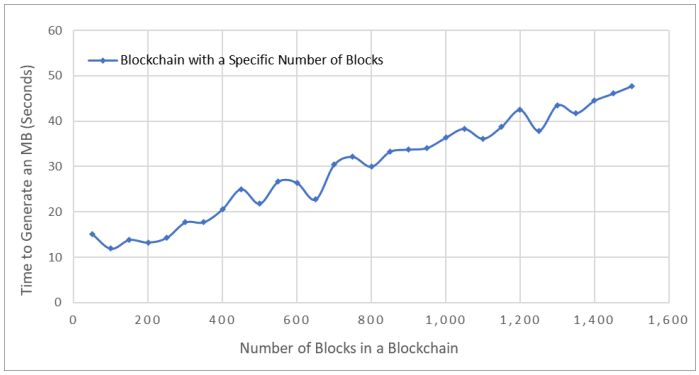


Figure 8: Time to Generate a Meta-Block

Note that Figure 8 only shows up to 1,500 blocks; however, based on the degree of linearity shown in the figure, we can infer how long it takes to generate an *MB* for a blockchain with a larger number of blocks. Additionally, the stochastic impact of the time to connect to a full-node peer is reduced when the number of blocks increases. This is because the connection time is centered on a 10-second average, so as the number of blocks increases, the impact of the connection time becomes smaller compared to the total generation time. For a blockchain with 1,500 blocks, the time required to generate an *MB* in less than a minute is reasonable,

and the fact that it only needs to be executed once demonstrates the effectiveness of this approach.

5.5. Analysis of Search and Retrieval Time

The search and retrieval (S&R) time for classifiers and their reviews is the main waiting time for regular users; therefore, we must minimize this time as much as possible. For each retrieval request made by a regular peer, several versions and their review comments must be retrieved and sent back. While the search time through an *MB* is expected to be efficient, regular peers that do not maintain the blockchain need time to connect to a full-node peer, which retrieves the requested data and returns results to the regular peer. Thus, the S&R time $T_{S\&R}$ consists of three components as shown in (12): the time required to connect to a full-node peer from the regular peer (*ConnPeer*), the time required for the full-node peer to search for each requested classifier and its review comments, and the time required to send this data back to the regular peer (i.e., the download time). Note that the download time is based on the total size of the classifiers *ClaSizes*, the total size of all reviews *RevSizes*, and the download rate (*DLR*) for downloading the classifiers and their reviews.

$$T_{S\&R} = ConnPeer + Search(n) + \frac{ClaSizes+RevSizes}{DLR} \quad (12)$$

The search time is measured on an Ethereum blockchain network. According to Algorithm 3, the expected search time *Search*(n) is approximately linear as the number of blocks n increases. The number of reviews for each model version follows the skew normal distribution defined in Table 2. The size of a classifier, the size of a review and the download rate are randomly generated following the normal distribution defined in Table 4.

Table 4: Parameters for Searching and Retrieving Classifiers

	Review Size (KB)	Classifier Size (KB)	Download Rate (KB/s)	Connection Time (Sec)
Mean	0.1	100	500	10
STD	0.03	50	300	2
Max	0.35	800	1,000	100
Min	0	10	200	1

As shown in Table 4, the size of each review and classifier is significantly different from those listed in Table 2 and Table 3, respectively. This is because in this experiment, we only need to download the reviews and classifiers themselves and not the transactions associated with them. It is important to note that transactions stored with Ethereum usually contain additional information that is not necessary for a regular peer to download. The time to download data is determined by the amount of data downloaded, defined as the last term in (12). In addition, we assume that *ConnPeer* is normally distributed, with a mean time of 10 seconds and a *STD* of 2.

Figure 9 shows the time required to search and retrieve three arbitrarily chosen classifiers and their reviews versus the number of blocks in a blockchain. To demonstrate the effectiveness of our approach, two different search methods are included in the figure: one is our approach using *MB*, and the other is a conventional blockchain method that does not use *MB*. The conventional approach searches for all transactions from the newest block to the oldest block and saves reviews and model data during the search. We conducted 250 experiments with each approach, where a data point represents a request from a regular peer. The S&R time for each data point is calculated using (12), where the search times were measured by simulations on Ethereum.

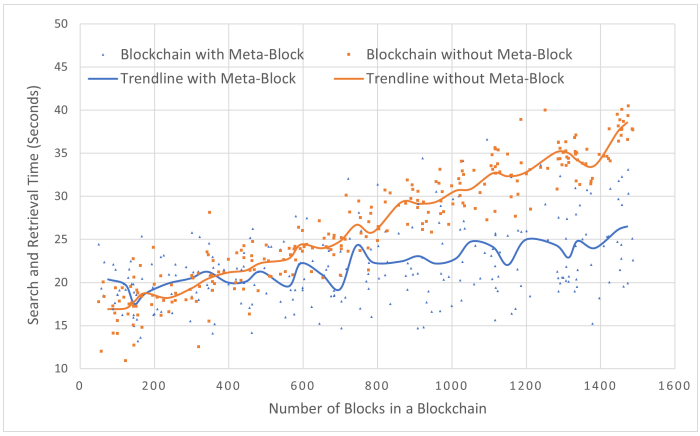


Figure 9: Search and Retrieval Time for Classifiers and Their Reviews

In order to clearly demonstrate the effectiveness of our approach, we show the trendlines for both methods. As can be seen from the trendlines, the average S&R time of the blockchains using *MB* remains relatively stable. The difference between a large number of blocks and a small number of blocks is about 5 seconds, while the conventional method without using *MB* takes significantly longer as the number of blocks increases. Thus, based on the experimental results, our approach greatly improves the ability to search a blockchain compared to a conventional approach without using *MB*.

It is worth noting that the experimental results are based on real-world data collected from the Ethereum blockchain network, as well as simulated data designed to demonstrate the practical implications of the case study. The inclusion of carefully selected distribution values enhances the credibility of the case study and increases confidence in the case study solution.

6. Conclusions and Future Work

As machine learning becomes an increasingly important part of many people's lives, the need to address the issue of trust in these algorithms grows every day. The approach proposed in this paper aims to completely eliminate the need for trust by combining machine learning models with a decentralized public blockchain network. Machine learning models are stored on a blockchain through smart contracts, which contain methods for accessing and validating the stored models. Transactions are used to track the deployment of models and record their reviews. Thus, model users do not need to trust individual organizations, but only peer reviews and their own audits of machine learning algorithms. To further enhance the user experience, we add a meta-block at the beginning of the blockchain to record indexing information for all models and reviews stored on the blockchain. Our experimental results on Ethereum show that our approach is effective and efficient when deploying predictive models on a public blockchain network.

For future work, we plan to develop dedicated public blockchain networks for deploying predictive models. Storage and computation costs can be reduced as the scale of a dedicated public blockchain network is more manageable. The impact of

congestion on the blockchain is also a key issue that needs to be addressed [36]. In our current approach, the approval of new blocks and the retrieval of classifiers and their reviews are handled by full-node peers. In future work, we may consider developing efficient load balancing mechanisms for full-node peers [34] to mitigate the impact of blockchain network congestion on system performance. Furthermore, combined with the trustless nature of blockchain, we can bring more systematic changes to the way the machine learning models are created. This could include publishing metrics for training data or goals for the training process on the blockchain. Additionally, training could take place entirely on the blockchain, rather than off-chain as in our approach. While an on-chain approach may lead to scalability issues, it allows users to trust the process of creating the model rather than the model itself. This can be important in situations where the trustworthiness of a model cannot be easily validated by regular users. To elaborate on this approach, a new smart contract could be designed to handle on-chain training, where full-node peers have the ability to define new methods and execute them to train predictive models. Finally, we can consider implementing automated auditing of new AI/ML models [37]. This automated approach helps eliminate undesirable AI/ML models from being stored on the blockchain, thus saving model storage space. Based on previous work [38], new models can be initially deployed in a temporary block, transitioning to a permanent block upon successful completion of auditing. This auditing procedure can be automated, facilitated by classifiers trained on historical data to discern undesirable models.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We thank the editors and all anonymous referees for spending their valuable time carefully reviewing this paper and making many suggestions for improvement. We also thank the University of Massachusetts Dartmouth for providing financial support to the first author to complete this work.

References

[1] Sarker, "AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems," SN Computer Science, **3**(158), 1-20, 2022, doi: 10.1007/s42979-022-01043-x.

[2] NYC 311, "Automated employment decision tools," The Official Website of the City of New York, July 2023. Retrieved on September 1, 2023 from https://portal.311.nyc.gov/article/?kanumber=KA-03552.

[3] S. Nakamoto, "Bitcoin: a peer-to-peer electronic cash system," White Paper, Bitcoin Project, October 2008. Retrieved on January 15, 2023 from https://bitcoin.org/bitcoin.pdf.

[4] V. Buterin, "Ethereum: a next-generation smart contract and decentralized application platform," Ethereum Whitepaper, 2014. Retrieved on May 15, 2023 from https://ethereum.org/en/whitepaper/.

[5] 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.

[6] H. Guo, X. Yu, "A survey on blockchain technology and its security," Blockchain: Research and Applications, **3**(2), February 2022, doi: 10.1016/j.bcra.2022.100067.

[7] 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.

[8] S. Kumar, A. K. Bharti, R. Amin, "Decentralized secure storage of medical records using blockchain and IPFS: a comparative analysis with future directions," Security and Privacy, **4**(5), 1-16, April 2021. doi: 10.1002/spy2.162.

[9] 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.

[10] S. Liu, H. Tang, "A consortium medical blockchain data storage and sharing model based on IPFS," In Proceedings of the 4th International Conference on Computers in Management and Business (ICCMB 2021), 147-153, Singapore, January 2021, doi: 10.1145/3450588. 3450944.

[11] Thamrin, H. Xu, "Hierarchical cloud-based consortium blockchains for healthcare data storage," 2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C), 644-651, Hainan, China, December 2021, doi: 10.1109/QRS-C55045.2021.00098.

[12] S. D. Ashwini, A. P. Patil, S. K. Shetty, "Moving towards blockchain-based solution for ensuring secure storage of medical images," In Proceedings of the 2021 IEEE 18th India Council International Conference (INDICON), 1-5, Guwahati, India, December 19-21, 2021, doi: 10.1109/INDICON52576.2021.9691516.

[13] S. Ballal, Y. Chandre, R. Pise, B. Sonare, S. Patil, "Blockchain-based decentralized platform for electronic health records management," In Proceedings of the 2023 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS), 1-5, New Raipur, India, October 06-08, 2023, doi: 10.1109/ICBDS58040.2023.10346392.

[14] S. Ajjarapu, S. K. Pasupuleti, "Blockchain based certificateless privacy preserving public auditing for cloud storage systems," In Proceedings of the 2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC), 286-291, Solan, Himachal Pradesh, India, November 25-27, 2022, doi: 10.1109/PDGC56933.2022.10053241.

[15] 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, January 2019, doi: 10.1109/ICOIN.2019.8718126.

[16] 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.

[17] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, "A Survey on bias and fairness in machine learning," ACM Computing Surveys, **54**(6), Article No. 115, 1-35, July 2022, doi:10.1145/3457607.

[18] V. N. Mandhala, D. Bhattacharyya, D. Midhunchakkaravarthy, "Need of mitigating bias in the datasets using machine learning algorithms," In Proceedings of the 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), 1-7, Chennai, India, 2022, doi: 10.1109/ACCAI53970.2022.9752643.

[19] M. Atay, H. Gipson, T. Gwyn, K. Roy, "Evaluation of gender bias in facial recognition with traditional machine learning algorithms," In Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence (SSCI), 1-7, Orlando, FL, USA, December 2021, doi: 10.1109/SSCI50451.2021.9660186.

[20] S. Rohani, R. Baeza-Yates, "Measuring bias," In Proceedings of the 2023 IEEE International Conference on Big Data (BigData), 1289-1298, Sorrento, Italy, 2023, doi: 10.1109/BigData59044.2023.10386679.

[21] H. Wang, S. Mukhopadhyay, Y. Xiao, S. Fang, "An interactive approach to bias mitigation in machine learning," In Proceedings of the 2021 IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), 199-205, Banff, AB, Canada, October 2021, doi: 10.1109/ICCICC53683.2021.9811333.

[22] M. C. Cohen, S. Miao, Y. Wang, "Dynamic pricing with fairness constraints," SSRN, September 2021. Retrieved on October 1, 2023 from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3930622.

[23] Y. Wang, H. Liu, "De-biasing methods in neural networks: a survey," In Proceedings of the 2023 International Conference on Machine Learning and Cybernetics (ICMLC), 458-463, Adelaide, Australia, July 2023, doi: 10.1109/ICMLC58545.2023.10327985.

[24] H. Maheshwari, U. Chandra, D. Yadav, A. Gupta, R. Kaur, "Machine learning and blockchain: a promising future," In Proceedings of the 4th International Conference on Intelligent Engineering and Management (ICIEM), 1-6, London, United Kingdom, May 09-11, 2023, doi: 10.1109/ICIEM59379.2023.10166343.

[25] X. Chen, J. Ji, C. Luo, W. Liao, P. Li, "When machine learning meets blockchain: a decentralized, privacy-preserving and secure design," In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), 1178-1187, Seattle, WA, USA, December 10-13, 2018, doi:10.1109/BigData.2018.8622598.

[26] T. Wang, "A unified analytical framework for trustable machine learning and automation running with blockchain," In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), 4974-4983, Seattle, WA, USA, 2018, doi: 10.1109/BigData.2018.8622262.

[27] S. Badruddoja, R. Dantu, Y. He, A. Salau, K. Upadhyay, "Scalable smart contracts for linear regression algorithm," International Conference on Blockchain Technology and Emerging Applications (BlockTEA 2022), Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, **498**, 19-31, Springer, Cham, April 2023, doi: 10.1007/978-3-031-31420-9_2.

[28] B. Gu, A. Singh, Y. Zhou, J. Fang, F. Nawab, "ML on chain: the case and taxonomy of machine learning on blockchain," In Proceedings of the 2023 IEEE International Conference on Blockchain and Cryptocurrency (ICBC), 1-18, Dubai, United Arab Emirates, May 1-5, 2023, doi: 10.1109/ICBC56567.2023.10174908.

[29] M. Folk, G. Heber, Q. Koziol, E. Pourmal, D. Robinson, "An overview of the HDF5 technology suite and its applications," In Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases, 36-47, Uppsala Sweden, March 25, 2011, doi: 10.1145/1966895.1966900.

[30] ConsenSys, "What is Ganache?" Overview - Truffle Suite, ConsenSys Software Inc., 2022. Retrieved on March 15, 2024 from https://archive.trufflesuite.com/docs/ganache/.

[31] al-Qerem, A. Hammarsheh, A. M. Ali, Y. Alslman, M. Alauthman, "Using consensus algorithm for blockchain application of roaming services for mobile network," International Journal of Advances in Soft Computing & its Applications, **15**(1), 99-112, March 2023, doi: 10.15849/IJASCA.230320.07.

[32] O'Hagan, T. Leonard, "Bayes estimation subject to uncertainty about parameter constraints," Biometrika, **63**(1), 201-203, 1976, doi: 10.1093/biomet/63.1.201.

[33] S. K. Ashour, M. A. Abdel-hameed, "Approximate skew normal distribution," Journal of Advanced Research, **1**(4), 341-350, October 2010, doi: 10.1016/j.jare.2010.06.004

[34] M. Felipe, H. Xu, "A scalable storage scheme for on-chain big data using historical blockchains," In 2022 IEEE 22nd International Conference on Software Quality, Reliability and Security Companion (QRS-C), 54-61, IEEE BSC 2022, Guangzhou, China, December 5-9, 2022, doi: 10.1109/QRS-C57518.2022.00017.

[35] Ethereum, "Blocks," Ethereum Documents, Feburary 27, 2024. Retrieved on March 15, 2024 from https://ethereum.org/developers/docs/blocks.

[36] S. Ahn, T. Kim, Y. Kwon, S. Cho, "Packet aggregation scheme to mitigate the network congestion in blockchain networks," In Proceedings of the 2020 International Conference on Electronics, Information, and Communication (ICEIC), 1-3, Barcelona, Spain, January 19-22, 2020, doi: 10.1109/ ICEIC49074.2020.9051158.

[37] J. R. Beckstrom, "Auditing machine learning algorithms: a white paper for public auditors," International Journal of Government Auditing, **48**(1), 40-41, Winter Edition, 2021.

[38] R. Ming, H. Xu, "Timely publication of transaction records in a private blockchain," In 2020 IEEE 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C), 116-123, IEEE BSC 2020, Macau, China, December 11-14, 2020, doi: 10.1109/QRS-C51114.2020.00030.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 1, 79-90 (2024)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Tracing the Evolution of Machine Translation: A Journey through the Myanmar (Burmese)-Wa (sub-group of the Austro-Asiatic language) Corpus

Florance Yune*, Khin Mar Soe

Natural Language Processing Lab, University of Computer Studies (Lecturer), Yangon, Myanmar

ARTICLE INFO

Article history:
Received: 02 December, 2023
Revised: 06 January, 2024
Accepted: 06 January, 2024
Online: 20 January, 2024

Keywords:
Machine Translation (MT)
Low-resource Languages
Myanmar-Wa Corpus
Language Preservation
Transfer Learning
Long Short-Term Memory
Recurrent Neural Network
Pivot-Based SMT

ABSTRACT

Machine Translation (MT) has come a long way toward reducing linguistic gaps. However, its progress in efficiently handling low-resource languages—such as the Wa language in the Myanmar-Wa corpus—has not received enough attention. This study begins with a thorough investigation of the historical development of MT systems, painstakingly following their development against the complex background of the Myanmar-Wa language region. Using an interdisciplinary methodology that integrates linguistics, technology, and culture, this investigation reveals the transformative journey of Machine Translation (MT) in its pursuit of overcoming linguistic barriers. It offers a thorough study that clarifies the opportunities and limitations present in MT's progress. More broadly, by clarifying the complex relationship between technology and linguistic diversity, our work not only advances our understanding of MT's evolutionary history but also supports the conservation of endangered languages, like Wa language. The research's conclusions have implications that go beyond machine translation to the larger conversation about language preservation and how technology development coexists harmoniously. Notably, this paper is an extension of work originally presented in "2023 IEEE Conference on Computer Applications (ICCA)", acknowledging its foundation and presenting substantial advancements.

1. Introduction

The critical role played by machine translation (MT) in eliminating linguistic barriers within the context of an expanding era of globalization cannot be overstated [1]. Essential instruments for fostering international collaboration, knowledge dissemination, and cross-cultural exchanges, MT systems have evolved to facilitate communication across linguistic and cultural boundaries. This paper aims to investigate the historical development of MT systems, with a specific focus on their adaptation to the intricate linguistic environment of the Myanmar-Wa corpus. The necessity for precise and efficient language translation is underscored by the accelerated pace of globalization, characterized by the seamless flow of ideas, information, and business across international borders [2]. Propelled by advancements in artificial intelligence and computational linguistics, machine translation has emerged as a pivotal element in the global communication paradigm. It has not only surmounted the constraints of traditional translation but has also enabled the bridging of languages previously considered challenging due to limited resources, as exemplified by the Wa language within the Myanmar-Wa corpus.

The ensuing methods are structured to fulfill the objectives of this study. Commencing with a historical overview of machine translation, encompassing its development and initial challenges, the analysis subsequently delves into the distinctive attributes of the Myanmar-Wa corpus, including its cultural significance and linguistic characteristics, along with the complexities associated with data acquisition. This research article proceeds to scrutinize early attempts at applying MT to the Myanmar-Wa corpus, elucidating their deficiencies and limitations. Moreover, it investigates innovative approaches and technological advancements that have enhanced MT for this specific language combination.

The subsequent section of the study involves linguistic analysis, dissecting the unique characteristics of the Wa language and examining the metrics and procedures employed to evaluate the efficacy of the MT system. Emphasizing the pivotal role that improved MT plays in cultural preservation and the broader objective of language revitalization, this analysis focuses on the practical applications and consequences of enhanced MT for the Myanmar-Wa language pair. Finally, the system's investigation concludes with an overview of major findings, a discussion of their

*Corresponding Author: Florance Yune, floranceyune@ucsy.edu.mm

implications, and concluding reflections on the dynamic field of machine translation and its critical role in bridging linguistic gaps.

The study begins with the acknowledgment that this paper is an extension of the work originally presented in the conference paper, "Myanmar-Wa Machine Translation based on LSTM-based Deep Learning Encoder-Decoder Model" [3].

2. Related Works

The evolution of machine translation (MT) stands as an enduring quest to surmount linguistic barriers, a journey that was inaugurated by the visionary insights of Warren Weaver. In [4] the author presented groundbreaking 1949 memo, Weaver introduced the concept of computer-assisted language translation, marking the genesis of MT as a distinct area of study and laying the foundation for subsequent developments. Early MT initiatives, notably the Georgetown-IBM experiment in the 1950s, played a pivotal role in establishing the rudiments of MT technology. Despite their nascent status, these early algorithms encountered formidable challenges, often yielding imprecise and unnatural translations. The intricacies of human language, encompassing idiosyncrasies, context-dependency, and cultural nuances, presented a significant hurdle. As linguistic theories matured and computational power increased, MT systems began to exhibit promise.

A pivotal advancement in machine translation materialized with the advent of rule-based MT in the 1960s and 1970s. These systems employed grammatical analysis and language rules to generate translations, exemplified by the Systran system initially designed for diplomatic and military communication [5].

The landscape of MT underwent transformative shifts with the introduction of statistical machine translation (SMT) systems in the 1980s, as seen in IBM's Candide and the European Union's EUROTRA project, as presented in [6]. Utilizing statistical models and extensive multilingual corpora, SMT systems revolutionized translation paradigms, ushering in a data-driven methodology that significantly improved translation accuracy and fluency. The 21st century heralded a new era in machine translation with the rise of neural machine translation (NMT) models, including recurrent neural networks (RNNs) and transformer models like the Transformer and BERT. Fueled by deep learning, these models exhibited remarkable proficiency in recognizing idiomatic expressions, context, and intricate linguistic patterns, with further enhancements through the integration of attention mechanisms [7].

In tandem with these advancements, MT systems grapple with challenges, particularly in handling low-resource languages like Wa in the Myanmar-Wa corpus. The absence of extensive bilingual corpora for low-resource languages hampers the development of precise and contextually appropriate translations. Additional complexities arise from linguistic nuances, such as tone systems, restricted written forms, and complex morphology. The lack of established grammatical rules and linguistic resources in these languages poses a challenge to the effective generalization of machine translation systems.

In response to these challenges, the field of machine translation (MT) has evolved with innovations such as transfer learning approaches, cross-lingual pre-training, and adaptations of existing models to low-resource languages. In [8] the author introduced a notable contribution to the field in "Cross-lingual Language Model Pretraining" (Advances in Neural Information Processing Systems, 2019). This concept emphasizes the pretraining of language models in a cross-lingual manner, enhancing their ability to transfer knowledge across languages. To address the issues faced by low-resource languages and enhance inclusivity, researchers are exploring novel approaches. This study traces the historical development of machine translation (MT) systems, emphasizing their adaptation to the Myanmar-Wa corpus. Through an examination of technological advancements, the study sheds light on strategies and innovations aimed at achieving efficient Wa language translation. Crucially, this work extends beyond the realm of machine translation, contributing to broader domains such as linguistics and language preservation. By scrutinizing the challenges and solutions in adapting MT to the Myanmar-Wa corpus, the study contributes valuable insights to the growing body of information on underrepresented languages. This effort plays a vital role in ensuring the survival and resurgence of languages in a globalized environment [9].

Furthermore, in [10], the authors introduced a novel approach to NMT by incorporating syntax-directed attention mechanisms. This research, presented at the AAAI conference on artificial intelligence, delves into the utilization of syntactic structures to enhance the attention mechanism in NMT systems. By integrating syntax-directed attention, the researchers aimed to improve the model's comprehension of sentence structures and syntactic relationships, potentially leading to more accurate and contextually relevant translations. In [10] the author found the valuable insights to the ongoing advancements in NMT, particularly in addressing syntactic complexities for improved translation quality.

3. Myanmar-Wa Corpus

The extraordinary linguistic tapestry that is the Myanmar-Wa corpus derives its significance from the merging of two separate languages, Wa and Myanmar. This corpus provides insights into the difficulties and advancements in machine translation in addition to serving as a monument to linguistic diversity [11].

The Wa language, which is spoken by the Wa people, an ethnic group that live in the border regions of China and Myanmar, is stored in the Myanmar-Wa corpus, which is located inside this complicated linguistic terrain. MT systems face significant hurdles due to the unique linguistic complexities of the Myanmar-Wa corpus, which include a limited written form and a rich tone system [12]. These difficulties include maintaining cultural context, grammatical structures, and subtle semantic differences—all essential components of accurate translation.

3.1. Myanmar Language

Spoken largely in Myanmar, the Burmese language, also called the Myanmar language, is linguistically rich and culturally significant. It has distinct linguistic characteristics that present intriguing opportunities and problems for machine translation systems. This section gives a general introduction to Burmese, emphasizing its unique vowels and consonants, and discusses its use in the larger field of machine translation. Three basic character groups—vowels, consonants, and medial—combine to form Myanmar writing, and each adds to the language's complex structure. In addition, there are 33 primary consonants in the Myanmar language, each with a unique phoneme.

The tonal and analytical characteristics of the Myanmar language contribute to its linguistic complexity [13]. Its script,

which is based on the Indian (Brahmi) prototype, has changed over the course of a millennium to reflect Myanmar's rich linguistic past. Significantly, the Myanmar language has functioned as a symbol of cultural identity within the nation as well as a means of communication. Since it serves as the foundation for translation into the Wa language, it is imperative to comprehend its subtleties in the context of machine translation.

A wide variety of vowels and consonants make up Burmese Voiced and voiceless nasals (m, n, ɲ, ŋ), stops and affricates (b, d, dʒ, ɡ, p, t, tʃ, k, ʔ), and fricatives (ð, z, θ, s, ʃ, sʰ, h) are among its consonantal inventory. Burmese also has voiceless approximants (l̥, ɹ̥) and voiced approximants (l, j, w). Burmese consonants are peculiar compared to those in many other languages, which presents an interesting linguistic problem for machine translation systems.

Example Consonants:

Voiced Nasals: m (မ), n(နှ)

Voiceless Stops: p (ပ), t (တ), k (က)

Fricatives: s (စ), h (ဟ)

Voiced Approximants: w (ဝ), j (ယ)

There are several vowel sounds in the Burmese vowel system, such as diphthongs and monophthongs. Along with a variety of low and mid vowels, the monophthongs include the high vowels /i/ and /u/, as well as the middle vowels /ɪ/ and /ʊ/. These vowel sounds can be combined to generate diphthongs, which give languages distinctive phonetic patterns and increased complexity.

Example Vowels:

High Vowels: /i/ (အီ), /u/ (ဥ)

Central Vowels: /ɪ/ (အိး), /ʊ/ (ဥု)

Low Vowels: /a/ (အ), /ɛ/ (ဧ)

3.2. Wa Language

The Wa language is still not as widely researched as the well-documented Myanmar language. Wa is a language belonging to the Mon-Khmer family that is spoken by about 950,000 people, mostly in northern Burma and some surrounding areas of Thailand and China. There are other varieties of Wa, such as Parauk, Vax, and Avax, each with a variety of dialects [14].

The Parauk language—also referred to as Phalok, Baroag, Praok, or Standard Wa—takes center stage in the Myanmar-Wa corpus. About 400,000 people in Burma speak it, mostly in the Shan States in the southeast, east, and northeast. Furthermore, the southern parts of China are home to the Parauk language. Wa is a language with unique linguistic characteristics, belonging to the Mon-Khmer family. An overview of the consonants in the Wa language is given in Table 1. Conversely, Table 2 displays the variety of vowels that add to the language's richness.

With limited English-language resources, the Wa language has not received much attention in grammatical research, despite its importance in the Myanmar-Wa corpus. The lack of thorough grammatical research emphasizes the importance of programs aimed at bridging the language divide between Wa and Myanmar [15].

Table 1: Consonants of Wa-Language

K ကꟾ[k]	KH ခꟾ[kʰ]	NG ငꟾ[ŋ]	S စꟾ[s]	SH ရꟾ[sʰ]	NY ညꟾ[ɲ]
T တꟾ[t]	TH ထꟾ[tʰ]	N နꟾ[n]	P ပꟾ[p]	PH ဖꟾ[pʰ]	M မꟾ[m]
Y ယꟾ[j]	R ရꟾ[r]	L လꟾ[l]	V ဗꟾ[v]	W ဝꟾ[ua]	H ဟꟾ[h]
X အꟾ[ʔ]	C ကျ[c]	CH ချ[cʰ]	D ဒꟾ[ⁿd]	G ဂꟾ[ᵑg]	Q ဂꟾ[ᵑg]
	B ဘ[ᵐb]	F ဖꟾ[vʱ]	J ဂျ[ᶮɟ]	Z ဇ[zə]	

Table 2: Vowels of Wa-Language

A အာ[a]	E အေ[ei]	IE အဲ[ɛ]	AW အော်[ɔ]	OI အွဲ[oi~ɔi]
AO အောင်း[au]	AU အော့ပ်[aɯ]	AI အိုင်[ai]	I အီ[i]	O အို[o]
	U အူ [u]	EE အီး[ɯ]	EU အေး[ɤ]	

Tables 1 and 2 present Wa's phonological features, encompassing its vowels and consonants, whereas Figure 1 depicts the Wa classification (Mon-Khmer Family Tree).

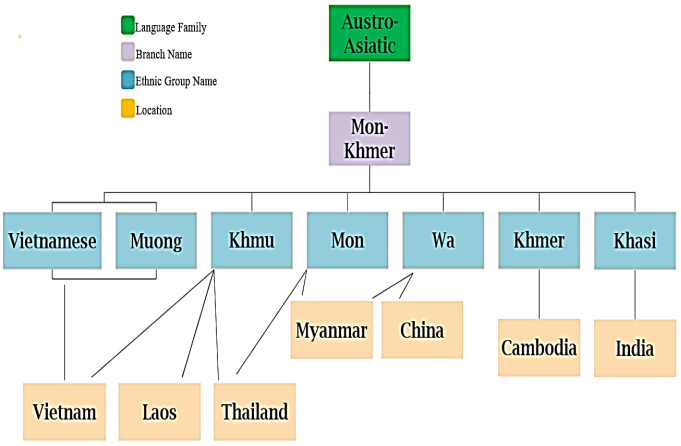


Figure 1: Mon-Khmer Family Tree [16]

3.3. Usage in Machine Translation

Burmese is becoming increasingly important in the worldwide environment, which makes it relevant in the field of machine translation. Accurate translation services between Burmese and other languages are in high demand due to Myanmar's recent political upheavals and growing involvement in regional and global issues.

Take, for instance, the translation of the straightforward phrase "Hello, how are you?" from English to Burmese. In this situation,

Burmese has unique phonological and grammatical patterns that machine translation systems must understand in order to provide a result that is accurate and suitable for the target culture.

Example Translation

"Hello, how are you?" (English)

"မင်္ဂလာ ပါ ၊ နေကောင်း ပါ သလား ။" (Burmese)

“Bawk yam pa mhawm, nyawm ot maix” (Wa)

(here “maix” means “you” in English and hidden word (နင်၊ သင်) in myanmar)

Here, the Burmese translation effectively conveys the subtleties of the target language as well as the customs of polite conversation, highlighting the value of precise translation in cross-cultural communication. In conclusion, there are many interesting opportunities and problems for machine translation research because of the Burmese language's distinctive morphology, phonology, and cultural relevance. Accurate translation between Burmese and other languages is becoming more and more important as Myanmar continues to interact with the outside world, highlighting the need for improvements in machine translation technology catered to this unique language.

Table 3: VSO-SVO word order variation observed in Wa clauses

ကျွန်တော်တို့ နိုင်ငံ သည် စိုက်ပျိုးရေး နိုင်ငံ တစ်ခု ဖြစ်သည် ။	Hak tiex ex mawh hak tiex sum van tix mu heue	Our country is an Agro-based country.
ငါတို့ ၁၀ နာရီ မှာ အစည်းအဝေး လုပ် တယ် ။	Ex yuh bang kum yam kao taix	We had a meeting at 10 o'clock.
မင်း ဘယ် နေရာ မှာလဲ ။	Maix dee mawx	Where were you?
ဝမ်းနည်း ပါတယ်	Voi rhawm heue	I'm sorry.
ငါ မေ့ သွား တယ် ထင်တယ် ။	Pi awm aux tix yong	I guess I forget.
ဘာလို့ မေ့ တာလဲ ။	Jao tix pi awm lie	How could you forget?
ငါ မသိ တော့ဘူး ။	Ang aux lai tawng	I don't know.

3.4. An Academic Exploration of the Wa Language

Southeast Asia is where most Wa speakers are found. There are typological characteristics in this language that call for more investigation. Wa is regarded as a head-initial language, which means that modifiers like adjectives, relative clauses, and numbers come after nouns and verbs come before objects. Wa's subject and object markings are logical given where they fall in the sentence. Wa's particular grammatical structures are further demonstrated by the placement of the negation before the verb.

A detailed examination of the Wa language reveals a number of fascinating conclusions. For example, in Wa, several adverbs from other languages are articulated as verbs. Wa also has two alternate clause word orders, SVO and VSO as shown in Table 3, and allows components to be moved out of noun phrases. Negation frequently involves a supplementary negation particle. Interestingly, this shift in word order seems to depend on the kind of clause rather than verb transitivity or semantics.

The Wa language has received comparatively little scholarly study, despite its importance within the Mon-Khmer linguistic family. research now in existence primarily concentrate on phonological characteristics; no thorough grammatical research have been published in English. Especially, it explores the unique VSO-SVO word order variation found in Wa clauses.

3.5. Data Collection and Corpus Size Analysis for Myanmar-Wa Machine Translation

The establishment of the Myanmar-Wa corpus marked a significant milestone as it represented our inaugural venture into crafting a corpus tailored for machine translation. Throughout its development, meticulous attention was dedicated to linguistic nuances, with the Myanmar-Wa-Chinese dictionary serving as a valuable point of reference. Despite our commitment to precision, the absence of parallel data for the Myanmar-Wa language pair posed a formidable challenge during the data gathering phase.

The unique nature of low-resource languages, exemplified by Wa, presented limited opportunities for text alignment compared to more commonly spoken language pairs like English-Spanish or English-French, which benefit from extensive bilingual corpora. In response to this hurdle, we employed innovative approaches such as crowdsourcing, cross-lingual transfer learning, and data augmentation to augment the corpus's size and diversity.

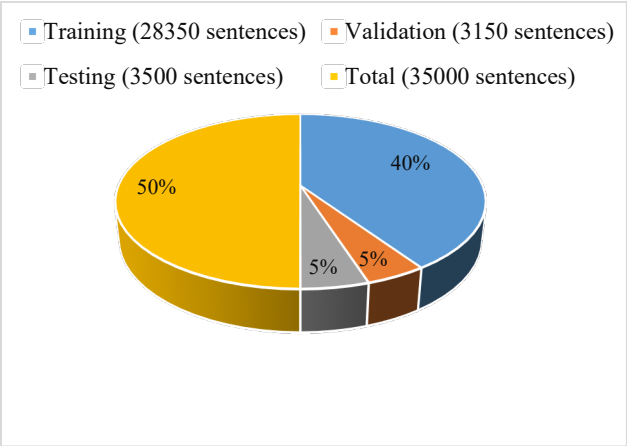


Figure 2: Corpus Size for Myanmar-Wa

It is crucial to note that the Myanmar-Wa corpus not only represents a pioneering effort in our machine translation endeavors but also holds immense value for future research initiatives. By addressing the specific data constraints inherent in low-resource languages, these initiatives aimed to lay the groundwork for the development of machine translation models characterized by heightened precision and contextual appropriateness.

Furthermore, the Figure 2 illustrates the corpus size chart provides a visual representation of the volume of data utilized for

training, validation, and testing phases in Myanmar-Wa language pair.

The Myanmar-Wa corpus, representing the pioneering effort in machine translation for this language pair, is strategically divided for training, validation, and testing. The training set comprises 28,350 instances, the validation set 3150, and the test set 3500, totaling 35,000 instances. This meticulous segmentation ensures a robust and comprehensive foundation for developing precise and contextually appropriate machine translation models for Myanmar-Wa.

4. Methodologies of the Proposed System

The evolution of machine translation (MT) has witnessed remarkable success, driven by the availability of massive parallel corpora and advancements in Natural Language Processing (NLP). Despite these strides, adapting MT to the unique linguistic challenges posed by the Myanmar-Wa language pair has been met with considerable difficulties. Early attempts faced restrictions and flaws, primarily stemming from the scarcity of aligned sentences in the Myanmar-Wa corpus, hindering the development of reliable translation models. The absence of parallel corpora also impeded the progress of Statistical Machine Translation (SMT) techniques [17].

The low-resource nature of the Wa language, characterized by peculiar grammar and a small vocabulary, further compounded the challenges. Traditional machine translation models, designed for high-resource languages, struggled to capture the nuances of Wa [18]. Linguistic distinctions between Wa and Myanmar, encompassing vocabulary, syntax, and word order, posed additional hurdles, leading to suboptimal translation quality. Cultural allusions and subtleties, integral to accurate translation, were often overlooked in early efforts, underscoring the need for creative solutions.

To overcome these challenges, scholars and industry professionals explored innovative approaches, leveraging cutting-edge NLP methods and modifying existing models to suit the Myanmar-Wa language combination. The journey towards enhancing machine translation for Myanmar-Wa unfolded through various stages, each contributing to the refinement and improvement of translation quality.

4.1. Statistical Machine Translation (SMT)

In the initial phases, Statistical Machine Translation (SMT) techniques were explored, but progress was hampered by the lack of parallel corpora. The limitations imposed by the absence of aligned sentences in the Myanmar-Wa corpus prompted the exploration of alternative methodologies.

4.2. LSTM-based Machine Translation Models

As a response to the challenges faced by SMT, Long Short-Term Memory (LSTM) models were introduced. LSTMs, with their ability to capture sequential dependencies, showed promise in handling the complexities of the Wa language. The LSTM architecture, though an improvement, still grappled with the unique linguistic characteristics of Wa, necessitating further advancements. Figure 3 describes the LSTM encoder-decoder architecture.

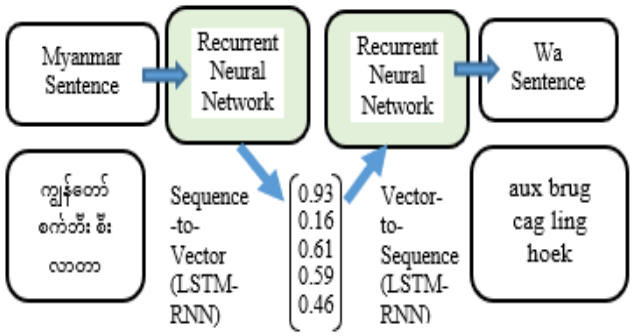


Figure 3: Architecture of LSTM Encoder-Decoder

4.3. Transformer-based Machine Translation Models

The advent of transformer-based models marked a pivotal shift in the trajectory of machine translation for Myanmar-Wa. The transformer architecture, with its attention mechanisms and positional embeddings, introduced a novel approach to handling sequential dependencies. The dynamic representation of the transformer's output, dependent on the input sequence's length, addressed issues related to varying word orders and contextual nuances, contributing to improved translation quality.

Figure 4 illustrates the architecture of the Transformer Encoder-Decoder, highlighting the interconnected layers and the flow of information between the encoder and decoder. The transformer's innovative design, incorporating attention mechanisms and positional embeddings, marks a significant advancement in machine translation architectures.

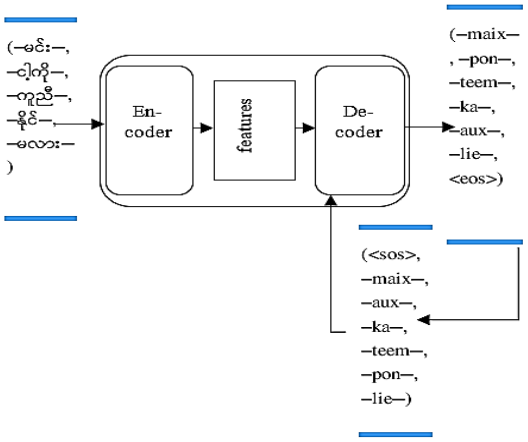


Figure 4: Architecture of Transformer Encoder-Decoder

4.4. Traditional Transfer Learning on MT5 Pretrained Model

The concept behind traditional transfer learning is rooted in enhancing a machine's performance in a related task by leveraging knowledge from a pre-existing model. In Neural Machine Translation (NMT), this involves training a new model (child) with initial parameters from a previously trained model (parent) and a sparse parallel corpus. The Myanmar-Wa corpus, our pioneering effort in machine translation, showcases the utilization of transfer learning. Despite challenges in aligning low-resource languages like Wa, innovative methods, including crowdsourcing and data augmentation, were employed to augment the corpus's size and diversity.

The corpus size chart (refer to Section 3.5) illustrates the division of the Myanmar-Wa corpus into training (28,000 instances), validation (5,600 instances), and test (7,000 instances) sets, totaling 35,000 instances. Transfer learning benefits machine translation in several ways: reducing the need for extensive data, improving translation quality, accelerating training, and enabling cost-effective multilingual translation.

$$\theta_{fine-tuned} = argmin_L(L(D_{Myanmar-Wa}, \theta) + \lambda L(D_{pretrained}, \theta))$$

where $D_{Myanmar-Wa}$ represents the Myanmar-Wa parallel corpus, $D_{pretrained}$ denotes the pre-trained multilingual corpus, θ signifies model parameters, and λ is the hyperparameter governing the pre-trained model's influence.

The use of the MT5 model significantly enhances translation quality for Myanmar-Wa. Transfer learning, coupled with SMT, addresses data scarcity, and LSTM-based models capture language dynamics.

4.5. A Pivotal Shift in Machine Translation with Myanmar and Wa Language Models

The pivotal role of machine translation (MT) in breaking linguistic barriers has been a relentless journey, marked by significant technological advancements. This section outlines the construction and training of a machine translation model, leveraging a pivot transfer learning approach for enhanced translation capabilities. The study focuses on the Myanmar (Burmese)-Wa corpus, representing a subgroup of the Austro-Asiatic language family.

The initial steps involve tokenization and encoding of the source and target sentences using state-of-the-art models. For the Myanmar language, a segmentation step is performed before tokenization. The source sentences are segmented and encoded using a Myanmar word segmentation tool [19], and then encoded using the Myanmar BERT model (UCSYNLP/MyanBERTa). The target sentences are encoded using the Myanmar word segmentation model, followed by encoding with the Facebook BART model.

Pivot transfer learning involves training an encoder-decoder architecture for machine translation. The Myanmar BERT model acts as the encoder. The Myanmar BERT model, initially designed for natural language understanding, serves as the encoder, extracting contextual information from the segmented source sentences. The language model, named MyanBERTa, is founded on the BERT architecture and tailored for the Myanmar language. MyanBERTa underwent a comprehensive pre-training phase encompassing 528,000 steps, utilizing a word-segmented dataset specific to Myanmar. This extensive corpus comprised 5,992,299 sentences, equivalent to 136 million words. The employed tokenizer is a byte-level Byte Pair Encoding (BPE) tokenizer with 30,522 subword units, learned subsequent to the application of word segmentation.The pre-trained Facebook BART model acts as the decoder, generating target translations based on the encoded information. Bidirectional AutoRegressive Transformer (BART), embodies a transformer architecture consisting of an encoder-decoder structure for sequence-to-sequence (seq2seq) tasks. The model exhibits notable efficacy when fine-tuned for text generation applications such as summarization and translation [20].

This two-step process facilitates effective transfer of knowledge from one model to another, enhancing the overall translation performance.

The training process spans multiple epochs, with each epoch iterating through the training dataset. The combined model is fine-tuned using the AdamW optimizer, and a step-wise learning rate scheduler ensures optimal convergence. To mitigate GPU memory constraints, gradient accumulation steps are employed, allowing the model to accumulate gradients over several batches before updating the parameters.

Validation is a crucial step to assess the model's generalization on unseen data. The validation dataset, comprising the Myanmar-Wa corpus, is utilized to compute the validation loss and perplexity. Perplexity, a measure of translation quality, is calculated as the exponential of the average validation loss. This metric provides insights into how well the model captures the complexity of the language and context.

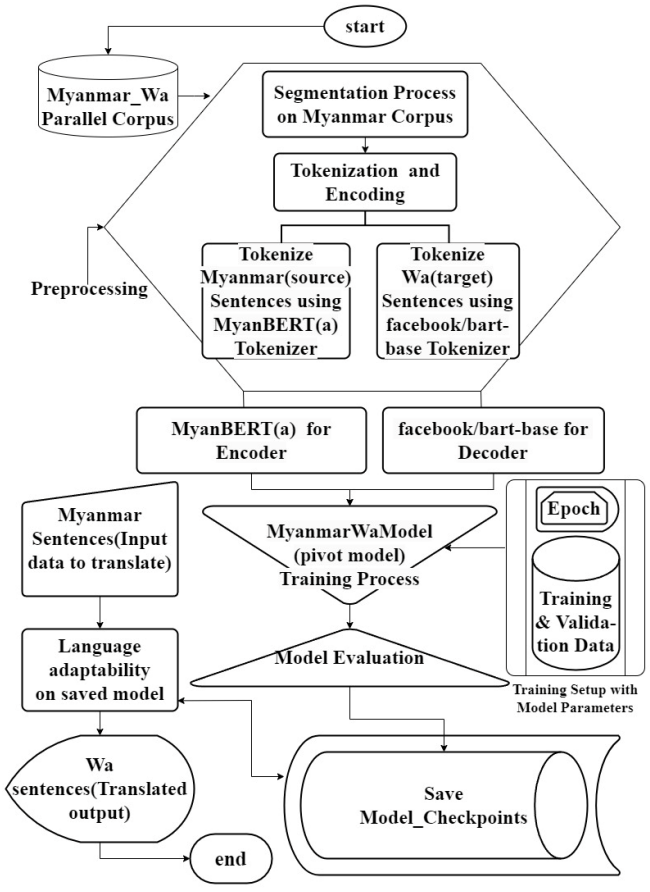


Figure 5: Architecture of the Pivot Transfer Learning Model for Myanmar-Wa Machine Translation

To capture the model's progression, snapshots are saved at the end of each epoch. The saved models serve as checkpoints and facilitate model reproducibility. The model's evolution is crucial for understanding its learning trajectory and identifying potential areas for improvement.

This model-building process encapsulates the essence of pivot transfer learning, employing two specialized models for optimal translation performance. The journey through the Myanmar-Wa corpus showcases the adaptability and effectiveness of the proposed approach. The continual evolution of machine translation models, driven by innovative strategies and adaptability to diverse linguistic challenges, contributes significantly to the broader domains of linguistics and language preservation.

Figure 5 illustrates the architecture of the pivot transfer learning model employed for machine translation in the Myanmar (Burmese)-Wa corpus. The model leverages Myanmar BERT as the encoder and Facebook BART as the decoder, demonstrating the pivotal role of transfer learning in enhancing translation capabilities. The training process, validation steps, and model evolution are highlighted, showcasing the adaptability and effectiveness of the proposed approach in breaking linguistic barriers.

Table 4: Explanation of System Design Symbols

Symbol	Description
	Represents the start of the process or system.
	Denotes a database or data storage component.
	Signifies a step involving data preparation or preprocessing.
	Represents a component responsible for displaying information.
	Indicates a general process or computation step.
	Represents a point where data is merged or stored.
	Denotes a step involving data extraction or measurement.
	Represents manual input by a user or external entity.
	Signifies the end or termination point of the process or system.
	Represents direct data flow between components.
	Denotes the use of stored data in a process.
	Represents a predefined or standardized process.
	Indicates a loop limit or iteration constraint.
	Denotes a time delay or waiting period in the process.

In Table 4, the comprehensive index of symbols utilized in the system design is presented. Each symbol corresponds to specific components or processes within the system, aiding in the visual representation and understanding of the system architecture.

4.6. Pivot-Based Transfer Learning NMT Models with MT5

Pivot-based machine translation is a dual-stage process involving the translation from the source language (SL) to the pivot language (PL) and subsequently from the pivot language (PL) to the target language (TL). Due to the scarcity of parallel data for Myanmar-Wa, pivot-based NMT with MT5 models has been meticulously explored. The architecture is rooted in transfer learning, meticulously tailored for non-English language pairs, particularly for Myanmar-Wa.

The operational framework encompasses three crucial stages. First is the pretraining phase, leveraging source-pivot and pivot-target parallel corpora. This initial step allows the model to grasp the nuances of both source and target languages in the context of the pivot language. The subsequent stage involves the integration of a source-target parallel corpus for iterative refinement, fine-tuning the model to enhance its translation capabilities for the specific task at hand. The final stage revolves around leveraging the pivot language for pretraining both source encoders and target decoders, facilitating a holistic understanding of language dynamics.

In this context, the MT5 model stands as a robust alternative to the traditional MT5 model. By strategically selecting Myanmar-English and English-Fil (Philippines) as pivot language pairs, chosen for their analogous sentence structures, the system capitalizes on the shared linguistic characteristics between Wa and Filipino [21]. This strategic choice significantly contributes to improving translation quality and captures the synergies between these languages.

Drawing parallels to the Python code provided, the transfer learning process mirrors the combining of parameters from two distinct models (`model1` and `model2`) into a new model (`new_model`). The encoder parameters from `model1` are seamlessly integrated into the new model, and likewise, the decoder parameters from `model2` are incorporated. This combination represents a fusion of linguistic knowledge from the two models into a unified architecture.

The use of pre-trained models, especially MT5, in this transfer learning paradigm underscores the flexibility and adaptability of this approach across diverse language pairs. This innovative methodology, akin to the Python code's model parameter merging, holds significant promise for advancing machine translation capabilities, particularly in low-resource language scenarios.

In this context, the MT5 model serves as a robust alternative to the MT5 model. By selecting Myanmar-English and English-Fil (Philippines) as pivot language pairs, which share similar sentence structures, the system capitalizes on the commonalities between Wa and Filipino, improving translation quality.

The utilization of pre-trained models, especially MT5, in transfer learning underscores the adaptability of this approach across diverse language pairs. This innovative methodology holds promise for advancing machine translation capabilities, particularly in low-resource language scenarios.

4.7. Myanmar Word Segmentation Tool

To further refine the Myanmar to Wa machine translation, a specialized Myanmar word segmentation tool, developed by the

NLP Lab at the University of Computer Studies, Yangon, Myanmar [22], was employed. This tool, tailored for Myanmar language segmentation, enhanced the accuracy of the translation process, providing a solid foundation for subsequent advancements.

The step-by-step progression from SMT to LSTM, transformer-based models, traditional transfer learning, and pivot-based transfer learning culminated in the development of our proposed model. Each iteration addressed specific challenges posed by the Myanmar-Wa language pair, resulting in a more robust and accurate machine translation system. This comprehensive approach showcases the continual evolution of machine translation methodologies, driven by a commitment to overcome linguistic barriers and contribute to the broader domains of linguistics and language preservation.

5. Linguistic Analysis and Evaluation

Linguistic analysis and evaluation form a critical aspect in gauging the efficacy of machine translation (MT) systems, especially in the intricate context of Myanmar-Wa language translation. This section delves into the multifaceted linguistic properties of the Wa language and elucidates the methodologies and metrics employed to comprehensively assess the performance of MT systems on the Myanmar-Wa corpus.

5.1. Analyzing Linguistic Features

The Wa language, belonging to the Mon-Khmer language family, introduces numerous linguistic intricacies that offer both potential and challenges for MT. A meticulous examination of these distinctive features is indispensable for a nuanced understanding of Wa-Myanmar translation. Key linguistic characteristics that demand in-depth analysis include:

- Word Order Flexibility: Wa exhibits both Verb-Subject-Object (VSO) and Subject-Verb-Object (SVO) structures, introducing considerable diversity in word order. Successful translation necessitates a profound understanding of context [23].
- Tonal Complexity: As a tonal language, Wa relies on intonation and pitch variations for meaning. Preserving intended meaning during translation hinges on capturing tonal subtleties accurately [24].
- Agglutinative Morphology: Wa's agglutinative morphology involves conveying different grammatical functions through affixes added to root words. Fluent translation requires familiarity with these affixes [25].
- Nominalization: Verbs or adjectives in Wa commonly undergo nominalization, a process crucial for accurate translation.
- Cultural Nuances: The Wa language is intertwined with Wa culture and identity. To maintain cultural integrity in translations, special attention is required for terms and phrases holding cultural value.

5.2. Evaluation Metrics

The evaluation of MT systems for Myanmar-Wa encounters unique challenges due to limited parallel corpora and linguistic distinctions. Performance assessment utilizes various metrics and techniques, including:

- BLEU Score: Measures the similarity between reference translations and MT output, though it may not fully capture linguistic subtleties in low-resource language pairs like Myanmar-Wa.
- METEOR Score: Considers word order, stemming, and synonyms, providing a more comprehensive assessment of translation quality.
- Manual Evaluation: Expert linguists or native speakers evaluate translations, offering insights into cultural and contextual components.
- Cross-lingual Embedding: Measures semantic similarity between source and target sentences for additional information on translation quality.
- Transfer Learning Assessment: Evaluates how well models optimize Myanmar-Wa translation through the utilization of pre-trained knowledge from other languages.

Combining these quantitative and qualitative measurements provides a comprehensive understanding of the strengths and weaknesses of MT systems for Myanmar-Wa, guiding future advancements in the field.

5.3. Sentence Length Analysis

In machine translation, sentence length significantly influences model performance, especially in languages like Wa with diverse syntactic structures. This sub-section scrutinizes the impact of sentence length on Myanmar-Wa machine translation systems. The analysis involves:

- Exploration of Distribution: A detailed examination of sentence length distribution in both source and target languages, revealing potential correlations with translation quality.
- Insights from Visualization: Figure 6 illustrates the distribution of sentence lengths in the Myanmar-Wa corpus, offering insights into the nature of sentences and emphasizing the complexity of translation between Myanmar and Wa.

Longer sentences in the source language pose challenges for maintaining context and meaning in the target language. Understanding sentence length distribution aids in adapting machine translation models effectively. This visualization serves as an initial exploration, paving the way for further quantitative analysis and enhancements in models tailored for the Myanmar-Wa language pair. Each bar in the chart signifies a sentence, with the y-axis denoting sentence length and the x-axis representing the number of sentences.

Longer sentences in the source language may pose challenges for translation, particularly in preserving context and meaning in the target language. Understanding the sentence length distribution aids in adapting machine translation models to effectively handle the linguistic variations present in the Myanmar-Wa corpus. This visualization serves as a foundational exploration of sentence length characteristics, laying the groundwork for further analysis and improvements in machine translation models for the Myanmar-Wa language pair. Each bar in the chart represents a sentence, with the y-axis indicating the length of the sentence and the x-axis representing the number of sentences.

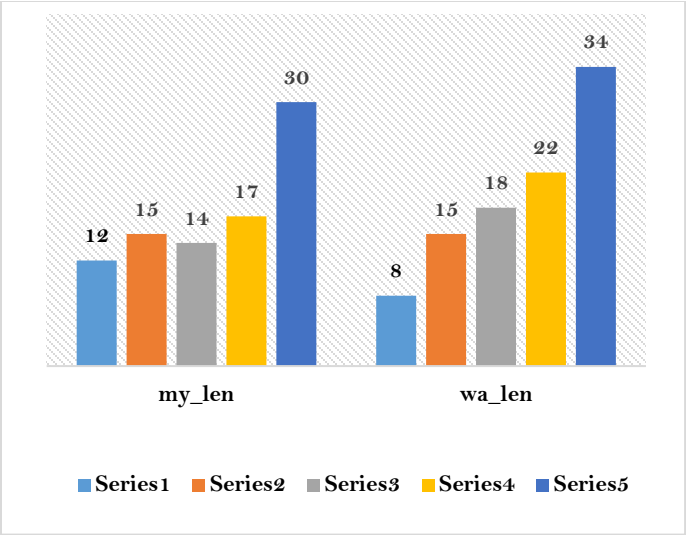


Figure 6: Sentence Length Analysis

6. Impacts and Future Directions

This section provides an in-depth exploration of the far-reaching implications and promising future avenues for studying improved Machine Translation (MT) in the Myanmar-Wa language pair. It examines the practical applications, language preservation aspects, and outlines key directions for future research.

6.1. Practical Applications

Improved MT for the Myanmar-Wa language pair has diverse practical applications, facilitating efficient communication, educational accessibility, and healthcare information translation. Additionally, it plays a crucial role in preserving Wa tradition and culture, as well as boosting tourism. Each application is detailed to highlight the significant implications.

6.2. Language Preservation

This subsection underscores the importance of enhanced MT as a tool for protecting and revitalizing endangered languages like Wa. It emphasizes the development of digital language resources, language learning support, and the preservation of dialectal variants and oral traditions. The broader cultural legacy of the Wa community is also discussed.

6.3. *Future Research Directions*

Identified research directions for MT in underrepresented languages like Wa are outlined, emphasizing the need for high-quality parallel corpora, domain-specific adaptations, and ethical considerations. The involvement of linguists, local communities, and technologists in collaborative projects is highlighted, along with potential enhancements through the incorporation of visual and audio modalities.

7. Comparative Analysis of Machine Translation Models

7.1. Previous Experiments on Myanmar-Wa MT System

The initial experiments on the Myanmar-Wa machine translation system utilized both Statistical Machine Translation (SMT) and Long Short-Term Memory (LSTM) models. The training data comprised 22,500 instances, with 700 instances for testing and 500 instances for validation. The analysis focused on the BLEU score, considering different batch sizes (64, 126, and 256). Results indicated that a higher batch size of 256 yielded the best accuracy in terms of BLEU score. These experiments, conducted with a limited dataset, underscored the need for a more extensive Myanmar-Wa parallel corpus in future endeavors. The SMT model achieved a BLEU score of 12.67.

7.2. Comparison with New Experiments and Added Corpus Size

Subsequent experiments saw the expansion of the Myanmar-Wa corpus, allowing for a comprehensive evaluation of machine translation models (Combined Model, T5, Transformer). The performance metrics for various models are summarized in Figure 7.

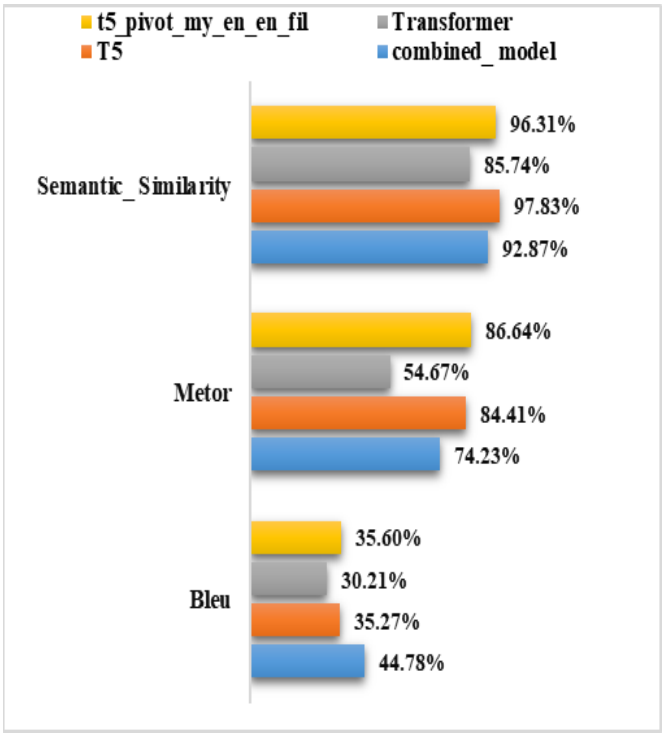


Figure 7: Performance Metrics for Various Models

7.3. Model Performance Analysis

- Combined Model (New): Stands out with the highest Bleu score, indicating strong n-gram overlap. The Metor score is commendable, reflecting alignment with human reference translations. High Semantic Similarity emphasizes proficiency in maintaining semantic accuracy during translation.
- T5: Demonstrates a strong Metor score, indicating correlation with human reference translations. Notably high Semantic Similarity suggests effective capture of semantic nuances. While the Bleu score is lower than the Combined Model, T5 excels across various metrics, showcasing versatility.
- Transformer: Shows respectable performance in Bleu scores and Semantic Similarity. Lagging Metor score indicates potential challenges in maintaining fidelity to human reference translations. Further optimization may enhance performance.
- T5 Pivot My-En En-Fil (New): Exhibits balanced performance across all metrics. Noteworthy Bleu and Metor scores, along with effective preservation of semantic content

indicated by Semantic Similarity. Strong contender in scenarios where a combination of metrics is crucial.

7.4. Observations and Insights

- Combined Model Outperforms: The newly introduced combined model excels across all metrics, with a Bleu score of 44.78, Metor score of 74.23, and high Semantic Similarity of 92.87.
- T5 Pivot My-En En-Fil: Demonstrates noteworthy improvements, particularly in Bleu (35.6) and Metor (86.64) scores. High Semantic Similarity (96.31) emphasizes effectiveness.
- T5 and Transformer Comparison: While T5 and Transformer models show respectable performance, T5 generally outperforms in terms of Bleu, Metor, and Semantic Similarity.
- Semantic Similarity Emphasis: Crucial for evaluating models' ability to capture meaning. The combined model excels, showcasing capability to preserve semantic nuances during translation.

7.5. Summary and Implications

The choice of the best model depends on specific translation task priorities. T5 and T5 Pivot My-En En-Fil exhibit robust performance in METEOR and semantic similarity, emphasizing versatility. Meanwhile, the Combined Model excels in Bleu scores and semantic similarity, showcasing linguistic precision. Notably, additional training data could further enhance models like the Combined Model.

These nuanced analyses should guide researchers in selecting models for machine translation tasks, recognizing that distinct aspects of performance are illuminated by different metrics. Strategies to enhance semantic preservation and leverage additional data for refining capabilities of all models could be explored in further investigations. The continuous evolution of machine translation models contributes to language understanding and communication.

It is imperative to note that the Myanmar-Wa corpus used in these experiments is the very first corpus created by the research team. The corpus size was expanded to enhance the machine translation system's performance and evaluation. Table 5 presents the sample result outputs on long sentences, middle sentences, and short sentences.

Table 5: Sample Result Outputs on Long Sentences, Middle Sentences, and Short Sentences

long sentences	
combined_model	source: ပြိုင်ပွဲ ဘယ်အချိန် စ ရ မလဲ ဆိုတာ မင်း ငါ့ ဆီ ဖုန်း ဆက် ပါ
	reference: you call me up when to start contest
	target: maix tah maox rhiem ka aux bang silah sang jah yam mawx
	prediction: maix tah maox rhiem ka aux sang pawk bawg yam mawx
Transformer	source: ပြိုင်ပွဲ ဘယ်အချိန် စ ရ မလဲ ဆိုတာ မင်း ငါ့ ဆီ ဖုန်း ဆက် ပါ
	reference: you call me up when to start contest
	target: maix tah maox rhiem ka aux bang silah sang jah yam mawx
	prediction: maix tah maox rhiem caw tix hoek yam mawx pa tah maox rhiem ka maix
T5	source: ပြိုင်ပွဲ ဘယ်အချိန် စ ရ မလဲ ဆိုတာ မင်း ငါ့ ဆီ ဖုန်း ဆက် ပါ
	reference: you call me up when to start contest
	target: maix tah maox rhiem ka aux bang silah sang jah yam mawx
	prediction: maix tah maox rhiem ka aux sang jah yam mawx
T5 with pivot logic	source: ပြိုင်ပွဲ ဘယ်အချိန် စ ရ မလဲ ဆိုတာ မင်း ငါ့ ဆီ ဖုန်း ဆက် ပါ
	reference: you call me up when to start contest
	target: maix tah maox rhiem ka aux bang silah sang jah yam mawx
	prediction: tah maox rhiem ka aux sang jah tix jah bang silah yam mawx
middle sentences	
combined_model	source: ဒါဟာ သိပ် ကောင်းတဲ့ စိတ်ကူး နဲ့ တူ ပါတယ်
	reference: that sounds like a good idea
	target: in awm krax keud pa kied sidaing mhawm
	prediction: in awm pa kied sidaing mhawm tete
Transformer	source: ဒါဟာ သိပ် ကောင်းတဲ့ စိတ်ကူး နဲ့ တူ ပါတယ်
	reference: that sounds like a good idea
	target: in awm krax keud pa kied sidaing mhawm
	prediction: in mawh grawng kaing pa sidaing mhawm
T5	source: ဒါဟာ သိပ် ကောင်းတဲ့ စိတ်ကူး နဲ့ တူ ပါတယ်
	reference: that sounds like a good idea
	target: in awm krax keud pa kied sidaing mhawm
	prediction: in awm mai krax keud pa kied mhawm
T5 with pivot logic	source: ဒါဟာ သိပ် ကောင်းတဲ့ စိတ်ကူး နဲ့ တူ ပါတယ်
	reference: that sounds like a good idea
	target: in awm krax keud pa kied sidaing mhawm
	prediction: in awm mai krax keud pa kied mhawm

short sentences	
combined_model	source: တို့ဘယ်မှာ ဆုံ ကြ မလဲ
	reference: where shall we meet
	target: ex sang pup dee mawx
	prediction: ex sang pup dee mawx
Transformer	source: တို့ဘယ်မှာ ဆုံ ကြ မလဲ
	reference: where shall we meet
	target: ex sang pup dee mawx
	prediction: ex sang sipup dee mawx
T5	source: တို့ဘယ်မှာ ဆုံ ကြ မလဲ
	reference: where shall we meet
	target: ex sang pup dee mawx
	prediction: ex sang sipup dee mawx
T5 with pivot logic	source: တို့ဘယ်မှာ ဆုံ ကြ မလဲ
	reference: where shall we meet
	target: ex sang pup dee mawx
	prediction: ex sang pup paox dee mawx

8. Conclusion

The journey through the evolution of machine translation within the context of the Myanmar-Wa corpus has been marked by iterative experimentation and continuous refinement. In the pursuit of refining translation capabilities, our study extensively evaluates various models, each contributing unique strengths and insights.

Among the models considered, the Combined Model (New) emerges as a standout performer, showcasing the highest Bleu score, a commendable Metor score, and exceptional Semantic Similarity. This comprehensive translation capability positions the Combined Model (New) as a leader across multiple metrics, highlighting its proficiency in maintaining semantic accuracy during translation.

T5, with its strong Metor score and notable Semantic Similarity, demonstrates versatility in translation tasks. While its Bleu score is slightly lower compared to the Combined Model, T5's performance remains commendable, underlining its effectiveness in capturing semantic nuances.

The Transformer model, although showing respectable performance in Bleu scores and Semantic Similarity, lags in Metor, suggesting potential challenges in fidelity to human reference translations. Further optimization may enhance its overall performance.

The T5 Pivot My-En En-Fil (New) variant exhibits balanced performance, particularly excelling in Bleu and Metor scores, along with high Semantic Similarity. This model stands as a strong contender, especially in scenarios where a combination of metrics is crucial.

In conclusion, our comprehensive evaluation indicates that the Combined Model (New) leads in overall performance, followed closely by T5 and T5 Pivot My-En En-Fil (New). These findings provide valuable insights for the advancement of machine translation in the Myanmar-Wa context, guiding future investigations and contributing to the broader landscape of machine translation research. The expanded corpus size, coupled with these model assessments, serves as a robust foundation for ongoing and future research endeavors.

Acknowledgment

The Divine Creator, who gave me the permission to start the "Tracing the Evolution of Machine Translation: A Journey through the Myanmar (Burmese)-Wa (sub-group of the Austro-Asiatic language) Corpus" study, is the reason this research exists. I would like to express my sincere gratitude to Dr. Mie Mie Khin, the distinguished Rector of the University of Computer Studies, Yangon, for providing me with this exceptional chance to explore this exciting field of study. I am also grateful to Dr. Win Le' Le' Phyu, Professor; she is a guiding light from the University of Computer Studies, Yangon, and her strong support has fostered my academic endeavors and given them the time they need to grow. I also express my gratitude to my group members and UCSY NLP Lab colleagues, who have kindly offered their personal and professional perspectives, enhancing my comprehension of scientific research and life in general. My profound gratitude is extended to Dr. Khin Mar Soe, my committed research mentor, whose function as a teacher and advisor has given me immeasurable wisdom. Dr. Win Pa' Pa, a renowned professor at NLP LAB (University of Computer Studies, Yangon), deserves special commendation for her acts, which embody the traits of a model scientist and kind person. I would especially want to thank Drs. Yi Mon Shwe Sin and Aye Mya Hlaing, whose contributions to this effort were deeply appreciated. Mr. AI MHAWM KHAM is a valuable contributor who deserves special recognition for his expert translation from Burmese to WA. This translation was a critical factor in making this work possible. During the execution of this project, my family provided steadfast support. I am truly grateful to my parents and siblings for being the best role models in my life and for their unending love and support. Throughout this journey, their constant belief in me gave me more drive and determination. Finally, I would want to express my appreciation to Pastor AI NAP TAO and Associate Pastor AI KHWAT PAN for their unwavering prayers and words of spiritual support.

References

[1] G. Salton, "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer," IEEE Transactions on Audio and Electroacoustics, **16**(3), 366-366, Sep. 1968.

[2] P. Hanks, "Computational Lexicography for Natural Language Processing," IEEE Computational Intelligence Magazine, **11**(3), 9-20, Aug. 2016.
[3] F. Yune and K. M. Soe, "Myanmar-Wa Machine Translation using LSTM-based Encoder-Decoder Model," 2023 IEEE Conference on Computer Applications (ICCA), Yangon, Myanmar, 2023, 1-5, doi: 10.1109/ICCA51723.2023.10181692.
[4] W. N. Locke and A. D. Booth, "Translation. Machine translation of languages", Fourteen essays, MIT Press, ISBN:9780262120029, 15 May, 1955.
[5] W. J. Hutchins, "Machine translation: Past, present, future", Ellis Horwood, 2005.
[6] P. Koehn, "Statistical machine translation", Cambridge University Press, 2010.
[7] A. Vaswani and et al. "Attention is all you need", Advances in neural information processing systems (NeurIPS), 2017.
[8] G. Lample, and et al. "Cross-lingual Language Model Pretraining", Advances in Neural Information Processing Systems, 2019.
[9] J. Smith, "Language Preservation in the Digital Age: Challenges and Strategies," IEEE Computer, **42**(10), 106-108, Oct. 2009.
[10] K. Chen, R. Wang, M. Utiyama, E. Sumita, and T. Zhao, "Syntax-directed attention for neural machine translation," in Proceedings of the AAAI conference on artificial intelligence, **32**(1), New Orleans, LA, USA, April 2018.
[11] Z. Zhizhi and Y. Qixiang. "A brief description of the Wa language", 1984.
[12] D. Wu, "Machine Learning for Sino-Tibetan Language Translation: Challenges and Progress," IEEE Transactions on Neural Networks and Learning Systems, **29**(11), 5616-5631, Nov. 2018.
[13] S. Duanmu, "The Phonology of Standard Chinese (2nd edition.)", Oxford University Press, 2007.
[14] S. Mai. "A descriptive grammar of Wa." (Master's thesis, Payap University, 2012). Retrieved from https://www.academia.edu/44390404/A_descriptive_grammar_of_wa_language.
[15] Y. Xiang and J. Ke, "Linguistic characteristics of the Wa language", A preliminary exploration. Language Documentation & Conservation, 15, 87-103,2021.
[16] Human Computation Institute, "Wa Dictionary Corpus". [Online]. Available: https://www.humancomp.org/wadict/wa_corpus.html
[17] S. Hochreiter and J. Schmidhuber,"Long Short-Term Memory", Neural Computation, **9**(8), pp - 1735-1780, 1997.
[18] T. Xie, "Burmese (Myanmar): An Introduction to the Script", Southeast Asian Language Resource Center, 2008.
[19] Aye Mya Hlaing, Win Pa Pa, "MyanBERTa: A Pre-trained Language Model For Myanmar", In Proceedings of 2022 International Conference on Communication and Computer Research (ICCR2022), November 2022, Seoul, Republic of Korea.
[20] A. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," CoRR, abs/1910.13461, 2019. [Online]. Available: http://arxiv.org/abs/1910.13461.
[21] C. Callison-Burch and et al., "The META-SHARE Metadata Schema for the Description of Language Resources", LREC, 2008.
[22] Win Pa Pa, Ni Lar Thein, "Myanmar Word Segmentation using Hybrid Approach", In Proceedings of 6th International Conference on Computer Applications, Yangon, Myanmar, 166-170, 2008.
[23] K, Papineni and et al., "A Method for Automatic Evaluation of Machine Translation", Association for Computational Linguistics, 2002.
[24] A. Lavie and A. Agarwal, "An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments", Association for Computational Linguistics, 2007.
[25] G. Doddington, "Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics", Association for Computational Linguistics, 2002.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 3, 84-91 (2024)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Automated Performance analysis E-services by AES-Based Hybrid Cryptosystems with RSA, ElGamal, and ECC

Rebwar Khalid Muhammed[1], Kamaran Hama Ali Faraj[2,3], Jaza Faiq Gul-Mohammed[4], Tara Nawzad Ahmad Al Attar[2], Shaida Jumaah Saydah[5], Dlsoz Abdalkarim Rashid[2]

[1]*Computer Science Institute, Sulaimani Polytechnic University, Sulaimani, Kurdistan Region, Iraq*

[2]*Department of Computer Science, University of Sulaimani, Sulaimani, Kurdistan Region, Iraq*

[3]*Department of Computer, Collage of Engineering and Computer Science, Lebanse Frence University, Erbil, Iraq*

[4]*Department of Computer Engineering University of Sulaimani, Sulaimani, Kurdistan Region, Iraq*

[5]*Ministry of education, Kirkuk Education Department of Kurdish Studies, Hawazen Preparatory School for Girls, Kirkuk, Iraq*

ARTICLE INFO

Article history:
Received: 24 February, 2024
Revised: 17 May, 2024
Accepted: 18 May, 2024
Online: 22 June, 2024

Keywords:
Hybrid AES-RSA
Hybrid AES-ECC
Hybrid AES- ElGamal

ABSTRACT

Recently Network safety has become an important or hot topic in the security society (i.e., Encryption and Decryption) developed as a solution of problem that have an important role in the security of information systems (IS). So protected/secure the shared data and information by many methods that require in all internet faciality, data health and the cloud computing that suggestively increased our data every in milliseconds unit. This performance analysis by two factors namely Encryption, Decryption and throughput time of three Hybrid Encryption schemes namely; Hybrid AES-RSA, Hybrid AES-ECC, and Hybrid AES-ElGamal which are based on Encryption and Decryption times by milliseconds unit in the form of throughput. The results evaluation shows clear distinctions schemes capabilities such as; Encryption and Decryption as well as throughput time consume. Nevertheless, the Hybrid AES-RSA emerges as the fastest types. Both encryption and decryption outcome with superior throughput. Hybrid AES-ECC and Hybrid AES-ElGamal results are slower processing times and making them more suitable for scenarios where performance is not the primary concern. The choice between these schemes should consider not only performance but also security requirements and specific application required for testing and realize to select Hybrid AES-RSA due to better performance in milliseconds. The programing language for proposed system is JAVA, this mean that all testing is by JAVA and discover that the Hybrid AES-RSA is better in performance. The security proposed is Hybrid AES-RSA for automated recruitment system is best.

1. Introduction

The rapidly growth of information networking technology (ICT) is main principals for common culture interchanging of the data very considerably. Since the huge amount of data transferred over the communication facility, data security modified the issue. The security requires in order to guard such data which communicates on unsecure channel [1]. The modification that occurred from past until now is create a generation in the security and changed to automated security [2] or online security. Attributable to the development of ICT to necessity reputation of data has directed to the mentioned secure data in several methods. The Insecure information as considered one of the difficulties at the present time with the development of ICT and the use of Internet networks through data correspondence between various sides such as organizations and users. The term security of information refers to a methodology that employed to satisfy the security requirements [3]. The Encryption method is one of the targets that make users to ensure the secrecy and access of the data to the receiving from side without affecting it from any third party. Also, the algorithms of Cryptography prevent third party or public reading of private messages [4]. The working of Encryption and

* Corresponding Author: Rebwar khalid, Sulaimani Polytechnic University, rebwar.khalid@spu.edu.iq

decryption is shown in Figure 1 [5]. The Figure 1 states that users Encrypts the message using secret key and send it through the communication channels. The Decrypt by Receiver the message using the secret key but Cryptography offers a number of security target to ensure the privacy of data with non-modification of data and so on. the great security advantages of Cryptography that widely in nowadays security [6].

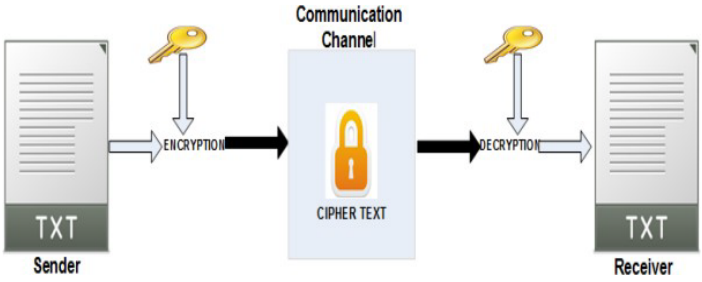


Figure 1: Working of Encryption and Decryption [1]

Cryptography algorithms can be intitled two important categories: Symmetric key Cryptography and Asymmetric key cryptography [7]. Figure 2 shows three types of Cryptography.

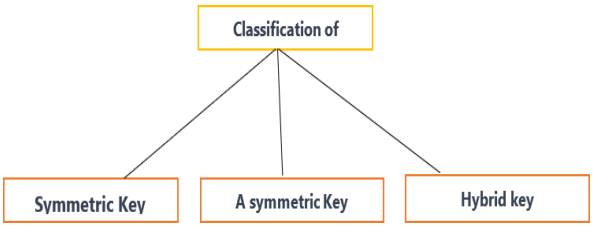


Figure 2: Shows three types of Cryptography [8]

2. Classification of Cryptography

2.1. Symmetric Key Cryptography

Symmetric algorithm is also called shared key Cryptography. During data transmission, the sender and the receiver share the same key for Encryption and Decryption [5]. The different types of Symmetric algorithms as same as to Data Encryption Standard (DES), Triple Data Encryption Standard (3DES), Advanced Encryption Standard (AES) and Blowfish [8]. Hence, AES is a type of symmetric key cryptography and very powerful and more reliable than the other.

2.2. Asymmetric Key Cryptography

Public key Cryptography is another name for the Asymmetric Algorithm with two keys element that namely 'Private key' and 'Public key'. While the data transmission the users encrypts the clear text with the help of public key known as the cipher text and the receiver Decrypts this cipher text with the help of its private key [8]. The different types of asymmetric algorithms are Rivest Shamir Adlemen (RSA), Diffie Hellman (DH) and Digital Signature Algorithm (DSA), ECC [8].

2.3. Hybrid Cryptography

The concept which combines the both shared key cryptography and public key Cryptographic techniques create new version of algorithms and mentioned as Hybrid Cryptography. A Hybrid Cryptosystem is a protocol using Symmetric and Asymmetric Cryptographic technics together, each to its best advantage. Hybrid Encryption is considered a highly secure type of encryption as long as the public and private keys are fully secure. Hybrid Cryptography is achieved through data transfer using unique session keys along with Symmetrical Encryption. public key Encryption is implemented for random Symmetric key Encryption. The recipient then uses the public key Encryption method to Decrypt the Symmetric key. Once the Symmetric key is recovered, it is then used to Decrypt the message [9]. The blended OR combined in between Symmetric and Asymmetric Cryptographic techniques and create Hybrid Cryptography with three key. However, the three key Hybrid must be more secure than the two key Asymmetric and one key Symmetric. The Table 1 below show the comparison between three different algorithms in respect of key type namely; Symmetric Cryptography, Asymmetric Cryptography and Hybrid Cryptography.

Table1: Shows the comparison between three different algorithms

Aspect	**Symmetric Cryptography**	**Asymmetric Cryptography**	**Hybrid Cryptography**
Key Type	Uses a single shared key	Uses a key pair (public and private keys)	Uses both shared and key pairs
Key Distribution	Securely sharing a single key can be challenging	Public keys can be freely distributed, private keys must be kept secret	Securely share a symmetric key, then use asymmetric encryption for key distribution
Speed	Generally, faster for encryption/decryption	Slower compared to symmetric encryption	Slower than symmetric but faster than pure asymmetric
Security	Vulnerable if the key is compromised	More secure due to the key pair, but still vulnerable if private key is compromised	Combines the strengths of both symmetric and asymmetric encryption for improved security

3. Related Work

In [7], a comparative theory of techniques Encryption in terms of (Symmetric and Asymmetric) keys to algorithms analyzed. In the Symmetric key Encryption (AES) algorithm is found to be better in terms of cost, security and implementation.

Nevertheless, the Asymmetric key Encryption (RSA) algorithm is much better in terms of speed and security. Nevertheless, in paper [7] show and discover that the two algorithms namely AES and RSA are good individually in term of the security and cost but didn't mention Hybrid techniques OR combined technique simultaneously. The AES with RSE are combined in our proposed system and came out with better results than before Paper [10] by Verma, P. Guha, and S. Mishra that comparative study of different key algorithms namely: AES, DES, 3DES, Blowfish and RSA are analyzed and compared with the results that found among the symmetric encryption algorithm, AES and Blowfish are the most secure and efficient algorithms. The performance and energy consumption of these algorithms are better compared to the others. In case of Asymmetric Encryption algorithm, RSA is secure and can be used for application in wireless network because of its good speed and security. However, in paper [10] show and compare between: AES, DES, 3DES, Blowfish and RSA algorithms individually but didn't mention Hybrid techniques OR combined technique simultaneously working between mentioned techniques. The AES, DES, 3DES, Blowfish and RSA algorithms are NOT combined and results came out individually.

In paper [11], ECC is explained in detail. Elliptic curve Cryptography (ECC) is a relatively newer form of public key Cryptography that provides more security per bit than other forms of Cryptography still being used today .

In paper [12], Encryption and Decryption of text using ECC is explained with mapping technique. It is concluded that ECC has low power consumption, less memory requirement, small key size and high security.

In paper [13], Hybrid Cryptography technique using AES and ECC is proposed. The system is intended to provide security to a variety of multimedia data ranging from text documents, images, audio, video. Proposed Hybrid system capable of Encrypting and Decrypting the sensitive data to protect it from unauthorized access and attacks.

In paper [14], Hybrid Cryptography approach implemented using AES and ECC. This system provides encryption to the multimedia data such as text, image, audio, video which resulted in an output with 100 percent accuracy without any loss of information.

In paper [15], Different text files are taken as input and encrypted using AES-ECC Hybrid approach. Analysis of AES Encryption with ECC is done on the basis of different parameters like storage requirement, Encryption time, Decryption time.

In paper [16], Hybrid approach for Encryption technique is implemented over a binary image, which provides more accuracy to the encryption process. The ECC and AES are combined in such a way that differentiates them from the usual manner of Encryption. These days with the increasing trend of security it becomes essential to protect the data and information in a better way.

4. Problem Definition

Cryptographic techniques provides the secure data transmission in automated performance analysis of E-services by AES-Based Hybrid Cryptosystems with RSA, ElGamal, and ECC which be a target for our proposed system. The complications existing systems such as a Cryptographic technique with time consume processes. Nevertheless, exactly techniques that isn't includes the reliability checks on transmitted data and another issue in key changing the presence of security. Thus, the implement an operational Cryptographic algorithm in all phases must be well thought-out to sort a strong method (i.e. AES-Based Hybrid Cryptosystems with RSA, ElGamal, and ECC) with complications in to the proposed method. The proposed Hybrid Cryptographic technique which uses the best features of Symmetric AES-Based Hybrid Cryptosystems with RSA, ElGamal, and ECC. The designed technique helps to reduce the time complexity in Encryption-time, Decryption-time and Throughput-time. This paper solves the performance in time consume also select fewer time response by throughput-time. Finally, our proposed algorithms AES-Based Hybrid Cryptosystems with RSA.

5. Description of Algorithms

Advanced Encryption Standard (AES), Rivest Shamir Adleman (RSA), Elliptic Curve Cryptography (ECC), and ElGamal are indeed powerful cryptographic procedures and use for different purposes and have different features. The AES is a Symmetric-key algorithm widely used for secure facts encryption while RSA operates as an Asymmetric-key algorithm for secure communication and digital signatures. ECC, another asymmetric-key algorithm, relies on elliptic curves for enhanced security with shorter key lengths. ElGamal, also asymmetric are utilized for public-key Encryption and digital signatures based on the discrete logarithm problem. In below explain in detail about all types of Algorithms.

5.1. Advanced Encryption Standard (AES)

The AES algorithm are operating as a Symmetric block cipher system that employs a replace or exchange network. The data block length and key length in AES can be adjusted based on specific requirements, with key lengths available in 128, 192, and 256 bits. The iteration cycle numbers for these round keys are 10, 12, and 14 rounds, respectively. The AES algorithm primarily comprises three components: round change, turns, and key expansion. Each round transformation consists of a non-linear layer, a linear mixture layer, and an add round key layer. The

Encryption process of AES is illustrated in Figure 3. Also, there are three different key lengths with round key iteration.

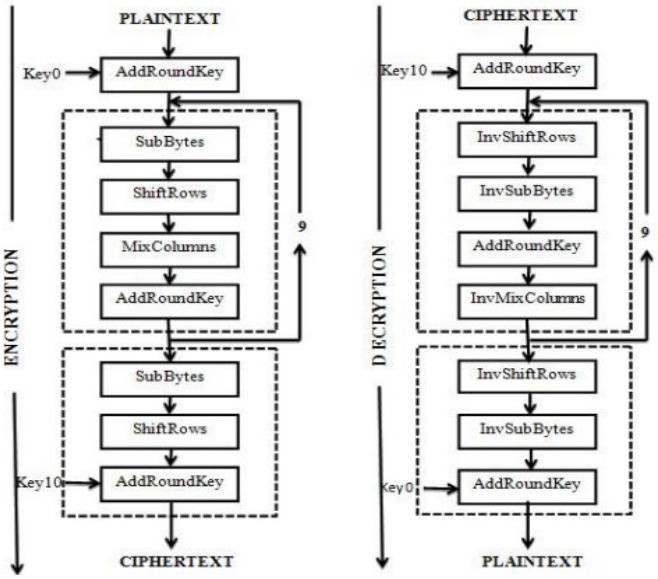


Figure 3: AES Encryption & Decryption process [17]

5.2. *Elliptic Curve Cryptography (ECC)*

In order to compare the public key cryptography that is comparatively more established by an elliptic curve cryptography (ECC) because offers more security per bit than other Cryptographic methods that are still in use nowadays technologies. Nevertheless the Mathematically an elliptic curves are cubic curves that are equivalent to tori topologically. Despite their name which is not closely related to the ellipse. The name of elliptic is integral. By Weierstrass normal equation form. The basic general elliptic curve used for cryptography is of the equation form

$$y2 = x3 + ax + b \quad (1)$$

which showed in Figure 4. The Curves of this form are defined by different values for a and b and modifying these values to visualize of the curve and expand, contract, or pinch off to be two separate pieces. Practically the Curves used for Cryptography to defined very large integer values for aa and bb respectively. The modification of a and b in the mentioned equation is modify the elliptic curve visualization. This mean that is a direct relation between a and b values and elliptic curve.

5.3. *RSA Algorithm*

The invented of the most widely Asymmetric key cryptosystem known as RSA algorithm is very powerful and user-friendly algorithms that Encryption and authentication Cryptosystem used since that time in many Cryptographic applications for instance an e-mail security, banking, e-commerce and digital signature over web operating systems over the internet facility.

The main security algorithm depends on the difficulty of finding prime numbers factor of large integers. RSA operation consists of three main stages key generation, Encryption and Decryption processes which are explained briefly in following [18].

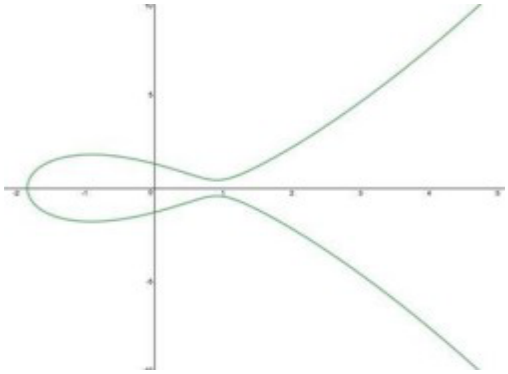


Figure 4: Simple elliptic curve visualization [11]

5.3.1 *Key Generation*

- Select two large prime integer numbers p and q to calculate the modulus n using the formula $n = p \times q$.
- Calculate $\emptyset$ using the formula $\emptyset(n) = (p - 1) \times (q - 1)$.
- Select an integer e which is the public exponent, such that GCD $(e, \emptyset(n)) = 1$,
- where GCD refers to greatest common divisor function between two numbers.
- Calculate d which is the private exponent, such that $e \times d = 1$ mod $\emptyset(n)$, where mod symbolizes to the modulus operation or the reminder after division.

Hence, (n,e) represents the public Encryption key, while (n, d) represents the private decryption key [18]. The two prime numbers are the target of the algorithms.

5.3.2 *Encryption/Decryption Processes*

Let m be a message that wanted to be encrypted, then the encrypted message c is calculated via the public key (n, e) using the equation: $c = m^e \, mod \, n$. To extract the original message m, the received encrypted message c is decrypted via the private key (n, d) using the equation: $m = c^d mod \, n$ [18].

5.4. *ElGamal algorithm*

Security mentioned algorithm depends over hard to calculating discrete logarithms of large prime numbers. If the same plaintext is encrypted using this cryptosystem, then a different cipher text is obtained in each time of Encryption. El-Gamal operation can be described in 3 main steps:

1) key generation, 2) Encryption and 3) Decryption processes which are explained briefly as follows.

5.4.1 Key Generation

- First, select a random prime number p and two other random numbers x and g, such that both of them are less than p.
- Calculate y using the formula: $y = gx \ mod \ p$.
- Thus, (p, g, y) represents the public key which can be shared between a group of users, while x represents the private key which should be kept secret [19].

5.4.2 Encryption/Decryption Processes

In order to Encrypt a message m, firstly, a random integer number k is selected, such that k is relatively prime with $(p - 1)$. Secondly, the cipher text pairs ($c1$, $c2$) is calculated using the equations: $c1 = g^k mod \ p$ and $c2 = (y^k \times m) \ mod \ p$. Finally, the cipher text ($c1$, $c2$) is transmitted to the recipient. To Decrypt the cipher text, pair ($c1$, $c2$), the private key x is employed to recover the original message m using the equation: $m = \frac{c2}{c1^x} \ mod \ p$ [19].

The following pseud codes for ElGamal Algorithms namely Key Generation, Encryption and Decryption. Figure 5: Pseudo-code key generation stage: encryption and decryption.

➢ **Encryption Stage**

ElGamal_Encryption (e_1, e_2, p, P) // P is the plaintext
{
Select a random integer r in the group $\mathbf{G} = < \mathbf{Z}_p^*, \times >$
$C_1 \leftarrow e_1^r \bmod p$
$C_2 \leftarrow (P \times e_2^r) \bmod p$ // C_1 and C_2 are the ciphertexts
return C_1 and C_2
}

➢ **Decryption Stage**

ElGamal_Decryption (d, p, C_1, C_2) // C_1 and C_2 are the ciphertexts
{
$P \leftarrow [C_2 (C_1^d)^{-1}] \bmod p$ // P is the plaintext
return P
}

➢ **Key Generation Stage**

ElGamal_Key_Generation
{
Select a large prime p
Select d to be a member of the group $\mathbf{G} = < \mathbf{Z}_p^*, \times >$ such that $1 \le d \le p - 2$
Select e_1 to be a primitive root in the group $\mathbf{G} = < \mathbf{Z}_p^*, \times >$
$e_2 \leftarrow e_1^d \bmod p$
Public_key $\leftarrow$ (e_1, e_2, p) // To be announced publicly
Private_key $\leftarrow$ d // To be kept secret
return Public_key and Private_key
}

Figure 5: Pseudo-code key generation stage: encryption and decryption

6. Methodology

Methodology of our proposed system is by Hybrid /blended of two algorithms namely: (i.e. AES with RSA) and (i.e. AES with ECC), also (i.e. AES with ElGamal). The enhancement of security performance for developing and comparing an excellent result in milliseconds unit by Java programing language. Hence, there are an indirect relation between the performance-security and time- consume in milliseconds unit. The Figure 6 show indirect relation between performance-security and time-consume. The creation of web Ecommerce and secured by one of the blended algorithms as a methodology and protect the website. As mentioned, that before the protection of website should be by one of the algorithms.

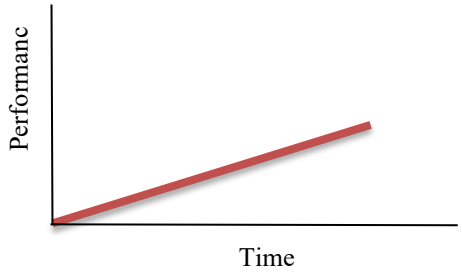


Figure 6: Time

Nevertheless, the performance analysis is an important part of our proposed system by Hybrid techniques between two algorithms and show that the best performance with less time unit and explained in Table below. The Table 2- show all file types with Hybrid algorithms of AES-RSA performance analysis. The main purpose of our proposed system is mixing two algorithms and test the six types of files such as: jpg, txt, pdf, pptx, mp3, mp4 with different file size over Hybrid AES-RSA. The Encryption time and throughput of all types that show it in Table 2 by blended AES and RSA

Table 2: Hybrid AES-RSA Encryption Performance Comparison: Encryption Time and Throughput

File Type	Text size (In Kb)	Encryption Time	Encryption Throughput
JPG	9328	120	71746.15
TXT	3765	55	68436.36
PDF	1136	25	45400.0
PPTX	1743	31	56193.54
MP3	57531	709	81142.45
MP4	4555	60	75900.0
Average Time		166.66	66469.75

The Hybrid techniques between two algorithms and show that the best performance with less time unit and explained in Table below. The Table 3 show all file types with Hybrid algorithms of AES-RSA performance analysis. The main purpose of our proposed system is mixing two algorithms and test the six types of files such as: jpg, txt, pdf, pptx, mp3, mp4 with different file size over Hybrid AES-RSA.

The decryption time and throughput of all types that show it in Table 3 by blended AES and RSA.

Table 3: Hybrid AES-RSA Decryption Performance Comparison: Decryption Time and Throughput

File Type	Text size (In Kb)	Decryption Time	Decryption Throughput
JPG	9328	179	52106.14
TXT	3765	75	50186.66
PDF	1136	50	22700.0
PPTX	1743	47	37063.82
MP3	57531	722	79681.44
MP4	4555	99	46000.0
Average Time		195.33	47956.34

The Hybrid techniques between two algorithms and show that the best performance with less time unit and explained in Table below. The Table 4 show all file types with Hybrid algorithms of AES- ECC performance analysis. The main purpose of our proposed system is mixing two algorithms and test the six types of files such as: jpg, txt, pdf, pptx, mp3, mp4 with different file size over Hybrid AES- ECC.

The Encryption time and throughput of all types that show it in Table 4 by blended AES and ECC.

Table 4: Hybrid AES- ECC Encryption Performance Comparison: Encryption Time and Throughput

File Type	Text size (In Kb)	Encryption Time	Encryption Throughput
JPG	9328	735	12689.79
TXT	3765	608	6190.78
PDF	1136	585	1940.17
PPTX	1743	595	2927.73
MP3	57531	1570	36643.31
MP4	4555	648	7027.77
Average Time		790.166	11236.59

The Hybrid techniques between two algorithms and show that the best performance with less time unit and explained in Table below. The Table 5 show all file types with Hybrid algorithms of AES- ECC performance analysis. The main purpose of our proposed system is mixing two algorithms and test the six types of files such as: jpg, txt, pdf, pptx, mp3, mp4 with different file size over Hybrid AES- ECC.

The Decryption time and throughput of all types that show it in Table 5 by blended AES and ECC

Table 5: Hybrid AES- ECC Decryption Performance Comparison: Decryption Time and Throughput

File Type	Text size (In Kb)	Encryption Time	Encryption Throughput
JPG	9328	186	50145.16
TXT	3765	100	37640.0
PDF	1136	65	17461.53
PPTX	1743	69	25246.37
MP3	57531	802	71733.16
MP4	4555	117	38923.07
Average Time		790.166	11236.59

The Hybrid techniques between two algorithms and show that the best performance with less time unit and explained in Table below. The Table 3 show all file types with Hybrid algorithms of AES-ElGamal performance analysis. The main purpose of our proposed system is mixing two algorithms and test the six types of files such as: jpg, txt, pdf, pptx, mp3, mp4 with different file size over Hybrid AES- ElGamal.

The Encryption time and throughput of all types that show it in Table 6 by blended AES and ElGamal.

Table 6: Hybrid AES- ElGamal Encryption Performance Comparison: Encryption Time and Throughput

File Type	Text size (In Kb)	Encryption Time	Encryption Throughput
JPG	9328	877	10635.11
TXT	3765	788	4776.64
PDF	1136	769	1475.942
PPTX	1743	763	2283.09
MP3	57531	1787	32193.62
MP4	4555	805	5657.14
Average Time		964.83	9503.59

The Hybrid techniques between two algorithms and show that the best performance with less time unit and explained in Table below. The Table 7 show all file types with Hybrid algorithms of AES- ElGamal performance analysis. The main purpose of our proposed system is mixing two algorithms and test the six types of files such as: jpg, txt, pdf, pptx, mp3, mp4 with different file size over Hybrid AES- ElGamal.

The Decryption time and throughput of all types that show it in Table 7 by blended AES and ElGamal.

Table 7: Hybrid AES- ElGamal Decryption Performance Comparison: Decryption Time and Throughput

File Type	Text size (In Kb)	Decryption Time	Decryption Throughput
JPG	9328	300	31090.0
TXT	3765	150	25093.33
PDF	1136	115	9869.56
PPTX	1743	127	13716.53
MP3	57531	1170	49170.94
MP4	4555	174	26172.41
Average Time		339.33	25852.12

In order to achieve more information from previous comparability between three performance which are: Hybrid algorithm of (AES-RSA, AES-ECC, AES- ElGamal). The Table below shows performance analysis for several Hybrid Encryption algorithm, in respect of the average time and throughput. encryption time and Encryption throughput are two parameters that complete and been blended together. In the context of the

presented data information. The three Hybrid /OR blended algorithms are explained below:

- Hybrid AES-RSA: The result blended AES-RSA Encryption algorithms by average time is 166.66 milliseconds. Nevertheless, an encryption time of throughput of blended AES-RSA is 66469.75 milliseconds. The encryption throughput for this Hybrid approach is much higher and more notable. this combination between two algorithms namely Hybrid AES-RSA, thus the Encryption Time and Encryption Throughput is much less than the other blend algorithms (Hybrid AES-ECC and Hybrid AES- ElGamal). However, the Hybrid AES-RSA is outcome is much better and fewer than the others blended algorithms, which are; Hybrid AES-ECC and Hybrid AES- ElGamal. All result is show in Table 8.

Table 8: Average Time Encryption and Throughput of AES-Based Hybrid Cryptosystems with RSA, ElGamal, and ECC in milliseconds

Average Time	Encryption Time	Encryption Throughput
Hybrid AES-RSA	166.66	66469.75
Hybrid AES-ECC	790.166	11236.59
Hybrid AES-ElGamal	964.83	9503.59

In order to achieve more information from previous comparability between three performance which are: Hybrid algorithm of (AES-RSA, AES-ECC, AES- ElGamal). The Table below shows performance analysis for several Hybrid decryption algorithm, in respect of the average time and throughput. decryption time and decryption throughput are two parameters that complete and been blended together. In the context of the presented data information. The three Hybrid /OR blended algorithms are explained below:

- Hybrid AES-RSA: The result blended AES-RSA decryption algorithms by average time is 195.33 milliseconds. Nevertheless, an Encryption time of throughput of blended AES-RSA is 47956.34 milliseconds. The decryption throughput for this Hybrid approach is much higher and more notable. this combination between two algorithms namely Hybrid AES-RSA, thus the Decryption Time and Decryption Throughput is much less than the other blend algorithms (Hybrid AES-ECC and Hybrid AES-ElGamal). However, the Hybrid AES-RSA is outcome is much better and fewer than the others blended algorithms, which are; Hybrid AES-ECC and Hybrid AES- ElGamal. All result is show in Table 9.

Table 9: Average Time Decryption and Throughput of AES-Based Hybrid Cryptosystems with RSA, ElGamal, and ECC in milliseconds

Average Time	Decryption Time	Decryption Throughput
Hybrid AES-RSA	195.33	47956.34
Hybrid AES-ECC	223.16	40191.54
Hybrid AES-ElGamal	339.33	25852.12

However, the two pi-chart below give more information regarding our proposed system average Time Encryption and throughput of AES-Based Hybrid Cryptosystems with RSA, ElGamal, ECC. the Figure 7 show the best performance Hybrid AES-RSA over (Hybrid AES-ECC and Hybrid AES- ElGamal).

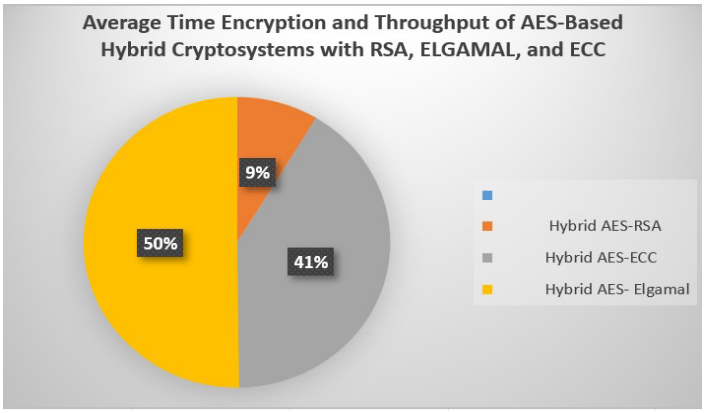


Figure 7: Average Time Encryption and Throughput (Hybrid AES-ECC and Hybrid AES- ElGamal)

The Hybrid AES-RSA Encryption method demonstrates an average Decryption time of 195.33 milliseconds, achieving a throughput of 47956.34 decrypts per second. In comparison, Hybrid AES-ECC exhibits a slightly longer decryption time of 223.16 milliseconds with a throughput of 40191.54 decrypts per second. Hybrid AES- ElGamal, on the other hand, has a higher Decryption time of 339.33 milliseconds and a throughput of 25852.12 Decrypts per second. which Mentioned and explained clearly in Table 9, by of information in Table 9 created a pie-chart Figure 8.

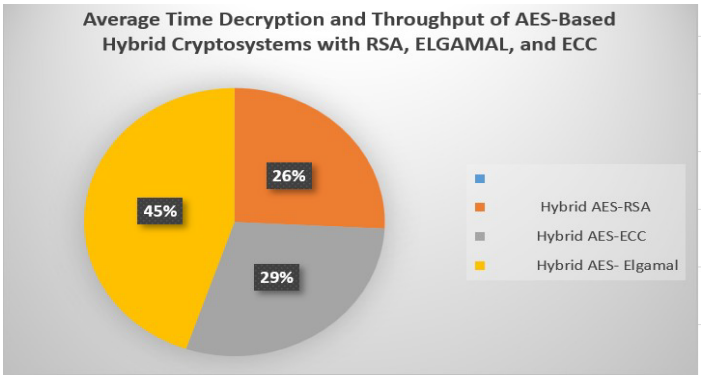


Figure 8: Average Time Decryption and Throughput (Hybrid AES-ECC and Hybrid AES- ElGamal)

7. Conclusion

In conclusion, the data analysis declared that the performance variations among the three Hybrid encryption methods namely

Hybrid AES-RSA, Hybrid AES-ECC, and Hybrid AES- ElGamal. The Hybrid AES-RSA proves superior speed in both schemes' encryption and decryption processes. This mean that making its an optimal choice for applications requiring swift data protection. Furthermore, it achieved the highest throughput for encryption and decryption and underlining its efficiency. Nevertheless, Hybrid AES-ECC and Hybrid AES-ElGamal exhibit slower encryption and decryption time-consume that is why making them less suitable for time-sensitive tasks. Although both are offer certain cryptographic advantages to reduced throughput may limit their applicability.

Eventually, the selection of an encryption scheme should be based on a careful consideration of security requirements and performance requirements. The speed is important for Hybrid AES-RSA stands out as the most favorable option, while Hybrid AES-ECC and Hybrid AES- ElGamal may be more appropriate in situations where performance is less critical, and specific security features are prioritized.

References

[1] P. Patil and R. Bansode, "Performance Evaluation of Hybrid Cryptography Algorithm for Secure Sharing of Text & Images," International Research Journal of Engineering and Technology 2020

[2] K. H. A. Faraj, A. B. Kanbar, J. Gul-Mohammed, W. M. Hmeed, and S. F. Karim, "Cloud Computing Loading Time Over Different Operating Systems," Science Journal of University of Zakho, **8**(4):154–159, 2020, DOI: https://doi.org/10.25271/sjuoz.2020.8.4.756

[3] Z. C. Oleiwi, W. A. Alawsi, W. C. Alisawi, A. S. Alfoudi, and L. H. Alfarhani, "Overview and Performance Analysis of Encryption Algorithms," J. Phys. Conf. Ser., **1664**(1), 2020, DOI: https://doi.org/10.1088/1742-6596/1664/1/012051

[4] P. Chinnasamy, S. Padmavathi, R. Swathy, and S. Rakesh, "Efficient Data Security Using Hybrid Cryptography on Cloud Computing," Lect. Notes Networks Syst., **145**(September):537–547, 2021, DOI: https://doi.org/10.1007/978-981-15-7345-3_46

[5] F. Maqsood, M. Ahmed, M. M. Ali, and M. A. Shah, "Cryptography: A Comparative Analysis for Modern Techniques," Int. J. Adv. Comput. Sci. Appl., **8**(6):442-448, 2017, DOI: https://doi.org/10.14569/IJACSA.2017.080659

[6] P. Verma, J. Shekhar, P. Preety, and A. Asthana, "A Survey for Performance Analysis Various Cryptography Techniques Digital Contents," International Journal of Computer Science and Mobile Computing, **4**(1):522–531, 2015

[7] N. Bisht and S. Singh, "A Comparative Study of Some Symmetric and Asymmetric Key Cryptography Algorithms," International Journal of Innovative Research in Science, Engineering and Technology, **4**(3):1028-1031, 2015, DOI: https://doi.org/10.15680/IJIRSET.2015.0403043

[8] S. Chandra and S. Paira, "A Comparative Survey of Symmetric and Asymmetric Key Cryptography," in 2014 International Conference on Electronics, Communication and Computational Engineering (ICECCE), 2014, 1-4, DOI: https://doi.org/10.1109/ICECCE.2014.7086640

[9] P. Kuppuswamy and S. Q. Y. Al-Khalidi, "Hybrid Encryption/Decryption Technique Using New Public Key and Symmetric Key Algorithm," Manage. Inf. Syst. Rev., **19**(2):1-13, 2014, DOI: https://doi.org/10.6131/MISR.2014.1902.01

[10] P. Verma, P. Guha, and S. Mishra, "Comparative Study of Different Cryptographic Algorithms," Int. J. Emerg. Trends Technol. Comput. Sci., **5**(2):58-63, 2016

[11] R. Harkanson and Y. Kim, "Applications of Elliptic Curve Cryptography: A Light Introduction to Elliptic Curves and a Survey of Their Applications," in 12th Annual Cyber and Information Security Research Conference, USA, 2017, 1-7

[12] K. Keerthi and B. Surendiran, "Elliptic Curve Cryptography for Secured Text Encryption," in International Conference on Circuits Power and Computing Technologies, India, 2017

[13] S. C. Iyer, R. R. Sedamkar, and S. Gupta, "A Novel Idea on Multimedia Encryption Using Hybrid Crypto Approach," in 7th International Conference on Communication, Computing and Virtualization, 2016, **79**:293-298

[14] S. C. Iyer, R. R. Sedamkar, and S. Gupta, "An Efficient Multimedia Encryption Using Hybrid Crypto Approaches," Int. J. Recent Trends Eng. Res., **2**:442-452, 2016

[15] S. Sharma and V. Chopra, "Analysis of AES Encryption with ECC," in Proceedings of International Interdisciplinary Conference on Engineering Science & Management, 2016, 195

[16] D. Ameta and S. Upadhyay, "A Hybrid Approach for Image Encryption Using Different Number Iterations in ECC and AES Techniques," Int. J. Comput. Appl., **175**(3), 2017, DOI: https://doi.org/10.5120/ijca2017915469

[17] N. Mathura and R. Bansode, "AES Based Text Encryption Using 12 Rounds with Dynamic Key Selection," in Proc. 7th Int. Conf. Commun., Comput., Virtualization, 2016, 131-135, DOI: https://doi.org/10.1016/j.procs.2016.03.131

[18] F. Yousif, "Encryption and Decryption of Audio Signal Based on RSA Algorithm," Int. J. Eng. Technol. Manage. Res., **5**(7):259-264, 2018, DOI: https://doi.org/10.29121/ijetmr.v5.i7.2018.259

[19] O. A. Imran, S. F. Yousif, I. S. Hameed, W. N. Al-Din Abed, and A. T. Hammid, "Implementation of El-Gamal Algorithm for Speech Signals Encryption and Decryption," Procedia Comput. Sci., **167**:1028–1037, 2020, DOI: https://doi.org/10.1016/j.procs.2020.03.402

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 1, 91-103 (2024)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Development and Usability Evaluation of Mobile Augmented Reality Contents for Railway Vehicle Maintenance Training: Air Compressor Case

Gil Hyun Kang[1], Hwi Jin Kwon[2], In Soo Chung[1], Chul Su Kim[*1]

[1]*Industry-Academic Cooperation Foundation, Korea National University of Transportation Uiwang, 16101, Korea*

[2]*Department of Railroad Convergence System Engineering, Korea National University of Transportation, Uiwang,16101, Korea*

ARTICLE INFO

Article history:
Received: 06 December, 2023
Revised: 11 January, 2024
Accepted: 11 January, 2024
Online: 20 January, 2024

Keywords:
Augmented reality
Railway vehicle maintenance
Air compressors
Usability evaluation
Mobile devices

ABSTRACT

The air compressor of a railroad vehicle is an important equipment that produces compressed air used in braking systems. New visual interaction techniques were proposed and evaluated to develop effective augmented reality content for maintenance support and training of this device. To this end, modeling techniques capable of fast animation, storyboard production to support light maintenance, and visualization algorithms that implement fluid flow have been developed. In the case of air compressors using compressed air, 2D-3D line matching distance calculation algorithm and modified flow generation algorithm were proposed to implement fluid flow for visualization. If two algorithms are used at the same time, the educational effect can be enhanced by visualizing the air flow in the 3D object simultaneously on a 2D air pipe diagram. In addition, the use of flow generation algorithm alone can visualize fluid flow or quantity control, such as air discharge or lubricant supplementation. As a result of usability evaluation of 100 users of the developed air compressor augmented reality content, the system usability scale score was 76.65, which was good. Similarly, the user experience score using six questionnaires was very good with an average of 4.12. Therefore, it was found that this training content could be used very effectively for air compressor maintenance support and training at the site.

1. Introduction

Since railway vehicles are a means of mass transportation, it is important to perform reliable maintenance to avoid causing traffic congestion to many passengers if a failure occurs during train operation. Therefore, vehicle failures have been prevented through periodic preventive maintenance, and the lifespan of important parts has been managed and maintenance cycles have been extended through Reliability-Centered Maintenance (RCM) based on failure statistics analysis [1]. Modern railway vehicles have developed into a complex system in which various technologies are fused with the development of electronic and control technologies. In order to fully carry out the maintenance of railway vehicles with such modern advanced technology, not only the improvement of maintenance technology but also the educational techniques of maintenance personnel need to be innovated. Recently, with the development of sensor technology and communication technology and Artificial Intelligence (AI) technology, the maintenance of railway vehicles has rapidly developed into predictive maintenance by Condition-Based Maintenance (CBM) and smart maintenance [2]-[4].

Most railway vehicles remained at the level where vehicle computers installed in intermediate vehicles and driver cabins, but event recording and information on major devices during train operation were displayed with driver front monitors and vehicle computers. Recently, however, more than 2,000 sensors have been installed and monitored to switch to a level that utilizes condition monitoring and AI technology during train operation to perform failure diagnosis as well as failure prediction and residual life prediction [5,6]. The core of this change is data-based analysis technology according to the development of computer and communication technology. Since these technologies essentially utilize AI and Internet of Things (IoT) technology using ground and onboard servers, it is expected that there will be a major change in the maintenance of the conventional human-dependent experience base. The Korean railway industry is also conducting a research project for smart maintenance of railway vehicles to overcome these technological changes as follows: (1) Predictive maintenance based on CBM by increasing the capacity

[*]Corresponding Author: Chul-Su Kim, School of Railroad Engineering, Korea National University of Transportation, Uiwang, 16101, chalskim@ut.ac.kr

https://dx.doi.org/10.25046/aj090109

of new sensors and data communication, (2) Automation of the acquisition of railway vehicle light maintenance data when a vehicle in operation through a large scale vehicle scanner installed on the ground passes through a specific point such as a deposit entrance or terminal station, (3) Digitization and automation of inspection and repair records for digitization of maintenance data and connection with mobile devices, and (4) Augmented Reality(AR)-based maintenance training to replace existing computer-aided electronic manuals and mobile device-linked maintenance education contents. Research to apply AR techniques in the industry was initially a study related to the development and production of automobiles and aircraft, which was intended to be used for crash analysis, assembly optimization of batch products, and assembly production technology of complex systems [7,8]. In the automobile industry, driving support using head-up displays, assembly, design, assembly, maintenance, and manufacturing were reported as examples in the fields of in-car systems, design, maintenance, and manufacturing, such as marker-based or markerless graphic modeling using head mounted displays and AR glasses [9]-[12].

Research which applies AR technology to maintenance is also gradually expanding in the field of automobiles, starting with the maintenance of specific devices of aircraft and ships with high operational safety priorities, and railway vehicles [13]-[15]. These technological advances can be confirmed through various research survey papers [16,17].

For more than a decade, the railway sector has tried to apply AR technology to the maintenance and repair of railway vehicles using mobile devices, but it has remained at a very slow development due to the complexity of railway vehicles and limitations of IT technology [18]. The main reason for this was that the method of training many people at the same time in theoretical education and vehicles without a special education system was simple and the cost of education was low. However, this field practice is suitable for disassembling and assembling at the level of looking at and explaining the physical appearance, but there was a clear limit to education to understand the specific operating principles of complex equipment consisting of many wiring pipe checkpoints and to train complete maintenance. In addition, there is a problem that on-site classes have a high risk of safety-related accidents. Therefore, research on alternative education using video or virtual reality has been continuously conducted because the actual effect of education and training cannot be high except for education and training performed with expensive practical equipment. Recently, AR technology has also developed rapidly, such as virtual simulation of automobiles, visual design, virtual protection, and AR self-serving to reduce development costs, with the development of high-performance CPUs and graphic cards, 3D scanners, and the development of the game industry where animation technology is important [19,20]. In addition, active research is being conducted by expanding not only building infrastructure but also various manufacturing industries [21,22]. Nevertheless, the development of AR technology in the maintenance field of railway vehicles led by railway operators was very limited.

Recently, most of the AR application cases for maintenance in the railway vehicle sector have been developed mainly on main mechanical equipment such as boogie and braking devices conducted by the research group of the National University of Transportation in Korea [23,24]. In the case of the air compressor conducted in this study, only some applications have been reported to manufacture industrial screw-type air compressors [25]. Therefore, this study aims to make it easy to use for train fleet maintenance by gradually converting important railway vehicle equipment, a collection of complex technologies such as mechanical engineering, electrical, communication, and electronic engineering, over several years to innovate existing manual-based maintenance methods in the railway industry.

Unlike previous studies such as object search and maintenance point search using head-ware devices and cameras, which are many AR research fields, this study aims to create a 3D AR model and connect this model with a database related to maintenance and a storyboard to enable reliable maintenance support and to effectively perform maintenance training with realistic content in 3D space with touch function on mobile devices [26]-[28]. This study aims to achieve the research objectives for supporting on-site maintenance of major railway vehicle equipment and effective education and training at low cost as follows: (1) Development of mobile-based maintenance support and educational training contents using AR technology, replacing the existing training methods of limited and inefficient document-based theory education and field practice, virtual reality video or Internet-based e-learning, (2) The development of contents that increase work efficiency by raising detailed information of major equipment and sub-components to the same level as actual field work. Therefore, the sub-component system is based on the part procurement BOM of the operator. Development of contents connected to tools for disassembly and assembly of parts, measuring devices or instruments for inspection, special equipment, and work procedures based on the maintenance manual, (3) Development of algorithms that enhance learning effect by implementing fluid or electric flow simultaneously with 3D AR models for operating principles and functions of equipment learned in 2D space from piping, wiring, and circuit diagrams to enhance maintenance support and education and training effect, (4) Analyze maintenance tasks to extend the user interface to inspect, measure, repair, and, Expanding practical education and training areas to provide a sense of immersion through physical equipment and interfaces, such as similar tools, etc., and (5) It increases the effectiveness of safety work through warnings related to safety work based on the operator's disaster database, such as serious injuries or deaths that may occur at the maintenance site.

The usability of AR contents developed for railway vehicle air compressor maintenance training was evaluated in two ways. The results of the System Usability Scale (SUS) survey and the User Experience (UX) experiment of 100 college students who are beginners in railway vehicle maintenance were as follows. The average SUS score of AR contents was 76.65, which was "Good." In addition, the results of the UX survey shows that the average usability score for six questionnaires was excellent at 4.12 out of 5. Therefore, AR content developed for Android-based mobile devices is expected to be conveniently used for maintenance training on a platform for maintenance efficiency for beginners in railway vehicle maintenance.

2. AR Contents and Developed Algorithm

2.1. Smart Maintenance Platform for Railway Vehicle

The Smart Maintenance Research Consortium Group, which includes the Korea Railroad Corporation, the Korea National University of Transportation, Hyundai Rotem, a railroad vehicle manufacturer, Seoul Metro, and AI software development companies, has been carrying out the railway vehicle smart maintenance research project with the support of the Korean government since 2019. Figure 1 shows the architecture of the final stage, the smart maintenance platform. The basic strategy is to install sensors on the main components of Electric Multiple Unit (EMU) and implement CBM and data analysis-based predictive maintenance in combination with AI. In addition, to improve maintenance efficiency, a maintenance decision-making system is established to support maintenance decisions in conjunction with CBM data by automating inspection and measurement and digitizing data. Finally, it is to establish a Mobile Maintenance Efficiency System (MMES) that promotes new innovation in automating maintenance records to support the implementation of new maintenance strategies and training employees using AR in mobile systems. This study introduces the development of AR contents that replace or support existing field training along with this MMES AR manual, and specifically relates to the process of developing AR contents for air compressors and the result of the user evaluation.

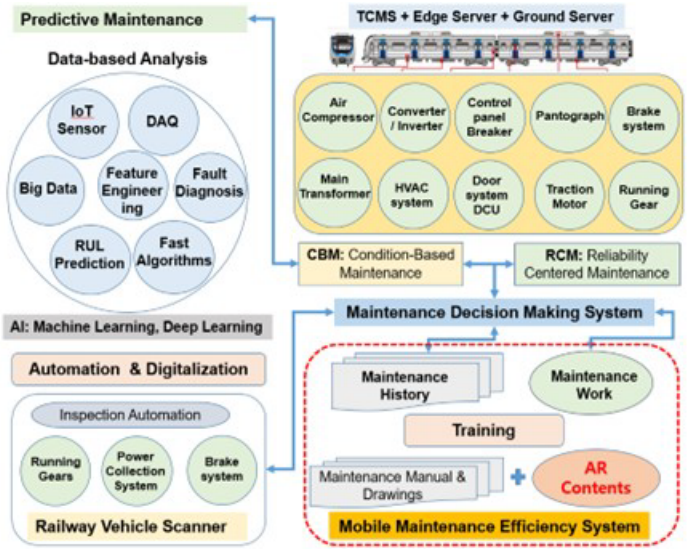


Figure 1: Architecture of Smart-maintenance platform for Railway vehicle

Train configurations vary from country to country due to population density, urban traffic characteristics, and traffic congestion. In the case of urban railroads in Seoul, a metropolitan city, 3 to 4 cars are built into one unit, and two units of 6 cars and 3 units of 10 car trainset is generally operated. Most of the major trunk lines are mass commuter transported by EMUs composed of 10 car trainset. The type of vehicle is also divided into a motor car with a propulsion device, a trailer car without power, a trailer car with cabin, and a motor car with cabin if there is a driver's room. However, most of the main equipment is installed in one unit, and when the number of unit increases, the device overlaps, but since it is the same equipment, AR contents can be manufactured for each major equipment. Most urban rail vehicles are direct current EMUs using DC 1500V, while long-distance trains operated to the suburban lines are AC EMUs using 25kV AC [29,30]. In addition, in the case of AC-DC multiple EMU, a converter is added to the propulsion system, including the main transformer, making the vehicle complicated by installing ultra-high voltage power devices and switches. Table 1 shows the AR manuals and contents of major equipment for EMU developed as a Korean government-funded research project until recently.

Table 1: Developed AR manual and contents

Equipment	EMU	AR manual	AR training
Pantograph	AC/DC	○	○
Main transformer	AC	○	
Air compressor	common	○	○
Propulsion system	AC/DC	○	
Traction motor	common	○	
Auxiliary power	common	○	
Battery box	common	○	
Bogie	common	○	
Wheelset	common	○	
Axle box	common		○
Access door	common	○	○
HVAC	common	○	
Brake operating unit	common	○	○
Wheel brake system	common		○
Disk brake system	common		○
Air spring	common		○
Front connector	common		○
Inter connector	common		○

The reason why AR manual and AR contents differ on the MMES platform is that different small groups have different development objectives and different intellectual property owners. Therefore, some equipment is developed and compared with each other, and some equipment has been developed independently, but plans to optimize in the future through upgrades.

2.2. AR contents for Mobile Maintenance Efficiency System

Figure 2 shows the MMES installed on the mobile phone and the drive screen. The left side shows the app screen installed on the mobile phone and the screen consisting of maintenance work, maintenance manual, maintenance record, and AR content in the center. The picture on the right shows the developed AR manual screen by pushing the menu. The MMES platform was developed for the main purpose of digitizing maintenance records for smart maintenance, and when users perform maintenance and move to designated routes, and store maintenance history on digital workbooks according to maintenance guidelines, data is transmitted to ground servers, and AR manuals were produced to support maintenance in this process.

Figure 3 shows the screen of AR contents developed by the KNUT research team for educational and training purposes. It can be used by using a separate AR contents app on the upper left side of the left screen of Figure 2.

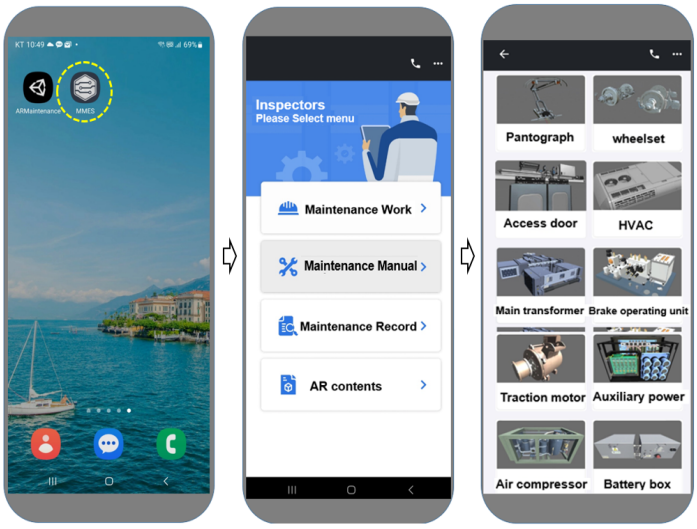


Figure 2: The AR manual screen for MMES

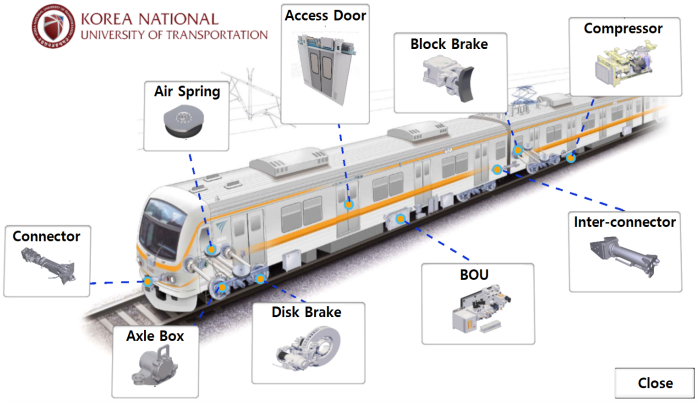


Figure 3: The AR training APP screen for MMES

3. Development of AR contents for Air Compressor

The air compressor of a railway vehicle is one of the essential equipment for starting, operating, and stopping trains. When starting an electric railway vehicle, compressed air is used to raise the current collecting pantograph to supply high-voltage electricity by contacting the power wire. During train operation, air is supplied to the braking device to stop the train at the station, and compressed air is supplied to the door system to enable passengers to get on and off the train. In the case of the 10-car electric vehicle of the Korea Urban Railroad, three sets are installed on the lower floor of the body, as shown in Figure 4. At least one air compressor is required for each train configuration unit. Modern air compressors are mainly reciprocated or screw types. In the operation, the motor is rotated to compress atmospheric pressure air to high pressure, store compressed air in the main air reservoir, and use compressed air to automate the brake system and opening and closing doors of the train. In the case of long-distance passenger trains, it is also used to supply water to the toilet or to provide water for handwashing. Thus, as with braking systems, air compressor maintenance training is important.



Figure 4: Air compressor of the Electric multiple unit

3.1. AR Contents Development Process

To produce AR content for maintenance, it is necessary to select and model equipment for content and organize and design content. Figure 5 shows the planning and design procedure of AR content.

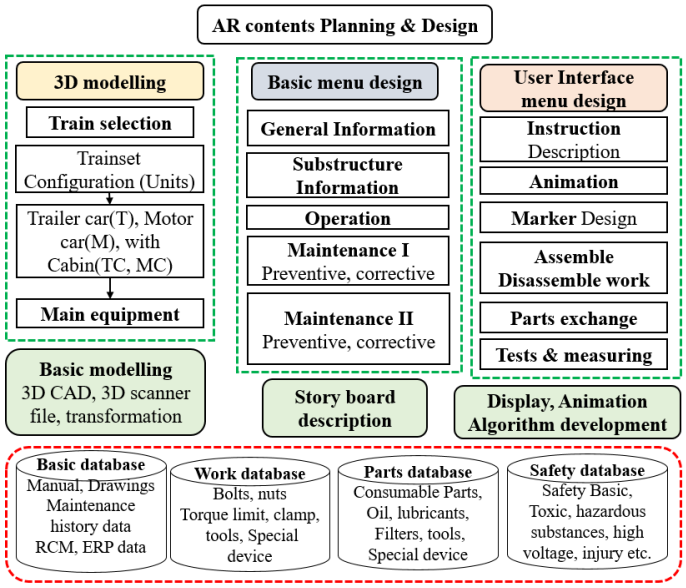
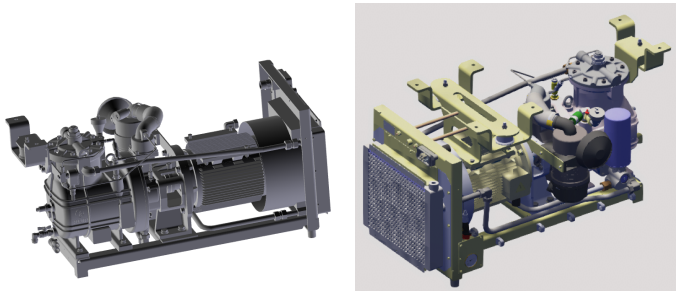


Figure 5: Procedures of AR contents planning and design

Modeling of an equipment assembly of a railway vehicle takes time and cost to model because the assembly is relatively large size and the number of parts is large, and there are typical methods as follows. (1) Using the parts manufacturer's 3D drawings to convert to a format for an AR production program: It uses 3D

CAD files, which is the most convenient method. (2) 3D scanning and measurement methods: 3D scanning parts are modeled by using scanners, and the rest are modeled by making cad drawings. (3) Model with similar dimensions: less precision, but with reference to drawings and real objects modeled with similar dimensions.

Figure 6 shows a 3D modeling technique that uses the Geom etric Design X and Unity program to convert air compressor larg e CAD raw files into FBX to read reduced capacity files and con vert them into APK format models, small files that run fast on m obile devices [31]. Depending on the type of equipment, 3D CA D format file formats are mostly large in capacity and cannot be used directly on mobile devices. It is also important to use a 3D s canner to create an initial device model file, since 3D CAD files are the most accurate models.

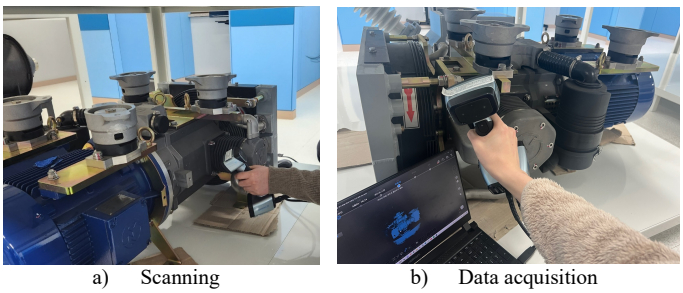


a) Compressor CAD file b) Unity model

c) Bogie 3D scanner file d) Unity model

Figure 6: 3D modelling for AR contents

Figure 7 shows an example of modeling an air compressor by acquiring external coordinate data using a 3D scanner. It is essential to read and check it in the CAD program, then convert it to FBX format so that Unity programs can import and work on AR content.



a) Scanning b) Data acquisition

Figure 7: Process of using a 3D scanner

Table 2 shows a comparison of file size of among the air compressor, a railway vehicle bogie and brake operation unit. The bogie case is a small capacity CAD-compatible file STP format using a 3D scanner, despite being a large structure. As shown in Table 2, 3D scanner files are only 6.8 M, which is advantageous for fast animation. Of course, the CAD program SolidWorks drawing file increases the file size to about 168MB only for the bogie frame.

Table 2: 3D modelling and file size

Equipment	cad file/3D scanner file size	Unity import file size (fbx)
Compressor	288MB (3D cad file)	68MB (77% reduced)
Bogie	168BM sdlprt file (frame only) 6,497KB stp file/3D scanner	9,427KB
BOU	115MB (3D cad file)	22MB (81% reduced)

Therefore, for single structures that do not require disassemble, the dimensional information file of the 3D scanner is very efficient in creating AR models. This seems to be due to the fact that the welding structure of the truck frame is perceived as a part on the scanner, unlike the 3D CAD file, whereas the air compressor 3D CAD file consists of many subsystems, and the complexity of the structure increases the size of the CAD file.

3.2. Creating Storyboard for Maintenance

Storyboards are an important process of determining the overall structure of content and determining the hierarchy of objects to be implemented as unity, and determining the validity of training to determine information to be delivered to users. (1) Basic menus such as device general information, level determination of sub-components of the device, and maintenance classification are defined through the content major classification of the selected equipment assembly. (2) In order to avoid unity hierarchy complexity in content production, sub-menus are defined by integrating the middle and sub-classification levels. Design various levels of administration that describe objects and other functions in drawings of device sub-parts at the level required by the manual or implement the operation. At this stage, various scripts are prepared in conjunction with databases related to parts and sub-parts. (3) Maintenance defines the scope of development of the AR manual implemented by the operating company and MMES, and various scenarios such as assembly/disassembly, measurement and inspection taking into account other tools, equipment, and special instruments in the degree of exchange repair, and the contents related to training, maintenance support, and safety work.

In this way, storyboard creation determines the quality and system efficiency of the entire content, so it should be written in collaboration with maintenance experts to understand the level that can be implemented with development programs or algorithms in a systematic approach and implementation virtual space, and to have a high training effect. Figure 8 shows basic architecture of storyboard of air compressor.

Table 3 shows that classifies detailed tasks that are frequently repeated in railway vehicle maintenance work by category, and classifies training purposes by type and methods implemented as

AR contents. A storyboard should be created by focusing on such work content and reviewing the maintenance manual database and maintenance regulations.

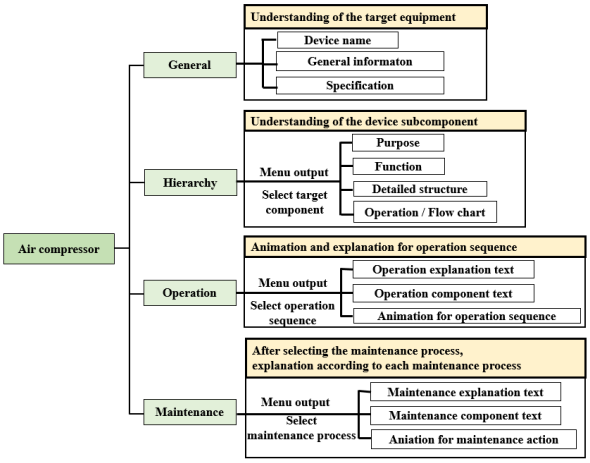
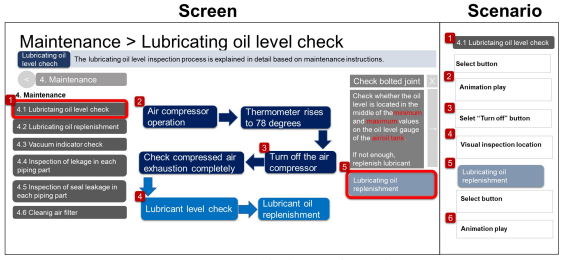
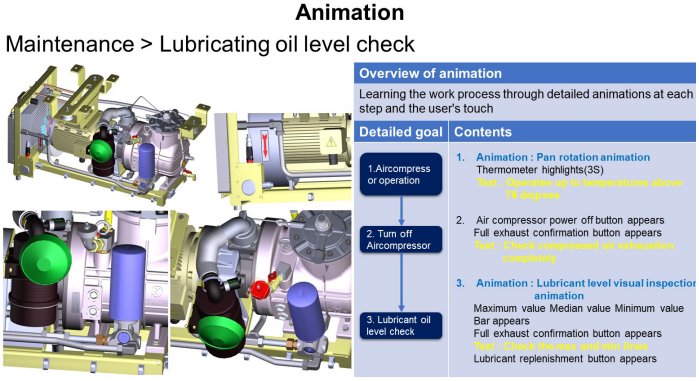


Figure 8: Storyboard creating for AR contents

Figure 9 is an example of a storyboard created during the development process. The left part of Figure 9(a) shows the screen configuration interface and text, sand the right part shows the content scenario that the user needs to learn. Figure 9(b) is an example of a visualized storyboard, and the progress of the animation, the setting by time, and the additional functions that appear on the screen are written in detail. The detailed storyboard is continuously updated and managed if a change is needed.



a) Detailed storyborad



b) Animation storyborad

Figure 9: Example of storyboard

3.3. Algorithm Development for Flow Expression

The main algorithm used in the air compressor training conte nt is as follows. Algorithm 1 is the 2D-3D Line Matching Distance Calculation (2D-3D LMDC) algorithm that calculates basic coor dinates and travel distances to implement electricity or fluid flow at the same time. The original purpose of this algorithm was dev eloped as a streamline matching variable calibration algorithm to make it easy to understand education and training of complex devices including circuit diagrams or pipes by displaying many 2D drawings widely used in railway vehicles in 3D space from real devices.

Table 3: Maintenance work definition for AR contents

Work	Detail task	Training target	Implementation
Assembly disassembly	Bolt, nut fastening loosening remove, insert pins seal	How to use tools, wrench and jig torque measurement assembling order procedure	Animation, scripts, button, movable window, Interaction algorithm requires
	Piping, wiring disconnecting	Connector fastener, How to use special tools, hazard understanding	Animation, Scripts, movable window
Cleaning washing Painting, oiling	Air filter, oil filter, case cleaning, painting, lubricating	How to clean the polluted filter or contam inated parts, How to select proper lubric ant and control quantity etc. safety from electric shock or injury	Select the appropriate range of contents based on manuals due to the diversity of work
Repair exchange	PCB, switch, breaker, relay, brake shoes, consumable part	Remove dust, how to repair relay or swit ch contact how to exchange consumable parts	Select the appropriate range of contents based on manuals due to the diversity of work
Inspection Measurement	Gauges, display panel, warni ng lamp, gap or clearance Insulation resistance	How to read temperature, pressure gaug es, how to measure voltage, current and to use tester, clearance measurement	Animation, Interaction algorithm requires
Drawings	Piping, wiring, electric and electronic circuit	Understanding piping, wiring and circuit diagrams	Interaction algorithm requires
Treble shooting	Device function, operation manipulation of switch and reset switch	Understanding function and operation manipulation of the parts. the reason of the failures and faults of the equipment	Animation, scripts, markers, Interaction algorithm requires

When used with Algorithm 2, Flow Generation (FG) algorith m, this algorithm can simultaneously display the electrical wirin g diagram of the control circuit and the display of electrical flow from the actual wire, or the 2D air flow diagram of the brake syst em and the flow of compressed air or control air from the actual 3D pipe. An animation duration is obtained by calculating the dis tance from spatial coordinates with different dimensions and calc ulating the position that can display the same flow in space at the same time. Figure 10 shows the same flow matched to each othe r in the 2D drawing and 3D model. Because it passes through the same location value, it is easy for users to learn by comparing ea ch other.

Algorithm 1: Code of the 2D-3D Line Matching Distance Calculation algorithm

Result: LD_i = Animation Duration $\cdot \frac{Distance_i}{TotalDistance}$}

Animation Duration ***AD***;

Vector3 Position List ***P***[$P_{1(x,y,z)}, P_2 \ldots P_x$];

Produce List of Distance ***D***;

for (i = 0, $P-1$, i++)

//Calculate Distance and $\boldsymbol{D_1(P_i - P_{i+1})} \ldots \boldsymbol{D_i}$

D_i=Distance(P_i, P_{i+1}))=

$\sqrt{(P_{i+1x} - P_{ix})^2 + (P_{i+1y} - P_{iy})^2 + (P_{i+1z} - P_{iz})^2}$

$LD_i = AD \cdot \frac{D_i}{TotalDistance}$

end

Algorithm 2: Code of the Flow Generation algorithm

Result: Particle System Quantity-Control

Object **Prefab;**

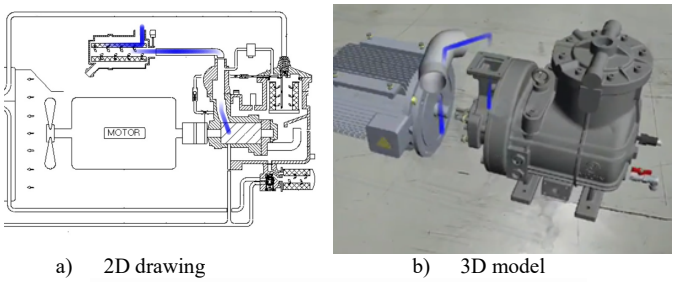
float **Qcontrol;** // Quantity-Control

for (i = 0, $P-1$, i++)

Instantiate Object **Prefab** As **a** on $\boldsymbol{P}_i$

Rotate **a** to look $\boldsymbol{P}_{i+1}$

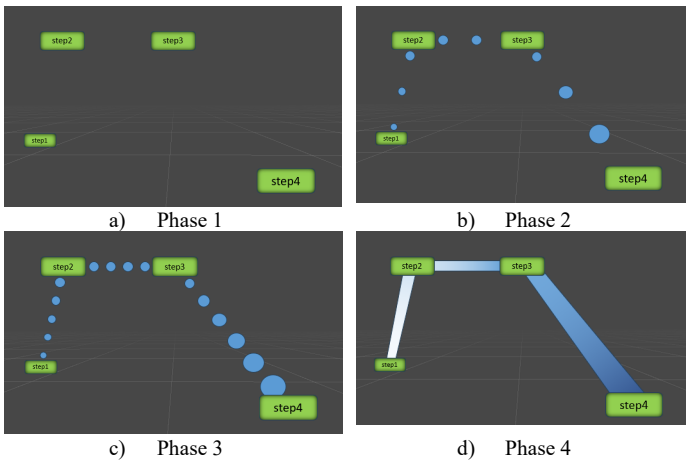
Set a.Particle System. Lifetime, Speed, Qcontrol, Emission.rate Over Time
Lifetime = L_i
Speed = $\frac{D_i}{L_i}$
Emission Rate over Time = Start Speed
Qcontrol += **Qcontrol**
end



a) 2D drawing b) 3D model

Figure 10: Flow displayed in 2D drawings and 3D models

The FG algorithm is an algorithm that makes it easy to express the flow of fluid in AR. This algorithm was originally developed to support the LMDC algorithm, allowing electricity or fluid to display the flow on the drawing by continuously flowing objects represented by dots. Since then, in the development of air compressor AR content, it has been expanded to display oil or air exhaust at the valve outlet. This called Prefab of Unity to enable the repetitive use of objects with specific attributes [32]. In order to express the pneumatic flow in 3D at the time of need, it is a method of instantiating the particle prefab with the properties of the particle system input. The module of the particle system was automatically modified once each attribute was entered to avoid the complexity of using a variable containing 24 properties that affect the entire system. When entered into the modified particle prefab, the prefab is instantiated, releasing the particles continuously, which has the effect of displaying the flow. It can also show arbitrary discharged airflow using random numbers and particles emitted by property changes. Figure 11 shows the detail process of the FG algorithm. It is released by each phase that increases the particles, giving a visual effect of creating a moving flow of electricity or compressed air.



a) Phase 1 b) Phase 2

c) Phase 3 d) Phase 4

Figure 11: The process of generating a flow as particles increase using the FG algorithm

Figure 12 shows a flow chart of the algorithms that combine two developed algorithms to display electricity or fluid flows simultaneously in a 2D drawing and a real 3D object on one screen, or to control the flow of fluid using a single fluid flow generation algorithm to generate a screen such as air discharge.

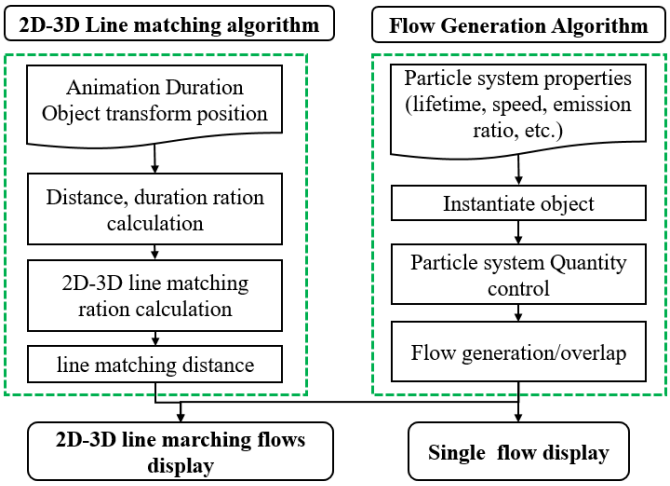


Figure 12: Flow chart of the algorithms for AR contents

Figure 13 shows the exhaust air at the outlet of the air compressor in a three-dimensional space using the FG algorithm alone. In this way, by showing the flow of fluid in various ways through Q-control, various functions of maintenance such as pressure control and exhaust air control can be actively shown to the trainee, thereby enhancing the training effect.

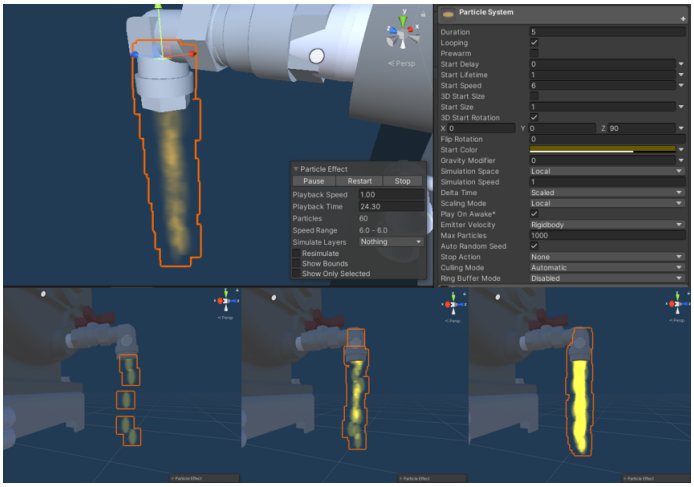


Figure 13: Fluid flow expression with quantity control for AR contents

3.4. Development of Menus for User Interaction

When the storyboard is completed, it is necessary to develop user menus to implement AR content on mobile device screens. In particular, menu arrangements that require using a limited narrow space were allocated in consideration of space use so that other menus were not pushed.

Figure 14 shows an example of menu configuration in a User Interface (UI) of air compressor AR content. The menu used for the content is located on the left and right sides of the screen so

that the device is displayed in the center. The menu consists of three types.

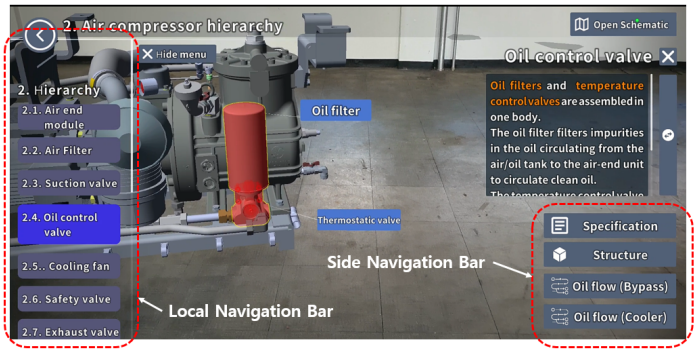
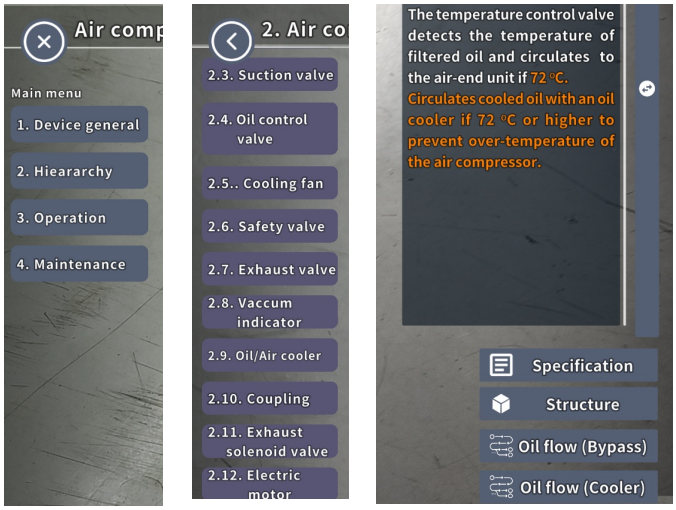


Figure 14: Menu design for sub-parts of air compressor

Figure 15 shows a detailed example of the three types of menus. The Global Navigation Bar (GNB) on the left, as shown in (a), is the main menu. This is a large category menu in GNB format, consisting of General, Hierarchy, Operation, and Maintenance, which are the highest levels of keywords for the maintenance process. When one of the GNB menus is touched, the Local Navigation Bar (LNB), a sub-menu of the GNB, appears, as shown in (b). Each LNB menu consists of maintenance process and necessary related information. The LNB is placed on the left side of the screen, and at the same time, a Side Navigation Bar (SNB) as shown in (c) is displayed on the right side. SNB is displayed with detailed information such as text and photos in a separate pop-up window [33]. In addition to this, there is "Hide menu" and "Open Scheme".



a) GNB b) LNB c) SNB

Figure 15: Detailed examples of menu types

In the early development version, GNB and LNB were placed on the left and right sides, but in the upgraded version, when the GNB is touched, the GNB menu disappears and the screen space can be used widely with a screen configuration replaced by LNB.

Figure 16 shows the enlarged screen space by moving the original right LNB to the left. Therefore, even a mobile phone with a small screen can use the screen widely for device screen.

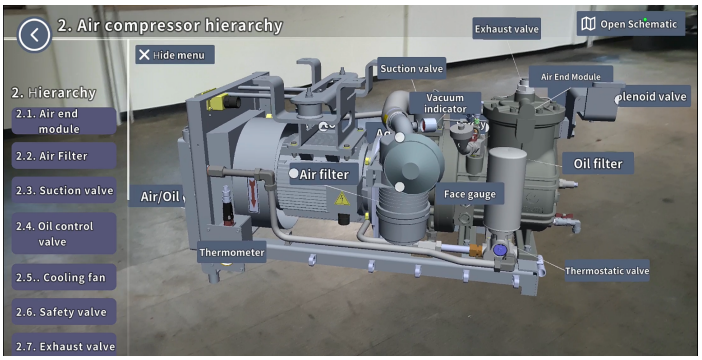


Figure 16: Enlarged screen space by moving the original right LNB to the left

3.5. Subcomponent Display Techniques

The developed model is assembled into sub-components, which are based on the maintenance manual and the replacement part procurement system using hierarchy and can identify individual parts from the assembly. It allows for a perceptive effect, and when a particular part is selected in the assembly, the assembly status can be easily identified by using a different color or by inserting an outline at the boundary between the equipment and sub-components. When the structure menu of SNB is touched at the component level, the sub-part is unfolded, and the name of the component is displayed with animation and buttons to make it easier for the user to understand the complex component, and if necessary, a separate SNB can be used to show the procurement product number of the component. This process is already detailed in the hierarchical design stage, but missing and insufficient functions or animations can be inserted additionally during use through content editing to increase completion. Therefore, through this content design, users can check the connection of device components and open the detailed structure between components in the center of the screen in detail, making it convenient to understand the entire device without disassembling the actual components one by one as shown in Figure 17.

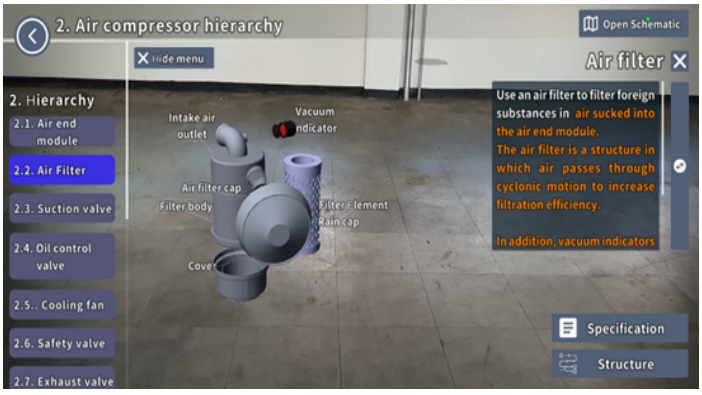


Figure 17: Subcomponent display screen of air filter assembly

The SNB-type menu on the right side of the window is mainly installed at the bottom of the movable window that can display detailed descriptions and other information about components. The SNB menu is used to show sub-parts in a disassembled structure, deliver operational information related to sub-parts, or animate compressed air, electricity, oil flow, etc. to help trainers understand and improve the convenience of maintenance training.

3.6. Maintenance Instruction Implementation for User Training

In the GNB menu, the maintenance menu mainly implements the maintenance items, contents, duration, procedures, inspection, repair and replacement methods of equipment by maintenance level required by the manufacturer in the manual approved by the vehicle operator, and the pre-learning support function for the replacement of spare parts and consumable parts. Sub-component contents such as oil amount, supplemental level, and exchange cycle related to the oil filter may be implemented as text in a pop-up window. In this way, the important information of the component is displayed visually to the user.

Figure 18 is a case screen showing lubricating oil filters and lubricating oil replenishment work among maintenance items. There is a maintenance support window built with a movable window that shows the process of replacing the lubricant filter. This content can train users by clearly presenting replacement of consumables, replacement of lubricating oil, supplementation procedures or criteria.

Maintenance AR content includes the ability to effectively pre-learn the replacement process of parts that have reached the replacement cycle. In the figure, the scene where the valve is opened and the lubricant is replenished from the lubricant plastic container to the refueling hole is realized as if it were real by algorithms that show fluid flow. The animation is also implemented with detailed procedures in the pop-up window, allowing users to effectively understand lubricating oil replenishment and learn how to do it.

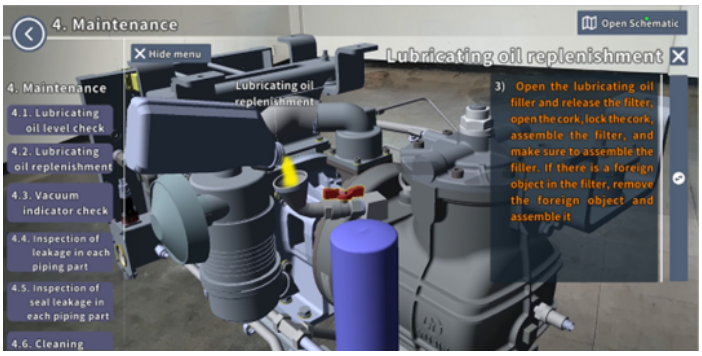


Figure 18: Maintenance instruction for lubricating oil replenishment

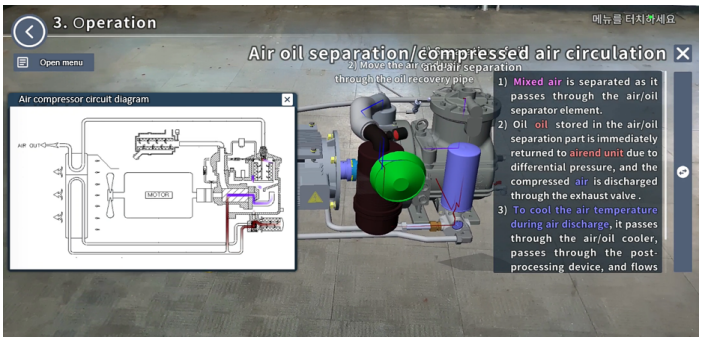


Figure 19: Implement fluid flow simultaneously in 2D pipe drawings and 3D model.

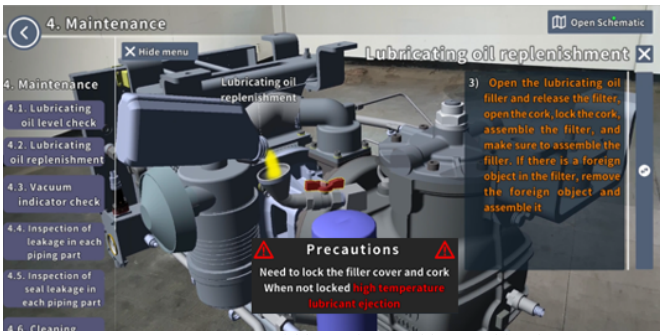
Figure 19 shows the function of implementing fluid flow simultaneously in a 2D pipe drawing and a 3D model. The algorithms described in the previous section are used to provide effective theory education as well as to allow users to easily understand the operating principles of complex devices.

3.7. Safety Warning Function for Hazardous Maintenance Work

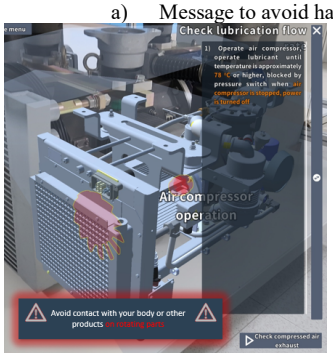
The maintenance work of railway vehicles includes various danger and hazard factors, which can lead to serious injury or even death of workers. Therefore, AR content needs to include the ability to proactively train these risks.

Figure 20 shows examples of safety warnings displayed in this content. (a) is a warning message to avoid burn hazard caused by touching hot areas during air compressor maintenance. (b) indicates a warning against body or other contact with cooling fans, coupling rotating parts, etc. (c) shows a safety warning to ensure that the air is exhausted in the pipe before starting work and to prepare for an unexpected situation where compressed air is ejected from the valve. This content reminds workers of things to pay attention to before and after starting work, and also makes workers aware of unexpected situations that may occur.

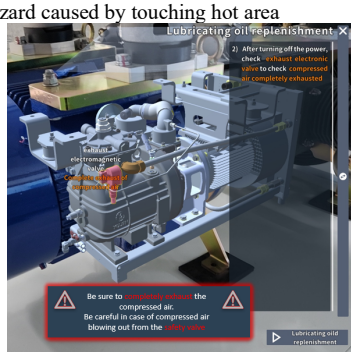
In addition, various risk factors such as electric shock caused by high-voltage electricity or serious injury caused by vehicle rolling are warned using appropriate warning text and sound effects. In particular, you can use red or yellow visually to highlight the hazard with highlights. Additionally, various pop-up windows can provide detailed explanations of maintenance hazards, such as handling toxic substances and pollutants.



a) Message to avoid hazard caused by touching hot area



b) Message on rotating parts



c) Message for unexpected situation

Figure 20: Caution and warning messages for safety

3.8. Finger Gestures on Mobile Device Touchscreen

This content can use a manipulation gesture that can control the location, size, and rotation of the 3D model using the mobile device's touch screen method. Table 4 shows the detailed functions of the 3D model manipulation interaction as an example [34]. Maintenance content users can open the app on the mobile device

screen, use the menu to load content on the screen, and use touch-type manipulation technology to perform training and maintenance work [35].

Table 4: Touch gesture for touchscreen manipulation

Value to change	Corresponding	Touch gesture
Position	swipe	
Scale	pinch/spread	
Rotation	twist	

Figure 21 is a photograph showing the inspection and work in front of the air compressor of a railway vehicle using the AR air compressor content of a mobile phone. The produced content is compatible with various screen sizes based on Android first. Using AR content, it overcomes the shortcomings of time, space, and manpower management, which are obstacles to existing document-based manual education, receives database information of a separately developed maintenance server, and is effectively digitized in conjunction with actual digital maintenance work records.

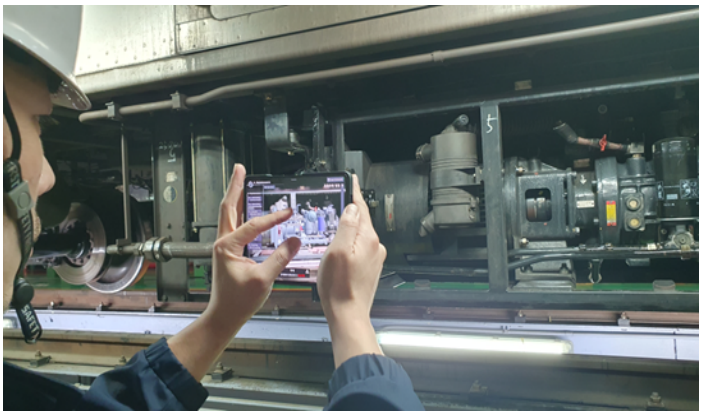


Figure 21: A photo of using mobile phone apps on the side of the vehicle at depot

4. Usability Evaluation

The usability of the developed AR content can be performed using various methods. A commonly used technique is to use the content developed by the user group, then conduct a survey and analyze the results. These techniques include the SUS, the Computer System Usability Survey (CSUQ), and the Usability Indicator for User Experience (UMUX) [36]-[38].

In order to evaluate the usability of the developed AR content, a SUS survey and a survey on UX techniques were conducted on 100 college students who majored in railway vehicle systems, respectively. Table 5 show the participants of the usability evaluation experiments [39].

Table 5: The participants of the experiments

Experiment participants	100 students
Average age	23.8 years
Male	73
Female	27

4.1. SUS

The SUS, as a quick and dirty usability scale is reliable tool for measuring the usability, originally created by John Brook in 1996 [40]. SUS survey consists of a 10 item questionnaire with 5 response options for participants of a scale of 1 to 5. Scale score 1 means strongly disagree and scale score 5 means strongly agree. Table 6 shows the modified SUS survey questions in this study [41]. The survey was carried out after using the AR contents.

Table 6: The modified SUS statements used in this study

No.	Questions
Q1	I think that I would like to use this AR content product frequently
Q2	I found the AR content unnecessarily complex
Q3	I thought the AR content was easy to use
Q4	I think that I would need the support of a technical person to be able to use this AR content
Q5	I found that the various functions in This AR content were well integrated
Q6	I thought that there was too much inconsistency in this AR content
Q7	I would imagine that most people would learn to use this AR content very quickly
Q8	I found the AR content very awkward to use
Q9	I felt very confident using the AR content.
Q10	I needed to learn a lot of things before I could get going with this AR content

The SUS score can be calculated using the following 3 equations [42].

$$Score_{Q1,Q3,Q5,Q7,Q9} = Scale\ Score - 1 \quad (1)$$

$$Score_{Q2,Q4,Q6,Q8,Q10} = 5 - Scale\ Score \quad (2)$$

$$SUS\ Scores = Sum\ of\ Scores \times 2.5 \quad (3)$$

Equation (1) is used to get the odd number question scores. The even number question scores are calculated by (2). Equation (3) is used to obtain the final value of the SUS survey. The scores have a range is from 0 to 100. It can also be mapped to the following adjective ratings. Figure 22 shows a comparison of mean SUS scores by quartile, adjective ratings and the acceptability of the overall SUS score [43].

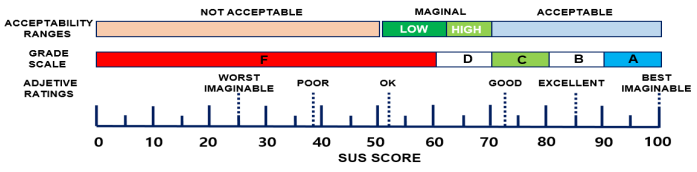


Figure 22: A comparison of mean SUS scores by quartile, adjective ratings, and the acceptability of the overall SUS score.

4.2. User Experience Experiment

Another technique, UX assessment, was conducted on AR contents. Six items related to usability were selected that were

reviewed as suitable for content evaluation for maintenance of specific equipment, such as air compressors. The survey tool referred to the UX evaluation, and the detailed survey contents of the survey are shown in Table 7 [44]. An experience survey using a questionnaire tool developed after the user used AR content for air compressors was conducted and analyzed.

Table 7: Indicators of Usability assessment

Categorized Items	Questionnaire details
Simplicity	Appearance of AR content, function provision method, simplicity of procedures.
Accessibility	User should be able to access AR content smoothly.
Efficiency	It should be possible to derive accurate results without unnecessary time consumption.
Informativeness	User should be able to obtain the necessary information through appropriate means and methods.
Learnability	User should be able to handle the interface of AR content proficiently
User Support	User should be able to respond well to problems that arise when using AR content.

4.3. SUS Survey Results

Figure 23 is a graph showing the score for each SUS question of the AR contents of the air compressor. The threshold reference value for the usefulness of SUS is 68 points. The results of this survey are as follows. The items that were relatively better than other questions were Q3 "ease of using content" (87.2) and Q1 "intention to use content" (87). On the other hand, Q4 "needs for technical helpers" (68.3) and Q10 "needs for pre-learning" (69) questions scored relatively lower than other questions. This seems to be because survey participants are not familiar with the use of AR content in the screen touch manipulation method, as shown in Table 3. The average score of 10 SUS questions for this content was 76.65, which was higher than the overall threshold of 68 points for each question. In addition, the usefulness rating is “B(Good)”.

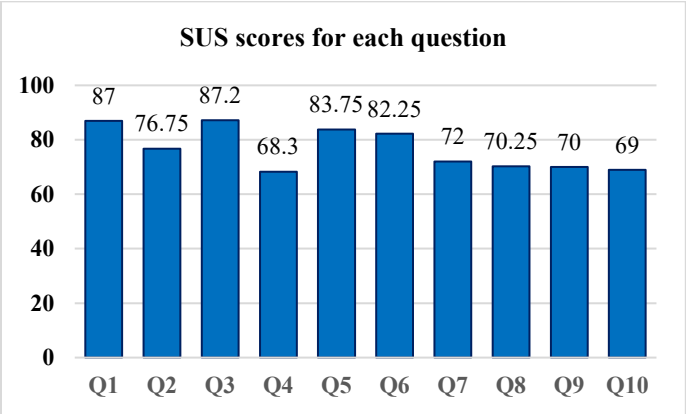


Figure 23: SUS score of the air compressor AR contents

Therefore, this content will be useful for beginners in railway vehicle maintenance. Therefore, it is considered suitable for use in the maintenance and repair training of air compressors for urban railway vehicles.

4.4. UX Survey Results

Figure 24 is a graph showing the UX evaluation results. This is the result of a survey after using AR contents developed four times. The overall average of the six questions was 4.12 points, and the score for the sense of reality item was the highest at 4.38 points. On the other hand, the score of the convenience item was 3.92, the lowest among the four items.

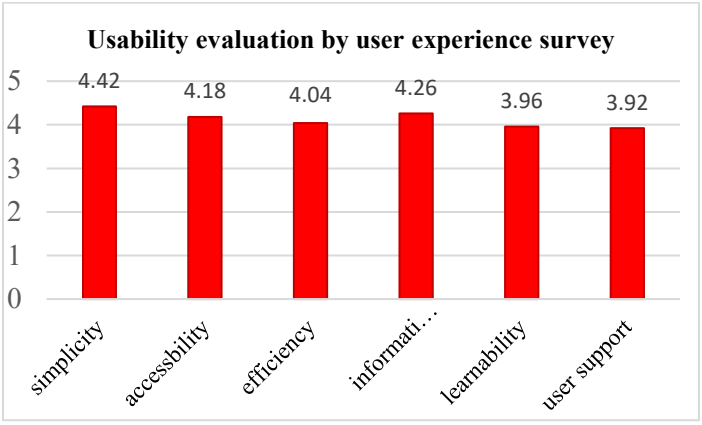


Figure 24: UX score of the air compressor AR contents

Table 8 shows the results of sub-items for each item of UX evaluation. The reason for the highest simplicity is that the maintenance procedure is simply and clearly structured to make it easier for even beginners to understand (4.58). It is presumed that various learning functions are also clearly organized based on keywords (4.41). The second highest informativeness item score is considered to be because the entire scenario of the content was systematically structured (4.38), and the text was written to make the maintenance procedure easier to understand (4.22). In comparison, the reason why the user support item is the lowest was that the description of the possible technical problems was insufficient (3.85), and the response method in case of an error was not described in detail (0.395).

Table 8: Scores of detail Questionnaire for UX

Categorized Items	Questionnaire details	score
Simplicity	Smooth output of content	4.27
	Simple configuration of various learning functions	4.41
	Simple and clear explanation of maintenance processes	4.58
Accessibility	Ease of running content	4.09
	Compatibility with mobile devices	4.20
	Clarity of description and usage	4.31
Efficiency	Comparison of efficiency with existing manuals	3.99
	Quickly acquire the information you want	4.13

Categorized Items	Questionnaire details	score
	Increased actual work efficiency	4
Informativeness	Provides accurate maintenance information	4.18
	Understand the entire maintenance process	4.22
	Systematicity of the entire scenario	4.38
Learnability	Seamless model touch availability	3.96
	Compatibility of UI and additional buttons	3.97
	Ease of screen transition	3.95
User Support	Description of response when error occurs	3.95
	Adequacy of help and error examples	3.96
	Examples of problems that may occur	3.85

4.5. Discussion

Two evaluation techniques, SUS and UX, were used to identify problems and derive improvements of developed AR content. As a result of the two experiments, a common issue is that participants have some difficulty using the content and take time to learn how to use it. They experienced some errors in the content during the experiment and expressed lack of help or countermeasures. The content was produced in consideration of the use of novice maintenance workers in the design stage, so this part was highly evaluated, but it is also considered that the use of maintenance workers who are relatively unfamiliar with digital content should be considered. In future contents, it will be necessary to produce it by receiving feedback from maintenance workers who have not been exposed to AR technology in the design stage. If additional functions such as inserting help for each function, producing tutorial version, and voice guidance are produced, high-quality content can be produced.

5. Conclusion

This study is a case of applying AR technology to the maintenance of railway vehicles and employee education and training as one of the smart maintenance research projects. AR contents that can be easily used in mobile devices for maintenance and training of air compressors for commuter electric vehicles have been developed. Two algorithms were developed and used to express fluid flows such as pressure air. In addition, in order to verify the usability of the developed content using a SUS survey and UX survey. The conclusions obtained from this study are as follows.

(1) By reducing the file size from 3D scanners or CAD original files, it is possible to implement fast animation on Android-based mobile devices. It was developed by linking training AR contents with the AR manual of the MMES, which is a high-level platform. Therefore, maintenance work and training are possible while watching actual equipment in the field depot, and maintenance support can be provided along with other equipment contents already developed or under development.

(2) The 2D-3D LMDC algorithm, which can implement flow simultaneously on a two-dimensional drawing or piping diagram and a three-dimensional equipment screen, was applied. In addition, the FG algorithm, which can control the flow rate of the fluid flow, was used to implement the movement or exhaust of compressed air, thereby enhancing the learning effect of the content.

(3) A simplified menu is placed on the left and right sides of the display screen, and the device is operated by user-friendly menu bar and touch screen operation technology on the central screen. In addition, through a systematic hierarchy structure when developing a storyboard, users can easily understand the operating principles and operations of the entire equipment, and perform reliable work by receiving detailed support for disassembly procedures and specific work methods

(4) The developed content can display components and sub-components according to the standardized BOM system of manufacturers and operators. This enables support for maintenance work using part hierarchies. In addition, through information on the replacement cycle and replacement process of spare parts based on the manufacturer's manual, it is possible to effectively train the replacement and process of spare parts and perform necessary inspections or repairs.

(5) Due to the nature of railway vehicle parts, there is a lot of heavy equipment, and there are maintenance tasks that involve workers' risks such as high voltage and high-pressure air. In order to eliminate safety accidents during work, safety items were visualized on a text display to generate a warning message, and a red highlight was applied to each inspection part for worker safety.

(6) The average SUS score for AR content of this air compressor content is 76.65, and the rating is “Good”. In addition, the results of the UX test showed that the average usability score for the six-item questionnaire was excellent at 4.12 out of 5 score, and this content is estimated to be quite useful for training beginners in railway vehicle maintenance.

Future research will need to expand content that has not yet been developed in the vehicle system, further develop algorithms for efficient content production, and continue to improve usability in conjunction with the AR manual of the maintenance efficiency system.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) by the Ministry of Land, Infrastructure and Transport under Grant RS-2021-KA162811.

References

[1] F.A. Marten, “Reliability Centered Maintenance: A Case Study of Railway Transit Maintenance to Achieve Optimal Performance,” Mineta Transportation Institute Report, 1006, 2010.

[2] E. Quatrini, F. Costantino, G.D. Gravio, R. Patriarca, “Condition-Based Maintenance An Extensive Literature Review,” Machines, **8**(2), 31, 2020, doi:10.3390/machines8020031.

[3] A.K.S. Jardine, D. Lin, D. Banjevic, “A review on machinery diagnostics and prognostics implementing condition-based maintenance,” Mechanical

systems and signal processing, **20**(7), 1483–1510, 2006, doi:10.1016/j.ymssp.2005.09.012.
[4] J. Campos, "Development in the application of ICT in condition monitoring and maintenance," Computers in Industry, **60**(1), 1–20, 2009, doi:10.1016/j.compind.2008.09.007.
[5] B. Zhang, Y. Sui, Q. Bu, X. He, "Remaining useful life estimation for micro switches of railway vehicles," Control Engineering Practice, **84**, 82–91,2019, doi:10.1016/j.conengprac.2018.10.010.
[6] C.P. Ward, P.F. Weston, E.J.C. Stewart, H. Li, R.M. Goodall, C. Roberts, T.X. Mei, G. Charles, R. Dixon, "Condition Monitoring Opportunities Using Vehicle-Based Sensors," Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, **225**(2), 202–218, 2011, doi:10.1177/09544097jrrt406.
[7] Z. Cujan, G. Fedorko, N. Mikusova, "Application of virtual and augmented reality in automotive," Open Engineering, **10**(1), 113–119, 2020, doi:10.1515/eng-2020-0022.
[8] T. Masood, J. Egger, "Augmented reality in support of Industry 4.0–Implementation challenges and success factors," Robotics and Computer-Integrated Manufacturing, **58**, 181–195, 2019, doi:10.1016/j.rcim.2019.02.003.
[9] L.F.D.S. Cardoso, F.C.M.Q. Mariano, E.R. Zorzal, "A Survey of Industrial Augmented Reality," Computers & Industrial Engineering, **139**, 106159, 2020, doi:10.1016/j.cie.2019.106159.
[10] R.T. Azuma, "A Survey of Augmented Reality," PRESENCE: Virtual and Augmented Reality Issues, **6**(4), 355–385, 1997, doi:10.1162/pres.1997.6.4.355.
[11] R. Palmarini, J.A. Erkoyuncu, R. Roy, H. Torabmostaedi, "A systematic review of augmented reality applications in maintenance," Robotics and Computer-Integrated Manufacturing, **49**, 215–228, 2018, doi:10.1016/j.rcim.2017.06.002.
[12] R.G. Boboc, F. Girbacia, E.V. Butila, "The Application of Augmented Reality in the Automotive Industry: A Systematic Literature Review," Applied Sciences, **10**(12), 4259, 2020, doi:10.3390/app10124259.
[13] H. Rios, E. Gonzalez, C. Rodriguez, H.R. Siller, M. Contero, "A Mobile Solution to Enhance Training and Execution of Troubleshooting Techniques of the Engine Air Bleed System on Boeing 737," Procedia Computer Science, **25**, 161–170, 2013, doi:10.1016/j.procs.2013.11.020.
[14] H. Eschen, T. Kotter, R. Rodeck, M. Harnisch, T. Schuppstuhl, "Augmented and Virtual Reality for Inspection and Maintenance Processes in the Aviation Industry," Procedia Manufacturing, **19**, 156–163, 2018, doi:10.1016/j.promfg.2018.01.022.
[15] M. Kans, D. Galar, A. Thaduri, "Maintenance 4.0 in railway transportation industry," in Proceedings of the 10th world congress on engineering asset management (WCEAM 2015), 317–331, 2016, doi:10.1007/978-3-319-27064-7_30.
[16] A. Muller, A.C. Marquez, B. Iung, "On the concept of e-maintenance: Review and current research," Reliability Engineering & System Safety, **93**(8), 1165–1187, 2008, doi:10.1016/j.ress.2007.08.006.
[17] P. Cipresso, I.A.C. Giglioli, M.A. Raya, G. Riva, "The Past, Present, and Future of Virtual and Augmented Reality Research: A Network and Cluster Analysis of the Literature," Frontiers in Psychology, **9**, 2086, 2018, doi:10.3389/fpsyg.2018.02086.
[18] W. Zhang, B. Han, P. Hui, "On the networking challenges of mobile augmented reality," in Proceedings of the Workshop on Virtual Reality and Augmented Reality Network, 24–29, 2017, doi:10.1145/3097895.3097900.
[19] J. Kiu, K. Zhang, "Design and Simulation Debugging of Automobile Connecting Rod Production Line Based on the Digital Twin," Applied Science, **13**(8), 4919, 2023, doi: 10.3390/app13084919.
[20] V. P. Gangwar, D, Reddy, "Hospitality Industry 5.0: Emerging Trends in Guest Perception and Experiences," Opportunities and Challenges of Business 5.0 in Emerging Markets, 185–211, 2023, doi: 10.4018/978-1-6684-6403-8.ch010.
[21] E.D.N. Orihuela, O.O.V. Villegas, V.G.C. Sanchez, R.I.B. Castillo, J.G.L. Solorzano, "Mobile Augmented Reality Prototype for the Manufacturing of an All-Terrain Vehicle," Advanced Topics on Computer Vision, Control and Robotics in Mechatronics, 49–75, 2018, doi:10.1007/978-3-319-77770-2_3.
[22] A. Chouchene, A.V. Carvalho, F.C. Santos, W. Barhoumi, "Augmented Reality-Based Framework Supporting Visual Inspection for Automotive Industry," Applied System Innovation, **5**(3), 48, 2022, doi:10.3390/asi5030048.
[23] H.J. Kwon, K.S. Kim, C.S. Kim, "Development and Evaluation of Augmented Reality Learning Content for Pneumatic Flow: Case Study on Brake Operating Unit of Railway Vehicle," IEEE Access, **11**, 46173–46184, 2023, doi:10.1109/ACCESS.2023.3273605.
[24] H.J. Kwon, S.I. Lee, J.H. Park, C.S. Kim, "Design of Augmented Reality Training Content for Railway Vehicle Maintenance Focusing on the Axle-Mounted Disc Brake System," Applied Science, **11**(19), 9090, 2021, doi:10.3390/app11199090.
[25] S. Scheffer, A. Martinetti, R. Damgrave, S. Thiede, L.V. Dongen, "How to Make Augmented Reality a Tool for Railway Maintenance Operations: Operator 4.0 Perspective" Applied Sciences, **11**(6), 2656, 2021, doi:10.3390/app11062656.
[26] N. Hagbi, O. Bergig, J.E. Sana, M. Billibghurst, "Shape Recognition and Pose Estimation for Mobile Augmented Reality," IEEE transactions on visualization and computer graphics, **17**(10), 1369–1379, 2010, doi:10.1109/TVCG.2010.241.
[27] D.P. Casssent, "Intelligent Robots and Computer Vision XV: Algorithms, Techniques, Active Vision, and Materials Handling," Intelligent Robots and Computer Vision XV: Algorithms, Techniques, Active Vision, and Materials Handling, 2904, 1996.
[28] D. Wagner, G. Reitmayr, A. Mulloni, T. Drummond, D. Schmalstieg, "Real-time detection and tracking for augmented reality on mobile phones," IEEE transactions on visualization and computer graphics, **16**(3), 355–368, 2009, doi:10.1109/TVCG.2009.99.
[29] M.S. Elbelkasi, E. Badaran, M.A Rahman, "Overview of DC and AC Electric Railway Systems Considering Energy Efficiency Enhancement Methods," Port-Said Engineering Research Journal, **24**(1), 128–140, 2020, doi: 10.21608/pserj.2019.18206.1012.
[30] M. Saeed, F. Briz, J.M. Guerrero, I. Larrazabal, D. Ortega, V. Lopez, J.J. Valera "Onboard energy storage systems for railway: present and trends," IEEE Open Journal of Industry Applications, **4**, 238–259, 2023, doi: 10.1109/OJIA.2023.3293059.
[31] 3D Systems, Geomagic Design X 2020 manual, 2020.
[32] Unity Asset Store, Easy Performant Outline 2.0, 2021.
[33] H.J. Kwon, J.H. Cha, Y.K. Park, C.S. Kim, "A study on the realistic content for air compressor maintenance of the urban railway vehicle," Journal of Korean Society for Urban Railway, **10**(1), 1163–1175, 2022.
[34] Unity Asset Store, Lean Touch+, 2021.
[35] M. Kim, J.Y. Lee, "Touch and hand gesture-based interactions for directly manipulating 3D virtual objects in mobile augmented reality," Multimedia Tools and Applications, **75**, 16529–16550, 2016, doi:10.1007/s11042-016-3355-9.
[36] J.R. Lewis, "IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use," International Journal of Human-Computer Interaction, **7**, 57–78, 1995, doi:10.1080/10447319509526110.
[37] J.R. Lewis, "Measuring Perceived Usability: The CSUQ, SUS, and UMUX" International Journal of Human-Computer Interaction, **34**(12), 1148–1156, 2018, doi:10.1080/10447318.2017.1418805.
[38] K. Finstad, "The usability metric for user experience," Interacting with Computers, **22**(5), 323–327, 2010, doi:10.1016/j.intcom.2010.04.004.
[39] K. Hornbæk, M. Hertzum, "Technology Acceptance and User Experience," ACM Transactions on Computer-Human Interaction (TOCHI), **24**(5), 1–30, 2017, doi:10.1145/3127358.
[40] J. Brooke, "Sus: a "quick and dirty'usability," Usability evaluation in industry, **189**(3), 189–194, 1996, doi:10.1016/j.ress.2007.08.006.
[41] A. Bangor, P,T, Kortum, J.T, Miller, "An Empirical Evaluation of the System Usability Scale," International Journal of Human–Computer Interaction, **24**(6), 574–594, 2008, doi:10.1080/10447310802205776.
[42] A.C. Wijaya, M.W.A. Munandar, F. Utaminingrum, "Usability Testing of Augmented Reality For Food Advertisement Based On Mobile Phone Using System Usability Scale," in 2019 International Conference on Sustainable Information Engineering and Technology (SIET), 266–269, 2019, doi:10.1109/SIET48054.2019.8986118.
[43] A. Bangor, P.T. Kortum, J.T, Miller, "Determining what individual SUS scores mean: Adding an adjective rating scale," Journal of usability studies, **4**(3), 114–123, 2009, doi:10.1080/10447310802205776.
[44] J. Shin, S.H. Lee, J.E. Yoo, C. Roh, J.A. Oh, S. Y. Kim, "Development and Evaluation of a Virtual Maintenance Training System for Thermal Power Plant Boiler," Journal of Digital Contents Society, **22**(5), 791–800, 2021.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).