ASTES

Advances in Science, Technology & Engineering Systems Journal

VOLUME 9-ISSUE 4|JUL-AUG 2024

www.astesj.com
ISSN: 2415-6698

Editorial

In an era where technology continues to redefine the boundaries of research and innovation, this issue of our journal presents a vibrant collection of interdisciplinary studies that reflect the diversity and dynamism of contemporary scientific inquiry. Spanning fields such as artificial intelligence, environmental monitoring, digital healthcare, education technology, and sustainable development, the 12 papers included in this edition showcase groundbreaking methodologies and practical applications that promise to shape the future of both academia and industry.

There is a critical challenge in aquaculture—the prediction of seawater temperatures with limited training data. By leveraging transfer learning, the authors demonstrate how accurate predictions can be achieved even with small datasets, offering a practical and efficient tool for farmers to mitigate the risk posed by abnormal temperature fluctuations. Their findings hold great promise for sustainable marine farming practices, particularly in data-scarce environments [1].

In the realm of medical imaging, there is an advanced method for the automatic 3D segmentation of brain tumors using optical scanning holography and active contour modeling. This innovative approach not only minimizes human error but also enhances diagnostic precision, making it a valuable asset for computer-aided diagnosis systems in neuro-oncology [2].

The architecture of RetePAIoT, is a public IoT network in the Emilia-Romagna region. By enabling seamless integration and sharing of data from thousands of sensors, the authors propose a robust and scalable system that empowers public administrations and private users alike. The work lays a foundational framework for region-wide IoT deployment with an emphasis on interoperability and data accessibility [3].

Focusing on medical informatics, there is an automated ICD coding system powered by deep learning and enhanced by GPT-4. The hierarchical classification model efficiently addresses the complexity and scale of ICD coding tasks, showing how AI can streamline healthcare documentation and improve clinical workflows with greater speed and accuracy [4].

Switch-mode power supplies a predictive maintenance model for aluminum electrolytic capacitors, incorporating functional data analysis and fuzzy logic. This research contributes significantly to the field of reliability engineering by offering a precise and adaptable approach to monitoring component degradation under variable conditions [5].

The human-robot interaction, proposed a multimodal system that integrates verbal commands with gesture recognition for precise task execution. By combining skeleton tracking and hand gesture detection with natural language processing, the system enables intuitive and efficient communication with robots, enhancing automation in service and industrial contexts [6].

By introducing two novel algorithms—one based on tree search and the other on trie-like graphs—the authors succeed in significantly reducing computational complexity, offering more practical solutions for personalized recommendation systems and customer satisfaction analysis [7].

The effectiveness of the MIMOSYS voice analysis system in detecting depression among Vietnamese speakers. By correlating voice-derived vitality scores with established psychological scales, the research offers compelling evidence for the system's potential in non-invasive mental health monitoring, paving the way for broader linguistic and cultural validation [8].

Addressing safety in domestic environments, there is a smart LPG monitoring system that merges IoT and business intelligence. Through automated data collection and real-time alerts, the model reduces the risk of gas leaks while demonstrating high user satisfaction. Its practical deployment across mobile and web platforms highlights its accessibility and real-world impact [9].

There is a big data maturity models tailored for SMEs, emphasizing the need for scalable and adaptable frameworks in the digital transformation journey. The study identifies existing gaps and proposes directions for developing new models that align with the resource constraints and operational realities of small and medium-sized enterprises [10].

The energy management landscape of ASEAN countries is analyzed with a focus on regional cooperation, policy frameworks, and sustainability initiatives. The authors identify critical challenges and propose policy recommendations to enhance energy security and support the region's transition toward a low-carbon economy, making it a timely contribution to environmental policy literature [11].

The integration of augmented reality with educational robotics through the ASCAT-AR system. Targeted at STEM education, the study shows that students using AR-assisted learning tools perform better and are more engaged compared to traditional methods. This paper exemplifies how immersive technologies can revolutionize learning experiences and boost student outcomes in engineering and programming education [12].

This edition brings together a multifaceted collection of research that addresses pressing global challenges through technological innovation, analytical rigor, and practical application. Whether it’s through the enhancement of healthcare diagnostics, the safeguarding of energy and environmental systems, or the reimagination of education through immersive technology, each paper contributes meaningfully to the advancement of its respective field. We hope these studies will inspire further research, spark interdisciplinary collaboration, and inform evidence-based practices in academia, industry, and policy-making

References:

[1] H. Murakami, T. Miwa, K. Shima, T. Otsuka, “Proposal and Implementation of Seawater Temperature Prediction Model using Transfer Learning Considering Water Depth Differences,” Advances in Science, Technology and Engineering Systems Journal, 9(4), 1–6, 2024, doi:10.25046/aj090401.

[2] A. El-Ouarzadi, A. Cherkaoui, A. Essadike, A. Bouzid, “Hybrid Optical Scanning Holography for Automatic Three-Dimensional Reconstruction of Brain Tumors from MRI using Active Contours,” Advances in Science, Technology and Engineering Systems Journal, 9(4), 7–13, 2024, doi:10.25046/aj090402.

[3] S. Nanni, M. Carboni, G. Mazzini, “From Sensors to Data: Model and Architecture of an IoT Public Network,” Advances in Science, Technology and Engineering Systems Journal, 9(4), 14–20, 2024, doi:10.25046/aj090403.

[4] J. Carberry, H. Xu, “GPT-Enhanced Hierarchical Deep Learning Model for Automated ICD Coding,” Advances in Science, Technology and Engineering Systems Journal, 9(4), 21–34, 2024, doi:10.25046/aj090404.

[5] D. Mallamo, M. Azarian, M. Pecht, “Early Detection of SMPS Electromagnetic Interference Failures Using Fuzzy Multi-Task Functional Fusion Prediction,” Advances in Science, Technology and Engineering Systems Journal, 9(4), 35–50, 2024, doi:10.25046/aj090405.

[6] S. Kumar Paul, M. Nicolescu, M. Nicolescu, “Integrating Speech and Gesture for Generating Reliable Robotic Task Configuration,” Advances in Science, Technology and Engineering Systems Journal, 9(4), 51–59, 2024, doi:10.25046/aj090406.

[7] Y. Chen, B. Chen, “On Mining Most Popular Packages,” Advances in Science, Technology and Engineering Systems Journal, 9(4), 60–72, 2024, doi:10.25046/aj090407.

[8] L.T. Vinh Phuc, M. Nakamura, M. Higuchi, S. Tokuno, “Effectiveness of a voice analysis technique in the assessment of depression status of individuals from Ho Chi Minh City, Viet Nam: A cross-sectional study,” Advances in Science, Technology and Engineering Systems Journal, 9(4), 73–78, 2024, doi:10.25046/aj090408.

[9] A.R. Espinoza de los Monteros, M.G. Tandazo Espinoza, B.I. Punina Cordova, R.E. Tandazo Vanegas, “IoT and Business Intelligence Based Model Design for Liquefied Petroleum Gas (LPG) Distribution Monitoring,” Advances in Science, Technology and Engineering Systems Journal, 9(4), 79–92, 2024, doi:10.25046/aj090409.

[10] D. Salian, S. Brown, R. Sbeit, “Digitalization Review for American SMEs,” Advances in Science, Technology and Engineering Systems Journal, 9(4), 93–101, 2024, doi:10.25046/aj090410.

[11] W.Y.L. Yie Leong, Y. Zhi Leong, W. San Leong, “Energy Management Policy and Strategies in ASEAN,” Advances in Science, Technology and Engineering Systems Journal, 9(4), 102–109, 2024, doi:10.25046/aj090411.

[12] W. Sawangnamwong, S. Charoenseang, “Assistive System for Collaborative Assembly Task using Augmented Reality,” Advances in Science, Technology and Engineering Systems Journal, 9(4), 110–118, 2024, doi:10.25046/aj090412.

Editor-in-chief

Prof. Passerini Kazmersk

ADVANCES IN SCIENCE, TECHNOLOGY AND ENGINEERING SYSTEMS JOURNAL

Volume 9 Issue 4 July-August 2024

CONTENTS

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°C.

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.

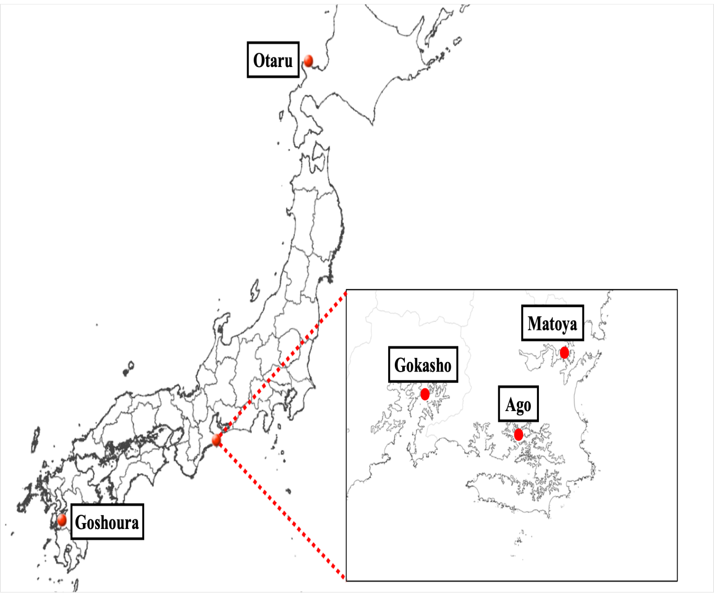


Figure 1: Five observation points of seawater temperature [16]

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°C, while that in summer ranges from 2 to 9°C.

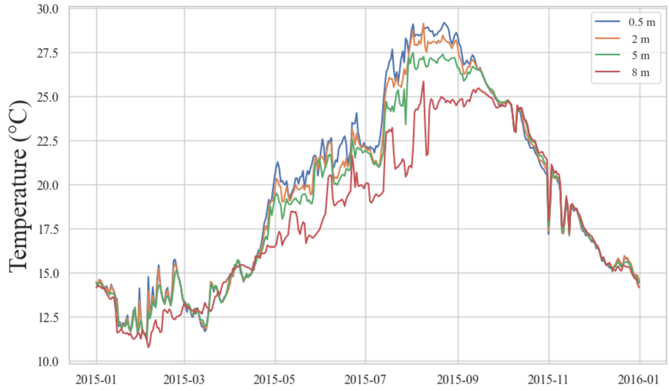


Figure 2: Mean daily seawater temperature at each water depth in Ago (2015).

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°C, while the water temperature difference in summer ranges from 1°C to 8°C. In Otaru, the water temperature difference in winter is within 2°C, while the water temperature difference in summer can be as high as 5°C.

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.

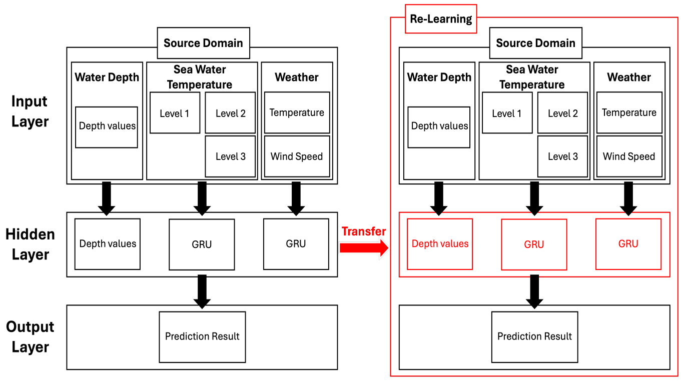


Figure 3: Overall diagram of the proposed model.

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°C. In this study, the standard value was set to 1°C to meet the needs of aquaculture farmers for an error less than 1°C.

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 [°C] and the values on the right represent the rate of errors above 1°C [%].

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 [°C] and the values on the right represent the rate of errors above 1°C [%].

Table 5: Results of Experiment 2 in Goshoura

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

Table 6: Results of Experiment 2 in Otaru

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

Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 4, 07-13(2024)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Hybrid Optical Scanning Holography for Automatic Three-Dimensional Reconstruction of Brain Tumors from MRI using Active Contours

Abdennacer El-Ouarzadi[1], Anass Cherkaoui[1], Abdelaziz Essadike[2], Abdenbi Bouzid[1]

[1] *Moulay Ismail University, Physical Sciences and Engineering, Faculty of Sciences, Meknès, 11201, Morocco.*

[2] *Hassan First University of Settat, Higher Institute of Health Sciences, Laboratory of Health Sciences and Technologies, Settat, 26000, Morocco.*

ARTICLE INFO

Article history:
Received: 03 May, 2024
Revised: 26 June, 2024
Accepted: 27 June, 2024
Online: 10 July, 2024

Keywords:
Bain tumor detection,
Fully Automatic Segmentation,
Active contour,
Optical Scanning Holography (OSH),
3D Automatic Reconstruction.

ABSTRACT

This paper presents a method for automatic 3D segmentation of brain tumors in MRI using optical scanning holography. Automatic segmentation of tumors from 2D slices (coronal, sagittal and axial) enables efficient 3D reconstruction of the region of interest, eliminating the human errors of manual methods. The method uses enhanced optical scanning holography with a cylindrical lens, scanning line by line, and displays MRI images via a spatial light modulator. The outgoing phase component of the scanned data, collected digitally, reliably indicates the position of the tumor.The tumor position is fed into an active contour model (ACM), which speeds up segmentation of the seeding region. The tumor is then reconstructed in 3D from the segmented regions in each slice, enabling tumor volume to be calculated and cancer progression to be estimated. Experiments carried out on patient MRI datasets show satisfactory results. The proposed approach can be integrated into a computer-aided diagnosis (CAD) system, helping doctors to localize the tumor, estimate its volume and provide 3D information to improve treatment techniques such as radiosurgery, stereotactic surgery or chemotherapy administration. In short, this method offers a precise and reliable solution for the segmentation and 3D reconstruction of brain tumors, facilitating diagnosis and treatment.

1. Introduction

A brain tumor is a mass of abnormal cells involved in a chaotic process of somatic driver mutations [1,2], where they cause various symptoms and increase the risk of brain damage. In fact, the secondary tumor infiltrates neighboring healthy tissues and proliferates within the brain or its membranes, making it critical to determine its shape and volume to ensure effective management of patients in the early stages of cancer. Magnetic resonance imaging (MRI) is the most commonly used non-invasive imaging modality for brain tumor detection [3–5]. MRI uses radio waves and a strong magnetic field to create a series of cross-sectional images of the brain. In other words, the 3D anatomical details of a tumor are represented as a set of parallel 2D cross-sectional images. Representing 3D data as projected 2D slices results in a loss of information and can raise questions about tumor prognosis. In addition, 2D slices do not accurately represent the complexity of brain anatomy. Therefore, interpretation of 2D images requires specialized training. Therefore, volume reconstruction from sequential parallel 2D cross-sectional slices is a necessity for 3D tumor visualization, In 2013, the authors in [6] proposed an improved interpolation technique to estimate missing slices and the Marching Cubes (MC) algorithm to mesh the tumor. For surface rendering, they applied the Phong shading and lighting model to better compute the tumor volume.

The 3D reconstruction of tumors first requires an appropriate segmentation of the region of interest. This 3D reconstruction helps radiologists to better diagnose patients and subsequently remove the entire tumor when surgical intervention is considered. Techniques presented in [7,8] are based on preprocessing, image enhancement, and contouring prior to reconstruction. In 2012, authors in [9] used a technique based on phase-contrast projection tomography to calculate the 3D density distribution in bacterial cells. In addition, an approach proposed by [10] in 2019 provides a technique for segmenting brain tumors in the 3D volume using a 2D convolutional neural network for tumor prediction. Authors in [11,12] conducted a comparison between conventional machine learning based techniques and deep learning based techniques. The latter are further categorized into 2D CNN and 3D CNN techniques. However, the results of techniques based on deep convolutional neural networks out perform those of machine

[*] Corresponding Author: Abdennacer EL-OUARZADI, Université Moulay Ismail, a.elouarzadi@edu.umi.ac.ma

https://dx.doi.org/10.25046/aj090402

learning techniques. As for the authors in [13,14], they introduced a two-stage optimal mass transport technique (TSOMT), which involve stransforming an irregular 3D brain image into a cube with minimal deformation, for segmenting 3D medical images. Automatic segmentation of a brain tumor from two-dimensional slices (coronal, sagittal, and axial), facilitated by convolutional neural networks [15], significantly aids in delineating the region of interest in 3D.

Conventional holography was invented by [16] in 1948 during his research to improve the resolution of electron microscopes. This invention evolved in the 1960s with the advent of lasers, and holograms were recorded on plates or photosensitive films based primarily on silver ions that darkened under the influence of light. With the progress of high-resolution matrix detectors, digital holography was generalized in 1994 by the authors in [17], paving the way for numerous applications: holographic microscopy [18–20], quantitative phase imaging [21–23], color holography [24–26], metrology [27–29], holographic cameras [30], 3D displays [31–34], and head-up displays [35,36]. Authors in [37] were the first to use phase-shifting holography to eliminate unwanted diffraction orders from the hologram. They introduced spatial phase shifting using a piezoelectric transducer with a mounted mirror [38], and slight frequency modifications of acousto-optic modulators (AOMs) [39–41]. The latter technique is closely related to heterodyne detection methods.

Optical Scanning Holography (OSH) is considered an intelligent application for processing the pupil interaction. The pupil interaction scheme was implemented using optical heterodyne scanning by Korpel and Poon in 1979, and the use of an interaction scheme in a scanning illumination mode was developed by Indebetouw and Poon in 1992. By modifying one of the twolenses relative to the other (specifically, one lensis an open mask and the other is a pinhole mask) and defocusing the optical system, in 1985 author in [42] developed an optical scanning system capable of holographically recording a scanned object. This technique led to the invention of optical scanning holography. OSH has various applications, including optical scanning microscopy, 3D shape recognition, 3D holographic TV, 3D optical remote sensing, and more.

Early work on preprocessing in the OSH system dates back to 1985. Later, it was shown that placing a Gaussian annular aperture instead of a flat lens is useful for recovering the edge of a cross-sectional image in a hologram [43,44]. In 2010, authors in [45] demonstrated that by choosing a pupil function such as the Laplacian of the Gaussian, the performance of the method is an efficient means to extract the edge of a 3D scanned object by the OSH system. The authors in [46] proposed a 1D image acquisition system for auto stereoscopic display consisting of a cylindrical lens, a focusing lens, and an imaging device. By scanning an object over a wide angle, the synthesized image can beviewed as a 3D stereoscopic image.

The aim of this paper is to propose and develop a fully automatic 3D segmentation of brain tumors from magnetic resonance images. The proposed model is based on the same principle as our previous articles [47,48] and specifically for the segmentation of high and low grade glioblastoma brain tumors. To achieve this goal, the following contributions are made:

- Proposing a fully automated 3D method for brain tumor segmentation from MRI image sequences.
- Improving the conventional optical scanning holography technique by exploiting the properties of the cylindrical lens to optimize the scanning process and shorten the holographic recording process.
- Transitioning active contour theory from semi-automatic to fully automatic status with reliable tumor detection.
- Perform tests to demonstrate the effectiveness of our method using the well-known BraTS 2019 and BraTS 2020 databases.

2. Materials And Methods

2.1. Data used

The database of brain tumor images used in this study was obtained from the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2019 and MICCAI 2020 Multimodal Brain Tumor Segmentation Challenges, organized in collaboration with B. Menze, A. Jakab, S. Bauer, M. Reyes, M. Prastawa, and K. Van Leemput. It includes 40 glioma patients, including 20 high-grade (HG) and 20 low-grade (LG) cases, with four MRI sequences (T1, T1C, T2, and FLAIR) available for each patient. The challenge data base contains fully anonymized images collected from the ETH Zurich, the University of Debrecen, the University of Bern, and the University of Utah. All images are linearly co-registered and craniocaudally oriented. Institutional review board approval was not required as all human subject data are publicly available and de-identified [1,3].

2.2. Methodology

Figure 1 shows the optical scanning holography system used in our method. A laser beam of frequency ω isshifted in frequency to ψ and $\psi+\Delta\psi$ by Acousto-Optic Modulators (AOM) 1 and 2, respectively. The beams from the AOMs are then collimated by collimators BE1 and BE2. The out going beam from BE2 is considered as a plane wave of frequency $\omega+\psi+\Delta\psi$, which is projected onto the object by the x-y scanner. Our novel method involves the integration of a cylindrical lens L1 into the chosen imaging system, which provides a cylindrical wave at $\omega+\Omega$ that is projected onto the object. A focusing lens is also used to capture a large number of elementary images containing extensive parallax data. These elemental images are transformed into a matrix of elemental images, where each captured elemental image corresponds to a vertical line in ray space [46]. Using this linear scanning technique, object images are captured in a single pass rather than point-by-point, and the shape of the surface is adjusted after each iteration, saving computational time. Accordingly, after appropriate sampling for viewing conditions, we achieved fully automatic segmentation through the improved algorithm and arrangement of color filters. This allowed the transformation of two-dimensional element images into Three-Dimensional (3D) images, as shown in Figures 5 and 6.

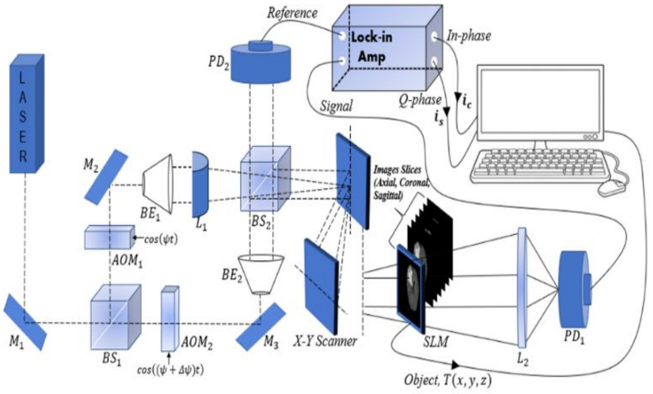


Figure 1: Schematic setup of the optical scanning holography (OSH).

The x-y scanner is used to scan the 3D object uniformly, line by line. As a result, each scan line of the object corresponds to a line in the hologram at the same vertical position. Along each scan line, Photo-Detectors PD1 and PD2 are used to capture the optical signal scattered by the object and the heterodyne frequency information ΔΩ as a reference signal, respectively, and convert them into electrical signals for the lock-in amplifier. The in-phase and quadrature-phase outputs of the lock-in amplifier circuit produce a sine hologram, H_sin (x, y), and a cosine hologram, H_cos (x, y), to achieve a complete 2D scan of the object, as shown below:

$$H(x,y) = H_{cos}(x,y) + jH_{sin}(x,y) = \sum_{k=0}^{N-1} H_k(x,y;\ z_k)$$

With :

$$i_c = \int \left\{ |T(x,y;z)|^2 * \frac{k_0}{2\pi z} \sin\left[\frac{k_0(x^2+y^2)}{2z}\right] \right\} dz = H_{sin\ (x,y)}$$

$$i_s = \int \left\{ |T(x,y;z)|^2 * \frac{k_0}{2\pi z} \cos\left[\frac{k_0(x^2+y^2)}{2z}\right] \right\} dz = H_{cos\ (x,y)}$$

2.3. Detection

Tumor contours are typically identified by fast pixel transitions, which indicate a significant change in information, while slow variations are eliminated by differentiation. Several methods are used for contour detection, including derivative methods based on evaluating the variation at each pixel by searching for maxima using a gradient or Laplacian filter.

The optical system in Figure 1 provides an output showing the distribution of both the phase and quadrature component of the heterodyne current. We will focus only on the phase component to extract the phase current after scanning the object images line by line. The maximum values characterizing this output are referred to as the peaks of the phase component.

Figure 2 illustrates how the initial contour C(i, j) is extracted from the peaks of the phase component maxima obtained from the optical scanning holography (OSH), i.e., the maximum of the intensity i_c with:

$$i_c = \int \left\{ |T(x,y;z)|^2 * \frac{k_0}{2\pi z} \sin\left[\frac{k_0(x^2+y^2)}{2z}\right] \right\} dz$$

Contours delineating these maxima are drawn, resulting in regions that designate the position of the tumor within the tumor tissue.

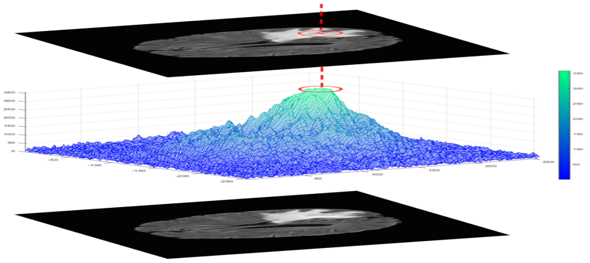


Figure 2: Preliminary extraction of the initial contour Ci inside the tumor tissue by OSH-based phase component peaks.

The output results of the Optical Scanning Holography(OSH) optical process are implemented numerically to extract the following parameters:

- C: the center of the tumor.
- L: the amplitude of the peak of the phase component.
- Ci: the initial contour formed using the principle illustrated in Figure 2.

Figures 3 and 4 provide examples of phase component peaks obtained using the OSH method.

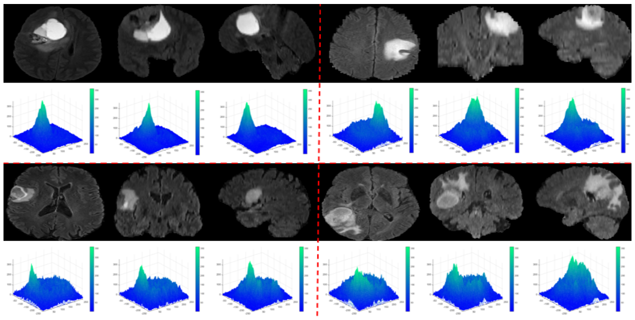


Figure 3: In-phase component peaks at the proposed OSH method's tumor position: examples of image slices (Axial, Coronal, Sagittal) from MICCAI 2019 database.

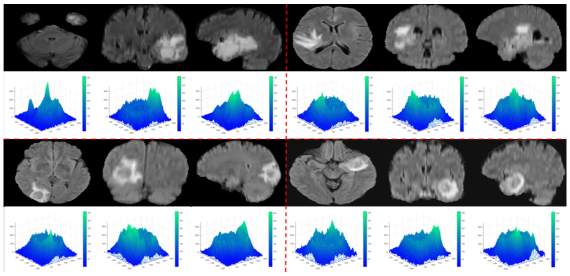


Figure 4: In-phase component peaks at the proposed OSH method's tumor position: examples of image slices (Axial, Coronal, Sagittal) from MICCAI 2020 database.

2.4. Segmentation and 3D Reconstruction

The segmentation of brain tumors using the active contour approach has received increasing attention. In this article, a fully automated active contour approach based on the OSH technique is proposed. This approach requires significant computational time. However, this technique allows a contour to be iteratively deformed to divide an image into several meaningful regions. Whitaker was able to reduce the computational time in his efficient

Sparse Field Method algorithm by accurately representing the target surface. In this method, each point of the target surface is processed, compared with the initialization curve, and updated after each iteration.

The novelty of our research is to apply the OSH technique by integrating a cylindrical lens into our system to improve the detection phase in terms of precision and acquisition time by processing the image of the region of interest line by line (line-by-line scanning).

It is true that the calculation of the terms of the active contour energy, as shown below, has allowed us to achieve an exact and precise segmentation of brain tumors. However, this segmentation is conditioned by the choice of an initial contour Ci,j, which gives it a semi-automatic nature :

$$E_{i,j} = \beta.\left|I - M_{i,j}\right|^2 + \gamma.\left|I - m_{i,j}\right|^2$$

In our work, we have proposed a model of active contours combined with the OSH approach, which enables us to address the issue of manual supervision in selecting the initial contour, especially for complex tumor shapes. We have achieved, through the use of the maximum phase current of the OSH, which corresponds to the initial contour of the active contour model, an improvement in the detection of tumor tissue contours. By adding the term"$\alpha.C_{i,j}$"derivedfrom the OSH technique, we were able to automate our model :

$$E_{i,j} = \alpha.C_{i,j} + \beta.\left|I - M_{i,j}\right|^2 + \gamma.\left|I - m_{i,j}\right|^2$$

This article builds upon previous work in the field of 3D reconstruction, primarily by combining it with enhanced detection and segmentation techniques. Indeed, this new technique improves the computational efficiency and precision in selecting the pixels that are crucial for reconstructing 3D object shapes.

The results of reconstructing 3D object images from a real patient dataset are presented in Figures 5 and 6.

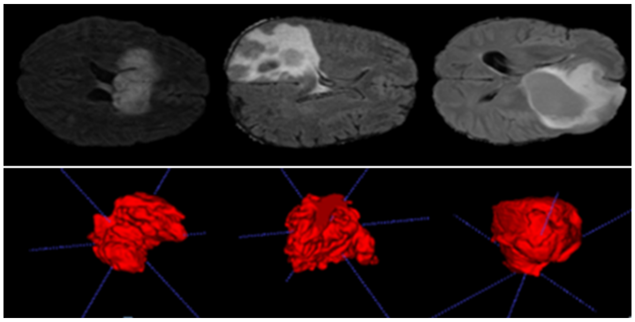


Figure 5: 3-D reconstruction results of real patient data from the MICCAI 2019 database.

3. Experimental Results

3.1. evaluation of detection phase

To extract brain tumors accurately, it is essential to reliably determine the parameter L (where L represents the maximum peak of the phase component). That's why we studied the values of L on different MRI images from Brats-2018 and Brats-2019 containing tumor tissues.

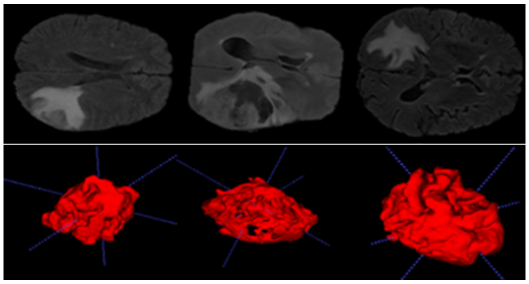


Figure 6: 3-D Reconstruction results of real patient data from the MICCAI 2020 database.

We observe that the values of L in the case of tumors are significantly separated from those of healthy brain tissue. We also notice that all the maxima of the phase components provided by the OSH process, used for tumor detection, fall within the range [300;350], while for healthy tissue, they are in the interval [100;150].

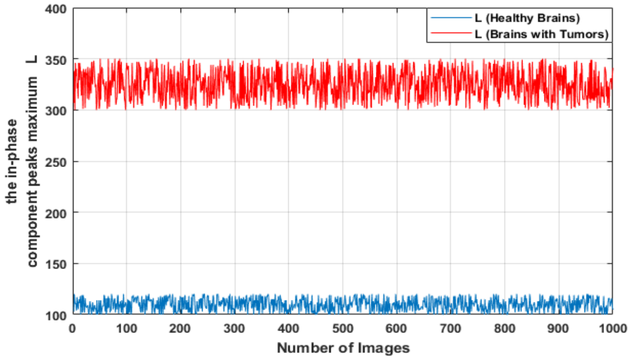


Figure 7: Distribution of the L parameter in the healthy and tumorous brain image slices (Axial, Coronal, Sagittal).

3.2. Evaluation of segmentation phase

Figure 8 displays box plots for the sensitivity, Dice score, specificity, and Hausdorff distance obtained by our method on the Brats-2019 and Brats-2020 datasets. The performance of our method is compared to that of the Geodesic Active Contour (GAC) model, Localized Active Contour (LAC), and Cuckoo-driven Active Contour (ACCS) models. We note that the performance of the proposed methodis the most satisfactory in terms of segmentation versus the other methods compared, particularly in terms of sensitivity, specificity and Dice score. The statistical computation time of the most relevant steps of our algorithm per MRI image is given in figure 9.

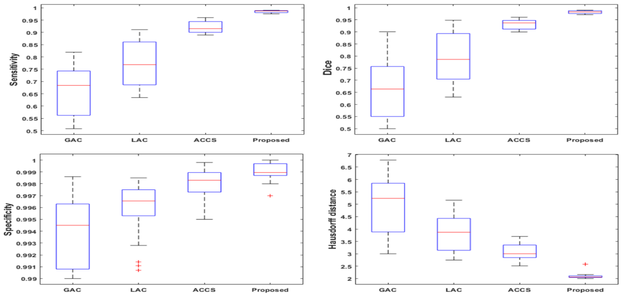


Figure 8: Boxplots of evaluation scores: Dice, sensitivity, specificity and hausdorff distance for the four approaches examined.

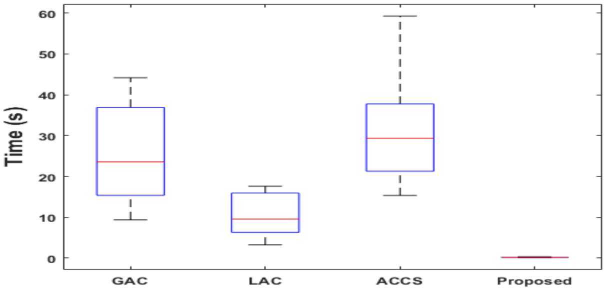


Figure 9: Boxplots of evaluation scores: evaluation time for the four approaches examined.

3.3. Evaluation of reconstruction phase

The data in tables (1 and 2) enable us to obtain the tumor volume for each patient with very respectable accuracy, making it easier to estimate the degree of cancer. These tables also provide useful information such as brain volume and mean intensity for each patient label (brain label and tumor label).

Table 1: 3-D Segmentation results of real patient data from the BRATS 2019 database.

Patients	Labels	Voxel count	Volume (mm^3)	Intensity Mean ±SD
Patient 1	Clear Label	8 812 673	8.812673 x10^6	32.4629 ± 76.7841
	Label with tumor	115 327	1.15327 x10^5	434.7898 ± 75.9586
Patient 2	Clear Label	8 908 742	8.908742 x10^6	31.0648 ± 66.3003
	Label with tumor	19 258	1.9258 x10^4	424.8549 ± 65.4917
Patient 3	Clear Label	8 896 112	8.896112 x10^6	24.4006 ± 649036
	Label with tumor	31 888	3.1888 x10^4	451.9312 ± 57.2040
Patient 4	Clear Label	8 893 509	8.893509 x10^6	36.5984 ± 90.5500
	Label with tumor	34 491	3.4491 x10^4	1137.5650 ± 202.0132
Patient 5	Clear Label	8 874 450	8.874450 x10^6	34.4983 ± 78.4227
	Label with tumor	53 550	5.3550 x10^4	432.8664 ± 75.4400
Patient 6	Clear Label	8 815 463	8.815463 x10^6	31.0447 ± 74.2778
	Label with tumor	112 537	1.12537 x10^5	451.1814 ± 65.6239
Patient 7	Clear Label	8 800 705	8.800705 x10^6	21.7560 ± 56.5531
	Label with tumor	127 295	1.27295 x10^5	305.4569 ± 57.3601
Patient 8	Clear Label	8 872 783	8.872783 x10^6	29.3676 ± 68.4357
	Label with tumor	55 217	5.5217 x10^4	334.1023 ± 48.2186
Patient 9	Clear Label	8 920 644	8.920644 x10^6	11.8151 ± 32.1527
	Label with tumor	7 356	7.356 x10^3	198.0174 ± 38.5634
Patient 10	Clear Label	8 911 678	8.911678 x10^6	23.9592 ± 59.2099
	Label with tumor	16 322	1.6322 x10^4	454.6847 ± 83.4070
Patient 11	Clear Label	8 900 834	8.900834 x10^6	35.1428 ± 75.6676
	Label with tumor	27 166	2.7166 x10^4	344.4761 ± 46.2303
Patient 12	Clear Label	8 833 615	8.833615 x10^6	64.3847 ± 39.4228
	Label with tumor	94 385	9.4385 x10^4	580.9401 ± 36.3711
Patient 13	Clear Label	8 876 743	8.876743 x10^6	68.8470 ± 153.2978
	Label with tumor	51 257	5.1257 x10^4	698.3368 ± 120.9774
Patient 14	Clear Label	8 698 867	8.698867 x10^6	25.9308 ± 57.5036
	Label with tumor	229 131	2.29131 x10^5	304.6895 ± 70.7478
Patient 15	Clear Label	8 778 460	8.778460 x10^6	44.1679 ± 112.9932
	Label with tumor	149 540	1.49540 x10^5	442.8497 ± 45.4427
Patient 16	Clear Label	8 801 474	8.801474 x10^6	18.6908 ± 43.6618
	Label with tumor	126 526	1.26526 x10^5	241.1445 ± 36.7579
Patient 17	Clear Label	8 690 176	8.690176 x10^6	16.6104 ± 39.7039
	Label with tumor	237 824	2.37824 x10^5	213.0597 ± 45.3580
Patient 18	Clear Label	8 833 752	8.833852 x10^6	16.8981 ± 43.7038
	Label with tumor	94 248	9.8248 x10^4	270.6864 ± 44.3378
Patient 19	Clear Label	8 699 808	8.699808 x10^6	16.7930 ± 41.8654
	Label with tumor	228 192	2.28192 x10^5	227.5470 ± 37.6771
Patient 20	Clear Label	8 902 965	8.902965 x10^6	47.3972 ± 115.0547
	Label with tumor	25 035	2.5035 x10^4	477.4638 ± 28.0403

Table 2: 3-D Segmentation results of real patient data from the BRATS 2020 database.

Patients	Labels	Voxel count	Volume (mm^3)	Intensity Mean ±SD
Patient 1	Clear Label	8796585	8.796585 x10^6	32.0395 ± 76.2065
	Label with tumor	131415	1.31415 x10^5	413.8778 ± 90.8952
Patient 2	Clear Label	8899672	8.899672 x10^6	30.7910 ± 65.7704
	Label with tumor	28328	2.8328 x10^4	384.8039 ± 81.0838
Patient 3	Clear Label	8880808	8.880808 x10^6	25.9012 ± 63.8240
	Label with tumor	47192	4.7192 x10^4	4079188 ± 80.5098
Patient 4	Clear Label	8875910	8.875910 x10^6	35.4049 ± 86.3327
	Label with tumor	52090	5.2090 x10^4	968.9609 ± 299.9311
Patient 5	Clear Label	8863102	8.863102 x10^6	34.2085 ± 78.1643
	Label with tumor	64898	6.4898 x10^4	402.7914 ± 81.2011
Patient 6	Clear Label	8794503	8.794503 x10^6	30.4320 ± 73.2910
	Label with tumor	133497	1.33497 x10^5	425.5790 ± 84.9031
Patient 7	Clear Label	8828307	8.828307 x10^6	22.3986 ± 57.6266
	Label with tumor	99693	9.9693 x10^4	327.0939 ± 44.3143
Patient 8	Clear Label	8777055	8.777055 x10^6	21.3723 ± 53.5396
	Label with tumor	150945	1.50945 x10^5	347.9547 ± 73.5984
Patient 9	Clear Label	8859507	8.859507 x10^6	29.0185 ± 67.8758
	Label with tumor	68493	6.8493 x10^4	320.2008 ± 54.0278
Patient 10	Clear Label	8856743	8.856743 x10^6	30.1615 ± 71.5588
	Label with tumor	71257	7.1257 x10^4	410.5989 ± 110.6990
Patient 11	Clear Label	8852777	8.852777 x10^6	41.6819 ± 100.7240
	Label with tumor	75223	7.5223 x10^4	568 0779 ± 103.8634
Patient 12	Clear Label	8825129	8.825129 x10^6	70.7885 ± 174.1385
	Label with tumor	102871	1.02871 x10^5	761.5829 ± 89.2313
Patient 13	Clear Label	8811475	8.811475 x10^6	19.5354 ± 43.5697
	Label with tumor	116525	1.16525 x10^5	226.5449 ± 29.4236
Patient 14	Clear Label	8829255	8.829255 x10^6	17.1114 ± 39.8459
	Label with tumor	98745	9.8745 x10^4	223.5896 ± 27.5731
Patient 15	Clear Label	8907363	8.907363 x10^6	41.1899 ± 100.3383
	Label with tumor	20637	2.0637 x10^4	405..570 ± 35.2389
Patient 16	Clear Label	8781794	8.781794 x10^6	19.6149 ± 45.3264
	Label with tumor	146206	1.46206 x10^5	243.5090 ± 39.3731
Patient 17	Clear Label	8852596	8.852596 x10^6	17.1913 ± 45.4955
	Label with tumor	75404	7.5404 x10^4	279.9906 ± 49.1193
Patient 18	Clear Label	8831696	8.831696 x10^6	37.5604 ± 103.4473
	Label with tumor	96304	9.6304 x10^4	390.8695 ± 40.2213
Patient 19	Clear Label	8823355	8.823355 x10^6	75.8865 ± 180.9212
	Label with tumor	104645	1.04645 x10^5	649.1030 ± 55.8957
Patient 20	Clear Label	8829763	8.829763 x10^6	9.3929 ± 32.4282
	Label with tumor	98237	9.8237 x10^4	229.8103 ± 44.2263

4. Conclusion

The aim of this article is to develop a 3D reconstruction and quantification approach to facilitate physician surgical planning and tumor volume calculation. The three-dimensional model of the brain tumor was reconstructed from a given set of two-dimensional brain slices (axial, coronal, and sagittal). In the slices containing tumors, the improved Optical Scanning Holography (OSH) method helped us to extract the maximum component in phase, and at the same time an Active Contour Model (ACM) was applied to this area of interest to perform a faster segmentation of the region corresponding to the tumors in each slice. We obtained satisfactory results with different active contour models with different similarity parameters on database images (BRATS 2019 and 2020), compared to several state-of-the-art brain tumor segmentation methods.

The application of this method does, however, present a number of limitations. Firstly, integration with existing diagnostic and processing systems can pose challenges, particularly with regard to the compatibility of data formats and communication protocols. In addition, the quality and resolution of MRI images are crucial to the success of the method. Low-quality images can result in imprecise segmentation and inaccurate 3D reconstruction, compromising the accuracy and clinical utility of the method. Therefore, high-quality images and resolution of integration issues are essential to maximize the effectiveness of this approach in a clinical setting.

To improve this method, it is necessary to validate its efficacy more widely and robustly in a variety of clinical settings. In the future, it will be necessary to refine the technology for greater accuracy, improve the quality of MRI images, and ensure seamless integration with other tools for diagnosing and treating brain tumors by resolving data compatibility issues. These efforts will open up new prospects for research and innovation in this field.

Conflict of Interest

All authors disclose any financial and personal relationships with other people or organizations that could inappropriately influence our work.

The authors declare no conflict of interest.

Acknowledgment

Authors thank Raphael Meier, University of Bern, Medical Image Analysis Institute for Surgical Technology and Biomechanics Switzerland, for securing available access to the Multimodal Brain Tumor Segmentation Benchmark (BRATS).

References

[1] T.J. Hudson, "The International Cancer Genome Consortium," Cancer Research, **176**(1), 139–148, 2009, doi:10.1038/nature08987.International.

[2] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, J.B. Freymann, K. Farahani, C. Davatzikos, "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," Scientific Data, **4**(1), 170117, 2017, doi:10.1038/sdata.2017.117.

[3] R. Ranjbarzadeh, A. Bagherian Kasgari, S. Jafarzadeh Ghoushchi, S. Anari, M. Naseri, M. Bendechache, "Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images," Scientific Reports, **11**(1), 10930, 2021, doi:10.1038/s41598-021-90428-8.

[4] R.P.J., "BRAIN TUMOR MRI IMAGE SEGMENTATION AND DETECTION IN IMAGE PROCESSING," International Journal of Research in Engineering and Technology, **03**(13), 1–5, 2014, doi:10.15623/ijret.2014.0313001.

[5] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B.B. Avants, N. Ayache, P. Buendia, D.L. Collins, N. Cordier, J.J. Corso, A. Criminisi, T. Das, H. Delingette, C. Demiralp, C.R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," IEEE Transactions on Medical Imaging, **34**(10), 1993–2024, 2015, doi:10.1109/TMI.2014.2377694.

[6] The Efficiency of 3D-Printed Dog Brain Ventricular Models from 3 Tesla (3T) Magnetic Resonance Imaging (MRI) for Neuroanatomy Education, The Pakistan Veterinary Journal, 2024, doi:10.29261/pakvetj/2024.168.

[7] M. Bartels, M. Priebe, R.N. Wilke, S.P. Krüger, K. Giewekemeyer, S. Kalbfleisch, C. Olendrowitz, M. Sprung, T. Salditt, "Low-dose three-dimensional hard x-ray imaging of bacterial cells," Optical Nanoscopy, **1**(1), 10, 2012, doi:10.1186/2192-2853-1-10.

[8] K. Pawar, Z. Chen, N. Jon Shah, G.F. Egan, An Ensemble of 2D Convolutional Neural Network for 3D Brain Tumor Segmentation, 359–367, 2020, doi:10.1007/978-3-030-46640-4_34.

[9] M. kamal Al-anni, P. DRAP, "Efficient 3D Instance Segmentation for Archaeological Sites Using 2D Object Detection and Tracking," International Journal of Computing and Digital Systems, **15**(1), 1333–1342, 2024, doi:10.12785/ijcds/150194.

[10] G. Gunasekaran, M. Venkatesan, "An Efficient Technique for Three-Dimensional Image Visualization Through Two-Dimensional Images for Medical Data," Journal of Intelligent Systems, **29**(1), 100–109, 2019, doi:10.1515/jisys-2017-0315.

[11] H. Khan, S.F. Alam Zaidi, A. Safi, S. Ud Din, A Comprehensive Analysis of MRI Based Brain Tumor Segmentation Using Conventional and Deep Learning Methods, 92–104, 2020, doi:10.1007/978-3-030-43364-2_9.

[12] V. Kumar, T. Lal, P. Dhuliya, D. Pant, "A study and comparison of different image segmentation algorithms," in 2016 2nd International Conference on Advances in Computing, Communication, & Automation (ICACCA) (Fall), IEEE: 1–6, 2016, doi:10.1109/ICACCAF.2016.7749007.

[13] M. Eliasof, A. Sharf, E. Treister, "Multimodal 3D Shape Reconstruction under Calibration Uncertainty Using Parametric Level Set Methods," SIAM Journal on Imaging Sciences, **13**(1), 265–290, 2020, doi:10.1137/19M1257895.

[14] W.-W. Lin, C. Juang, M.-H. Yueh, T.-M. Huang, T. Li, S. Wang, S.-T. Yau, "3D brain tumor segmentation using a two-stage optimal mass transport algorithm," Scientific Reports, **11**(1), 14686, 2021, doi:10.1038/s41598-021-94071-1.

[15] P. Li, W. Wu, L. Liu, F. Michael Serry, J. Wang, H. Han, "Automatic brain tumor segmentation from Multiparametric MRI based on cascaded 3D U-Net and 3D U-Net++," Biomedical Signal Processing and Control, **78**, 103979, 2022, doi:10.1016/j.bspc.2022.103979.

[16] D. GABOR, "A New Microscopic Principle," Nature, **161**(4098), 777–778, 1948, doi:10.1038/161777a0.

[17] U. Schnars, W. Jüptner, "Direct recording of holograms by a CCD target and numerical reconstruction," Applied Optics, **33**(2), 179, 1994, doi:10.1364/AO.33.000179.

[18] W. Xu, M.H. Jericho, I.A. Meinertzhagen, H.J. Kreuzer, "Digital in-line holography for biological applications," Proceedings of the National Academy of Sciences, **98**(20), 11301–11305, 2001, doi:10.1073/pnas.191361398.

[19] P. Marquet, B. Rappaz, P.J. Magistretti, E. Cuche, Y. Emery, T. Colomb, C. Depeursinge, "Digital holographic microscopy: a noninvasive contrast imaging technique allowing quantitative visualization of living cells with subwavelength axial accuracy," Optics Letters, **30**(5), 468, 2005, doi:10.1364/OL.30.000468.

[20] J. Garcia-Sucerquia, W. Xu, S.K. Jericho, P. Klages, M.H. Jericho, H.J. Kreuzer, "Digital in-line holographic microscopy," Applied Optics, **45**(5), 836, 2006, doi:10.1364/AO.45.000836.

[21] E. Cuche, F. Bevilacqua, C. Depeursinge, "Digital holography for quantitative phase-contrast imaging," Optics Letters, **24**(5), 291, 1999, doi:10.1364/OL.24.000291.

[22] B. Rappaz, P. Marquet, E. Cuche, Y. Emery, C. Depeursinge, P.J. Magistretti, "Measurement of the integral refractive index and dynamic cell morphometry of living cells with digital holographic microscopy," Optics Express, **13**(23), 9361, 2005, doi:10.1364/OPEX.13.009361.

[23] B. Kemper, G. von Bally, "Digital holographic microscopy for live cell applications and technical inspection," Applied Optics, **47**(4), A52, 2008, doi:10.1364/AO.47.000A52.

[24] N. Demoli, D. Vukicevic, M. Torzynski, "Dynamic digital holographic interferometry with three wavelengths," Optics Express, **11**(7), 767, 2003, doi:10.1364/OE.11.000767.

[25] J. Zhao, H. Jiang, J. Di, "Recording and reconstruction of a color holographic image by using digital lensless Fourier transform holography," Optics Express, **16**(4), 2514, 2008, doi:10.1364/OE.16.002514.

[26] P. Tankam, P. Picart, D. Mounier, J.M. Desse, J. Li, "Method of digital holographic recording and reconstruction using a stacked color image sensor," Applied Optics, **49**(3), 320, 2010, doi:10.1364/AO.49.000320.

[27] F. Charrière, J. Kühn, T. Colomb, F. Montfort, E. Cuche, Y. Emery, K. Weible, P. Marquet, C. Depeursinge, "Characterization of microlenses by digital holographic microscopy," Applied Optics, **45**(5), 829, 2006, doi:10.1364/AO.45.000829.

[28] J.-M. Desse, P. Picart, P. Tankam, "Digital three-color holographic interferometry for flow analysis," Optics Express, **16**(8), 5471, 2008, doi:10.1364/OE.16.005471.

[29] P. Tankam, Q. Song, M. Karray, J. Li, J. Michel Desse, P. Picart, "Real-time three-sensitivity measurements based on three-color digital Fresnel holographic interferometry," Optics Letters, **35**(12), 2055, 2010, doi:10.1364/OL.35.002055.

[30] K. Kim, H. Yoon, M. Diez-Silva, M. Dao, R.R. Dasari, Y. Park, "High-resolution three-dimensional imaging of red blood cells parasitized by Plasmodium falciparum and in situ hemozoin crystals using optical diffraction tomography," Journal of Biomedical Optics, **19**(01), 1, 2013, doi:10.1117/1.JBO.19.1.011005.

[31] F. Yaraş, H. Kang, L. Onural, "Real-time phase-only color holographic video display system using LED illumination," Applied Optics, **48**(34), H48, 2009, doi:10.1364/AO.48.000H48.

[32] P.-A. Blanche, A. Bablumian, R. Voorakaranam, C. Christenson, W. Lin, T. Gu, D. Flores, P. Wang, W.-Y. Hsieh, M. Kathaperumal, B. Rachwal, O. Siddiqui, J. Thomas, R.A. Norwood, M. Yamamoto, N. Peyghambarian, "Holographic three-dimensional telepresence using large-area photorefractive polymer," Nature, **468**(7320), 80–83, 2010, doi:10.1038/nature09521.

[33] J. Geng, "Three-dimensional display technologies," Advances in Optics and Photonics, **5**(4), 456, 2013, doi:10.1364/AOP.5.000456.

[34] M. Kujawinska, T. Kozacki, C. Falldorf, T. Meeser, B.M. Hennelly, P. Garbat, W. Zaperty, M. Niemelä, G. Finke, M. Kowiel, T. Naughton, "Multiwavefront digital holographic television," Optics Express, **22**(3), 2324, 2014, doi:10.1364/OE.22.002324.

[35] H. Mukawa, K. Akutsu, I. Matsumura, S. Nakano, T. Yoshida, M. Kuwahara, K. Aiki, M. Ogawa, "8.4: Distinguished Paper : A Full Color Eyewear Display Using Holographic Planar Waveguides," SID Symposium Digest of Technical Papers, **39**(1), 89–92, 2008, doi:10.1889/1.3069819.

[36] C. Jang, C.-K. Lee, J. Jeong, G. Li, S. Lee, J. Yeom, K. Hong, B. Lee, "Recent progress in see-through three-dimensional displays using holographic optical elements [Invited]," Applied Optics, **55**(3), A71, 2016, doi:10.1364/AO.55.000A71.

[37] I. Yamaguchi, Phase-Shifting Digital Holography, Springer US: 145–171, 2006, doi:10.1007/0-387-31397-4_5.

[38] I. Yamaguchi, J. Kato, S. Ohta, J. Mizuno, "Image formation in phase-shifting digital holography and applications to microscopy," Applied Optics, **40**(34), 6177, 2001, doi:10.1364/AO.40.006177.

[39] F. Le Clerc, L. Collot, M. Gross, "Numerical heterodyne holography with two-dimensional photodetector arrays," Optics Letters, **25**(10), 716, 2000, doi:10.1364/OL.25.000716.

[40] E. Absil, G. Tessier, M. Gross, M. Atlan, N. Warnasooriya, S. Suck, M. Coppey-Moisan, D. Fournier, "Photothermal heterodyne holography of gold nanoparticles," Optics Express, **18**(2), 780, 2010, doi:10.1364/OE.18.000780.

[41] B. Samson, F. Verpillat, M. Gross, M. Atlan, "Video-rate laser Doppler vibrometry by heterodyne holography," Optics Letters, **36**(8), 1449, 2011, doi:10.1364/OL.36.001449.

[42] T.-C. Poon, Three-dimensional image processing and optical scanning holography, 329–350, 2003, doi:10.1016/S1076-5670(03)80018-6.

[43] Y. Zhang, R. Wang, P. Tsang, T.-C. Poon, "Sectioning with edge extraction in optical incoherent imaging processing," OSA Continuum, **3**(4), 698, 2020, doi:10.1364/OSAC.383473.

[44] G. Indebetouw, W. Zhong, D. Chamberlin-Long, "Point-spread function synthesis in scanning holographic microscopy," Journal of the Optical Society of America A, **23**(7), 1708, 2006, doi:10.1364/JOSAA.23.001708.

[45] X. Zhang, E.Y. Lam, "Edge detection of three-dimensional objects by manipulating pupil functions in an optical scanning holography system," in 2010 IEEE International Conference on Image Processing, IEEE: 3661–3664, 2010, doi:10.1109/ICIP.2010.5652483.

[46] Y. Momonoi, K. Taira, Y. Hirayama, "Scan-type image capturing system using a cylindrical lens for one-dimensional integral imaging," in: Woods, A. J., Dodgson, N. A., Merritt, J. O., Bolas, M. T., and McDowall, I. E., eds., in Stereoscopic Displays and Virtual Reality Systems XIV, 649017, 2007, doi:10.1117/12.703170.

[47] A. Cherkaoui, A. El-Ouarzadi, A. Essadike, Y. Achaoui, A. Bouzid, "Brain Tumor Detection and Segmentation Using a Hybrid Optical Method by Active Contour," in 2023 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA), IEEE: 132–138, 2023, doi:10.1109/ICDATA58816.2023.00032.

[48] A. EL-OUARZADI, A. Cherkaoui, A. Essadike, A. Bouzid, "Brain Tumor Detection and Segmentation Using a Hybrid Optical Method by Active Contour," Available at SSRN 4629062, doi:10.2139/ssrn.4629062.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 4, 14-20 (2024)

www.astesj.com

ASTESJ

ISSN: 2415-6698

From Sensors to Data: Model and Architecture of an IoT Public Network

Stefania Nanni[1], Massimo Carboni[1], Gianluca Mazzini[2]

[1] *LepidaScpa, Research & Prototypes, Bologna, 40128, Italy*

[2] *LepidaScpa, CEO, Bologna, 40128, Italy*

ARTICLE INFO

Article history:
Received: 03 April, 2024
Revised: 28 May,2024
Accepted: 29 June, 2024
Online: 11 July, 2024

Keywords:
IoT network
Decoder
Data extraction

ABSTRACT

RetePAIoT of Emilia-Romagna region is an IoT Public Network, financed by Emilia-Romagna Region and developed by Lepida Scpa, where citizens, private companies and Public Administrations can integrate free of charge their own sensors of any type and anywhere in the region. The main objective of the project is to provide a facility to implement the IoT paradigm, receiving data potentially from thousands of different sensors from the territory and to make them available to their owners and, in aggregate or anonymous form, to Public Administrations for their institutional purposes. In this context the interpretation of payloads sent by sensors, i.e. the extraction of the values measured by sensors, as well sharing them with all authorized subjects. are fundamental aspects that present a significant complexity due to the variety and unplannable context of the project. This paper illustrates the model and the architecture of a solution for the automatic extraction of values potentially coming from thousands of different sensors, which therefore requires a high level of flexibility, robustness and scalability as well as different methods for sharing them with third parties, depending on purposes and technical level required.

1. Introduction

As illustrated in the original paper [1], in 2019 LepidaScpa launched a project, financed by Emilia-Romagna Region, called retePAIoT, aimed at covering the entire regional territory with a LoRaWan network [2], offering an IoT infrastructure to both Public Administrations and to private citizens or companies to connect their own sensors.

Several actors are participating to this project, with different roles:

- LepidaScpA is in charge of the development and the maintenance of the IoT infrastructure: gateways, server network, included the web platform for the management of sensors, their data and users;
- Municipalities have to provide free of charge places of installation of gateways and to advertise the initiative on their own territory;

- Municipalities have to provide free of charge places of installation of gateways and to advertise the initiative on their own territory;

- End users (both public and private) have to purchase sensors. Type and model of sensors can be any, but users have to provide LepidaScpa with information on those not yet present in the catalog, to be update. Users, in fact, have also to register their own sensors into retePAIoT network, through retePAIoT web interface [3], specifying the correct brand and model of sensors, from which decoding rules of payload of sensors messages depend.

The main objective of the retePAIoT project is to make available a free of charge IoT regional network to Public Administrations as well as private users or companies, to facilitate the installation of their own sensors, collecting data from the territory and making them available to owners of the sensors and to Public Administrations for their institutional purposes [4]. This means that the measurements sent by the sensors integrated in the

* Corresponding Author: Stefania Nanni, Lepida ScpA; stefania.nanni@lepida.it

https://dx.doi.org/10.25046/aj090403

public IoT network, retePAIoT, can be consulted both by owners of the sensors, but also, anonymously or in aggregate form, by the Public Administrations interested in exploiting the data coming from the regional territory. This important goal is possible only if the retePAIoT project, in addition to facilitate the installation and the collection of data sent by the sensors, also provides their decoding, i.e. the extraction of the measured values, making them usable, through appropriate interfaces, both to the owners of the sensors and to the Public Administrations.

To encourage the use of the retePAIoT project by all interested public and private users, no limits have been set to the types, models and brands of sensors that can be used. For this reason, the catalogue of the sensors that can be registered in retePAIoT is constantly updated upon users request with new models and types.

It is important to underline that other LoRaWan networks exist in Europe and around the world which are public and free and which allow different types of sensors to be integrated by different users. An exemplary case is the TTN network, The Thing Network [5], which, in addition to being widespread in Europe, especially in France, and in the world, constitutes a global network for the integration of both sensors and LoRaWan gateways. But while both networks, TTN and retePAIoT, share the primary purpose of providing a network infrastructure that is accessible to anyone at no cost, promoting innovation and widespread adoption of IoT technology, retePAIoT has the additional and fundamental objective to use the data collected by the sensors.

This purpose is not only peculiar to the retePAIoT network, compared to other public ones, which normally delegate it to single users, but it involves a series of onerous activities such as knowledge, cataloguing, validation, description of the measurements and the corresponding units within the Data Base of all integrated sensors, as well as the implementation of a new architecture for the automatic extraction of data from the messages sent by the sensors, which constitutes the real added value and aspect innovative of the solution presented in the original paper [1] and in this one.

In this paper it will be highlighted the importance, the critical issues and the automatic solution adopted for the decoding of the payloads and for the extraction of the data within a complex scenario such as an IoT public network like retePAIoT and some different methods and interfaces to share them to different users, according to their needs.

For this purpose, this paper starts form a brief overview of the state of the art of some useful management features offered by ChirpStack [6], that is the open LoRaWan server network used to manage retePAIoT network, and on which the model and architecture of the solution proposed in this paper is based. The rest of this paper is organised as follows: the third section briefly illustrates the main architecture of the Internet of Things public network, retePAIoT; section IV describes the structure of the centralized database, with particular reference to the fundamental extension of sensors registry tables introduced to resolve criticalities and to guarantee the requirements of payload decoding service; section V focuses on the new architecture implemented to automatically decode payloads sent by sensors, based on a new feature made available by ChirpStack, its logical flow, closely based on sensors database model extended, and the relevant goals and advantages that it achieves respect the original one [1]; section VI describes different interfaces made available by retePAIoT network to share data to different users and objectives; section VII describes a significant use data case, and the last one summarises the main results achieved by retePAIoT network.

2. The State of the Art

ChirpStack is an open-source platform for managing and monitoring LoRaWan networks. It provides a modular software suite that enables network operators, developers and end users to build and manage scalable and reliable LoRaWan networks. ChirpStack Application Server is a module of ChirpStack that manages IoT applications, allowing developers to create and manage custom applications to analyze and interact with data from LoRaWan devices. In particular, the 'device profile' and the 'application' are two entities, managed by the Application server module, which respectively allow to specify the communication and configuration settings of the devices and to manage a specific IoT application or use case within the ChirpStack platform.

Starting from ChirpStack version 3.0, decoding rules of payloads devices have been associated to the 'device profile' entity instead of the ‘application’ one, making it possible to avoid duplication of decoding rules in the very common cases in which devices of the same type were used in applications associated to different users.

The new positioning of the payload decoding function constitutes the opportunity to provide the decoding of the payloads of the various devices directly within the ChirpStack, in place of an external module, using directly the javascript code, normally provided by the devices manufacturers, instead of requiring the development of different software modules, as implemented in the first release of retePAIoT [1]. ‘Application' is now only a logical entity that allows network operators to create and to manage customized IoT applications according to the specific needs of different use cases and users, relatively, for example, to different modalities to share data of devices associated with other external systems. The two entities ‘device profile’ and ‘application’ managed by ChripStack in last releases are at the base of improved model and architecture of retePAIoT platform, for the extraction and the sharing of the data, that are two fundamental aspects for the use of sensors data, which constitutes the final objective of retePAIoT and, in general of all IoT networks.

3. retePAIoT Network Architecture

The basic architecture of a public network for the internet of things like retePAIoT is shown in Fig.1, where main components, sensors, gateways and network server, and types of connections are highlighted: the black line represents an internet connection, the violet line a Lepida fiber connection, the dotted line a communication based on wireless LoraWan protocol, whose specifications can be found in [7].

One or more LoraWan gateway are installed in each participating municipality, and represent the meeting point between two types of communication: on the one hand, through the LoraWAN protocol, they receive the data coming from the users sensors, on the other they are connected to one of the fiber points of access of the Lepida Network, allowing the sensors messages to reach the LoraWan server.

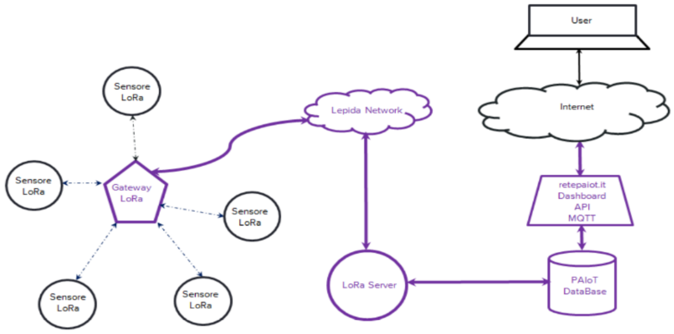


Figure 1: Architecture of internet of things public network, retePAIoT

This server is a virtual machine hosted in one of the four Regional Data Centers managed by LepidaScpA [8], and it undertakes the registration of sensors in the network and the subsequent establishment of encrypted sessions through which payloads are transmitted. The software installed for the management of the protocol is the open source LoRa Server ChirpStack [6].

Every user must register himself and the sensors, for which he is responsible or owner, through the web portal, another virtual machine in the same Datacenter where the LoRaServer is hosted.

The lack of a communication standard shared between all IoT devices is a problem whose solution strongly depends on the technology and the scope in which you are working [9], [10], [11]. The major criticality addressed in this paper arises from the need to postpone the decoding process with respect to that of messages reception, in order to better comply with the nature of the retePAIoT network which is free, open to any brand and model that users may require and mindful of the decoupling between raw and processed information.

In the following paragraphs, the architectural and functional aspects to save sensors payloads and to interpret and extract their data and different methods to share them will be examined in depth, because, as already highlighted, they constitute essential services of the project that present some elements of complexity which require not only a critical and in-depth analysis, but it also deserves a specific focus, that this paper intends to highlight, together with the solution devised and implemented to resolve them.

4. Extended DataBase

The database of retePAIoT network is an Oracle relational database that allows to manage both sensors and users data and to store the payloads and the values of the measured quantities, as shown in Figure 2, as already illustrated in a previous paper [3].

The next one [1], instead, focused on extension of some tables of the database provided for the description of the sensors, 'sensor', and for saving of their messages, ‘sensor-value’, which became necessary to deal with some cricities related to the decoding process. In this paragraph it is considered useful to summarise the critical aspects that made it necessary and above all the advantages that derived from it.

The problems that managers of retePAIoT network have to face and have to solve in retePAIoT scenario are basically three:

1. check of the correctness of the brand and the model specified by the user for a sensor, through the comparison between the format of expected payload and that of payloads actually received
2. description of all measures, provided by a sensor, in ‘sensor-measure’ table, according to sensor brand and model
3. implementation of the rules for extracting and storing the values from payloads for every new type, brand a model of a sensor

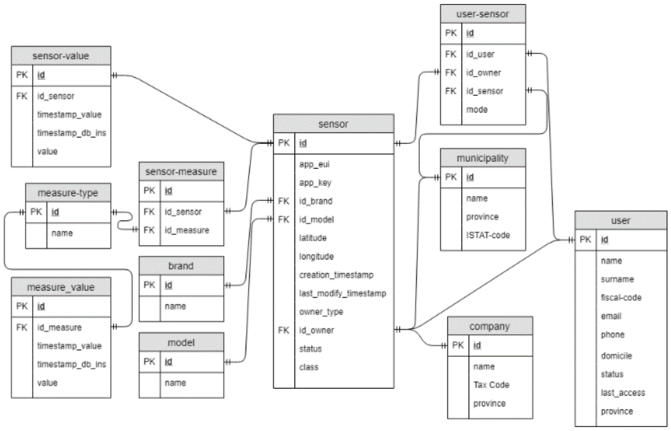


Figure 2: DataBase schema of retePAIoT

These problems imply that it is not possible to ensure that a sensor has been completely described within the database, through the specification of the corresponding measurements neither the payload decoding service is active at the very moment in which its payloads start to arrive.

To cope with these problems and guarantee the correct decoding of the payloads at any later time without any loss of information, an extension of two tables, 'sensor' and 'sensor-value' tables, which respectively contain the details of all the sensors registered on the PAIoT network and all payloads received by sensors, was provided:

‘sensor’ table:

- instance (integer, default = 0)
- interpretation (boolean, default = false)

‘sensor-values’ table:

- interpreted (boolean, default = false)

The 'sensor.instance= 0' field indicates that the correctness of the model and brand specified by the user for the sensor has not yet been verified and that, in particular, within the database, in the 'sensor-measure' table the sensor corresponding measures have not been described yet.

The 'sensor.interpretation = false' field indicates that the sensor payloads cannot be decoded yet, either because the sensor has not yet been instantiated, 'sensor.istance = 0', or because the decoding rules of the specific sensor have not been implemented yet.

The sensor-value.interpreted=false' field keeps track of the payloads already received from the network server and stored within the database in the 'sensor-value' table, but not interpreted yet.

The extension of the 'sensor.instance' field allows to complete the description of the corresponding measurements of a sensor at any time after the registration and only thereafter the validation of the corresponding model.

The extension of the 'sensor.interpretation' field allows to postpone the decoding of the payloads until all the conditions that make it possible are verified:

- consistent description of the sensor corresponding measures within the database in the 'measure' table, corresponding to 'sensor.istance' = 2
- availability of the rules for extracting the corresponding values in the decoding module 'sensor.interpretation = true'

Finally, the extension of the 'sensor-values' table with the 'interpreted' field allows to keep track of the received and stored payloads, which have not yet been decoded and to be able to do that at any subsequent time, without losing any information.

5. Decoding Modules Improvement

As already explained in the original paper [1], in retePAIoT platform the process of receiving, decrypting, saving and interpreting messages sent by sensors via the LoRaWan network is carried out by ChirpStack network server and two main modules: 'archiver payload' and 'decoder'. In the original release [1], however, the 'archiver payload' module had only the task of saving the messages decrypted by the network server into 'sensor-value' table, completely delegating the extraction of the corresponding values to the asynchronous 'decoder' module, when all the conditions necessary for decoding payloads i.e. description of the sensor measurements within the database (sensor.istance = 'true') and the implementation of the payload decoding rules (sensor.interpretation='true'), had been verified. The need to provide an asynchronous 'decoder' module capable of decoding payloads in a phase following the reception of the messages, is intrinsic into the complex and unplannable scenario in which the public network RetePaIoT operates and has been also maintained in the advanced solution illustrated in this paper. The possibility offered by the new versions of ChirpStack of inserting the payload decoding code into the ChirpStack 'device_profile' for each different type of sensor has the big advantage of exploiting the javascript code normally provided by the sensors manufacturers, optimising its development and maintenance.

The advanced solution presented in this paper provides the possibility of decoding the payloads of the messages both in synchronous mode by the ChirpStack server, once the configuration of the new sensors in retePAIoT is fully operational, and in asynchronous mode, as a robust and flexible mechanism for recovering all the messages received and not yet decoded or even for the re-execution of the message decoding phase in case of need to correct any decoding errors.

It is important to underline that, for each type of sensor, the advanced solution presented in this paper uses the same decoding code implemented in the correspondent 'device profile' both in the synchronous and asynchronous mode. To this end, the asynchronous 'decoder' module has been modified so that, for each record in the 'sensor-value' table to be decoded, it recalls the decoder code foreseen for the 'device_profile' assigned to the device sending the message. In the second release, therefore, the asynchronous decoding of the messages process has been improved regarding the following two aspects:

- entrusting the network server with the real-time decoding of the messages coming from the sensors already associated with a profile and already validated and described (sensor.instance = 2 and sensor.interpretation = true)
- taking charge of the decoding only of those messages not still interpreted by the network server ('sensor-value.interpreted = 'false'), but using the same javascript code specified in the ChirpStack devices profiles, once it has been made available

In the new release of retePAIoT the process of managing messages received by ChirpStack network server works as follow (Fig.3):

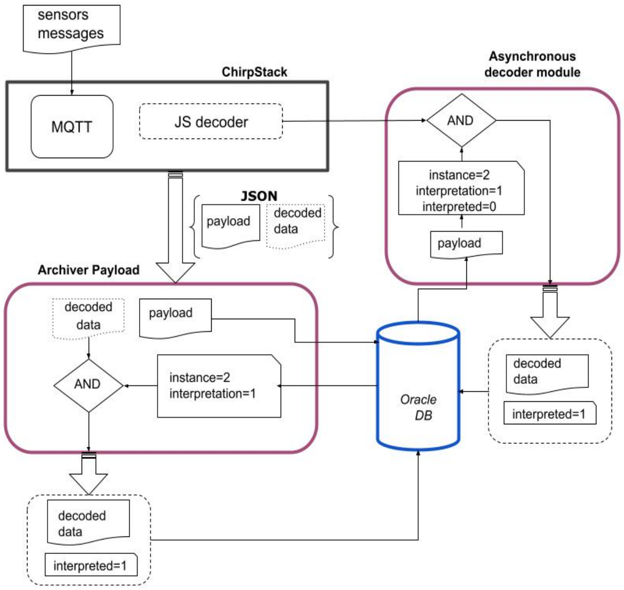


Figure 3: Decoding improved modules data flow

5.1. "archiver payload" module

This module, always running, via the mqtt protocol performs a subscribe on the queue with topic "application" of the LoRaWan ChirpStack network server.

When the network server receives a frame from a device registered on its network database, it decrypts the message, it packs all the information of the frame, both the payload and transmission metadata (i.e. SF, RSSI, SNR, ect.) into a json format and it publishes them on the internal mqtt server. If for the device, that has sent the message, is associated a 'profile device' for which is provided also a decoder java script code, that knows how to analyse, parse and extract the values from the hexadecimal data of the payload, the server network also extracts the values contained in the payload, adding them to the published json data.

The callback function registered for the 'onMessage' event of the topic 'application' is invoked in the "archiver module", that receives the json of the message just published from the server network as an argument and it provides to save the content, i.e. payload and metadata transmission of the message in 'sensor-value' table of the database and, if present and sensor.istance = 2 and sensor.interpretation=true, also moves the decoded data, in 'measure_value' table. In this case the field 'sensor-value.interpreted' of the record just saved in table 'sensor-value' is set to 'true', meaning that the payload has already been interpreted, otherwise to 'false'.

5.2. "decoder" module

'Decoder' module is an asynchronous module that queries the database by joining the 'sensor' and 'sensor-value' tables, in order to retrieve all and only payloads which still need to be interpreted, for which the decoder is now available and whose measurements are now described within the database, that satisfy, therefore, the following correspondents conditions, expressed in the ' where' clause:

- sensor-value.interpreted=0
- sensor.interpretation=1
- sensor.istance=2

It should be highlighted that, in case of crash of the 'decoder' module, the 'sensor-value.interpreted' flag remains set to zero: this does not involve loss of decoded data, since the payloads can always be reprocessed asynchronously at any later time. The same applies in case one of the above conditions (instance, interpretation) is not satisfied at the moment of first processing but changes afterwards. Additionally it becomes possible to reprocess the payloads as many many times as needed in case of errors or updates in the decoder javascript.

5.3. The transition to the new architecture

The design of the data flow in steps controlled by flags set in the database for each sensor and payload was crucial also to allow for a seamless transition from the previous to the new system architecture. With the previous design, a php decoder script was always alert to catch newly stored payloads into the database with 'sensor-value.interpreted=false' flag. This software was not terminated but is being kept running to process the data from those sensors that do not have a decoder profiled in the ChirpStack server yet. At the same time, those payloads arriving from sensors with a profile equipped with a decoder javascript are given a 'sensor-value.interpreted=true' flag before being written to the database in such a way that the php script of the old architecture will not catch them up. Conversely, the payloads that have not been decoded directly by the network server because no code was added to their sensor profile, are given an 'sensor-value.interpreted=0' flag allowing the payload module to store the raw payload in the database without attempting to recover the decoded information and leaving to the php script the decoding task.

The two systems, old and new architecture, are therefore running at the same time: the old php decoders are progressively dismissed while new js-decoders are added to the ChirpStack sensor profiles. No sudden switch from the old to the new system was necessary, allowing the administrators to test the new decoders with ease one at the time.

5.4. Derived measures

It is often useful to compute derived measurements from the data that has been decoded, for example changes in the unit of measure or collection of cumulated values. If a description of the new measures is provided for a sensor in the 'sensor-measure' table, the relative information is accessed during the processing of the payload and a new attribute 'sensor-values.to_be_derived' can be set for the payload and stored in the database. This technique allows to compute derived measurements for flagged payloads at any subsequent time employing asynchronous scripts with no waste of computing resources.

6. Data sharing interfaces

There are different ways to provide users of the retePAIoT network with access to the transmitted data, both in real-time and at a later time. They consist in taking advantage of the mqtt functionalities, in creating push integrations and offering restful API interfaces.

To exploit the mqtt features, a new instance of mqtt server has been created and made available on a public ip address. The public mqtt server is tightly and bidirectionally linked with the private mqtt instance of ChirpStack by means of the mqtt bridge (Fig.4):

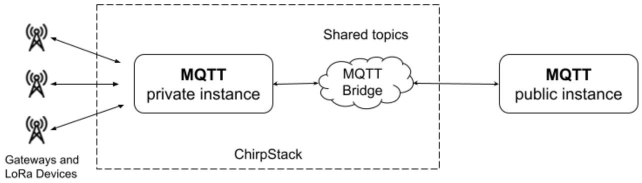


Figure 4: Public and Private instance of MQTT server

The mqtt bridge configuration requires to specify the topics that are to be shared; in this way, any message published on any of these topics on one of the mqtt instances is immediately sent to the other. The user, who has previously been provided with personal login credentials (username, password), has simply to connect to the public instance and subscribe to any desired topic. The login credentials are needed to profile the users in order to restrict the access only to their own data and keep all data private. Besides, the implementation design of dividing the devices into groups of 'applications', as defined on the ChirpStack application server, allows for simple and immediate management of data sharing with all authorized users. This solution highlights the advantage of using a public mqtt, that is to allow data sharing in a transparent and secure manner with all and only the authorized users.

It is worth noting that data are published only after having being decrypted, since the retePAIoT network is the owner of the devices and knows the communication keys for each device; that must be done immediately upon receiving a message because the keys are subject to change, ie the device can eventually renegotiate them at any time both in otaa and abp connection modes. Moreover, the private mqtt instance of the ChirpStack will receive all messages from any device that is in communication range of a gateway belonging to the retePAIoT network, so it is not unlikely that it will receive a message from an unknown device. However, the public mqtt will receive all and only the messages managed by the retePAIoT network that must be shared. . The availability of a public MQTT directly connected to the private one allows to share even very large quantities of data in real time and in row form as if they were arriving directly from the private MQTT ChirpStack network server.

In our specific case of the retePAIoT network, an additional integration with an external InfluxDB database has been set up. For this integration to work, the user is required to notify the retePAIoT administrators for the settings to use. Once the server configuration is done - which in our case only necessitated editing a configuration file - the ChirpStack server automatically handles all the communication steps with the database, dramatically simplifying the process of acquiring and storing data on the database side.

InfluxDB is specialized in managing time series, which makes it a perfect match for IoT applications. When exporting the sensor data, the only attention should be paid to the building of the json data block with the prescription given by InfluxDB. In our specific case the aim of the external application is to monitor and analyse the information concerning the signal transmission quality, therefore only this part of data is being shared outside the retePAIoT and no specific adjustment was required to our json format. With this feature of the ChirpStack server the whole process of sharing the information with an external source was essentially effortless.

The retePAIoT project also offers access to data by means of the well-known restful APIs. Since APIs work on stored data, they can fetch measurements only after all of the previous modules have finished processing the incoming payloads. So far, the retePAIoT projects offers APIs to access the raw data (ie. the unprocessed message from the sensor) enriched with specific LoRaWan parameters like the RSSI, the SNR, the spreading factor and the used fPort as well as APIs to access the decoded value of any of the measurements of the sensor. The output is generated in the common json format for ease of use. For security reasons, moreover, when using the APIs the user must provide his personal access key in the body of the request. No data will be extracted without the auth key or if the requested sensor is not owned by the user identified by the auth key. The use of APIs for data retrieval is especially indicated for application purposes in the case of a limited number of sensors involved to exploit the availability of already decoded measurement values.

7. A significant data use case

Currently retePaIoT network is composed by seventy LoRa gateways, installed in fifty municipalities of Emilia-Romagna region; it manages almost one hundred type of sensors and their correspondents decoder through as many 'device profile' and ten 'application' to share data with mqtt interface to as many users; it integrates more than two-thousand sensors.

In this section a simple but significant use case of data from a single type of sensor integrated into retePAIoT is presented, which highlights the ultimate purpose of retePAIoT i.e. the use of sensors data which, in the specific case, concern the support to the rationalization of water resource.

Figure.5, in particular, shows a graphic relating to a water meter installed in a school immediately downstream of the multi-utility one, which accounts for daily water usage inside the building. Graphed data immediately highlight any losses, especially the large ones as in the case focused: the figure shows that during the first two week-ends, in the base line of the graph, when schools are closed, the daily consumption of water results more than seven cubic meters, evidently due to a loss in the piping system, and corresponding to almost 50% of the actual average daily consumption. The prompt detection of extra consumptions led to its subsequent resolution without further unnecessary waste, as shown in the graphic during following week-ends.

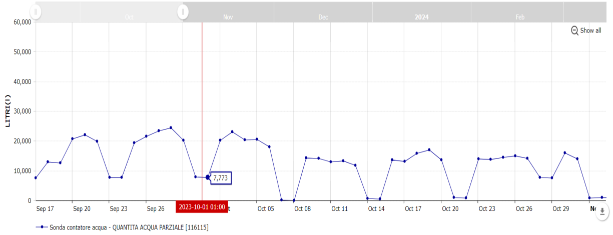


Figure 5: Graph of water consumption in a school

Figure.6 shows last, but not least, the installation of the water meter in the school, which by exploiting the LoRaWan interface and being battery-based, is particularly simple as well as the use of its data.

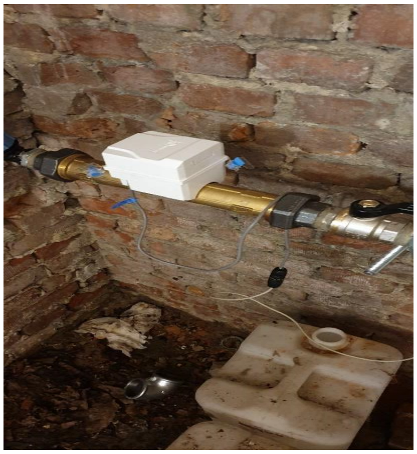


Figure 6: Installation of the meter water in the school

8. Conclusions

RetePAIoT is the public IoT network of Emilia-Romagna region, based on LoRaWan protocol, available for free to all public and private users interested in installing their own sensors of any type and for any purpose within the region.

Data generated by the sensors and collected by retePAIoT are made available through different interfaces, both real-time value and historical series, to the owners of the sensors, but, in aggregate and anonymized form, also to the Public Administrations for their own institutional purposes.

The original paper [1] had already highlighted the critical issues that the decoding process must manage in a public IoT network and the flexibility, robustness and scalability of the solution implemented to solve them. Its evolution, presented in this paper, adds, however, a significant improvement of the decoding management of messages, because it based on a new native functionality of ChirpStack network server, on which retePAIoT is based, and on the javascript code normally provided by sensors manufacturers, without the need for new development for each type of new sensor and in favour of a greater efficiency. In particular, while the flexibility and robustness aspects of the solution presented derive mainly from having made the messages decoding phase asynchronous with respect to their reception, the evolution presented in this document has a positive impact especially with regards to the scalability of the service understood as the ability to manage new types of sensors efficiently in response to the increase of the market offers, maintaining unchanged performance, reliability and quality of service. The scalability is ensured, in fact, not only because, as in the original version, the management of the decoding modules grows linearly with the number of sensor types, and not with that of the sensors, but also because the possibility of using the java script code enormously reduces the time for the update of the service.

The improved model and architecture of data extraction and the different interfaces with which they are made available to all users and to third party platforms not only highlight the main goal of retePAIoT but they also provide an efficient and effective solution replicable for all IoT platforms which, like retePAIoT, have not only the objective of providing the collection, transport and storage of the messages of the sensors, but also that of the extraction of their data and the sharing for their use.

The data use case described at the end of the paper it's a demonstration of the importance of the use of data and therefore of the process of their extraction and sharing illustrated in this paper. It also shows how retePAIoT is an IoT infrastructure that can enable efficient and low-cost monitoring of various phenomena, processes and infrastructures, through which it is possible to detect and understand certain problems and act consequently to resolve them.

References

[1] S. Nanni, M. Carboni, G. Mazzini, “Flexible, Robust, Scalable Solution to Extract Information from IoT Public Network Sensors”, Softcom 2023 – Conference

[2] Vangelista, Lorenzo et al. “Long-Range IoT Technologies: The Dawn of LoRa™.” FABULOUS (2015)

[3] Web interface for managing an Internet of Things Public Network Elisa Benetti (LepidaScpA, Italy); Gian Paolo Jesi (Lepida ScpA, Italy); Gianluca Mazzini (LepidaSpA & UniFe, Italy), Sensornets 2019 – Technical Workshop.

[4] S. Nanni, M. Carboni, G. Mazzini, “PAIoT Network: a unique regional IoT network for very different applications ”, Sensornets 2021 – Conference

[5] https://www.thethings network.org

[6] https://www.chirpstack.io (May 2022)

[7] LoRa specification provided by LoRa Alliance (2015). [Online]. Available: https://lora-alliance.org/about-lorawan, last retrieved 10 May 2019

[8] Benetti, E., Bonino, S., Odorizzi, A., Mazzini, G. (2014). Design of Data Centers for Public Administration. In SoftCom 2014 (pp. 1-5). IEEE.

[9] https://www.chirpstack.io/application-server/use/device-profiles/#custom-javascript-codec-functions

[10] P.P. Ray, A survey on Internet of Things architectures, Journal of King Saud University - Computer and Information Sciences, vol. 30, no. 3, pp 291-319, 2018, doi: 10.1016/j.jksuci.2016.10.003

[11] Jararweh, Y., Al-Ayyoub, M., Darabseh, SDIoT: a software defined based internet of things framework., J Ambient Intell Human Comput, vol. 6, pp 453–461, 2015, doi: 10.1007/s12652- 015-0290-y.

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.

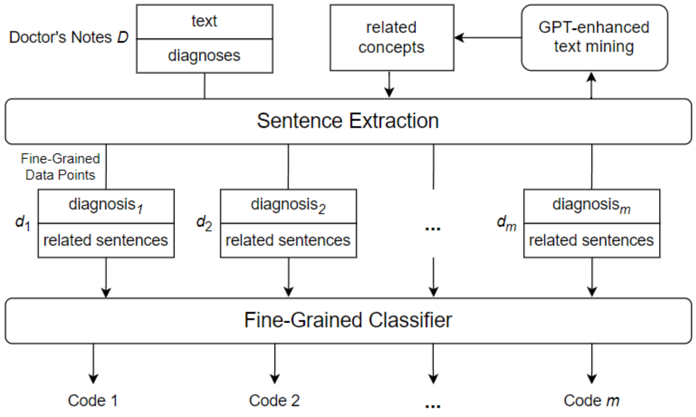


Figure 2: Overview of a fine-grained approach enhanced by GPT-4.

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.

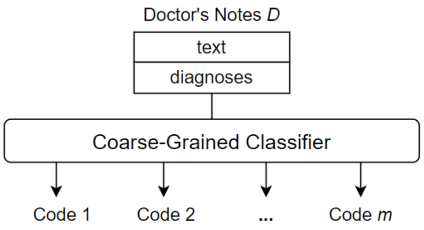


Figure 3: Automated ICD coding using a coarse-grained approach.

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 $F_{\text{fine}}$
 Output: a set of $m$ predicted ICD codes $codeSet$, where
  $m = |dNotes.DIAG|$

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

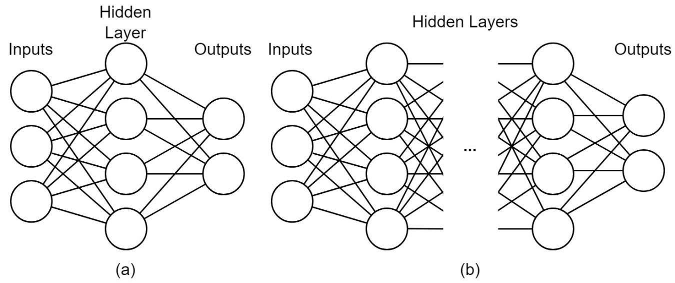


Figure 5: (a) A traditional neural network with one hidden layer. (b) A deep neural network with more than one hidden layer.

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.

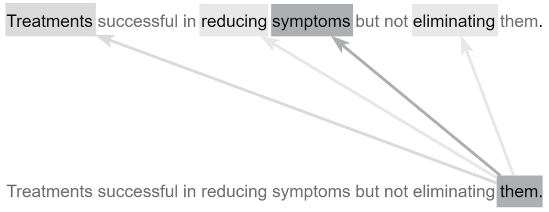

Treatments successful in reducing symptoms but not eliminating them.
Treatments successful in reducing symptoms but not eliminating them.

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.

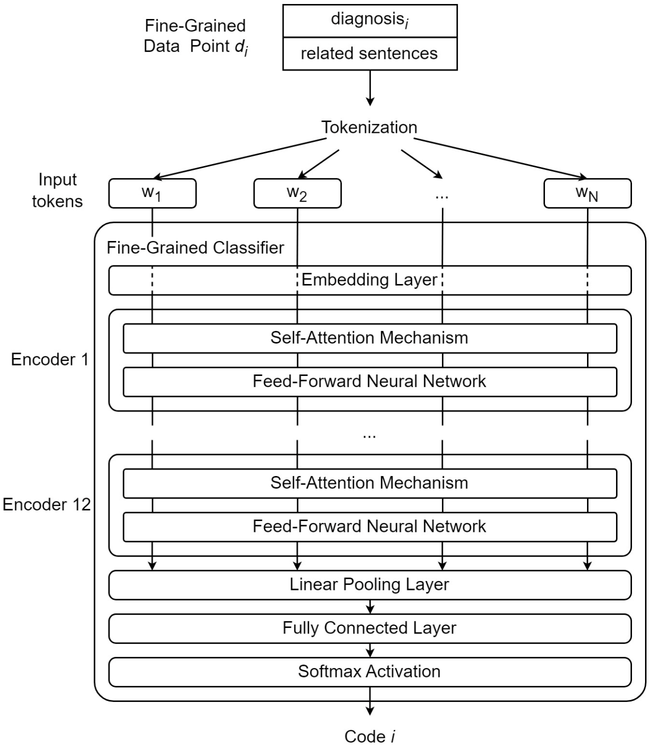


Figure 7: The transformer-based fine-grained deep learning classifier used to generate final ICD code predictions under the fine-grained approach.

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.

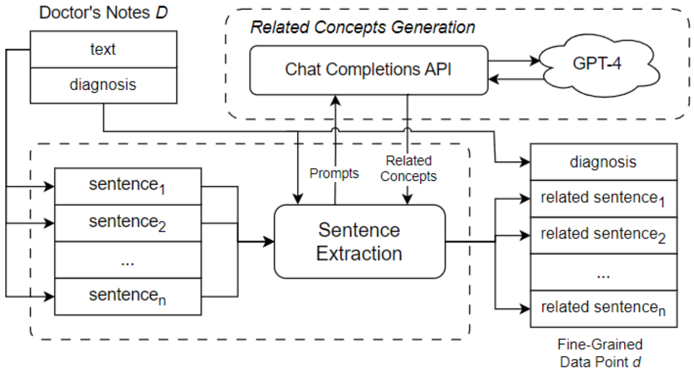


Figure 8: The process of sentence extraction using GPT-4.

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 $\alpha$, free text $\xi$ of doctor's notes dNotes.
Output: a set of related sentences $\Psi$

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

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.

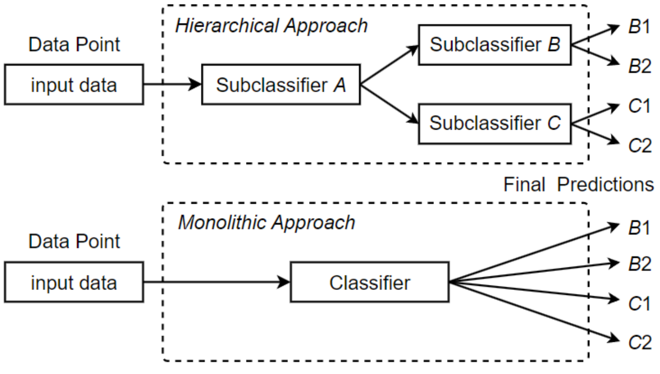


Figure 9: Comparison of hierarchical and monolithic classification approaches.

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 $F_{\text{current}}$ be the current single-label multiclass subclassifier
Initialize $F_{\text{current}} = F_{\text{root}}$
while $F_{\text{current}} \neq null$:
    $confidences = F_{\text{current}}.FSUB(dp)$
    $class = argmax(confidences)$
    if $F_{\text{current}}.CHID \neq \emptyset$ then $F_{\text{current}} = F_{\text{current}}.CHID[class]$
    else $F_{\text{current}} = null$
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.

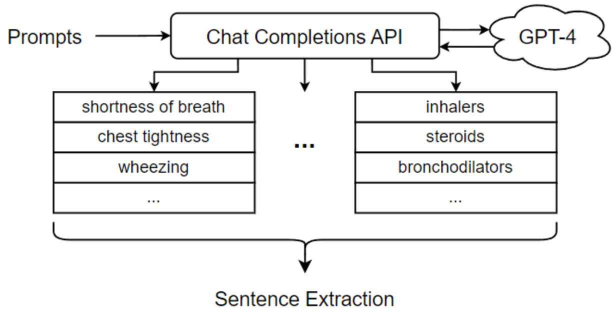
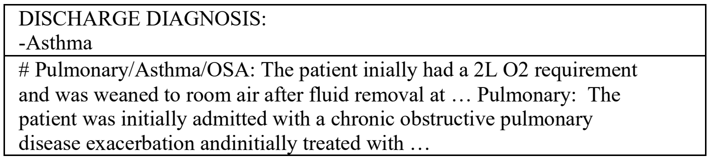


Figure 10: Procedure for prompting GPT-4 and collecting the responses.

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.

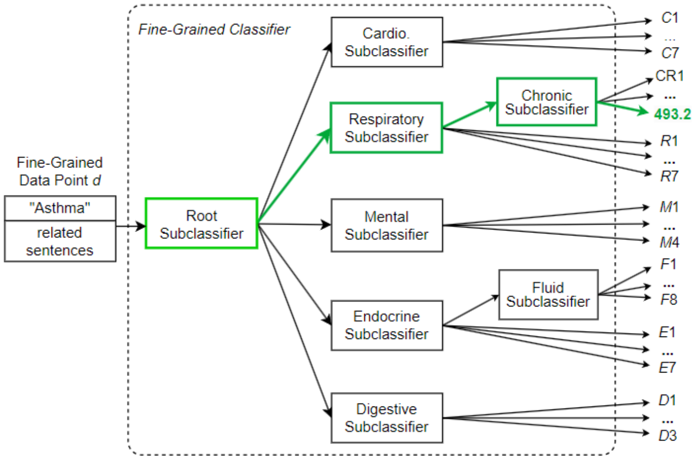


Figure 12: Classification pathway for the diagnosis "Asthma."

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.

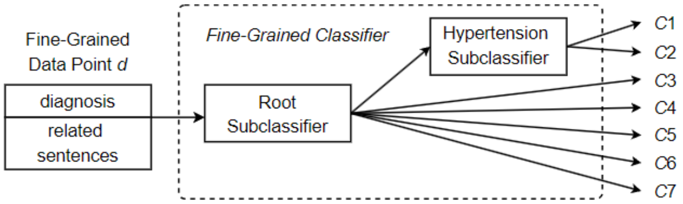


Figure 13: Hierarchical classification with a small code set.

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.

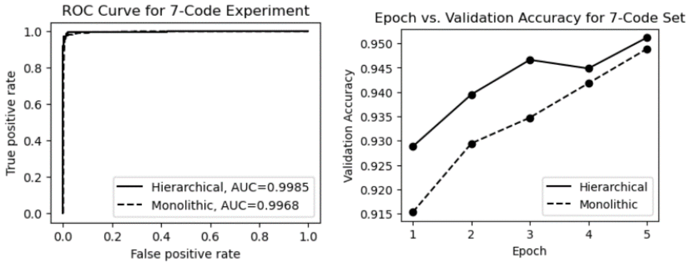


Figure 14: ROC curves and training epoch vs. validation accuracy plots for hierarchical and monolithic classification approaches on the 7-code set.

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 $\{E1, E2, \ldots, E7\}$. 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 $\{F1, F2, \ldots, F8\}$ 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.

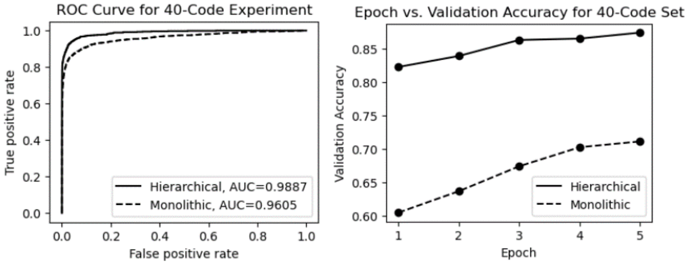


Figure 15: ROC curves and training epoch vs. validation accuracy plots for hierarchical and monolithic classification approaches on the 40-code set.

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.

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.

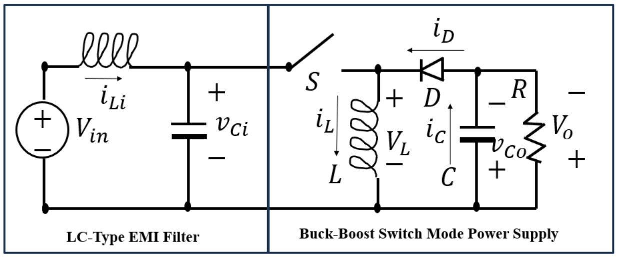


Figure 1: Buck-boost switch mode power supply with a general low pass LC-input filter.

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.

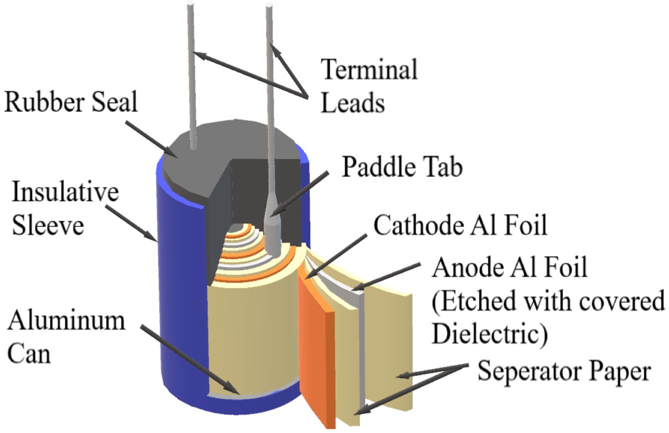


Figure 2: Exploded view of an Aluminum electrolytic capacitor, showing the different layers of the wound jellyroll which is immersed in the electrolytic solution.

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.

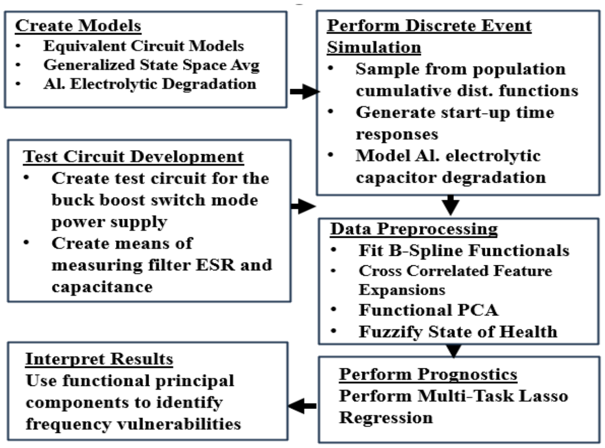

Create Models
• Equivalent Circuit Models
• Generalized State Space Avg
• Al. Electrolytic Degradation
Perform Discrete Event Simulation
• Sample from population cumulative dist. functions
• Generate start-up time responses
• Model Al. electrolytic capacitor degradation
Test Circuit Development
• Create test circuit for the buck boost switch mode power supply
• Create means of measuring filter ESR and capacitance
Data Preprocessing
• Fit B-Spline Functionals
• Cross Correlated Feature Expansions
• Functional PCA
• Fuzzify State of Health
Interpret Results
Use functional principal components to identify frequency vulnerabilities
Perform Prognostics
Perform Multi-Task Lasso Regression

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).

$$\begin{aligned}\dot{x} &= A \cdot x + B \cdot u \\ y &= C \cdot x + D \cdot u\end{aligned} \qquad (1)$$

$$x = \begin{bmatrix} i_{Li} \\ i_L \\ v_{Ci} \\ v_{Co} \end{bmatrix}, u = \begin{bmatrix} v_{in} \\ v_M \\ v_D \\ i_O \end{bmatrix}, y = \begin{bmatrix} v_o \\ i_{Li} \\ i_L \\ v_{Ci} \\ v_{Co} \end{bmatrix}$$

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).

$$\dot{x} = 0 \rightarrow X = -A^{-1} \cdot B \cdot U \qquad (2)$$

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).

$$\tilde{\delta} = K_p\left(\tilde{v}_{ref} - \tilde{v}_o\right) + K_i \int \left(\tilde{v}_{ref} - \tilde{v}_o\right)dt + K_d \frac{d}{dt}\left(\tilde{v}_{ref} - \tilde{v}_o\right) \quad (3)$$

The PID controller is formulated to manage output voltage deviations as follows in (4).

$$\frac{d}{dt}\tilde{\delta} = K_p \frac{d}{dt}\tilde{v}_O + K_i\left(\tilde{v}_{ref} - \tilde{v}_O\right) - K_d \frac{d}{dt}\frac{d}{dt}\tilde{v}_O \qquad (4)$$

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:

$$x(t) = \int_{t_0}^{t_0+T} A \cdot x(t) + B \cdot u(t)\, dt + x(t_0)$$

$$= \sum_{k=-n}^{n} < x >_k (t)\, e^{i\omega kt}$$

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.

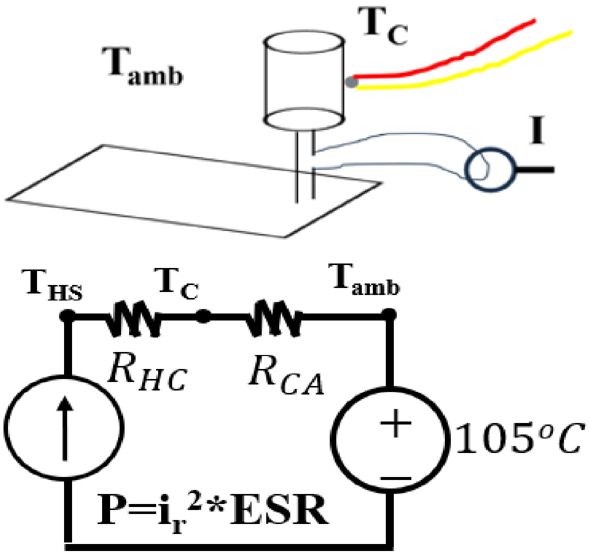


Figure 4: Testing configuration to measure case-to-ambient thermal resistance assuming a negligible difference between case temperature, T_C vs hotspot temperature T_{HS}.

$$R_{CA} = \frac{T_C - T_{amb}}{I^2 * ESR}$$

$$T_{max} = R_{CA} * I_{max}^2 * ESR + T_{amb}$$

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.

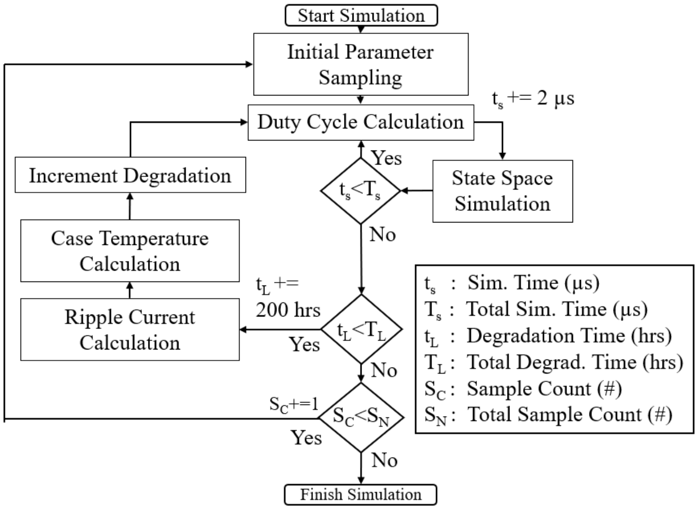


Figure 5: Discrete event simulation routine used to simulate the small time scale responses and the large time scale capacitor degradation responses.

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.

$$\Sigma = \frac{1}{n-1} X X^T \tag{10}$$

Eigen analysis is used to create synthetic variables that are linear combinations of the original features.

$$\Sigma \text{v} = \lambda \text{v} \tag{11}$$

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.

$$Softmax(x_i) = \frac{e^{x_i}}{\Sigma_{j=1}^{n} e^{z_i}} \quad (Output) \tag{13}$$

The PCA/DNN pipeline for the AC analysis feature data can be seen in Figure 6.

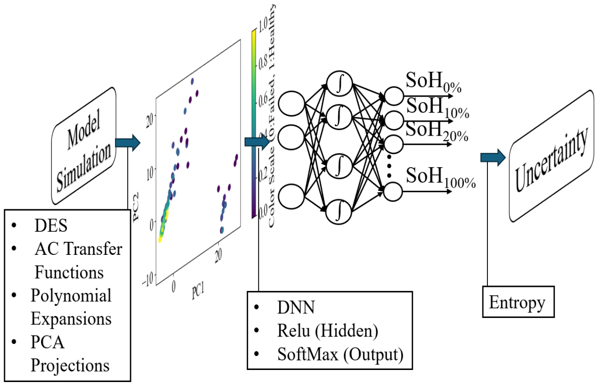


Figure 6: Diagram of the PCA-DNN framework, where high dimensional frequency response data is projected using the truncated discrete principal components and used as inputs to a deep neural network.

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:

$$Objective\ = \frac{1}{2}\left|\left|w_d\right|\right|^2 \tag{15}$$

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:

$$y_i(w^T x_i + b) \geq 1,\ \forall\ i \tag{16}$$

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.

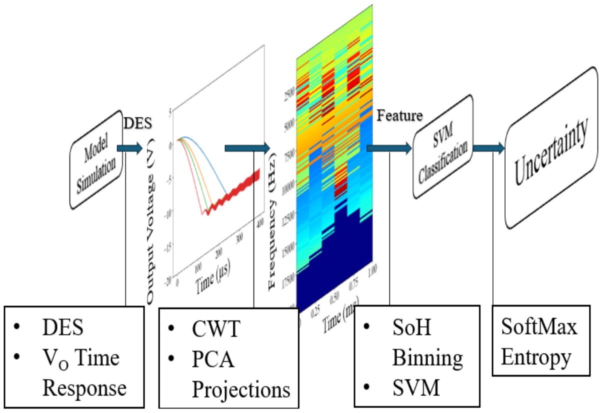


Figure 7: Diagram of the CWT-SVM framework, where high dimensional time-frequency response data is projected using PCA and then used as input to a support vector machine algorithm used for classification of the state of health.

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:

$$y_i\left(\sum_{j=1}^{n} \alpha_j y_j K(x_i, x_j) + b\right) \geq 1,\ \forall\ i \tag{17}$$

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).

$$f(t) = \sum_{i=0}^{n} P_i B_{i,k}(t)$$

$$B_{i,1}(t) = \begin{cases} 1, & if, t_i \leq t \leq t_{i+1} \\ 0, & \text{otherwise} \end{cases}$$

$$B_{i,k}(t) = \frac{t - t_i}{t_{i+k-1} - t_i} B_{i,k-1}(t) + \frac{t_{i+k} - t}{t_{i+k} - t_{i+1}} B_{i+1,k-1}(t) \tag{18}$$

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.

$$f^{\prime}(x) = \frac{f(x+h) - f(x)}{h} \tag{19}$$

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:

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x) \cdot g(x - \delta)\, d\delta \tag{20}$$

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.

$$\int C(s,t)\phi_k(t)dt = \lambda_i\phi_k(s) \qquad (21)$$

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.

$$C(s,t) = \frac{1}{n}\Sigma_{i=1}^{n}\big(X_i(s)-\mu(s)\big)\big(X_i(t)-\mu(t)\big) \qquad (22)$$

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 ith-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:

$$\hat{X}_i(s) = \mu(s) + \Sigma_{k=1}^{K}\int \big(X_i(s)-\mu(s)\big)\phi_k(s)ds\ \phi_k(s) \qquad (23)$$

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.

$$Bin(x) = \lfloor SoH * N_{bins} \rfloor \qquad (24)$$

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:

$$Objective\ = \min_{W,b}\frac{1}{N}\Sigma_{i=1}^{N}||Y-(W\Theta+b)||_F^2 + \lambda\Sigma_{j=1}^{D}||W||_1 \qquad (25)$$

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.

$$SoftMax(x_i) = \frac{e^{x_i}}{\Sigma_{j=1}^{n}e^{x_{ij}}} \qquad (26)$$

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.

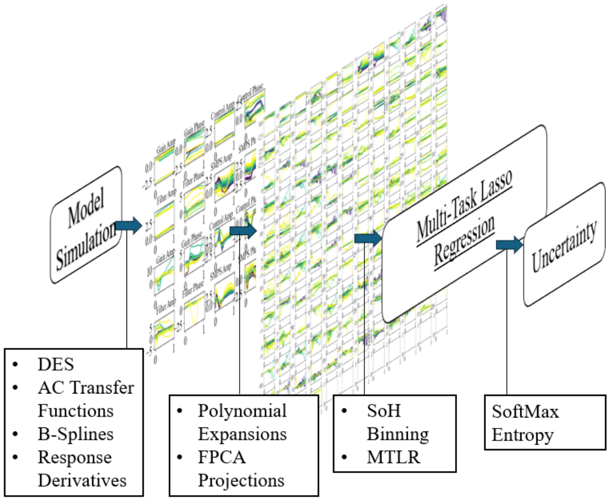


Figure 8: Diagram of the FMT-FFP framework, where high dimensional frequency response data is smoothed using functional b-splines and extended using functional derivatives and polynomial expansion using convolution integrals. Multi-Task LASSO regression is used for feature selection and produce a probabilistic assessment of state-of-health.

$$entropy(x_i) = \Sigma_{j=1}^{n}x_{ij}\log(x_{ij}) \qquad (27)$$

A scaling is applied to the entropy function to create a more discernable measure of uncertainty,

$$uncertainty(x_i) = \frac{e^{1-\xi * entropy(x_i)}}{e^1} \qquad (28)$$

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).

$$RUL(SoH,UT) = a * e^{-b*SoH} + c * UT^d + e \qquad (29)$$

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):

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\Sigma_{\mathrm{i=1}}^{\mathrm{n}}\big(\mathrm{V_{out,i}} - \mathrm{V_{ref,i}}\big)^2} \qquad (30)$$

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.

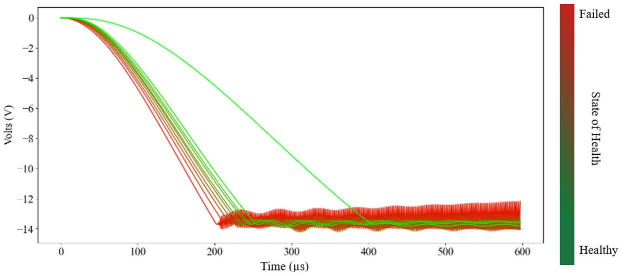


Figure 9: Output voltage start-up transient response for the buck boost SMPS from the generalized state space average model with input EMI filter exhibiting aluminum electrolytic capacitor degradation over time.

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.

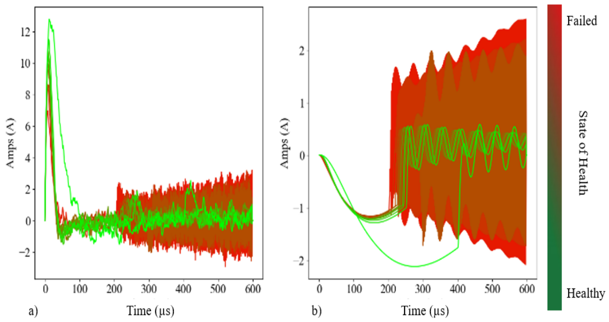


Figure 10: Current waveforms input EMI filter exhibiting aluminum electrolytic capacitor degradation over time.

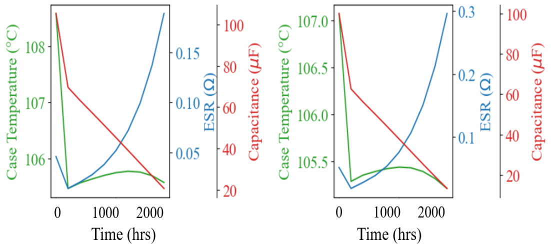


Figure 11: Output voltage start-up transient response for the buck boost SMPS from the generalized state space average model with input EMI filter exhibiting aluminum electrolytic capacitor degradation over time.

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.

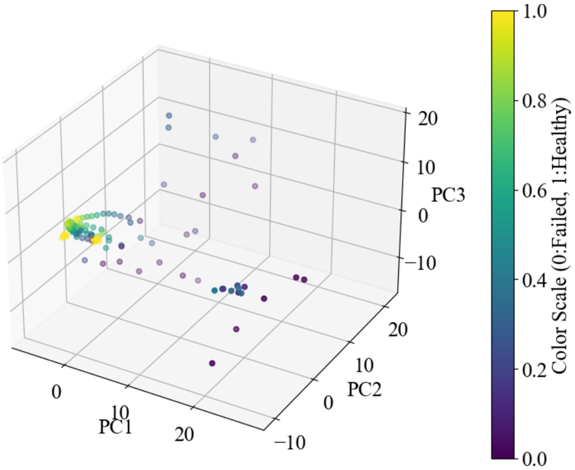


Figure 12: Current waveforms input EMI filter exhibiting aluminum electrolytic capacitor degradation over time.

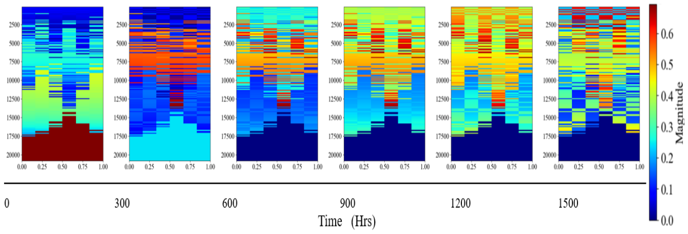


Figure 13: Sequential CWT heatmaps from 0 to 1600 hours, showing the evolution of signal frequencies over time.

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).

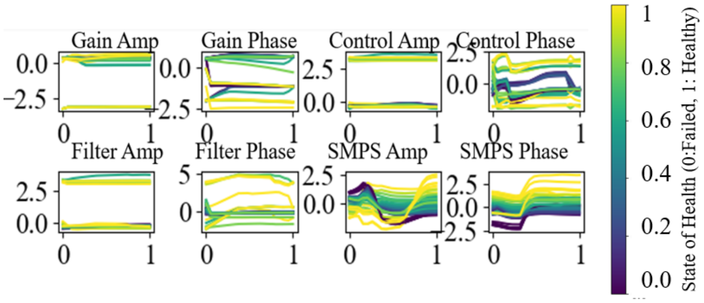


Figure 14: Current waveforms input EMI filter exhibiting aluminum electrolytic capacitor degradation over time.

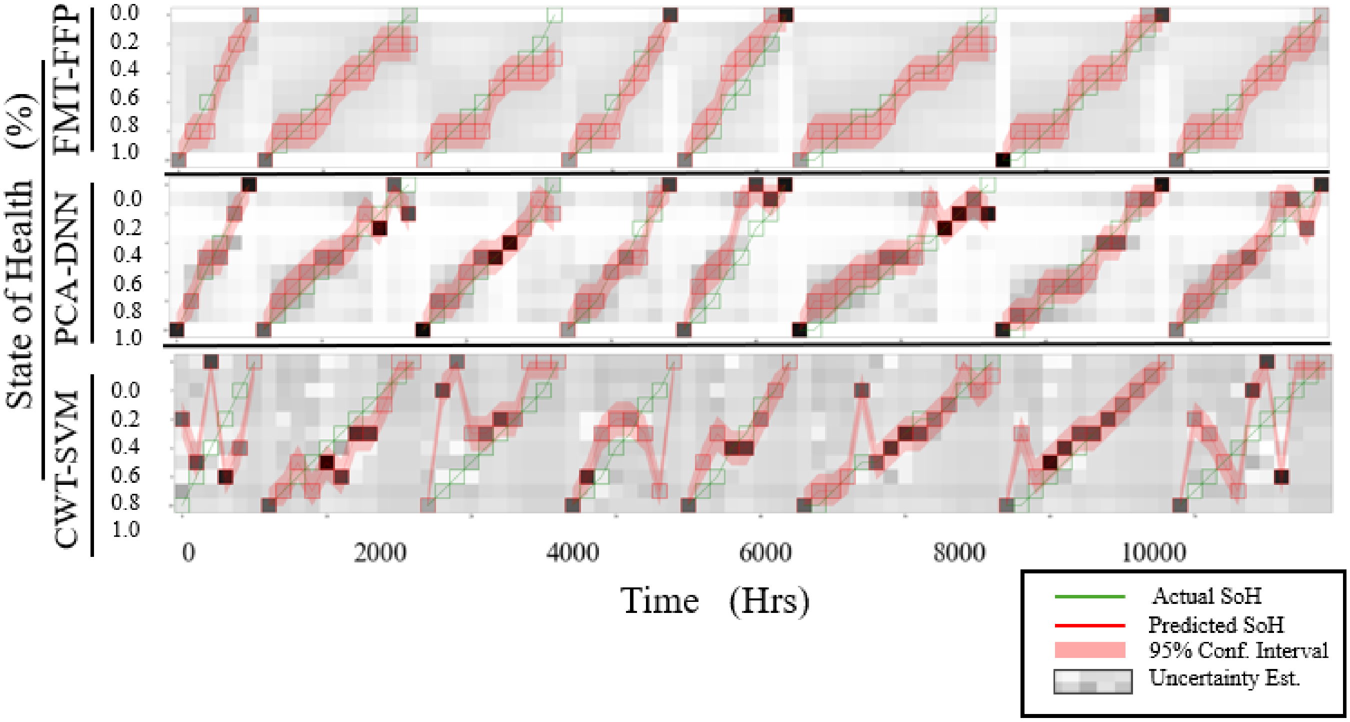


Figure 15: State-of-health prediction from the multi-task LASSO regression for the testing data.

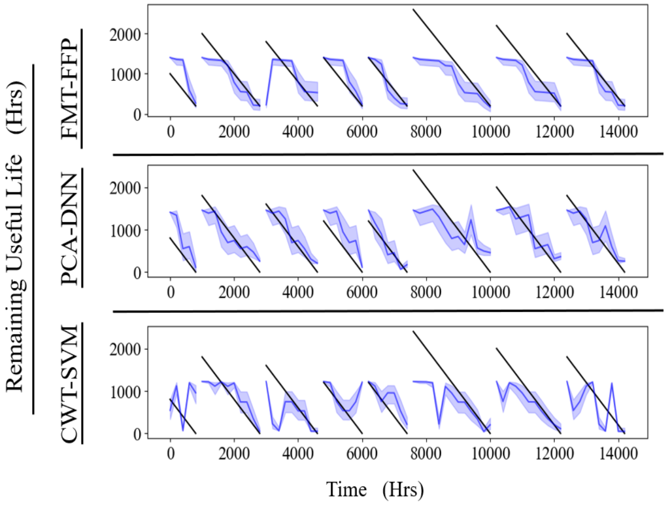


Figure 16: Remaining useful life prediction from the composite model contained in (30).

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.

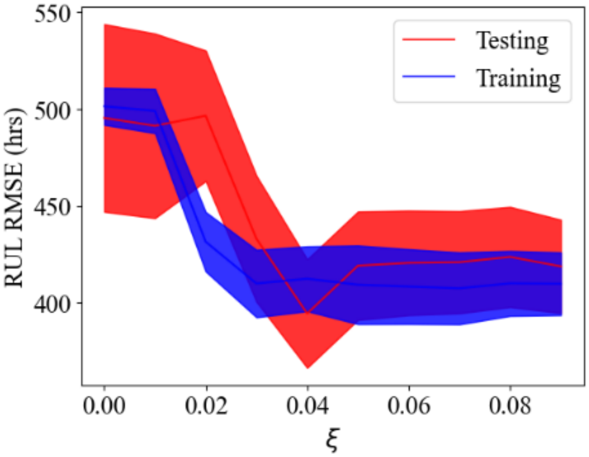


Figure 17: k-fold cross validation results used to find an optimal value of ξ=0.04 for the entropy scaling parameters in the SoH uncertainty calculation.

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.

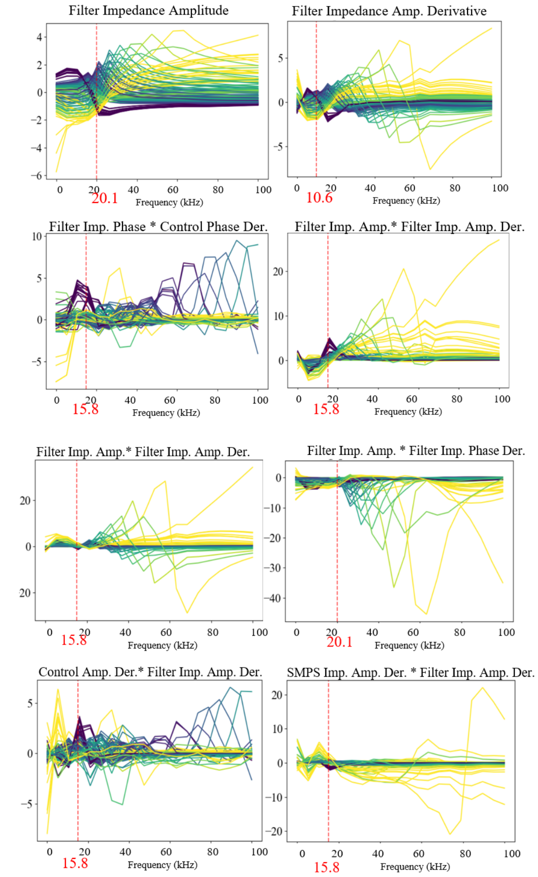


Figure 18: Feature importance visualizations for the down selected features identified in the FMT-FFP method.

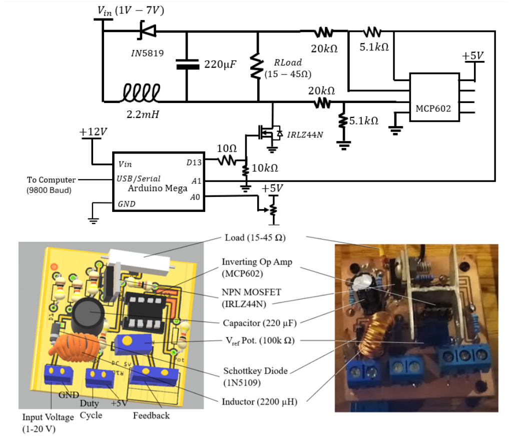


Figure 19: A schematic of the test circuit topology, a 3D model and actual image of the test circuit.

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.

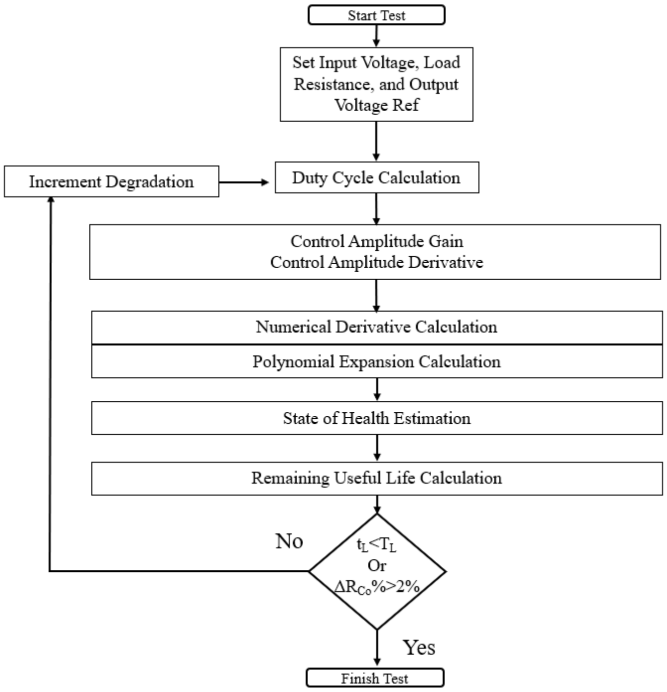

Start Test
Set Input Voltage, Load Resistance, and Output Voltage Ref
Increment Degradation
Duty Cycle Calculation
Control Amplitude Gain
Control Amplitude Derivative
Numerical Derivative Calculation
Polynomial Expansion Calculation
State of Health Estimation
Remaining Useful Life Calculation
No
t_L<T_L
Or
ΔR_{Co}%>2%
Yes
Finish Test

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.

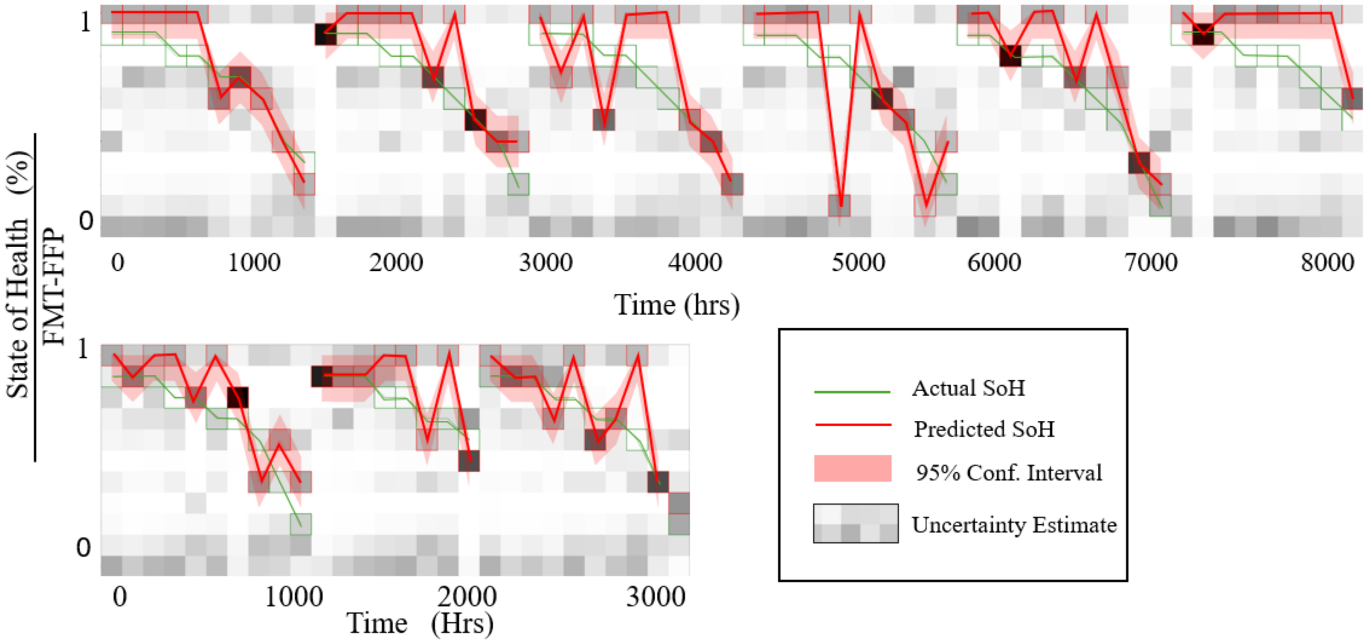


Figure 21: Test circuit results using the down selected features identified in the FMT-FFP method.

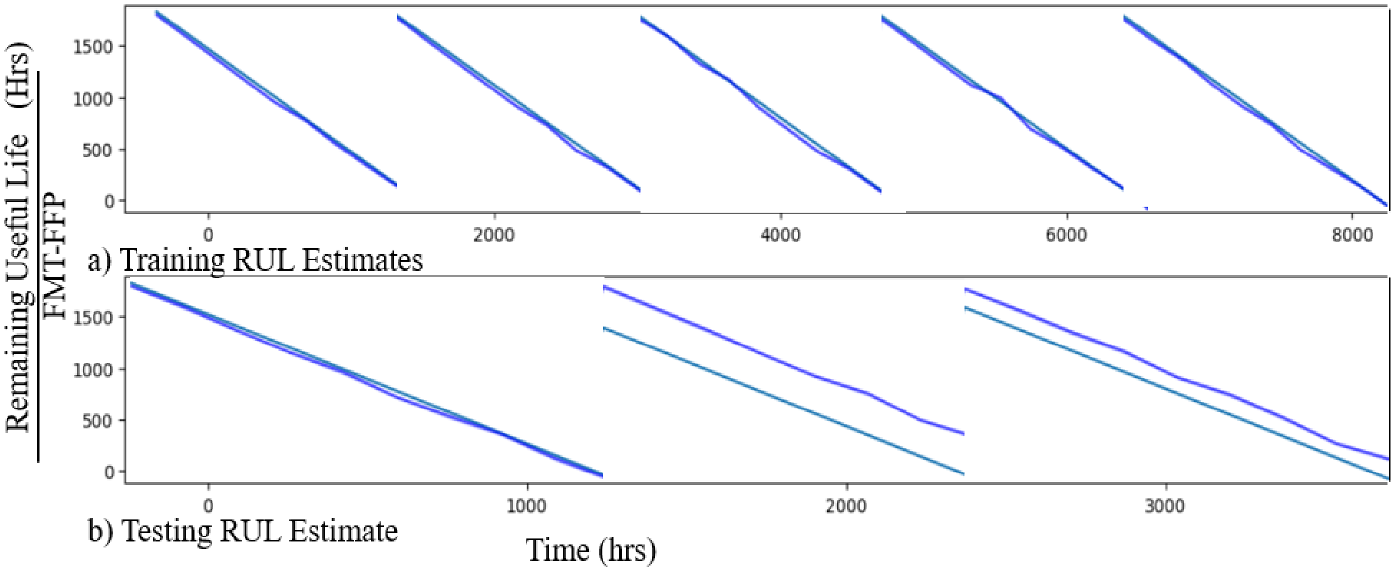


Figure 22: Test circuit results using the down selected features identified in the FMT-FFP method.

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

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.

$$A=\begin{bmatrix} A_{11} & A_{12} & A_{13} & 0 & A_{15}\\ A_{21} & A_{22} & A_{23} & A_{24} & A_{25}\\ A_{31} & A_{32} & 0 & 0 & A_{35}\\ 0 & A_{42} & 0 & A_{44} & A_{45}\\ A_{51} & A_{52} & A_{53} & A_{54} & 0 \end{bmatrix}$$

$A_{11}=-\frac{r_{C_i}+r_{L_i}}{L_i}$; $A_{12}=\frac{D\,r_{C_i}}{L_i}$; $A_{13}=-\frac{1}{L_i}$; $A_{15}=\frac{I_L r_{C_i}}{L_i}$;

$A_{21}=-\frac{D\,r_{C_i}}{L}$; $A_{22}=K_1$; $A_{23}=\frac{D}{L}$; $A_{24}=\frac{R(D-1)}{L(R+r_{C_o})}$;

$A_{25}=\frac{1}{L(R+r_{C_o})}(I_L(R\,r_{C_o}+R\,r_D+R\,r_L+r_{C_o}r_D+r_{C_o}r_L-(R+r_{C_o})(r_{C_i}+r_L+r_M))+R(I_O r_{C_o}+V_{C_o})+(R+r_{C_o})(I_{L_i}r_{C_i}+V_{C_i}-V_D+V_M)$;

$A_{31}=\frac{1}{C_i}$; $A_{32}-\frac{D}{C_i}$; $A_{35}=-\frac{I_L}{C_i}$; $A_{42}=\frac{R(1-D)}{C_o(R+r_{C_o})}$; $A_{44}=-\frac{1}{C_o(R+r_{C_o})}$; $A_{45}=\frac{R(-I_L+2\,I_O)}{C_o(R+r_{C_o})}$;

$$A_{51}=\frac{DK_dR\,r_{C_i}r_{C_o}(1-D)}{L\,L_i(R+r_{C_o})}+\frac{DK_pR\,r_{C_i}r_{C_o}(1-D)}{L(R+r_{C_o})}+\frac{Dr_{C_i}}{L}\left(\frac{K_dK_1R\,r_{C_o}(1-D)}{(R+r_{C_o})}+\frac{K_dR^2(1-D)}{C_o(R+r_{C_o})^2}\right)+\frac{DK_dR\,r_{C_o}(1-D)}{C_i\,L_i(R+r_{C_o})};$$

$$A_{52}=\frac{D^2K_dR\,r_{C_i}^2r_{C_o}(1-D)}{L\,L_i(R+r_{C_o})}+\frac{K_iR\,r_{C_o}(1-D)}{(R+r_{C_o})}+\frac{K_pK_1Rr_{C_o}(1-D)}{(R+r_{C_o})}+K_1\left(\frac{K_dK_1R\,r_{C_o}(1-D)}{(R+r_{C_o})}+\frac{K_dR^2(1-D)}{C_o(R+r_{C_o})^2}\right)+\frac{K_pR^2(1-D)}{C_o(R+r_{C_o})^2}+\frac{R(1-D)\left(\frac{-K_dR^2\,r_{C_o}(1-D)^2}{L(R+r_{C_o})^2}-\frac{K_dR}{C_o(R+r_{C_o})}\right)}{C_o(R+r_{C_o})}-\frac{K_dD^2R\,r_{C_o}(1-D)}{C_iL(R+r_{C_o})^2};$$

$$A_{53}=\frac{DK_dR\,r_{C_i}r_{C_o}(1-D)}{L\,L_i(R+r_{C_o})}+\frac{DK_pR\,r_{C_i}r_{C_o}(1-D)}{L(R+r_{C_o})}+\frac{D\left(\frac{K_dK_1R\,r_{C_o}(1-D)}{(R+r_{C_o})}+\frac{K_dR^2(1-D)}{C_o(R+r_{C_o})^2}\right)}{L};$$

$$A_{54}=\frac{K_iR}{R+r_{C_o}}-\frac{K_pR}{C_o(R+r_{C_o})}-\frac{K_pR^2r_{C_o}(1-D)^2}{L(R+r_{C_o})^2}-\frac{-\frac{K_dR^2r_{C_o}(1-D)^2}{L(R+r_{C_o})^2}+\frac{K_pR}{C_o(R+r_{C_o})^2}}{C_o(R+r_{C_o})}+\frac{R(D-1)\left(\frac{K_dK_1R\,r_{C_o}(1-D)}{(R+r_{C_o})}+\frac{K_dR^2(1-D)}{C_o(R+r_{C_o})^2}\right)}{L(R+r_{C_o})};$$

$$B=\begin{bmatrix} B_{11} & 0 & 0 & 0 & 0\\ 0 & B_{22} & B_{23} & B_{24} & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & B_{42} & 0 & 0 & 0\\ B_{51} & B_{52} & B_{53} & B_{54} & B_{55} \end{bmatrix}$$

$B_{11}=\frac{1}{L_i}$;$B_{22}=-\frac{R\,r_{C_o}(1-D)}{L(R+r_{C_o})}$;$B_{23}=-\frac{D}{L}$; $B_{24}=-\frac{1-D}{L}$;

$B_{42}=\frac{R\,(1-2D)}{C_o(R+r_{C_o})}$; $B_{51}=\frac{DK_dR\,r_{C_i}r_{C_o}(1-D)}{L\,L_i(R+r_{C_o})}$;

$$B_{52}=\frac{K_iRr_{C_o}(1-2D)}{R+r_{C_o}}+K_pR^2r_{C_o}^2\cdot\frac{(1-D)(D-1)}{L(R+r_{C_o})2}+R\frac{r_{C_o}(D-1)\left(K_dR\frac{r_{C_o}(1-D)\left(-D(R+r_{C_o})(r_{C_i}+r_L+rM)+(D-1)(RrCo+RrD+RrL+rCorD+rCorL)\right)}{L(R+r_{C_o})^2}+\frac{K_d^2R(1-D)}{C_o(R+r_{C_o})^2}\right)}{L(R+r_{C_o})}+\frac{K_pR^2(1-2D)}{C_o(R+r_{C_o})^2}-\frac{R(1-2D)\left(\frac{K_dR^2r_{C_o}(1-D)^2}{L(R+r_{C_o})^2}-\frac{K_dR}{C_o(R+r_{C_o})^2}\right)}{C_o(R+r\_C_o)};$$

$$B_{53}=-\frac{D\,K_pR\,r_{C_o}(1-D)}{L(R+r_{C_o})}-\frac{D}{L}\left(\frac{K_dRr_{C_o}(1-D)\left(-D(R\,r_{C_o})(r_{C_i}+r_L+r_M)+(D-1)(Rr_{C_o}+Rr_D+R\,r_L+r_{C_o}r_D+r_{C_o}r_L)\right)}{L(R+r_{C_o})^2}+\frac{K_dR^2(1-D)}{C_o(R+r_{C_o})^2}\right);$$

$$B_{54}=-\frac{K_pR\,r_{C_o}(1-D)^2}{L(R+r_{C_o})}-\frac{(1-D)}{L}\left(\frac{K_dRr_{C_o}(1-D)\left(-D(R+r_{C_o})(r_{C_i}+r_L+r_M)+(D-1)(Rr_{C_o}+Rr_D+R\,r_L+r_{C_o}r_D+r_{C_o}r_L)\right)}{L(R+r_{C_o})^2}+\frac{K_dR^2(1-D)}{C_o(R+r_{C_o})^2}\right);$$

$B_{55}=-K_i$;

$$C=\begin{bmatrix} 0 & \frac{Rr_{C_o}(1-D)}{R+r_{C_o}} & 0 & \frac{R}{R+r_{C_o}} & \frac{Rr_{C_o}(-I_L-2\,I_O)}{R+r_{C_o}}\\ 1 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$D=\begin{bmatrix} 0 & \frac{Rr_{C_o}(1-2D)}{R+r_{C_o}} & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

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

https://dx.doi.org/10.25046/aj090406

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.

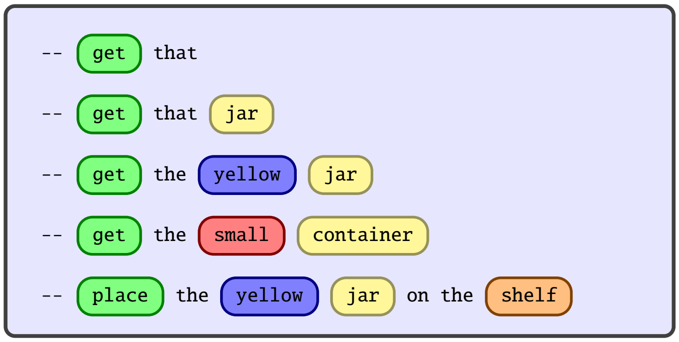
3. Methodology

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.



-- get that

-- get that jar

-- get the yellow jar

-- get the small container

-- place the yellow jar on the shelf

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 $W \in \mathbb{R}^{1\times V}$. 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.

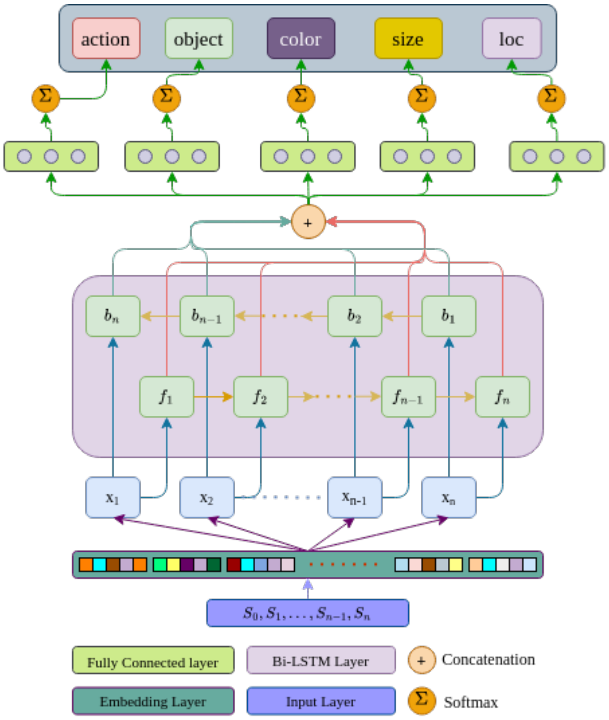


Figure 2: Neural Network (NN) model for parameter extraction from verbal commands.

3.2.1. Deriving θ_a from the positions of the wrist and elbow

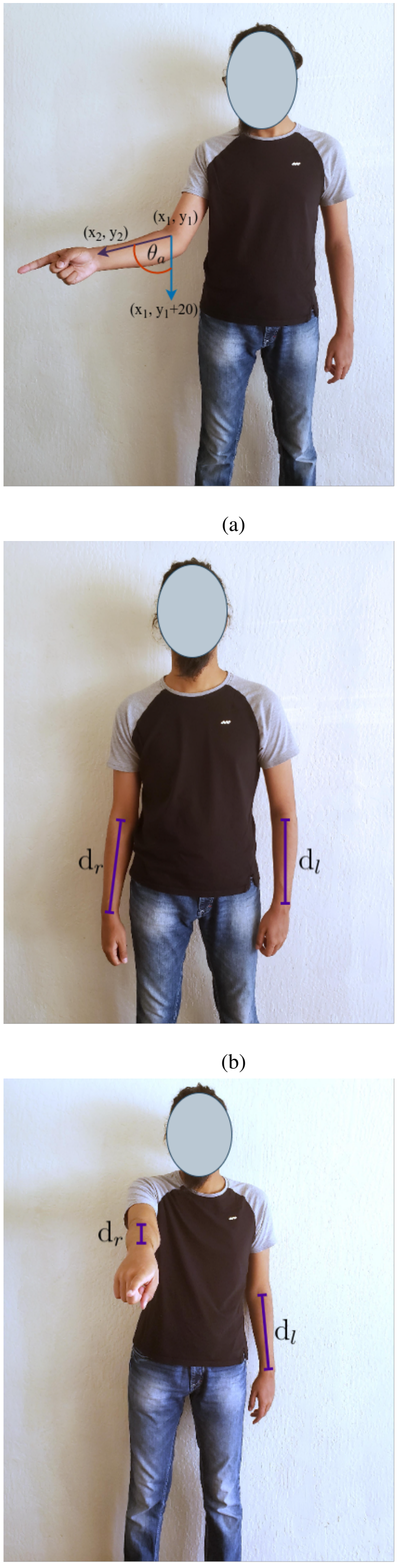
We specifically need the locations of certain skeletal joints. These are locations of the left elbow, left wrist, right elbow, and the right wrist. This ensures our method remains effective even if some body parts are obscured, as long as the pointing hand's joints are detected. Let (x_1, y_1) denote 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 $\vec{a} = (x_2 - x_1, y_2 - y_1)$ and using $\vec{v} = (0, 1)$ as the reference vertical vector, we can calculate the pointing angle θ_a using Equation 1:

$$\theta_a = \cos^{-1}\frac{\vec{a}\cdot\vec{v}}{|\vec{a}||\vec{v}|} \tag{1}$$

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

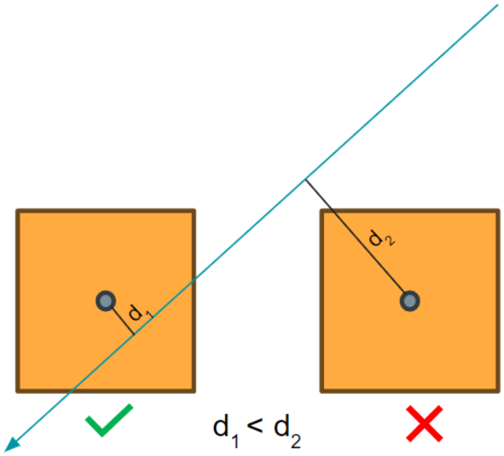
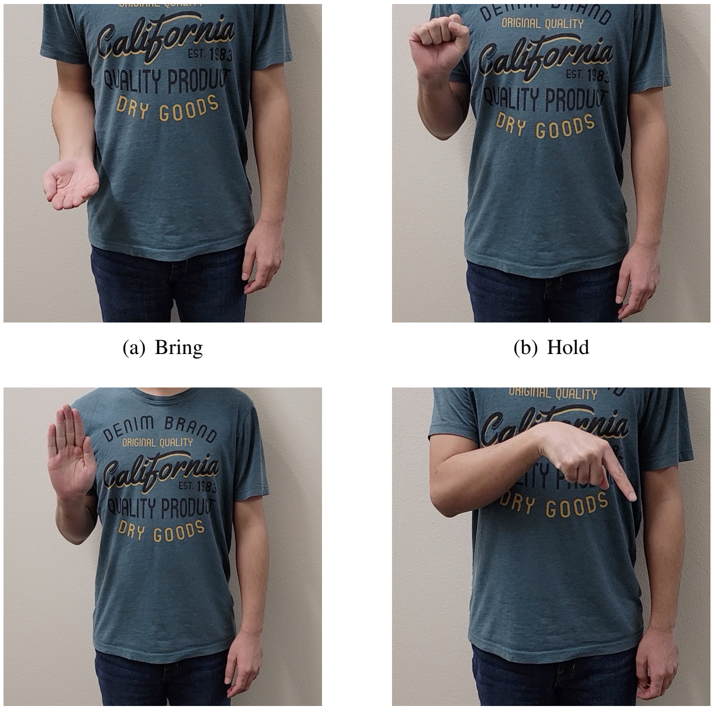


Figure 4: Determining Object of Interest (OOI) from pointing direction.

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

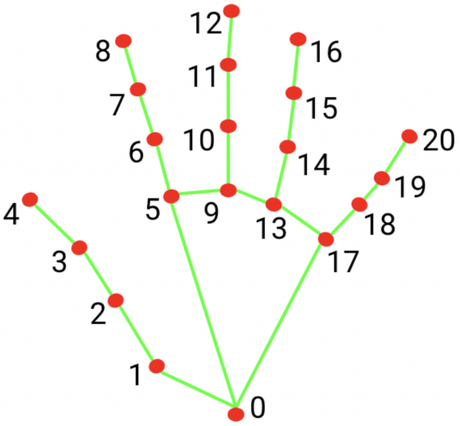


Figure 6: Extracted hand landmark [16]

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.

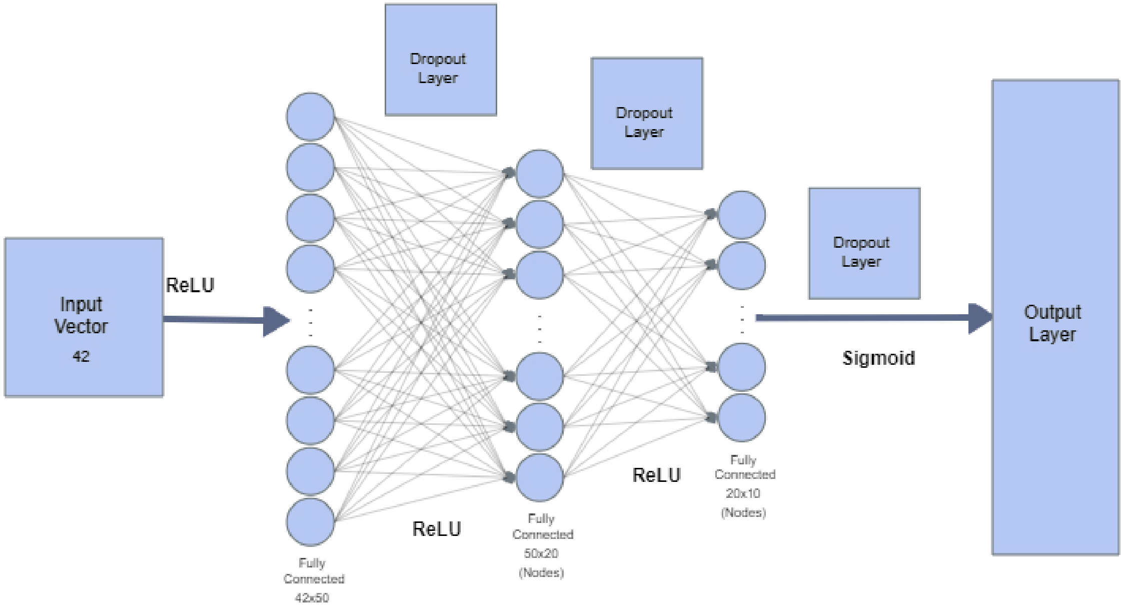


Figure 7: Gesture recognition model architecture.

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

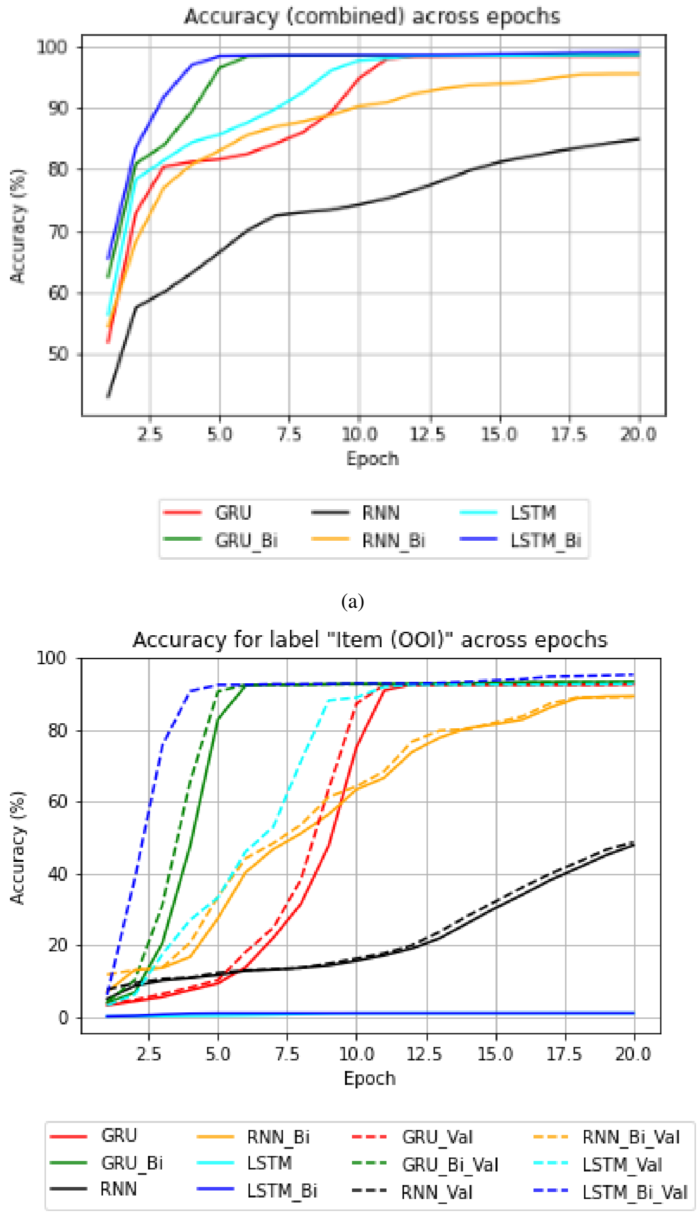
Verbal command: ”bring me the jar”
Object: jar — Action: bring — Attributes: None — Location: None

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)

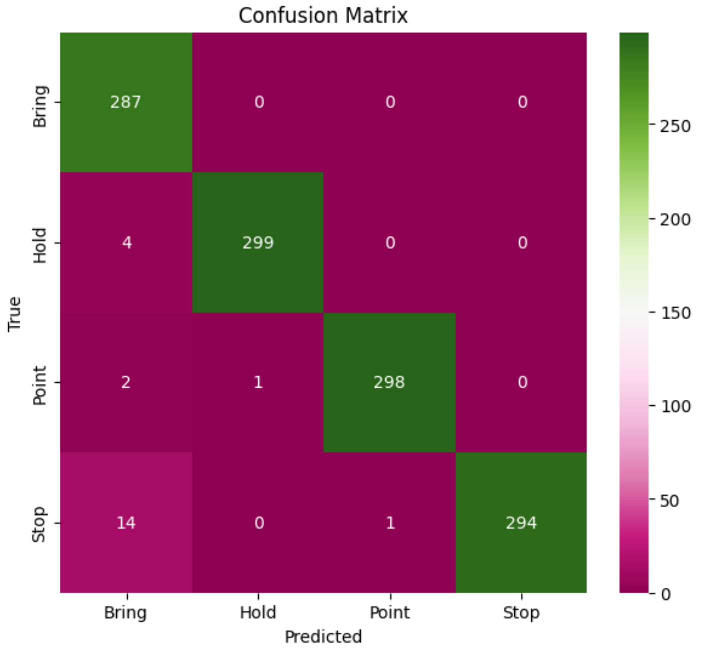
(b)

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.



	Bring	Hold	Point	Stop
Bring	287	0	0	0
Hold	4	299	0	0
Point	2	1	298	0
Stop	14	0	1	294

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.

Advances in Science, Technology and Engineering Systems Journal
Vol. 9, No. 4, 60-72 (2024)
www.astesj.com

ASTES Journal
ISSN: 2415-6698

On Mining Most Popular Packages

Yangjun Chen[*], Bobin Chen

Department of Applied Computer Science, University of Winnpeg, Manitoba, R3B 2E9, Canada

ARTICLE INFO

Article history:
Received: 22 April, 2024
Revised: 11 June, 2024
Accepted: 31 July, 2024
Online: 07 August, 2024

Keywords:
Data Mining
Most popular packages
NP-complete
Priority-first tree search
Tries
Trie-like graphs

ABSTRACT

In this paper, we will discuss two algorithms to solve the so-called package design problem, by which a set of queries (referred to as a query log) is represented by a collection of bit strings with each indicating the favourite activities or items of customers. For such a query log, we are required to design a package of activities (or items) so that as many customers as possible can be satisfied. It is a typical problem of data mining. For this problem, the existing algorithm requires at least $O(n2^m)$ time, where m is the number of activities (or items) and n is the number of queries. We try to improve this time complexity. The main idea of our first algorithm is to use a new tree search strategy to explore the query log. Its average time complexity is bounded by $O(nm^2 + m2^{m/2})$. By our second algorithm, all query bit strings are organized into a graph, called a trie-like graph. Searching such a graph bottom-up, we can find a most popular package in $O(n^2m^3(log_2\ nm)^{log_2\ nm})$ time. Both of them work much better than any existing strategy for this problem.

1. Introduction

Frequent pattern mining plays an important role in mining associations [1, 2, 3, 4], which are quite useful for decision making. For instance, for the supermarket management, the association rules are used to decide, how to place merchandise on shelves, what to put on sale, as well as how to design coupons to increase the profit, etc.

In general, by the frequent pattern mining [5, 6, 7, 8], we are required to find a frequent pattern, which is in fact a subset of items supported (or say, contained) by most of transactions. Here, by a transaction, we mean a set of attributes or items.

In this paper, we study a more challenging problem, the so-called *single package design* problem (SPD for short [2, 3, 9, 10]), defined below:

- A set of attributes (items, or activities):

  $A = \{a_1, ..., a_m\}$,

- A query log:

  $Q = \{q_1, \ldots\ q_n\}$,

  where each $q_i = c_{i1}c_{i2}\ \ldots\ c_{im}$ with each $c_{ij} \in \{0, 1, *\}$ ($i = 1, \ldots, n,\ j = 1, \ldots, m$).

- In q_i, whether a_j is chosen, depends on the value of c_{ij}. That is, if $c_{ij} = 1$, a_j is selected; if $c_{ij} = 0$, a_j is not selected; otherwise, c_{ij} = '*' means 'don't care'.

Our purpose is to find a bit string $\tau = \tau_1\ ...\ \tau_m$ that satisfies as many queries q_i's in Q as possible. We say, τ satisfies a $q_i = c_{i1}c_{i2}\ \ldots\ c_{im}$ if for each j ($1 \leq j \leq m$) the following conditions are satisfied:

$c_{ij} = 1 \rightarrow \tau_j = 1$,
$c_{ij} = 0 \rightarrow \tau_j = 0$,
c_{ij} = '*' $\rightarrow \tau_j = 1$ or 0.

A τ is referred to as a package. If it is able to satisfy a maximum subset of queries, we call it a *most popular* package. For instance, for the above vacation package, a query in a log can be created by specifying yes, no, or 'don't care' for each activity by a client. Then, the design of a most popular package is essentially to decide a subset of such activities to satisfy as many queries' requirements (normally according to a questionnaire) as possible. It is a kind of extension to mining association rules in data mining [5], but more general and therefore more useful in practice.

This problem has been investigated by several researchers [9, 10]. The method discussed in [10] is an approximation algorithm, based on the reduction of SPD to MINSAT [11], by which we seek to find a truth assignment of variables in a logic formula (in conjunctive normal form) to minimize the number of satisfied clauses. This is an optimization version of the satisfiability problem [12]. In [9], a kind of binary trees, called *signature trees* [13, 14, 15] for *signature files* [16, 17, 18, 19, 20], is contructed to represent query logs. Its worst-case time complexity is bounded by $O(n2^m)$.

In this paper, we address this issue and discuss two different algorithms to solve the problem. By the first method, we will con-

[*] Corresponding Author: Department of Applied Computer Science, University of Winnpeg, Manitoba, Canada, R3B 2E9 & y.chen@uwinnipeg.ca

struct a binary tree over a query log in a way similar to [9], but establishing a kind of heuristics to cut off futile branches. Its average time complexity is bounded by O($nm^2 + m2^{m/2}$). The second algorithm is based on a compact representation of the query log, by which all the query bit strings are organized into a trie-like graph G. Searching G bottom-up recursively, we can find a most popular package in O($n^2m^3(log_2\ nm)^{log_2\ nm}$) time.

The remainder of the paper is organized as follows. First, we show a simple example of the SPD problem in Section 2. Then, Section 3 is devoted to the discussion of our first algorithm for solving the SPD problem, as well as its time complexity analysis. Next, in Section 4, we discuss our second algorithm in great detail. Finally, we conclude with a summery and a brief discussion on the future work in Section 5.

2. An example of SPD

In this section, we consider a simple SPD shown in Table 1, which contains a query log with $n = 6$ queries, and $m = 6$ attributes (activities), created based on a questionnaire on customers' favourites. For example, the query $q_6 = a_{11}a_{12} \ldots\ a_{16}$ = (*, 1, *, 0, *, 1) in Table 1 shows that *ride* and *boating* are q_6's favourites, but *hike* is not. In addition, q_6 does not care about whether *hot spring*, *glacier* or *airline* is available or not.

Table 1: A query log Q

query	hot spring	ride	glacier	hike	airline	boat-ing
q_1	1	*	0	*	1	*
q_2	1	0	1	*	*	*
q_3	*	0	0	1	1	*
q_4	0	*	1	*	1	*
q_5	*	0	0	*	*	0
q_6	*	1	*	0	*	1

For this small query log, a most popular package can be found, which contains three items: *hot spring*, *hiking*, *airline*, and is able to satisfy a maximum subset of queries: q_1, q_3, q_5.

3. The First Algorithm

In this section, we discuss our first algorithm. First, in Section 3.1, we give a basic algorithm to provide a discussion background. Then, we describe this algorithm in great detail in Section 3.2. The analysis of the average running time is conducted in Section 3.3.

3.1. Basic algorithm

We first describe a basic algorithm to facilitate the subsequent discussion, which is in fact an extension of an algorithm discussed in [15]. The main idea behind it is the construction of a binary tree T over a query log Q. The algorithm works in two steps. In the first step, a signature-tree-like structure is built up, referred to as a *search-tree*. Then, in the second step, the search-tree is explored to find a most popular package.

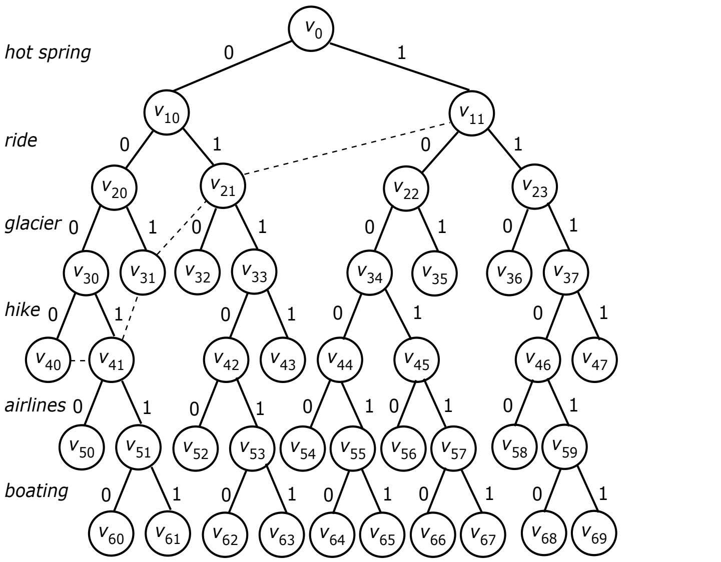
Given a set of attributes: $A = \{a_1, a_2, \ldots, a_m\}$ and a query log: $Q = \{q_1, \ldots, q_n\}$ over A. Denote by $q_i[j]$ the value of the jth attribute a_j in q_i ($i = 1, \ldots, m$). Then, the binary tree T can be constructed as follows.

1. First, for the whole Q, create *root* of T. $j := 1$.

2. For each leaf node v of the current T, denote by s_v the subset of queries represented by v. For query q_i ($\in s_v$), if $q_i[1]$ = '0', we put q_i into the left branch. If $q_i[1]$ = '1', it is put into the right branch. However, if $q_i[1]$ = '*', we will put it in both left and right branches, showing a quite different behavior from a traditional signature tree construction [15].

3. $j := j + 1$. If j ¿ m, stop; otherwise, go to (2).

For example, for the query log given in Table 1, we will construct a binary search tree as shown in Fig. 1.


hot spring
ride
glacier
hike
airlines
boating
v_0
v_{10} v_{11}
v_{20} v_{21} v_{22} v_{23}
v_{30} v_{31} v_{32} v_{33} v_{34} v_{35} v_{36} v_{37}
v_{40} v_{41} v_{42} v_{43} v_{44} v_{45} v_{46} v_{47}
v_{50} v_{51} v_{52} v_{53} v_{54} v_{55} v_{56} v_{57} v_{58} v_{59}
v_{60} v_{61} v_{62} v_{63} v_{64} v_{65} v_{66} v_{67} v_{68} v_{69}
0
1

$s_0 = \{q_1, q_2, q_3, q_4, q_5, q_6\}$ $s_{01} = \{q_1, q_2, q_5, q_6\}$ $s_{11} = \{q_3, q_4, q_5, q_6\}$
$s_{20} = \{q_1, q_2, q_3, q_5\}$ $s_{21} = \{q_1, q_5\}$ $s_{22} = \{q_3, q_4, q_5\}$ $s_{23} = \{q_4, q_6\}$
$s_{30} = \{q_3, q_5\}$ $s_{31} = \{q_4\}$ $s_{32} = \{q_6\}$ $s_{33} = \{q_4, q_6\}$ $s_{34} = \{q_1, q_3, q_5\}$
$s_{35} = \{q_2\}$ $s_{36} = \{q_1, q_6\}$ $s_{37} = \{q_6\}$
$s_{40} = \{q_1\}$ $s_{41} = \{q_4, q_5\}$ $s_{42} = \{q_4, q_6\}$ $s_{43} = \{q_4\}$ $s_{44} = \{q_1, q_5\}$
$s_{46} = \{q_1, q_6\}$ $s_{45} = \{q_1, q_3, q_5\}$ $s_{47} = \{q_1\}$
$s_{50} = \{q_3\}$ $s_{51} = \{q_3, q_5\}$ $s_{52} = \{q_6\}$ $s_{53} = \{q_4, q_6\}$ $s_{54} = \{q_3\}$ $s_{55} = \{q_1, q_5\}$
$s_{56} = \{q_5\}$ $s_{57} = \{q_1, q_3, q_5\}$ $s_{58} = \{q_1\}$ $s_{59} = \{q_1\}$
$s_{60} = \{q_3, q_5\}$ $s_{61} = \{q_3\}$ $s_{62} = \{q_4\}$ $s_{63} = \{q_4, q_6\}$ $s_{64} = \{q_1, q_3\}$
$s_{65} = \{q_1\}$ $s_{66} = \{q_1, q_3, q_5\}$ $s_{67} = \{q_1, q_3\}$ $s_{68} = \{q_1\}$ $s_{69} = \{q_1, q_6\}$

Figure 1: A search tree.

In Fig. 1, we use s_u to represent the subset of queries associated with u. In terms of the corresponding attribute a, s_u is decomposed into two subsets: $s_u(a)$ and $s_u(\neg a)$, where for each $q \in s_u(a)$ we have $q[a] = 1$ and for each $q' \in s_u(\neg a)$ $q'[a] = 0$. In general, $s_u(\neg a)$ is represented by u's left child while $s_u(a)$ is represented by u's right child.

For example, the subset of queries associated with v_{11} is s_{11} = {q_3, q_4, q_5, q_6}. According to attribute '*ride*', s_{11} is split into two subsets respectively associated with its two children (v_{22} and v_{23}): s_{22} = {q_3, q_4, q_5} and s_{23} = {q_4, q_6}. In addition, we can also see that among all the leaf nodes the subset s_{66} (= {q_1, q_3, q_5}) associated with v_{66} is of the largest size. Then, the labels along the path from the root to it spell out a string 100110, representing a most popular package: {*hot spring*, *hiking*, *airlines*}.

The computational complexity of this process can be analyzed as follows.

First, we notice that in the worst case an search-tree can have O(2^m) nodes. Since each node is associated with a subset of queries, we need O(n) time to determine its two children. So, the time for constructing such a tree is bounded by O($n2^m$). The space requirement can be slightly improved by keeping only part of the search tree in the working process. That is, we need only to maintain the bottom frontier (i.e., the last nodes on each path at any time point during the construction of T.) For example, nodes v_{40}, v_{41}, v_{31}, and v_{21} (see the dashed lines in Fig. 1) make up a bottom frontier at a certain time point. At this point, only these nodes are kept around. However, for each node v on a bottom frontier, we need to keep the bit string along the path from *root* to v to facilitate the recognition of the corresponding best package. In the worst case, the space overhead is still bounded by O($n2^m$).

3.2. Algorithm based on priority-first search

The basic algorithm described in the previous section can be greatly improved by defining a partial order over the nodes in the search tree T to cut off futile paths. For this purpose, we will associate a key with each node v in T, which is made up of two values: <$|s_v|$, l_v>, where $|s_v|$ is the subset of queries associated with v and l_v is the level of v. (Here, we note that the level of the root is 0, the level of the root's children is 1, and so on.) In general, we say that a pair <$|s_v|$, l_v> is larger than another pair <$|s_u|$, l_u> if one of the following two conditions is satisfied:

- $|s_v| > |s_u|$, or
- $|s_v| = |s_u|$, but l_v ¿ l_u.

In terms of this partial order, we define a *max-priority queue* H for maintaining the nodes of T to control the tree search, with the following two operations supported:

- *extractMax*(H) removes and returns the node of H with the largest pair.
- *insert*(H, v) inserts the node v into the queue H, which is equivalent to the operation $H := H \cup \{v\}$.

In addition, we utilize a kind of heuristics for efficiency, by which each time we expand a node v, the next attribute a chosen among the remaining attributes should satisfy the following conditions:

1. $|s_v(a)|$ - $|s_v(\neg a)|$ is maximized.
2. In the case that more than one attributes satisfy condition (1), choose a from them such that the number of queries q in s_v with $q[a]$ = '*' being minimized (the tie is broken arbitrarily.)

Using the above heuristics, we can avoid as many useless branches as possible.

By using the priority queue, the exploration of T is not a *DFS* (depth-first search) any more. That is, the search along a current path can be cut off, but continued along a different path, which may lead to a solution quickly (based on an estimation made according to the pairs associated with the nodes.) This is because by *extractMax*(H) we always choose a node with the largest possibility leading to a most popular package.

Algorithm 1: *PRIORTY-SEARCH*(Q, A)

```
Input  : a query log $Q$.
Output : a most popular package $P$.
the pair for root is set to be <$|Q|$, 0>; $i$ := 0;
inset($H$, root); (*root represents the whole $Q$.*)
while $i \leq m$ do
    $v$ := extractMax($H$); (*recall that the pair associated
      with $v$ is <$|s_v|$, $l_v$>.*)
    if $i = m$ then
        return the package represented by the path from
          root to $v$;
    recognize a next attribute $a$ from $A$ according to
      heuristics;
    generate left child $v_l$ of $v$, representing $s_v(\neg a)$;
    create left child $v_r$ of $v$, representing $s_v(a)$;
    the pair of $v_l$ is set to be <$|s_v(\neg a)|$, $l + 1$>;
    the pair of $v_r$ is set to be <$|s_v(a)|$, $l + 1$>;
    insert($H$, $v_l$); insert($H$, $v_r$);
    $i$ := $l_v + 1$;
```

The procedure is given in Algorithm 1. This is in fact a tree search controled by using a priority queue, instead of a stack. First, the *root* is inserted into the priority queue H and its key (a pair of values) is set to be <$|Q|$, 0>. Then, we will go into a **while**-loop, in each iteration of which we will extract a node v from the priority queue H with the largest key value (line 4), that is, with the largest number of queries represented by v and at the same time on the deepest level (among all the nodes in H). Then, the subset of queries represented by v will be split (see lines 7 and 8) according to a next attribute chosen in terms of the heuristics described above (line 7). Next, two children of v, denoted respectively as v_l and v_r, will be created (see lines 8 and 9). and their keys are calculated (lines 10 and 11). In line 12, these two children are inserted into H. Finally, we notice that i is used to record the level of the currently encountered node. Thus, when $i = m$, we must get a most popular package.

The following example helps for illustration.

Example 1. *1 In this example, we show the first three steps of computation when applying the algorithm PRIORITY-SEARCH( ) to the query log given in Table 1. For simplicity, we show only the nodes in both H and T for each step. In addition, the priority queue is essentially a max-heap structure [21], represented as a binary tree.*

In the first step, (see Fig. 2(a)), the *root* (v_0) of T is inserted into H, whose pair is <6, 0>. It is because the *root* represents the whole query log Q which contains 6 queries and is at level 0. Then, in terms of attribute *glacier* (selected according to the heuristics), Q

is split into two subsets (which may not be disjoint due to possible '*' symbols in queries), which are stored as v_0's two child nodes: v_{10} and v_{11} with $s_{v_{10}} = \{q_1, q_3, q_5, q_6\}$ and $s_{v_{11}} = \{q_2, q_4, q_6\}$. Then, v_{10} with pair <4, 1> and v_{11} with <3, 1> will be inserted into H.

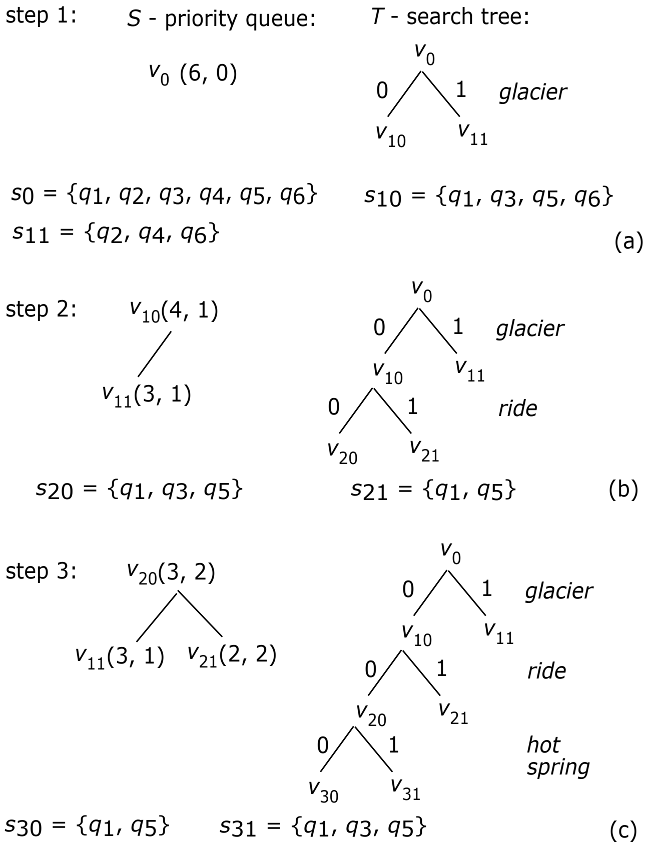


Figure 2: A sample trace.

In the second step (see Fig. 2(b)), v_{10} with pair <4, 1> will be extracted from H. This time, in terms of the heuristics, the selected attribute is *ride*, and $s_{v_{10}}$ will accordingly be further divided into two subsets, represented by its two children: v_{20} with $s_{v_{20}} = \{q_1, q_3, q_5\}$ and v_{21} with $s_{v_{21}} = \{q_1, q_5\}$. Their pairs are respectively <3, 2> and <2, 2>.

In the third step (see Fig. 2(c)), v_{20} with pair <3, 2> will be taken out from H. According to the selected attribute *hot spring*, $s_{v_{20}}$ will be divided into two subsets, represented respectively by its two children: v_{30} with $s_{v_{30}} = \{q_1, q_5\}$ and v_{31} with $s_{v_{31}} = \{q_1, q_3, q_5\}$.

The last step of the computation is illustrated in Fig. 3, where special attention should be paid to node v_{60}, which is associated with a subset of queries: $\{q_1, q_3, q_5\}$, larger than any subset in the current priority queue H. Then, it must be one of the largest subset of queries which can be found since along each path the sizes of subsets of queries must be non-increasingly ordered. Now, by checking the labels along the path from the root to v_{60}, we can easily recognize all the attributes satisfying the queries in this subset. They are {*hot spring*, *hiking*, *airline*}. Since $\{q_1, q_3, q_5\}$ is a maximum subset of satisfiable queries, this subset of attributes must ba a most popular package.

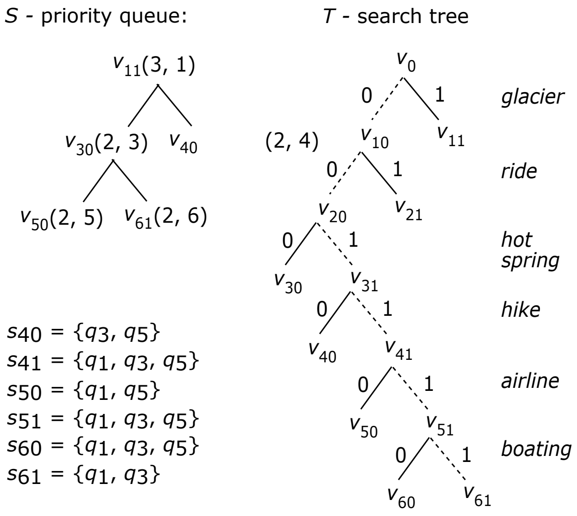


Figure 3: A sample trace.

By this example, a very important property of T can be observed. That is, along each path, the sizes of subsets of queries (represented by the nodes) never increase since the subset of queries represented by a node must be part of the subset represented by its parent. Based on this property, we can easily prove the following proposition.

Proposition 1. *Let Q be a query log. Then, the subset of attributes found in Q by applying the algorithm PRIORITY-SEARCH( ) to Q must be a most popular package.*

Proof. Let v be the last node created (along a certain path). Then, the pair associated with v will have the following properties:

- $l_v = m$, and
- $|s_v|$ is the largest among all the nodes in the current H.

Since the sizes of subsets of queries never increase along each path in T, s_v must be a maximum subset of queries which can be satisfied by a certain group of attributes. One such a group can be simply determined by the labels on the path from *root* to v. This completes the proof. □

The computation shown in Example 1 is super efficient. Instead of searching a binary tree of size O(2^m) (as illustrated in Fig. 1), the algorithm *PRIORITY-SEARCH*( ) only explores a single root-to-leaf path, i.e., the path represented by dashed edges in Fig. 3). However, in general, we may still need to create all the nodes in a complete binary tree in the worst case.

Then, we may ask an interesting question: how many nodes need to be generated on average?

In the next subsection, we answer this question by giving a probabilistic analysis of the algorithm.

3.3. Average time complexity

From Fig. 2 and 3, we can see that for each internal node encountered, both of its child nodes will be created. However, only for some of them, both of their children will be explored. For all the others, only one of their children is further explored. For ease of explanation, we call the former 2-nodes while the latter 1-nodes.

We also assume that in T each level corresponds to an attribute. Let $a = a_1a_2 \ldots\ a_m$ be an attribute sequence, along which the tree T is expanded level-by-level. For instance, in the tree shown in Fig. 3 the nodes are explored along an attribute sequence: *glacier – ride – hot spring – hiking – airline - boating*. For simplicity, we will use $a', a'', a''' \ldots$ to designate the strings obtained by circularly shift the attributes of a. That is,

$a' = a_2 \ldots\ a_m a_1,$
$a'' = a_3 \ldots\ a_m a_2 a_1,$
$\ldots$ ...
$a^{(m)} = a = a_1a_2 \ldots\ a_m.$

In addition, we will use $\aleph_a(T)$ to represent the number of nodes created when applying *PRIORITY-SEARCH*( ) to Q, along a path from top to bottom.

$$\aleph_a(T) = 1 + \aleph_{a'}(T_1) + \aleph_{a'}(T_2) \qquad (1)$$

where T_1 and T_2 represent the left and right subtree of *root*, respectively.

However, if the root of T is a 1-node, we have

$$\aleph_a(T) = 1 + \aleph_{a'}(T_1) \quad \text{or} \quad \aleph_a(T) = 1 + \aleph_{a'}(T_2) \qquad (2)$$

depending on whether $s_{v_l} \geq s_{v_r}$ or $s_{v_l} < s_{v_r}$, where v_l and v_r stand for the left and right child of the root.

Now we consider the probability that $|T_1| = p$ and $|T_2| = \aleph - p$, where $\aleph$ is the number of all nodes in T. This can be estimated by the *Bernouli probabilities*:

$$\binom{\aleph}{p}\left(\frac{1}{2}\right)^p\left(\frac{1}{2}\right)^{\aleph-p} = \frac{1}{2^\aleph}\binom{\aleph}{p} \qquad (3)$$

Let $c_{a,\aleph}$ denote the expected number of nodes created during the execution of *PRIORITY-SEARCH*( ) against Q. In terms of (1), (2) and (3), we have the following recurrences for $\aleph \geq 2$:

$$\text{if root is 2-node, } c_{a,\aleph} = 1 + \frac{2}{2^\aleph}\Sigma_p\binom{\aleph}{p}c_{a',p} \qquad (4)$$

$$\text{if root is 1-node, } c_{a,\aleph} = 1 + \frac{1}{2^\aleph}\Sigma_p\binom{\aleph}{p}c_{a',p} \qquad (5)$$

Let $\gamma_1 = 1$ if *root* is a 1-node, and $\gamma_1 = 2$ if root is a 2-node. Then, (4) and (5) can be rewritten as follows:

$$c_{a,\aleph} = 1 + \frac{\gamma_1}{2^\aleph}\Sigma_p\binom{\aleph}{p}c_{a',p} - \delta_{\aleph,0} - \delta_{\aleph,1} \qquad (6)$$

where $\delta_{\aleph,j}$ ($j = 0, 1$) is equal to 1 if $\aleph = j$; otherwise, equal to 0.

To solve this recursive equation, we consider the following exponential *generating* function of the average number of nodes searched during the execution of *PRIORITY-SEARCH*( ).

$$C_a(z) = \Sigma_{\aleph\geq 0}c_{a,\aleph}\frac{z^\aleph}{2^\aleph} \quad (0 \leq z \leq 1) \qquad (7)$$

In the following, we will show that the generating function satisfies a relation given below:

$$C_a(z) = \gamma_1 e^{z/2}C_{a'}(\frac{z}{2}) + e^z - 1 - z. \qquad (8)$$

In terms of (6), we rewrite $C_a(z)$ as follows:

$$\begin{aligned} C_a(z) &= \Sigma_{\aleph\geq 0}(1 + \gamma_1(\tfrac{1}{2})^\aleph\Sigma_p\tbinom{\aleph}{p} - \delta_{\aleph,0} - \delta_{\aleph,1})\tfrac{z^\aleph}{2^\aleph} \\ &= \Sigma_{\aleph\geq 0}\tfrac{z^\aleph}{2^\aleph} + \Sigma_p\gamma_1(\tfrac{1}{2})^\aleph\Sigma_{\aleph\geq}\tbinom{\aleph}{p}c_{a',\aleph}\tfrac{z^\aleph}{2^\aleph} \\ &\quad -\Sigma_{\aleph\geq 0}\delta_{\aleph,0}\tfrac{z^\aleph}{2^\aleph} - \Sigma_{\aleph\geq 0}\delta_{\aleph,1}\tfrac{z^\aleph}{2^\aleph} \qquad (9) \\ &\leq 2 + \gamma_1\Sigma_p\tfrac{(z/2)^p}{2^p}\Sigma_{\aleph\geq 0}c_{a',p}\tfrac{(z/2)^{\aleph-p}}{2^{\aleph-p}} - 1 - z \\ &= \gamma_1 e^{z/2}C_{a'}(\tfrac{z}{2}) + e^z - 1 - z. \end{aligned}$$

Next, we need to compute $C_{a'}(z), C_{a''}(z), \ldots, C_{a^{(m-1)}}(z)$. To this end, we define γ_i for $i \geq 2$ as follows:

- $\gamma_i = 1$, if all the nodes at level i are 1-nodes.
- $1 < \gamma_i \leq 2$, if at least one node at level i is a 2-node.

Concretely, γ_i is calculated as below:

$$\gamma_i = \frac{2 \times num(\text{2-nodes at level } i) + num(\text{1-nodes at level } i)}{num(\text{nodes at level } i)} \qquad (10)$$

where num(node at level i) represents the number of nodes at level i.

In the same way as above, we can get the following equations:

$$\begin{aligned} C_a(z) &= \gamma_1 e^{z/2}C_{a'}(\tfrac{z}{2}) + e^z - 1 - z, \\ C_{a'}(z) &= \gamma_2 e^{z/2}C_{a''}(\tfrac{z}{2}) + e^z - 1 - z, \\ &\ldots\ldots \qquad (11) \\ C_{a^{(m-1)}}(z) &= \gamma_m e^{z/2}C_a(\tfrac{z}{2}) + e^z - 1 - z, \end{aligned}$$

These equations can be solved by successive transportation, as done in [22]. For example, when transporting the expression of $C_{a'}(z)$ given by the second equation in (11), we will get

$$C_a(z) = b(z) + \gamma_1 e^{z/2}b(\frac{z}{2}) + \gamma_1\gamma_2 e^{z/2}e^{z/2^2}C_{a''}(\frac{z}{2^2}), \qquad (12)$$

where $b(z) = e^z$ - 1 - z.

In a next step, we transport $C_{a''}$ into the equation given in (12). Especially, this equation can be successively transformed this way until the relation is only on $C_a(z)$ itself. (Here, we assume that in this process a is circularly shifted.) Doing this, we will eventually get

$$\begin{aligned} C_a(z) &= \gamma_1...\gamma_m exp[z(1 - \tfrac{1}{2^m})]C_a(\tfrac{z}{2^m}) + \\ &\quad \Sigma_{j=1}^{m-1}\gamma_1...\gamma_m exp[z(1 - \tfrac{1}{2^j})](exp(\tfrac{z}{2^j}) - 1 - \tfrac{z}{2^j}) \qquad (13) \\ &\leq 2^{m-k}exp[z(1 - \tfrac{1}{2^m})]C_a(\tfrac{z}{2^m}) \\ &\quad +\Sigma_{j=1}^{m-1}\gamma_1...\gamma_m exp[z(1 - \tfrac{1}{2^j})](exp(\tfrac{z}{2^j}) - 1 - \tfrac{z}{2^j}) \end{aligned}$$

where k is the number of all those levels each containing only 1-nodes.

Let $\alpha = 2^{m-k}, \beta = 1$ - $\frac{1}{2^m}$, and

$$B(z) = \Sigma_{j=0}^{m-1}\gamma_1\gamma_2 \ldots \gamma_m exp[z(1 - \tfrac{1}{2^j})](exp(\tfrac{z}{2^j} - 1 - \tfrac{z}{2^j}).$$

We have

$$C_a(z) = \alpha e^{\beta z} C_a(\gamma z) + B(z). \tag{14}$$

Solving the equation in a way similar to the above, we get

$$\begin{aligned} C_a(z) &= \Sigma_{j=0}^{\infty} \alpha^j exp(\beta \tfrac{1-\gamma^j}{1-\gamma} z) B(\gamma^j z) \\ &= \Sigma_{j=0}^{\infty} 2^{j(m-k)} \Sigma_{h=0} z^{m-1} \gamma_1 \gamma_2 ... \gamma_h [exp(z) \\ &\quad - exp(z(1 - \tfrac{1}{2^h 2^{mj}}))(1 + \tfrac{z}{2^h 2^{mj}})] \end{aligned} \tag{15}$$

Finally, using the $Taylor$ formula to expand $exp(z)$ and $exp(z(1-\frac{1}{2^h 2^{mj}}))(1+\frac{z}{2^h 2^m j})$ in the above equation, and then extracting the $Taylo$r coefficients, we get

$$C_{a,\aleph} = \Sigma_{h=0}^{m-1} \gamma_1 \gamma_2 ... \gamma_h \Sigma_{j \geq 0} 2^{j(m-k)} D_{jh}(\aleph) \tag{16}$$

where $D_{00}(\aleph) = 1$ and for $j > 0$ or $h > 0$,

$$\begin{aligned} D_{jh}(\aleph) &= 1 - (1 - 2^{-mj-h})^{\aleph} \\ &\quad - \aleph 2^{-mj-h} (1 - 2^{-mj-h})^{\aleph - 1}. \end{aligned} \tag{17}$$

$C_{a,\aleph}$ can be estimated by using the Mellin transform [23] for summation of series, as done in [22]. According to [22], $C_{a,\aleph} \sim \aleph^{1-k/m}$. If $k/m \geq 1/2$, $C_{a,\aleph}$ is bounded by O($\aleph^{0.5}$).

It can also be seen that the priority queue can have up to O($2^{m/2}$) nodes on average. Therefore, the running time of *extractMax*( ) and *insert*( ) each is bounded by log $2^{m/2} = m/2$. Thus, the average cost for generating nodes during the process should be bounded by O($m\aleph^{0.5}$) ≤ O($m2^{m/2}$). In addition, the whole cost for selecting an attribute to split s_v for each internal node v into its two child nodes: s_{v_l} and s_{v_r} is bounded by O(nm^2) since the cost for splitting all the nodes at a level is bounded by O(nm) and the height of T is at most O(m). Here, v_l and v_r represent the left and right child nodes of v, respectively.

So we have the following proposition.

Proposition 2. *Let n = |Q| and m be the number of attributes in Q. Then, the average time complexity of Algorithm PRIORITY-SEARCH(Q) is bounded by O($nm^2 + m2^{m/2}$).*

4. The Second Algorithm

In this section, we discuss our second algorithm. First, we describe the main idea of this algorithm in Section 4.1. Then, in Section 4.2, we discuss the algorithm in great detail. Next, we analyze the algorithm's time complexity in Section 4.3.

4.1. Main idea

Let $Q = \{q_1, ..., q_n\}$ be a query log and $A = \{a_1, ..., a_m\}$ be the corresponding set of attributes. For each $q_i = c_{i1}c_{i2} \ldots c_{im}$ ($c_{ij} \in \{0, 1, *\}$, $j = 1, \ldots, m$), we will create another sequence: $r_i = d_{j_1} ... d_{j_k}$ ($k \leq m$), where $d_{j_l} = a_{j_l}$ if $c_{ij_l} = q_i[j_l] = 1$, or $d_{j_l} = (a_{j_l}, *)$ if $c_{ij_l} = q_i[j_l] = *$ ($l \in \{1, ..., k\}$). If $c_{ij_l} = q_i[j_l] = 0$, the corresponding a_{j_l} will not appear in r_i at all. Let s and t be the numbers of 1s and *s in q_i, respectively. We can then see that $k = s + t$.

For example, for q_1 = (1, *, 0, *, 1, *) in Table 1, we will create a sequence shown below:

$$r_1 = hot\text{-}spring.\ (ride, *).\ (hiking, *).airline.(boating, *).$$

Next, we will order all the attributes in Q such that the most frequent attribute appears first, but with ties broken arbitrarily. When doing so, (a, *) is counted as an appearance of a. For example, according to the appearance frequencies of attributes in Q (see Table 2), we can define a global ordering for all the attributes in Q as below:

$$A \rightarrow Hs \rightarrow H \rightarrow B \rightarrow R \rightarrow G,$$

where A stands for *airline*, Hs for *hot spring*, H for *hiking*, B for *boating*, R for *ride*, and G for *glacier*.

Following this general ordering, we can represent each query in Table 1 as a sorted attribute sequense as demonstrated in Table 3 (in this table, the second column shows all the attribute sequences while the third column shows their sorted versions). In the following, by an attribute sequence, we always mean a sorted attribute sequence.

In addition, each sorted query sequence in Table 3 is augmented with a start symbol # and an end symbol \$, which are used as sentinels for technical convenience.

Finally, for each query sequence q, we will generate a directed graph $\mathcal{G}$ such that each path from the root to a leaf in $\mathcal{G}$ represents a package satisfying q. For this purpose, we first discuss a simpler concept.

Definition 4.1. *(p-graph) Let $q = a_0 a_1 ... a_k a_{k+1}$ be an attribute sequence representing a query as described above, where a_0 = #, a_{k+1} = \$, and each a_i ($1 \leq i \leq k$) is an attribute or a pair of the form (a, *); A p-graph over q is a directed graph, in which there is a node for each a_j (j = 0, ..., k + 1); and an edge for (a_j, a_{j+1}) for each j ∈ {0, ..., k}. In addition, there may be an edge from a_j to a_{j+2} for each j ∈ {0, ..., k - 1} if a_{j+1} is a pair (a, *), where a is an attribute.*

As an example, consider the p-graph for q_1 = #.A.Hs.(H, *).(B, *).(R, *).\$, shown in Fig. 4(a). In this graph, besides a main path going through all the attributes in q_1, we also have three off-line *spans*, respectively corresponding to three pairs: (H, *), (B, *), and (R, *). Each of them represents an option. For example, going through the span for (H, *) indicates that 'H' is not selected while going through 'H' along the main path indicates that 'H' is selected.

In the following, we will represent a span by the sub-path (of the main path) covered by it. Then, the above three spans can be represented as follows:

(H, *) - <v_2, v_3, v_4>

(B, *) - <v_3, v_4, v_5>

(R, *) - <v_4, v_5, v_6>.

Here, each sub-path p is simply represented by a set of contiguous nodes <v_{i_1}, ..., v_{i_j}> ($j \geq 3$) which p goes through. Then, for the graph shown in Fig. 4, <v_2, v_3, v_4> stands for a sub-path from v_2 to v_4, <v_3, v_4, v_5> for a sub-path from v_3 to v_5, and <v_4, v_5, v_6> for a sub-path from v_4 to v_6.

Table 2: Appearance frequencies of attributes.

attributes	*Hs*	*R*	*G*	*H*	*A*	*B*
appearance frequencies	5/6	6/6	5/6	5/6	5/6	5/6

Table 3: Queries represented as sorted attribute sequences.

query ID	attribute sequences*	sorted attribute sequences
q_1	Hs.(R, *).(H, *).A.(B, *).	#.A.Hs.(H, *).(B, *).(R, *).$
q_2	Hs.G.(H, *).(A, *).(B, *).	#.(A, *).Hs.(H, *).(B, *).G.$
q_3	(Hs, *).H.A.(B, *).	#.A.(Hs, *).H.(B, *).$
q_4	(R, *).G.(H, *).A.(B, *).	#.A.(H, *).(B, *).(R, *).G.$
q_5	(Hs, *).(H, *).(A, *).	#.(A, *).(Hs, *).(H, *).$
q_6	(Hs, *).R.(G, *).(A, *).B.	#.(A, *).(Hs, *).B.R.(G, *).$

*Hs: *hot spring*, R: *ride*, G: *glacier*, H: *hiking*, A: *airline*, B: *boating*.

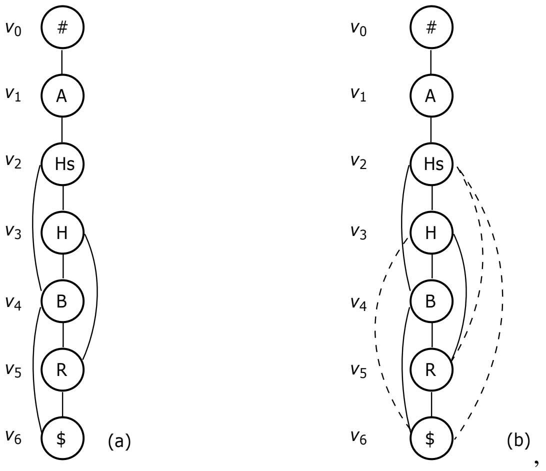


Figure 4: A *p*-path and a *p*∗-path.

In fact, what we want is to represent each packages for a query q as a root-to-leaf path in such a graph. However, p-graph fails for this purpose. It is because when we go through from a certain node v to another node u through a span, u must be selected. If u itself represents a (c, *) for some variable name c, the meaning of this '*' is not properly rendered. It is because (c, *) indicates that c is optional, but going through a span from v to (c, *) makes c always selected. So, (c, *) is not well interpreted.

For this reason, we will introduce another concept, the so-called p*-graph, to solve this problem.

Let $w_1 = <x_1, ..., x_k>$ and $w_2 = <y_1, ..., y_l>$ be two spans in a same p-graph. We say, w_1 and w_2 are overlapped, if one of the two following conditions is satisfied:

- $y_1 = x_j$ for some $x_j \in \{x_2, ..., x_{k-1}\}$, or
- $x_1 = y_{j'}$ for some $y_{j'} \in \{y_2, ..., y_{l-1}\}$

For example, in Fig. 4(a), $<v_3, v_4, v_5>$ (for (B, *)) are overlapped with both $<v_2, v_3, v_4>$ (for (H, *)) and $<v_4, v_5, v_6>$ (for (R, *)). But $<v_2, v_3, v_4>$ and $<v_4, v_5, v_6>$ are only connected, not overlapped. Being aware of this difference is important since the ovetlapped spans correspond to consecutive *'s while the connected overspans not. More important, the overlapped spans are *transitive*. That is, if w_1 and w_2 are two overlapped spans, the $w_1 \cup w_2$ must be a new, but bigger span. Applying this union operation to all the overlapped spans in a p-graph, we will get their '*transitive closure*'. Based on this discussion, we can now define graph $\mathcal{G}$ mentioned above.

Definition 4.2. *(p*-graph) Let p be the main path in a p-graph P and W be the set of all spans in P. Denote by W* the 'transitive closure' of W. Then, the p*-graph $\mathcal{G}$ with respect to P is defined to be the union of p and W*, denoted as $\mathcal{G} = p \cup W$*.*

See Fig. 4(b) for illustration, in which we show the p*-graph $\mathcal{G}$ of the p-graph P shown in Fig. 4(a). From this, we can see that each root-to-leaf path in $\mathcal{G}$ represents a package satisfying q_1.

In general, in regard to p*-graphs, we can prove the following important property.

Lemma 1. *Let P* be a p*-graph for a query (attribute sequence) q in Q. Then, each path from # to $ in P* represents a package, under which q is satisfied.*

Proof. (1) Corresponding to any package σ, under which q is satisfied, there is definitely a path from # to $. First, we note that under such a package each attribute a_j with $q[j] = 1$ must be selected, but with some don't cares it is selected or not. Especially, we may have more than one consecutive don't cares that are not selected, which are represented by a span equal to the union of the corresponding overlapped spans. Therefore, for σ we must have a path representing it.

(2) Each path from # to $ represents a package, under which q is satisfied. To see this, we observe that each path consists of several edges on the main path and several spans. Especially, any such path must go through every attribute a_j with $q[j] = 1$ since for each of them there is no span covering it. But each span stands for a '*' or more than one successive '*'s. □

4.2. Algorithm based on trie-like graph search

In this subsection, we discuss how to find a packaget that maximizes the number of satisfied queries in Q. To this end, we will first construct a *trie-like* structure G over Q, and design a recursive algorithm to search G bottom-up to get results.

Let $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_n$ be all the p*-graphs constructed for all the queries q_i (i = 1, ..., n) in Q, respectively. Denote by p_j and W_j* (j = 1, ..., n) the main path of $\mathcal{G}_j$ and the corresponding transitive closure of spans. We will build up G in two steps. In the first step, we will establish a *trie*, denoted as $T = trie(R)$ over $R = \{p_1, ..., p_n\}$ as follows.

If $|R| = 0$, $trie(R)$ is, of course, empty. For $|R| = 1$, $trie(R)$ is a single node. If $|R| > 1$, R is split into m (possibly empty) subsets $R_1, R_2, \ldots, R_m$ so that each R_i ($i = 1, \ldots, m$) contains all those sequences with the same first attribute name. The tries: $trie(R_1)$, $trie(R_2), \ldots, trie(R_m)$ are constructed in the same way except that at the kth step, the splitting of sets is based on the kth attribute (along the global ordering of attributes). They are then connected from their respective roots to a single node to create $trie(R)$.

In Fig. 5(a), we show the trie constructed for the attribute sequences shown in the third column of Table 2. In such a trie, special attention should be paid to all the leaf nodes each labeled with \$, representing a query (or a subset of queries). Each edge in the trie is referred to as a *tree edge*.

The advantage of tries is to cluster common parts of attribute sequences together to avoid possible repeated checking. (Then, this is the main reason why we sort attribute sequences according to their appearance frequencies.) Especially, this idea can also be applied to the attribute subsequences (as will be seen later), over which some dynamical tries can be recursively constructed, leading to an efficient algorithm.

In the following discussion, the attribute c associated with a node v is referred to as the label of v, denoted as $l(v) = c$.

In addition, we will associate each node v in the trie T with a pair of numbers (*pre*, *post*) to speed up recognizing ancestor/descendant relationships of nodes in T, where *pre* is the order number of v when searching T in preorder and *post* is the order number of v when searching T in postorder.

These two numbers can be used to check the ancestor/descendant relationships in T as follows.

- Let v and v' be two nodes in T. Then, v' is a descendant of v iff $pre(v') > pre(v)$ and $post(v') < post(v)$.

For the proof of this property of any tree, see Exercise 2.3.2-20 in [24].

For instance, by checking the label associated with v_3 against the label for v_{12} in Fig. 5(a), we get to know that v_3 is an ancestor of v_{12} in terms of this property. Particularly, v_3's label is (3, 9) and v_{12}'s label is (11, 6), and we have $3 < 11$ and $9 > 6$. We also see that since the pairs associated with v_{14} and v_6 do not satisfy this property, v_{14} must not be an ancestor of v_6 and *vice versa*.

In the second step, we will add all W_i* ($i = 1, ..., n$) to the trie T to construct a trie-like graph G, as illustrated in Fig. 5(b). This trie-like graph is in fact constructed for all the attribute sequences given in Table 2. In this trie-like graph, each span is associated with a set of numbers used to indicate what queries the span belongs to. For example, the span $<v_2, v_3, v_4>$ is associated with three numbers: 3, 5, 6, indicating that the span belongs to 3 queries: q_3, q_5 and q_6. In the same way, the labels for tree edges can also be determined. However, for simplicity, the tree edge labels are not shown in Fig. 5(b).

From Fig. 5(b), we can see that although the number of satisfying packages for queries in Q is exponential, they can be represented by a graph with polynomial numbers of nodes and edges. In fact, in a single p*-graph, the number of edges is bounded by $O(n^2)$. Thus, a trie-like graph over m p*-graphs has at most $O(n^2m)$ edges.

In a next step, we will search G bottom-up level by level to seek all the possible largest subsets of queries which can be satisfied by a certain package.

First of all, we call a node in T with more than one child a *branching* node. For instance, node v_1 with two children v_2 and v_{13} in G shown in Fig. 5(a) is a branching node. For the same reason, v_3, v_4, and v_5 are another three branching nodes.

Node v_0 is not a branching node since it has one child in T (although it has more than one child in G.)

Around the branching node, we have two very important concepts defined below.

Definition 4.3. *(reachable subsets through spans) Let v be a branching node. Let u be a node on the tree path from root to v in G (not including v itself). A reachable subset of u through spans are all those nodes with a same label c in different subgraphs in G[v] (the subgraph rooted at v) and reachable from u through a span, denoted as* $RS_s^{v,u}[c]$*, where s is a set containing all the labels associated with the corresponding spans.*

For $RS_s^{v,u}[c]$, node u is called its *anchor* node.

For instance, v_3 in Fig. 5(a) is a branching node (which has two children v_4 and v_{11} in T). With respect to v_3, node v_2 on the tree path from *root* to v_3, has a reachable subset:

- $RS_{\{1,5\}}^{v_3,v_2}[\$] = \{v_8, v_{12}\}$,

We have this *RS* (reachable subset) due to two spans $v_2 \xrightarrow{1} v_8$ and $v_2 \xrightarrow{5} v_{12}$ going out of v_2, respectively reaching v_5 and v_8 on two different p*-graphs in $G[v_3]$ with $l(v_5) = l(v_{12}) =$ '\$'. Then, v_2 is the anchor node of $\{v_8, v_{12}\}$.

In general, we are interested only in those *RS*'s with $|RS| \geq 2$ (since any *RS* with $|RS| = 1$ only leads us to a leaf node in T and no larger subsets of queries can be found.) So, in the subsequent discussion, by an *RS*, we always mean an *RS* with $|RS| \geq 2$.

The definition of this concept for a branching node v itself is a little bit different from any other node on the tree path (from *root* to v). Specifically, each of its *RS*s is defined to be a subset of nodes reachable from a span or from a tree edge. So for v_3 we have:

- $RS_{\{1,3,5\}}^{v_3,v_3}[\$] = \{v_8, v_{11}, v_{11}\}$,

due to two spans $v_3 \xrightarrow{1} v_8$ and $v_3 \xrightarrow{3} v_{11}$, and a tree edge $v_3 \rightarrow v_{12}$, all going out of v_3 with $l(v_8) = l(v_{11}) = l(v_{12}) =$ '\$'. Then, v_3 is the anchor node of $\{v_8, v_{11}, v_{11}\}$.

Based on the concept of reachable subsets through spans, we are able to define another more important concept, *upper boundaries*. This is introduced to recognize all those p*-subgraphs around a branching node, over which a trie-like subgraph needs to be constructed to find some more subsets of queries satisfiable by a certain package.

Definition 4.4. *(upper boundaries) Let v be a branching node. Let* $v_1, v_2, ..., v_k$ *be all the nodes on the tree path from root to v. An upper boundary (denoted as upBounds) with respect to v is a largest subset of nodes* $\{u_1, u_2, ..., u_f\}$ *with the following properties satisfied:*

1. Each u_g ($1 \leq g \leq f$) appears in some $RS_{v_i}[c]$ ($1 \leq i \leq k$), where c is a label.

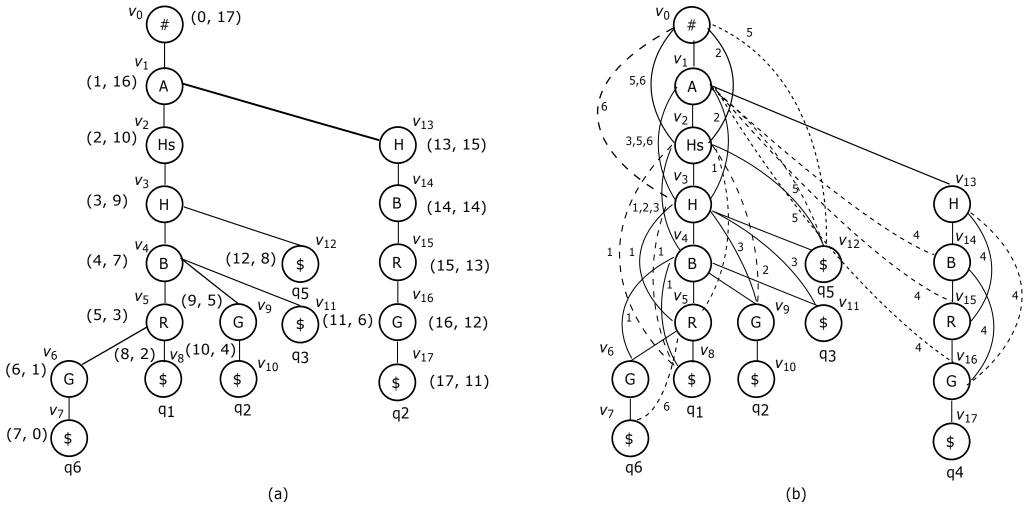


Figure 5: A trie and a trie-like graph.

2. For any two nodes u_g, $u_{g'}$ ($g \neq g'$), they are not related by the ancestor/descendant relationship.

Fig. 6 gives an intuitive illustration of this concept.

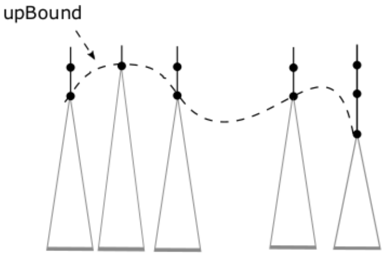


Figure 6: Illustration for upBounds.

As a concrete example, condider branching node v_4 in 5(b). With repect to v_4, we have

- $RS^{v_4,v_3}_{\{1,3\}}[\$] = \{v_8, v_{11}\}$,
- $RS^{v_4,v_4}_{\{1,2\}}[\mathrm{G}] = \{v_6, v_9\}$,

Since all the nodes in these two RSs (v_8, v_6, v_9, and v_{11}) are not related by ancstor/descendant relationsh, they make up an upBound with respect to v_4 (a branching node), as illustrated in 7(a).

Then, we will construct a trie-like graph over all the four p^*-subgraphs, each starting from a node on the upBound. See Fig. 7(b) for illustration, where v_{6-9} stands for the merging v_6 and v_9, v_{7-10} for the merging v_7 and v_{10}, and v_{8-11} for the merging v_8 and v_{11}.

Obviously, this can be done by a recursive call of the algorithm itself.

In addition, for technical convenience, we will add the corresponding branching node (v_4) to the trie as a virtual *root*, and $v_4 \rightarrow v_{6-9}$ and $v_3 \rightarrow v_{8-11}$ as two virtual edges, each associated with the corresponding RSs to facilitate the search of all those packages satisfying corresponding queries. This is because to find such packages we need to travel through this branching node to the *root* of T. See Fig. 9(c) for illustration.

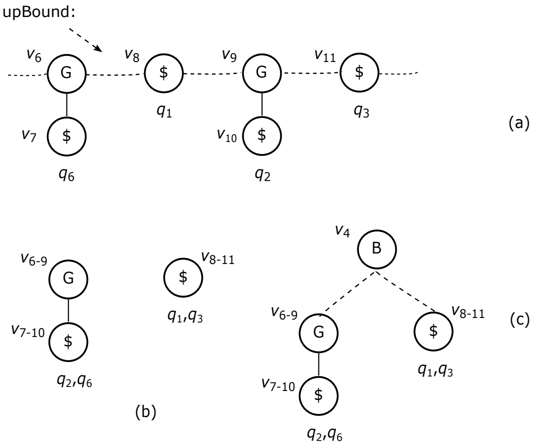


Figure 7: Illustration for upBounds and recursive contruction of trie-like subgraphs.

Specifically, the following operations will be carried out when encountering a branching node v.

- Calculate all RSs with respect v.
- Calculate the upBound in terms of RSs.
- Make a recursive call of the algorithm over all the subgraphs within $G[v]$ each rooted at a node on the corresponding upBound.

In terms of the above discussion, we design a recursive algorithm to do the task, in which R is used to accommodate the result, represented as a set of triplets of the form:

$<\alpha, \beta, \gamma>$,

where α stands for a subset of conjunctions, β for a truth assignment satisfying the conjunctions in α, and γ is the size of α. Initially, $R = \emptyset$.

Algorithm 2: *popularPack*(Q)

```
Input  : a query log $Q$.
Output : a most polular package.
let $Q = \{q_1, ..., q_n\}$;
for $i = 1$ to $n$ do
    construct a $p^*$-graph $P_i^*$ for $q_i$;
construct a trie-like graph $G$ over $P_1^*, ..., P_n^*$;
return $SEARCH(G, \emptyset)$;
```

The input of *popularPack*( ) is a query log $Q = \{q_1, ..., q_n\}$. First, we will build up a p^*-graph for each q_i ($i = 1, ..., n$), over which a trie-like graph G will be constructed (see lines 2 - 4). Then, we call a recursive algorithm *SEARCH*( ) to produce the result (see line 5), which is a set of triplets of the form $<\alpha, \beta, \gamma>$ with the same largest γ value. Thus, each β is a popular package.

Algorithm 3: *SEARCH*(G, R)

```
Input  : a trie-like subgraphs $G$.
Output : a largest subset of conjunctions satisfying a certain
         truth assignment.
if $G$ is a single $p^*$-graph then
    $R'$ := subset associated with the leaf node;
    $R := merge(R, R')$;
    return $R$;
for each leaf node $v$ in $G$ do
    let $R'$ be the subset associated with $v$;
    $R := merge(R, R')$;
let $v_1, v_2, ..., v_k$ be all branching nodes in postorder;
for $i = 1$ to $k$ do
    let $P$ be the tree path from root to $v_i$;
    for each $u$ on $P$ do
        calculate $RS$s of $u$ with respect to $v_i$;
    create the corresponding upBound $L$;
    construct a trie-like subgraph $D$ over all those
      subgraphs each rooted at a node on $L$;
    $D' := \{v_i\} \cup D$;
    $R' := SEARCH(D', R)$;
    $R := merge(R, R')$;
return $R$;
```

SEARCH( ) works recurcively. Its input is a pair: a trie-like subgraph G' and a set R' of triplets $<\alpha, \beta, \gamma>$ found up to now. Initially, $G' = G$ and $R = \emptyset$.

First, we check whether G is a single p^*-graph. If it is the case, we must have found a largest subset of queries associated with the leaf node, satisfiable by a certain package. This subset should be merged into R (see lines 1 - 4).

Otherwise, we will search G bottom up to find all the branching nodes in G. But before that, each subset of queries associated with a leaf node in R will be first merged into R (see line 5 - 7).

For each branching node v encountered, we will check all the nodes u on the tree path from *root* to v and compute their RSs (reachable subsets through spans, see lines 8 - 12), based on which we then compute the corresponding upBound with respect to v (see line 13). According to the upBound L, a trie-like graph D will be created over a set of subgraphs each rooted at a node on L (see line 14). Next, v will be added to D as its root (see line 15). Here, we notice that $D' := \{v\} \cup D$ is a simplified representation of an operation, by which we add not only v, but also the corresponding edges to D. Next, a recursive call of the algorithm is made over D' (see linee 16). Finally, the result of the recursive call of the algorithm will be merged into the global answer (see line 17).

Here, the *merge* operation used in line 3, 7, 17 is defined as below.

Let $R = \{r_1, ..., r_t\}$ for some $t \geq 0$ with each $r_i = <\alpha_i, \beta_i, \gamma_i>$. We have $\gamma_1 = \gamma_2 = ... = \gamma_t$. Let $R' = \{r'_1, ..., r'_s\}$ for some $s \geq 0$ with each $r'_i = <\alpha'_i, \beta'_i, \gamma'_i>$. We have $\gamma'_1 = \gamma'_2 = ... = \gamma'_s$. By $merge(R, R')$, we will do the following checks.

- If $\gamma_1 < \gamma'_1$, $R := R'$.
- If $\gamma_1 > \gamma'_1$, R remains unchanged.
- If $\gamma_1 = \gamma'_1$, $R := R \cup R'$.

In the above algorithm, how to figure out β in a triple $<\alpha, \beta, \gamma>$ is not specified. For this, however, some more operations should be performed. Specifically, we need to trace each chain of recursive calls of *SEARCH*( ) for this task.

Let $SEARCH(G_0, R_0) \rightarrow SEARCH(G_1, R_1) \rightarrow ... \; SEARCH(G_k, R_k)$ be a consecutive recursive call process during the execution of *SEARCH*( ), where $G_0 = G$, $R_0 = \emptyset$, and G_i is a trie-like subgraph constructed around a branching node in G_{i-1} and R_i is the result obtained just before $SEARCH(G_i, R_i)$ is invoked ($i = 1, ..., k$).

Assume that during the execution of $SEARCH(G_k, R_k)$ no further recursive call is conducted. Then, R_k can be a single p^*-graph, or a trie-like subgraph, for which no RSs with $|RS|$ ¿ 1 for any branching node can be found. Denote by r_j the root of G_j and by v_j the branching node around which $SEARCH(G_{j+1}, R_{j+1})$ is invoked ($j = 0, ..., k$ - 1). Denote by T_j the trie in G_j.

Then, the labels on all the paths from r_j in T_j for $j \in \{1, ... \; k\}$ (connected by using the corresponding anchor nodes) consist of β in the corresponding triple $<\alpha, \beta, \gamma>$ with α being a subset of queries associated with a leaf node in G_k. As an example, consider the the trie-like subfraph G' shown in Fig. 7(c) again. In the execution, we will have a chains of recursive calls as below.

$$SEARCH(G, ...) \rightarrow SEARCH(G', ...)$$

Along this chain, we will find two query subset $Q_1 = \{q_2, q_6\}$ (associated with leaf node v_{7-10}) and $Q_2 = \{q_1, q_3\}$ (associated with leaf node v_{7-10}). To find the package for Q_1, we will trace the path in G' bottom from v_{7-10} to v_{6-10}, the reverse edge $v_{6-9} \xrightarrow{1,2} v_4$ (recognized according to $RS^{v_4,v_4}_{\{1,2\}}[G]$), and then the path from $v_4 \rightarrow$ to v_0 in G. Since $\{1, 2\}$ does not contain 6, q_6 should be removed from Q_1. The package is {A, Hs, H, B, G}. In the same way, we can find the package {A, Hs, H, B} for $Q_2 = \{q_1, q_3\}$.

The following example helps for illustrating the whole working process.

Example 2. *When applying SEARCH( ) to the p^*-graphs constructed for all the attribute sequences given in Table 1, we will first construct a trie-like graph G shown in Fig. 5(b). Searching G bottom up, we will encounter four branching nodes: v_5, v_4, v_3 and v_1.*

For each branching node, a recursive call of *SEARCH*( ) will be carried out. But we show here only the recursive call around v_1 for simplicity. With respect to v_1, we have only one RS with $|RS| > 1$:

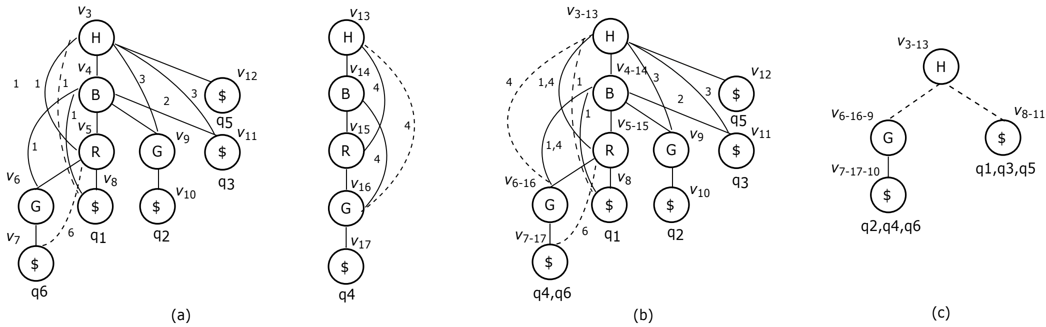


Figure 8: Illustration for Example 2.

- $RS^{v_1,v_1}_{\{1,3,4,5\}}$[H] = $\{v_3, v_{13}\}$,

Due to span $v_1 \xrightarrow{3,5,6} v_3$ and tree edge $v_1 \rightarrow v_{13}$.

Therefore, the corresponding upBound is $\{v_3, v_{13}\}$. Then, a new trie-like subgraphs (see Fig. 8(b)) will be constructed by merging two subgraphs shown in Fig. 8(a).

In Fig. 8(b), the node v_{3-13} represents a merging of two nodes v_3 and v_{13} in Fig. 8(a). All the other merging nodes v_{4-14}, v_{5-15}, v_{6-16}, and v_{7-17} are created in the same way.

When applying *SEARCH*( ) to this new trie-like subgraph, we will check all its branching nodes v_{5-15}, v_{4-14}, and v_{3-13} in turn. Especially, with respect to v_{3-13}, we have

- $RS^{v_{3-13},v_{3-13}}_{\{3,4\}}$[G] = $\{v_{6-16}, v_9\}$,
- $RS^{v_{3-13},v_{3-13}}_{\{1,3,5\}}$[\$] = $\{v_8, v_{11}, v_{12}\}$,

According to these *RS*s, we will contruct a trie-like subgraph as shown in Fig. 8(c). From this subgraph, we can find another two query subsets $\{q_2, q_4, q_6\}$ and $\{q_1, q_3, q_5\}$, respectively satisfiable by two packages {A, H, G} and {A, Hs, H}.

In the execution of *SEARCH*( ), much redundancy may be conducted, but can be easily removed. See Fig. 9(a) for illustration.

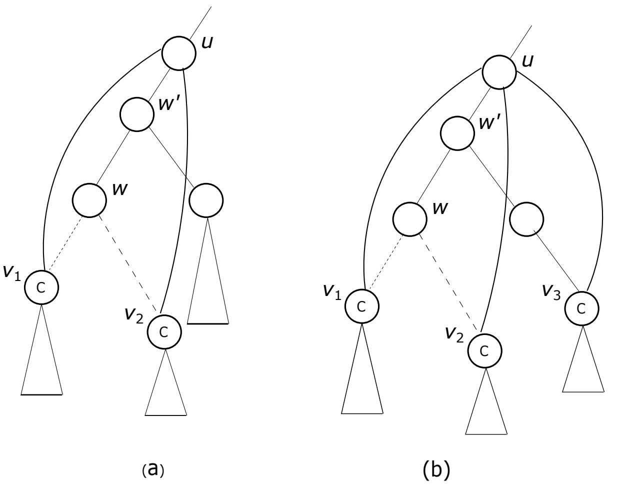


Figure 9: Illustration for redundancy.

In this figure, w and w' are two branching nodes. With respect to w and w', node u will have the same *RS*s. That is, we have

$$RS^{w,u}_s[\mathrm{C}] = RS^{w',u}_s[\mathrm{C}] = \{v_1, v_2\}.$$

Then, in the execution of *SEARCH*( ), the corresponding trie-like subgraph will be created two times, but with the same result produced.

However, this kind of redundancy can be simply removed in two ways.

In the first method, we can examine, by each recursive call, whether the input subgraph has been checked before. If it is the case, the corresponding recursive call will be suppressed.

In the second method, we create *RS*s only for those nodes appearing on the segment (on a tree path) between the current branching node and the lowest ancestor branching node in T. In this way, we may lose some answers. But the most popular package can always be found. See Fig. 9(b) for illustration. In this case, the *RS* of u with respect to w is different from the *RS* with respect to w'. However, when we check the branching node w, $RS^{w,u}_s$[C] will not be computed and therefore the corresponding result will not be generated. But $RS^{w',u}_s$[C] must cover $RS^{w,u}_s$[C]. Therefore, a package satisfying a larger, or a same-sized subset of queries will definitely be found.

4.3. Time complexity analysis

The total running time of the algorithm consists of three parts.

The first part, denoted as $\tau_{\cdot}$, is the time for computing the frenquencies of attribute appearances in Q. Since in this process each attribut in a q_i is accessed only once, τ_1 = O(nm).

The second part, denoted as τ_2, is the time for constructing a trie-like graph G for Q. This part of time can be further partitioned into three portions.

- τ_{21}: The time for sorting attribute sequences for q_i's. It is obviously bounded by O($nm\log_2 m$).
- τ_{22}: The time for constructing p^*-graphs for each q_i (i = 1, ..., n). Since for each variable sequence a transitive closure over its spans should be first created and needs O(m^2) time, this part of cost is bounded by O(nm^2).
- τ_{23}: The time for merging all p^*-graphs to form a trie-like graph G, which is also bounded by O(nm^2).

The third part, denoted as τ_3, is the time for searching G to find a maximum subset of conjunctions satisfied by a certain truth assignment. It is a recursive procedure. To analyze its running time, therefore, a recursive equation should established. Let $l = nm$. Assume that the average outdegree of a node in T is d. Then the average time complexity of τ_4 can be characterized by the following recurrence:

$$\Gamma(l) = \begin{cases} O(1), & \text{if } l \leq \text{a constant}, \\ \sum_{i=1}^{\lceil log_d\ l \rceil} d^i \Gamma(\frac{l}{d^i}) + O(l^2 m), & \text{otherwise.} \end{cases} \tag{18}$$

Here, in the above recursive equation, O(l^2m) is the cost for generating all the reachable subsets of a node through spans and upper boundries, together with the cost for generating local trie-like subgraphs for each recursive call of the algorithm. We notice that the size of all the RSs together is bounded by the number of spans in G, which is O(lm).

From (4), we can get the following inequality:

$$\Gamma(l) \leq d \cdot log_d\ l \cdot \Gamma(\frac{l}{d}) + O(l^2 m). \tag{19}$$

Solving this inequality, we will get

$$\begin{aligned} \Gamma(l) &\leq d \cdot log_d\ l \cdot \Gamma(\tfrac{l}{d}) + O(l^2m) \\ &\leq d^2 (log_d\ l)(\ log_d \tfrac{l}{d}) \Gamma(\tfrac{l}{d^2}) +\ (log_d\ l)\ l^2 m +\ l^2 m \\ &\leq \ldots \ldots \\ &\leq d^{\lceil log^l_d \rceil} (log_d\ l)\ (log_d(\tfrac{l}{d}))\ \ldots\ (\ log_d \frac{l}{d^{\lceil log^l_d \rceil}}) \\ &+ l^2 m ((log_d\ l)(log_d(\tfrac{l}{d}))\ \ldots\ (log_d \frac{l}{d^{\lceil log^l_d \rceil}}) + \ldots + log_d\ l + 1) \\ &\leq O(l(log_d\ l)^{log_d\ l} +\ O(l^2 m (log_d\ l)^{log_d\ l}) \\ &\sim O(l^2 m\ (log_d\ l)^{log_d\ l}). \end{aligned} \tag{20}$$

Thus, the value for τ_3 is $\Gamma(l) \sim$ O($l^2 m\ (log_d\ l)^{log_d\ l}$).

From the above analysis, we have the following proposition.

Proposition 3. *The average running time of our algorithm is bounded by*

$$\begin{aligned} \Sigma_{i=1}^4 \tau_i &= O(nm) + (O(nm\ log_2\ m) + 2 \times O(nm^2)) \\ &\quad + O(l^2 m\ (log_d\ l)^{log_d\ l}) \\ &= O(n^2 m^3 (log_d\ nm)^{log_d\ nm}). \end{aligned} \tag{21}$$

But we remark that if the average outdegree of a node in T is < 2, we can use a brute-force method to find the answer in polynomial time. Hence, we claim that the worst case time complexity is bounded by O($l^2 m (log_2\ l)^{log_2\ l}$) since ($log_d\ l)^{log_d\ l}$ decreases as d increases.

5. Conclusions

In this paper, we have discussed two new method to solve the problem to find a most popular package in terms of a given questionnaire. The first method is based on a kind of tree search, but with the priority queue structure being utilized to control the tree exploration. Together with a powerful heuristics, this approach enables us to cut off a lot of futile branches and find an answer as early as possible in the tree search process. The average time complexity is bounded by O($nm^2 + m2^{m/2}$), where $n = |Q|$ and m is the number of attributes in the query log $|Q|$. The main idea behind the second method is to construct a graph structure, called p*-graph. In this way, all the queries in Q can be represented as a trie-like graph. Searching G bottom up, we can find the answer efficiently. the average time complexity of the algorithm is bounded by O($n^2 m^3 (log_2\ nm)^{log_2\ nm}$).

As a future work, we will make a detailed analysis of the impact of the heuristics discussed in Section 4.2 to avoid any repeated recursive calls. If it is the case, the number of recursive calls for each branching node will be bounded by O(m) since the height of the trie-like graph G is bounded by O(m). Thus, the worst-case time complexity of our algorithm should be bounded by O(n^2m^4). It is because we have at most O(nm) branching nodes, and for each recursive call we need O(nm^2) time to construct a dynamical trie. So, the total running time will be O(nm) × O(m) × O(nm^2) = O(n^2m^4).

Conflict of Interest The authors declare no conflict of interest.

References

[1] R. Agrawal, T. Imieliński, A. Swami, "Mining association rules between sets of items in large databases," in Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207–216, Association for Computing Machinery, 1993, doi:10.1145/170035.170072.

[2] B. Gavish, D. Horsky, K. Srikanth, "An Approach to the Optimal Positioning of a New Product," Management Science, **29**, 1277–1297, 1983, doi:10.1287/mnsc.29.11.1277.

[3] T. S. Gruca, B. R. Klemz, "Optimal new product positioning: A genetic algorithm approach," European Journal of Operational Research, **146**, 621–633, 2003, doi:https://doi.org/10.1016/S0377-2217(02)00349-1.

[4] J. Resig, A. Teredesai, "A framework for mining instant messaging services," in In Proceedings of the 2004 SIAM DM Conference, 2004.

[5] J. Han, J. Pei, Y. Yin, "Mining frequent patterns without candidate generation," SIGMOD Rec., **29**, 1–12, 2000, doi:10.1145/335191.335372.

[6] M. Miah, G. Das, V. Hristidis, H. Mannila, "Standing Out in a Crowd: Selecting Attributes for Maximum Visibility," in 2008 IEEE 24th International Conference on Data Engineering, 356–365, 2008, doi:10.1109/ICDE.2008.4497444.

[7] J. C.-W. Lin, Y. Li, P. Fournier-Viger, Y. Djenouri, L. S.-L. Wang, "Mining High-Utility Sequential Patterns from Big Datasets," in 2019 IEEE International Conference on Big Data (Big Data), 2674–2680, 2019, doi:10.1109/BigData47090.2019.9005996.

[8] A. Tonon, F. Vandin, "gRosSo: mining statistically robust patterns from a sequence of datasets," Knowledge and Information Systems, **64**, 2329–2359, 2022, doi:10.1007/s10115-022-01689-2.

[9] Y. Chen, W. Shi, "On the Designing of Popular Packages," in 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 937–944, 2018, doi:10.1109/Cybermatics_2018.2018.00180.

[10] M. Miah, "Most popular package design," in 4th Conference on Information Systems Applied Research, 2011.

[11] R. Kohli, R. Krishnamurti, P. Mirchandani, "The Minimum Satisfiability Problem," SIAM Journal on Discrete Mathematics, **7**, 275–283, 1994, doi:10.1137/S0895480191220836.

[12] S. A. Cook, The Complexity of Theorem-Proving Procedures, 143–152, Association for Computing Machinery, 1st edition, 2023.

[13] Y. Chen, "Signature files and signature trees," Information Processing Letters, **82**, 213–221, 2002, doi:https://doi.org/10.1016/S0020-0190(01)00266-6.

[14] Y. Chen, "On the signature trees and balanced signature trees," in 21st International Conference on Data Engineering (ICDE'05), 742–753, 2005, doi:10.1109/ICDE.2005.99.

[15] Y. Chen, Y. Chen, "On the Signature Tree Construction and Analysis," IEEE Transactions on Knowledge and Data Engineering, **18**, 1207–1224, 2006, doi:10.1109/TKDE.2006.146.

[16] F. Grandi, P. Tiberio, P. Zezula, "Frame-sliced partitioned parallel signature files," in Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 286–287, Association for Computing Machinery, 1992, doi:10.1145/133160.133211.

[17] D. L. Lee, Y. M. Kim, G. Patel, "Efficient signature file methods for text retrieval," IEEE Transactions on Knowledge and Data Engineering, **7**, 423–435, 1995, doi:10.1109/69.390248.

[18] D. L. Lee, C.-W. Leng, "Partitioned signature files: design issues and performance evaluation," ACM Trans. Inf. Syst., **7**, 158–180, 1989, doi:10.1145/65935.65937.

[19] Z. Lin, C. Faloutsos, "Frame-sliced signature files," IEEE Transactions on Knowledge and Data Engineering, **4**, 281–289, 1992, doi:10.1109/69.142018.

[20] E. Tousidou, P. Bozanis, Y. Manolopoulos, "Signature-based structures for objects with set-valued attributes," Information Systems, **27**, 93–121, 2002, doi:https://doi.org/10.1016/S0306-4379(01)00047-3.

[21] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, Introduction to algorithms, MIT press, 2022.

[22] P. Flajolet, C. Puech, "Partial match retrieval of multidimensional data," J. ACM, **33**, 371–407, 1986, doi:10.1145/5383.5453.

[23] L. Debnath, D. Bhatta, Integral transforms and their applications, Chapman and Hall/CRC, 2016.

[24] E. K. Donald, "The art of computer programming," Sorting and searching, **3**, 4, 1999.

Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 4, 73-78 (2024)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Effectiveness of a voice analysis technique in the assessment of depression status of individuals from Ho Chi Minh City, Viet Nam: A cross-sectional study

Le Truong Vinh Phuc [1,2], Mituteru Nakamura[3], Masakazu Higuchi[3], Shinichi Tokuno[1,3]

[1]*Graduate School of Health Innovation, Kanagawa University of Human Services, Kanagawa, 241-0815, Japan.*

[2]*University of Medicine and Pharmacy at Ho Chi Minh City, Ho Chi Minh City, 749000, Vietnam.*

[3]*Voice Analysis and Measurement of Pathophysiology, Department of Bioengineering, Graduate School of Engineering, The University of Tokyo, Tokyo, 113-0033, Japan.*

ARTICLE INFO

Article history:
Received:04 June, 2024
Revised: 29 July, 2024
Accepted: 01 August, 2024
Online: 03 August, 2024

Keywords:
Mind Monitoring System
Mental health assessment
Depression
Voice analysis
Cross-sectional study

ABSTRACT

The Mind Monitoring System (MIMOSYS) is a novel voice analysis technique for mental health assessment that has been validated in some languages; however, no research has been conducted on the Vietnamese yet. This study aimed to examine the ability of the Vitality score extracted from the MIMOSYS system to assess depression status based on the standards of the Patient Health Questionnaire-9 items (PHQ-9) and Beck's Depression Inventory (BDI) questionnaire using the Vietnamese language. In this cross-sectional study conducted between August 1, 2022 and September 4, 2022, students and staff from a university, and patients from a hospital were recruited. Participants were asked to complete the self-administered depression questionnaires (PHQ-9 and BDI), and their voice data was collected by reading designated sentences. MIMOSYS extracted the Vitality score from the participant's voice data. One hundred and twenty-two participants with a mean age of 31 years were included in the study; 72.4% of them were female. After adjusting for age and sex, negative correlations between the Vitality score and psychological test scores were found. For discriminating individuals with a high risk of depression, using the BDI score as a standard, the area under the curve of the Vitality score was 0.72. Sensitivity, specificity, and accuracy evaluations also reported a moderate discrimination ability of the Vitality score on the risk of depression by the BDI. In conclusion, voice analysis can be a viable technique for depression assessment in Vietnamese; however, further investigations are necessary to confirm our findings.

1. Introduction

Despite the fact that mental health disorders are recognized as an important global public health problem, mental health care remains a low priority in low- and middle-income countries [1], [2]. The Vietnamese Ministry of Health reported that the prevalence of depression and anxiety in the population is 2.8% and 2.6%, respectively [3]. Conventional methods, such as the use of self-administered questionnaires, are still effective for assessing depression [4] and have served as the gold standard for evaluating mood and emotional issues [5]. However, since patients can provide incorrect information in interviews and questionnaires, objective indicators are perceived to be more useful for diagnostic support [6]. Voice analysis was gradually applied to health monitoring as a non-invasive technique, which requires no specialized equipment and can be performed easily and remotely. In several languages, such as Japanese, English, and German, physical characteristics of voice such as pitch rate, jitter, shimmer, and Harmony to Noise Ratio were examined for their capability to discriminate between healthy individuals and those with mental disorders [7].

[*]Le Truong Vinh Phuc, Graduate School of Health Innovation, Kanagawa University of Human Services, Kanagawa, 241-0815, Japan, +81-80-9279-7409 & vinhphuc0104@gmail.com

https://dx.doi.org/10.25046/aj090408

Recently, a voice analysis system named the Mind Monitoring System (MIMOSYS) is in practical use to assess and monitor an individual's mental health status, particularly in Japan [8], [9], [10]. The significant difference in mean Vitality score, the short-term indicator of MIMOSYS, between the high-risk and low-risk depression groups was also confirmed by research on the Romanian and Russian languages (p-value < 0.05) [11]. Therefore, this voice analysis technique can be a potential tool to assess depression because it has the advantages of overcoming the barrier of languages, eliminating the bias of psychological tests, and being simple for implementation and management, which may help reduce the burden on the screening and diagnostic process. To our knowledge, no similar research has been conducted on the Vietnamese language yet. This study aimed to 1) examine the correlation between the Vitality score extracted from the MIMOSYS system and the depression scores measured by the Patient Health Questionnaire-9 items (PHQ-9), and the Beck's Depression Inventory (BDI), using the Vietnamese language, and 2) evaluate the performance of the Vitality score in discriminating individuals with a high risk of depression using the thresholds of the PHQ-9 and BDI as standards.

2. Methods

2.1. Ethics statements

The study protocol was approved by the Ethics Committee of the Graduate School of Health Innovation, Kanagawa University of Human Services (No. SHI-03), and by the Ethics Committee in Biomedical Research of the University of Medicine and Pharmacy at Ho Chi Minh City, Viet Nam (No. 397/HDDD-DHYD). All methods were performed in accordance with the guidelines stipulated in the Declaration of Helsinki. Written informed consent was obtained from each participant before data collection. The STROBE guidelines were followed in reporting the study.

2.2. Participants

This cross-sectional study was conducted from August 1, 2022 to September 4, 2022 and recruited individuals from the Faculty of Public Health, University of Medicine and Pharmacy at Ho Chi Minh City, and outpatients visiting the Department of Clinical Psychology, Le Van Thinh Hospital in Ho Chi Minh City, Viet Nam. To be eligible for study inclusion, individuals must be 18 years or older at the time of the study, have Vietnamese nationality, and use Vietnamese as their native language. Additionally, at the hospital, outpatients had to be diagnosed with major depressive disorder according to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) criteria by a psychologist. Individuals with speech problems, such as apraxia, dysarthria, orofacial myofunctional disorders, stuttering, or aphasia, those with congenital hearing impairment, or those with confirmed organic brain diseases, were excluded from our study. Data was collected in a separate room, with one participant interviewed each time.

Altogether, 125 potential eligible individuals were approached, including 100 students and staff from the school and 23 depressive patients from the hospital. Two patients were younger than 18 years old; thus, they were excluded from our study. A total of 123 participants were finally included in our analysis (Figure 1).

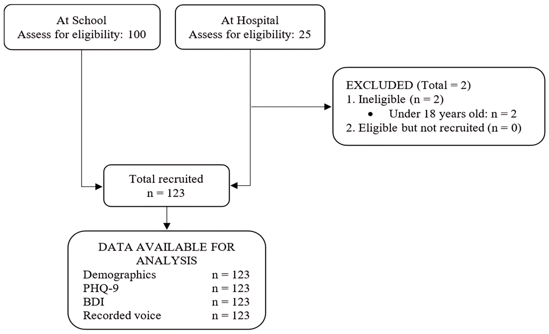


Figure 1: Recruitment of participants and data for analysis

2.3. Voice data

A sheet of sixteen designated phrases used for recording voice was provided. The participant's voice was recorded by reading out loud the phrases one by one that followed the signal of the data collector. Phrases were read in a constant order for all participants. Phrases used in this study were conventionally translated into Vietnamese based on the meaning of the original Japanese phrases and in comparison with the English version of the phrases. Phrases were also categorized into two groups: reading phrases (phrases with full sentences) and vowel sounds (/Ah/, /Eh/, and /Uh/ phrases). Participant's study ID number was also included in the recording to match their response in paper-based questionnaires.

Voice recordings were acquired at 24-bit and 96 kHz resolution using a TASCAM DR-100MK3 portable digital audio recorder (Tokyo, Japan) and an Olympus ME52W lavalier microphone (Tokyo, Japan). For the speech analysis, we used MIMOSYS (PST Inc., Kanagawa, Japan), a system used for monitoring health status through voice based on the sensibility technology [12]. For the requirements of the MIMOSYS data input format, the recorded speech was converted into 11 kHz, 16-bit audio data for analysis. The Adobe® Audition™ software version 1.5 (Adobe Inc., San Jose, California) was used for the audio conversion.

In brief, using MIMOSYS, the degree of intensity of each of the four emotional components ("calmness", "anger", "joy", and "sorrow") is firstly calculated from the voiced speech on an eleven-point scale of zero to 10. On a ten-point scale of one to ten, MIMOSYS also calculates the intensity of "excitement." The intensity of these five variables is then used to calculate "vivacity" and "relaxation". "vivacity" is determined by the components of "joy" and "sorrow", whereas "relaxation" is determined by the components of "calmness" and "excitement". Finally, "vivacity" and "relaxation" are used to calculate "Vitality" [7]. The data calculation flow of MIMOSYS is presented in Figure 2 [8]. The output, Vitality score, is a real numeric value ranging from zero to one. A higher Vitality score indicates a better mental state. In addition, the mean and variation in Vitality over two weeks can be used to generate the "Mental Activity" indicator [13].

In this study, the Vitality was characterized as a scale on which individuals suffering from depression got a lower score, and healthy individuals got a higher score. The Vitality score of each phrase and each combination of phrases was used to examine the study objectives. The combination that provided the overall best results was reported.

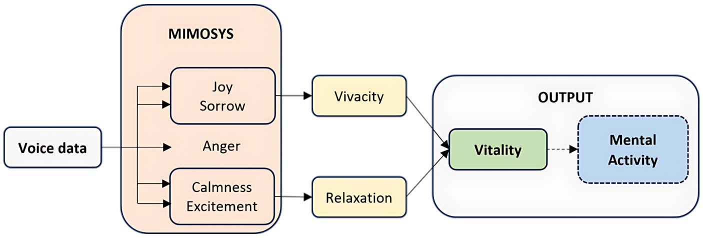


Figure 2: Diagram for Vitality calculation by MYMOSIS

2.4. Self-administered questionnaires

Participants were asked to complete self-administered questionnaires to determine their depressive symptoms. The questionnaires used were the PHQ-9 and BDI in Vietnamese. These questionnaires are frequently used assessment instruments for depression, freely accessible, and the Vietnamese version of these tools is also available [14], [15]. The sum scores of the PHQ-9 and BDI tests were included in our analysis as dependent variables. The score ranges from 0 to 27 for the PHQ-9 and 0 to 63 for the BDI; a lower score indicates a milder depression status. The PHQ-9 and BDI cutoff scores of 12 [8] and 19 [15], respectively, were used to define an individual with a high risk of depression. The factors considered as potential effect modifiers were age and sex.

2.5. Data analysis

Pearson's correlation analysis was applied to evaluate the correlation between the Vitality score and PHQ-9 or BDI score. The strength of the correlation was categorized as follows based on the absolute correlation coefficient: 0.5–1, strong correlation; 0.3–<0.5, moderate correlation; and <0.3, weak correlation. T-test analysis was applied to evaluate the difference in Vitality score between the low-risk and high-risk groups of depression categorized by the PHQ-9 or BDI sum score.

The area under the curve (AUC) from the receiver-operating characteristics (ROC) model using logistic regression was applied to examine the classification performance of the Vitality score. The performance was judged according to the AUC value: 0.5–< 0.7 as low accuracy, 0.7–< 0.9 as moderate accuracy, and 0.9–1 as high accuracy [6]. Our ROC models used a BDI score threshold of 19 and a PHQ-9 score threshold of 12. The ROC curve was plotted with sensitivity as the vertical axis and 1–specificity as the horizontal axis by altering the threshold of the Vitality score.

The sensitivity, specificity, and accuracy of the optimal Vitality cutoff score using the PHQ-9 or BDI as the standard were also reported. The optimal cutoff point for the Vitality score was specified with 1) a sensitivity value greater than 0.5, 2) a specificity value greater than 0.5, and 3) the value of Youden's J index (sensitivity + specificity – 1) being the highest.

All the statistical tests were considered significant at α level of 5%. The online version of the SAS program was used to analyze and visualize the data.

3. Results

The participants' mean age was 31 years, and 89 of them were female (72%). Most of the participants lived in Ho Chi Minh City (93%) (Table 1).

Table 1: Paritipants' baseline characteristics (n=123)

Characteristics	Results
Age (years), mean ± sd	31 ± 12
Sex (female), n (%)	89 (72)
Residence (Ho Chi Minh City), n (%)	114 (93)
History of depression (Yes), n (%)	30 (24)

sd: standard deviation

The average scores in the PHQ-9 and the BDI tests among the participants were 6.9 ± 5.5 and 10.5 ± 9.4 points, respectively. The number of participants at high risk of depression was 19 and 20 based on the PHQ-9 cut and BDI cutoff scores (Table 2).

Table 2: Results of the PHQ-9 and BDI tests (n=123)

Psychological tests	Results
PHQ-9 score, mean ± sd	6.9 ± 5.5
High risk of depression, n (%)	19 (15.5)
BDI score, mean ± sd	10.5 ± 9.4
High risk of depression, n (%)	20 (16.3)

sd: standard deviation

3.1. Correlation between the Vitality and test scores

The vitality score and correlation coefficient of each phrase are presented in Table 3. Mean Vitality score ranged from 0.36 to 0.56, Not all phrases had a negative correlation and/or significant result with the psychological tests (Table 3).

Table 3: Vitaltiy score and correlation of each phrase (n=123)

Phrase	Vitality	Correlation	
	mean (sd)	PHQ-9	BDI
A B C D E F G	0.36 (0.12)	−0.049	−0.099
Alpha beta gamma delta	0.44 (0.13)	−0.021	−0.011
It is a sunny day today	0.46 (0.16)	−0.020	0.005
Once upon a time	0.42 (0.13)	−0.038	0.012
Galapagos Islands	0.46 (0.14)	−0.049	−0.092
I'm very fine	0.46 (0.15)	−0.189*	−0.173
Yesterday I had a good sleep	0.47 (0.18)	−0.107	−0.218*
I have an appetite	0.52 (0.19)	−0.200*	−0.184*
I feel calm	0.48 (0.15)	0.082	0.042
I feel irritable	0.49 (0.17)	0.023	0.011
I feel tired and exhausted	0.56 (0.17)	0.116	0.168
I always looked ahead	0.40 (0.15)	−0.078	−0.129
Come on, myself!	0.39 (0.14)	−0.108	−0.149
/Ah/ sound	0.41 (0.15)	−0.179*	−0.193*
/Eh/ sound	0.38 (0.13)	−0.061	−0.111
/Uh/ sound	0.38 (0.13)	−0.138	−0.208*

*: *p-value < 0.05*

Vitality scores from 65,519 phrase combinations were also examined for the study objectives. Among them, the Vitality score from a combination of six phrases showed the highest overall results and was chosen for the report.

The correlation between the Vitality and test scores, adjusted for age and sex, is shown in Figure 2. The adjusted correlation between the Vitality and BDI scores was negative with a value of −0.272 and was significant (p-value = 0.003) (Figure 3). Similar result of correlation was found between the Vitality and PHQ-9 score (r = −0.215, p-value = 0.018) (Figure 4).

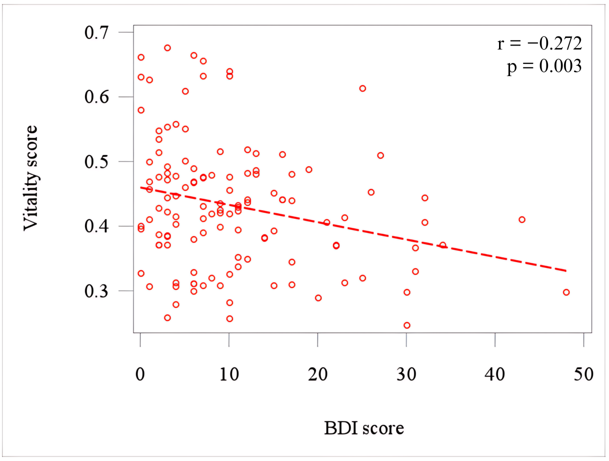


Figure 3: Scatter plot and regression line of the Vitality and BDI score

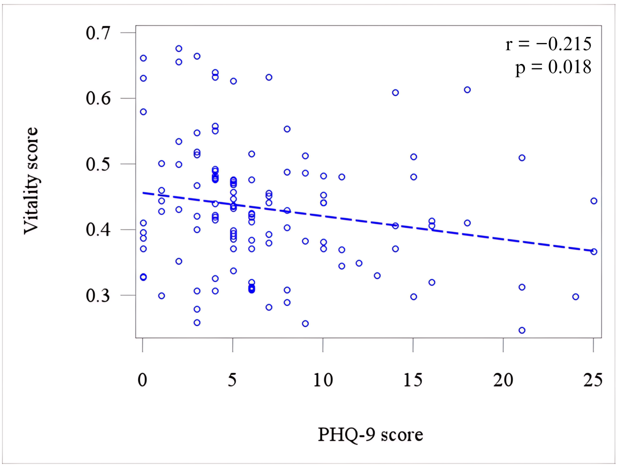


Figure 4: Scatter plot and regression line of the Vitality and PHQ-9 score

The box and whisker plots of the Vitality scores among those with a low and a high risk of depression are shown in Figure 5. When comparing the participants categorized according to their BDI score, the high-risk group showed a significantly lower Vitality score than the low-risk group (p-value at 0.024). However, there was no significant difference in the Vitality score between the groups categorized according to the PHQ-9 score (Figure 6).

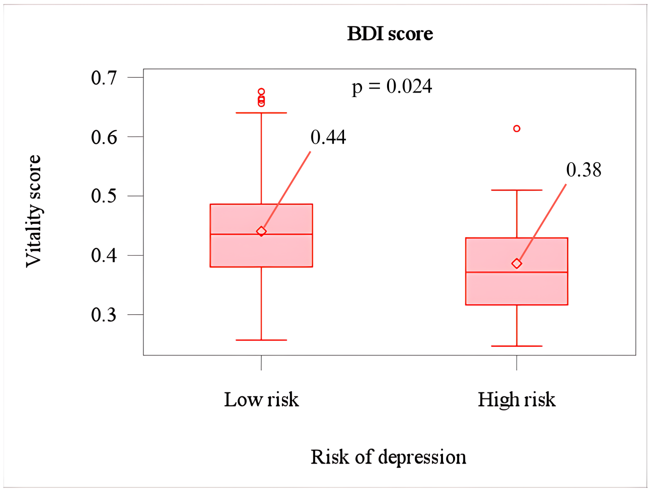


Figure 5: Box and whisker plot of the Vitality score on low-risk and high-risk group of depression by BDI score

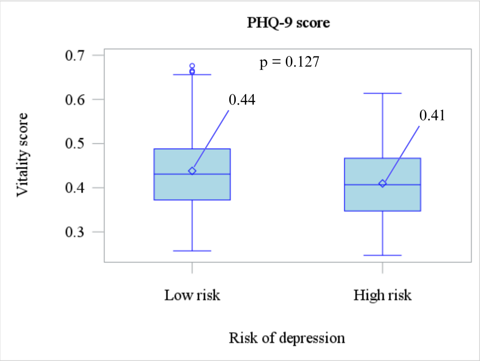


Figure 6: Box and whisker plot of the Vitality score on low-risk and high-risk group of depression by BDI score

3.2. *MIMOSYS performance in discriminating individuals with a high risk of depression*

Using the BDI as a standard, the AUC of the Vitality score controlled for age and sex was 0.717, and the ROC model is significant (p-value = 0.001) (Figure 7). Using the PHQ-9 as standard, the AUC was 0.623; however, the ROC model is not significant (p-value > 0.05) (Figure 8). The sensitivity, specificity, and accuracy of the "optimal cutoff point" in discriminating individuals with a low and a high risk of depression were reported in Table 4. A comparison regarding the effectiveness of the Vitality score in identifying depressive disorder between this study and previous studies was additionally shown in Table 5.

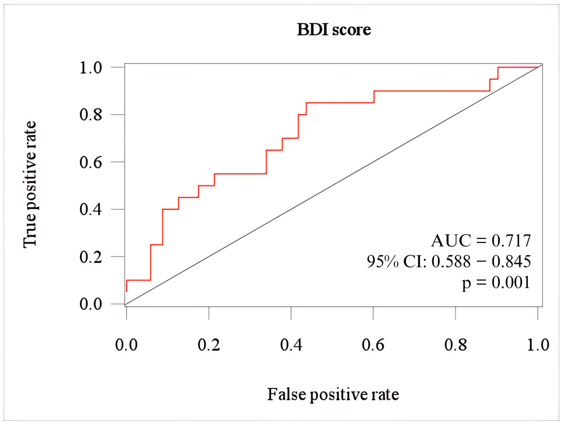


Figure 7: The ROC curve of Vitality score using the BDI score as standard

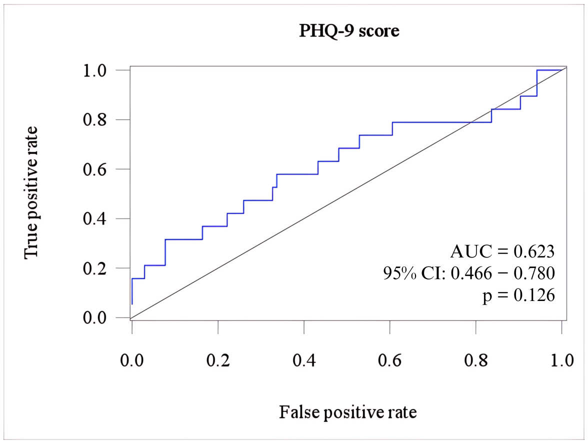


Figure 8: The ROC curve of Vitality score using the PHQ-9 score as standard

Table 4: Discriminating performance of the Vitality score (n=123)

Vitality cutoff = 0.4139	PHQ-9 score cutoff = 12		Sen.	Spec.	Acc.
	High (n)	Low (n)			
High risk	13	42	0.68	0.60	0.61
Low risk	6	62			
Vitality cutoff = 0.4139	**BDI score cutoff = 19**		**Sen.**	**Spec.**	**Acc.**
	High (n)	**Low (n)**			
High risk	15	40	0.75	0.61	0.63
Low risk	5	63			

Sen.: Sensitivity; Spec.: Specificity; Acc.: Accuracy

Table 5: Performance of the Vitality score in different studies using BDI (n=123)

Author	Hagiwara [8]	Higuchi [16]	Shinohara [17]	This study
Year	2017	2017	2017	2023
Language	Jap.	Jap.	Jap.	Vie.
AUC	0.78	0.64	0.80	0.72
Sensitivity	0.80	NR	0.94	0.75
Specificity	0.64	NR	0.64	0.61
Accuracy	0.66	NR	NR	0.63

Jap.: Japanese; Vie.: Vietnamese; NR: not reported

4. Discussion

Our study evaluated the effectiveness of the Vitality, a biomarker retrieved from voice, to assess the depression status of Vietnamese individuals. A lower Vitality score denotes a more severe state of depression [17], while the PHQ-9 and the BDI questionnaires interpret a lower score as milder depressive status. In other words, the Vitality score and these psychological test scores are expected to have a negative correlation.

In this study, voice data were divided into utterances; each utterance corresponded to a voice uttered for a designated phrase. It is worth noting that not all of the recorded phrases showed a negative correlation with psychological tests; a similar issue was also reported in the research conducted in the Japanese language [10]. Among all combinations of phrases generated and examined from our study, the Vitality score from the combination of six utterances was chosen for analyzing the study objectives. The combination consists of Phrase 1 (A-B-C-D-E-F-G), Phrase 6 (I am very fine), Phrase 7 (Yesterday I had a good sleep), Phrase 8 (I have an appetite), Phrase 12 (I always look ahead), and Phrase 13 (Come on, myself!). The vitality score extracted from each of these phrases was negatively correlated with both the PHQ-9 score and the BDI score (as shown in Table 3). Our proposed combination also met the minimum number of utterances recommended for the MIMOSYS to achieve sufficient accuracy of the Vitality score [9].

4.1. Correlation between the Vitality score and self-administered psychological test scores

In our study, a weak negative correlation was found between the Vitality and the PHQ-9 score as well as the BDI score. A similar trend of correlation was also reported in a previous study using the BDI test and voice data recorded via phone, with a smaller magnitude of 0.208 [8]. The difference in results can possibly be due to the difference in the way the voice data was collected. In our study, the participant's voice was recorded by a specialized recorder, in a noise-controlled environment. Since the interference of noise from the surrounding environment may decrease the precision of utterance detection, it can lead to an unproperly analysis of the Vitality score [18] Another factor that may affect the result is whether the voice data is reading speech or spontaneous speech. As the Vitality score is calculated based on changes in emotion, it is likely that reading a fixed sentence would generate less variance in emotional expression than, for example, a phone conversation, which resulted in a lower value of the Vitality score [8].

Regarding the difference in the Vitality score between groups categorized by the test's cutoff, a significant result was found only when using the BDI test. Previous studies using the same questionnaire in other languages such as Japanese, Romanian, or Russian also reported similar results [11], [17]. Result using the PHQ-9 test, however, showed no difference in Vitality score between the low-risk group and high-risk group. Since the PHQ-9 includes 9 items with a relatively narrow sum score, it does not necessarily reflect an individual's personal depression experiences in their life [19], [20]. The questionnaire is also commonly used to measure outcomes in primary care rather than being used as a diagnostic tool [21].

4.2. Performance in discriminating individuals with high risk of depression

Using the BDI test as a standard, a moderate accuracy AUC of Vitality score was found in our study. Our result was higher than a similar study conducted on Japanese seniors [22], but lower than the study involving Japanese employees [8]. The result of the AUC may be interfered by other un-examined factors such as the homogeneity of the study population, number of participants, or the voice data collecting method.

Results on the sensitivity, specificity, and accuracy from our study were relatively close to, but lower than the report numbers from the previous studies conducted in Japanese [8], [17]. Enhancing the performance of Vitality, especially the specificity result, remains a future research issue for the applicability of this voice analysis technique in Vietnamese. This result also suggests the potential effect of acoustic differences between the Japanese and Vietnamese. Further studies focused on addressing this difference may help clarify the mechanism of pathophysiological voice analysis in different languages.

4.3. Limitations

Due to the small sample size and convenient sampling method, the generalizability of our findings was limited. Compared to the Vitality score, the MIMOSYS's Mental Activity score, which was not generated and evaluated in our study, was reported to have a higher performance both in the correlation and discrimination evaluations with the BDI test [8]. Besides the emotion-based technique, voice data can be analyzed by its acoustical characteristics such as shimmer, jitter, or harmonic-to-noise Ratio. However, this technique was not included and parallel examined in our study; thus, the performance of MIMOSYS may not have been comprehensively compared yet. Spontaneous or free speech is reported to better classify individuals with a major depressive disorder than reading designated sentences [7]. Therefore, further comparison of the results from different types of voice data in the Vietnamese language is necessary.

5. Conclusion

Although our findings may support the use of the voice analysis technique in the Vietnamese language by its acceptable results, further investigations are necessary to confirm and enhance the performance of depression assessment of this method.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We would like to express our sincere appreciation to the Faculty of Public Health, University of Medicine and Pharmacy at Ho Chi Minh City, and to the Department of Clinical Psychology, Le Van Thinh Hospital, Ho Chi Minh City for their substantial support in conducting the research. We also acknowledge Enago (www.enago.com) for English language review.

References

[1] F.T. Alloh, P. Regmi, I. Onche, E.v. Teijlingen, S. Trenoweth, "Mental Health in low-and middle income countries (LMICs): Going beyond the need for funding," Health Prospect, **17**(Special Issue), 12–17, 2018, doi:10.3126/hprospect.v17i1.20351.

[2] World Health Organization, World Mental Health Report; Transforming Mental Health for all. WHO: Geneva, 2022.

[3] D.A. Vuong, E.v. Ginneken, J. Morris, S.T. Ha, R. Busse, "Mental health in Vietnam: burden of disease and availability of services," Asian J Psychiatr, **4**(1), 65–70, 2011, doi:10.1016/j.ajp.2011.01.005.

[4] S. Shinohara, H. Toda, M. Nakamura, Y. Omiya, M. Higuchi, T. Takano, T. Saito, M. Tanichi, S. Boku, S. Mitsuyoshi, et al., "Evaluation of emotional arousal level and depression severity using voice-derived sound pressure change acceleration," Scientific Reports, **11**(1), 13615, 2021, doi:10.1101/2020.08.19.20177048.

[5] N.K. Patel, L. Nivethitha, A. Mooventhan, "Effect of a yoga based meditation technique on emotional regulation, self-compassion and mindfulness in college students," Explore, **14**(6), 443–447, 2018, doi:10.1016/j.explore.2018.06.008.

[6] Y. Omiya, T. Takano, T. Uraguchi, M. Nakamura, M. Higuchi, S. Shinohara, S. Mitsuyoshi, M. So, S. Tokuno, An attempt to estimate depressive status from voice – in book: Pervasive Computing Paradigms for Mental Health, Springer, 2019.

[7] S. Shinohara, Y. Omiya, M. Nakamura, N. Hagiwara, M. Higuchi, S. Mitsuyoshi, S. Tokuno, "Multilingual evaluation of voice disability index using pitch rate," Advances in Science, Technology and Engineering Systems Journal, **2**(3), 765–772. 2017, doi:10.25046/aj020397.

[8] N. Hagiwara, Y. Omiya, S. Shinohara, M. Nakamura, M. Higuchi, S. Mitsuyoshi, H. Yasunaga, S. Tokuno, "Validity of Mind Monitoring System as a mental health indicator using voice," Advances in Science Technology and Engineering Systems, **2**(3), 338–344, 2017, doi:10.25046/aj020343.

[9] M. Higuchi, M. Nakamura, S. Shinohara, Y. Omiya, T. Takano, S. Mitsuyoshi, S. Tokuno, "Effectiveness of a voice-based mental health evaluation system for mobile devices: prospective study," JMIR Formative Research, **4**(7), e16455, 2020, doi:10.2196/16455.

[10] M. Higuchi, N. Sonota, M. Nakamura, K. Miyazaki, S. Shinohara, Y. Omiya, T. Takano, S. Mitsuyoshi, S. Tokuno, "Performance evaluation of a voice-based depression assessment system considering the number and type of input utterances," Sensors, **22**(67), 1–12, 2022, doi:10.3390/s22010067.

[11] T. Uraguchi, S. Shinohara, M. Taicu, G. Savoiu, Y. Omiya, M. Nakamura, M. Higuchi, T. Takano, N. Hagiwara, S. Mitsuyoshi, et al., "Evaluation of Mind Monitoring System (MIMOSYS) by subjects with Romanian and Russian as their native language," in 2018 IEEE International Conference in Medicine and Biology Society (EMBC2018).

[12] S. Mitsuyoshi, F. Ren, Y. Tanaka, S Kuroiwa, "Non-verbal voice emotion analysis system," International journal of innovative computing, information & control, **2**(4), 819–830, 2006.

[13] S. Shinohara, M. Nakamura, Y. Omiya, M. Higuchi, N. Hagiwara, S. Mitsuyoshi, H. Toda, T. Saito, M. Tanichi, A. Yoshino, et al., "Mental health assessment method based on emotion level derived from voice," Preprints, 2020, doi:10.21203/rs.2.18354/v1.

[14] N.H.B. An, N.T.T. Ngoc, V.Q. Nghia, H.T.N. Ngoc, B.C. Minh, L.V. My, C.T. Nha, N.H. Dung, "Application of the Beck Depression Inventory – II, Zung Self-Rating Anxiety Scale and Pittsburgh Sleep Quality Index on student of School of Medicine – Vietnam National University Ho Chi Minh City: a pilot study," (article in Vietnamese with an abstract in English), Journal of Science and Technology Development – Health Sciences, **2**(2), 323–329, 2021, doi:10.32508/stdjhs.v2i2.490.

[15] N.T. Nguyen, P.L. An, N.D.P. Tien, "Reliability and validity of Vietnamese version of patient health questionnaire 9 items (PHQ-9) among UMP medical freshmen," in 8th International Conference on the Development of Biomedical Engineering in Vietnam, 901–923, 2019, doi:10.1007/978-3-030-75506-5_72.

[16] M. Higuchi, S. Shinohara, M. Nakamura, Y. Omiya, N. Hagiwara, S. Mitsuyoshi, S. Tokuno, "Study on depression evaluation indicator in the elderly using sensibility technology," in 3rd International Conference on Information and Communication Technologies for Ageing Well and e-Health, 70–77, 2017, doi:10.5220/0006316700700077.

[17] S. Shinohara, Y. Omiya, N. Hagiwara, M. Nakamura, M. Higuchi, T. Kirita, T. Takano, S. Mitsuyoshi, S. Tokuno, "Case studies of utilization of the Mind Monitoring System (MIMOSYS) using voice and its future prospects," Econophysics, Sociophysics & other Multidisciplinary Sciences Journal, **7**, 7–12, 2017.

[18] M. Higuchi, S. Shinohara, M. Nakamura, Y. Omiya, N. Hagiwara, T. Takano, S. Mitsuyoshi, S. Tokuno, "Accuracy evaluation for mental health indicator based on vocal analysis in noisy environments," Journal of Information and Communication Engineering, **4**(1), 217–222 2018, doi:10.5281/zenodo.4300086.

[19] C. Dowrick, G.M. Leydon, A. McBride, A. Howe, H. Burgess, P. Clarke, S. Maisey, T. Kendrick, "Patients' and doctors' views on depression severity questionnaires incentivised in UK quality and outcomes framework: Qualitative study," British Medical Journal, **338**(b663), 2009, doi:10.1136/bmj.b663.

[20] A. Malpass, A. Shaw, D. Kessler, D. Sharp, "Concordance between PHQ-9 scores and patients' experiences of depression: a mixed methods study," British Journal of General Practice, **60**(575), e231–e238, 2010, doi:10.3399/bjgp10X502119.

[21] J. Ford, F. Thomas, R. Byng, R. McCabe, "Use of the Patient Health Questionnaire (PHQ-9) in practice: interactions between patients and physicians," Qualitative Health Research, **30**(13), 2146–2159, 2020, doi:10.1177/1049732320924625.

[22] Y. Omiya, N. Hagiwara, M. Nakamura, S. Shinohara, M. Higuchi, T. Takano, "音声を用いたメンタルヘルス状態計測における年齢及び性別の影響の検証," Journal of Industrial Hygiene, **60**(Special Issue), 487, 2018.

Advances in Science, Technology and Engineering Systems Journal
Vol. 9, No. 4, 79-92 (2024)
www.astesj.com

ASTES Journal
ISSN: 2415-6698

IoT and Business Intelligence Based Model Design for Liquefied Petroleum Gas (LPG) Distribution Monitoring

Amalia Rodriguez Espinoza de los Monteros[*], Maximo Giovani Tandazo Espinoza[*], Byron Ivan Punina Cordova, Ronald Eduardo Tandazo Vanegas

Universidad Politécnica Salesiana, Computer Science, Guayaquil, Ecuador

ARTICLE INFO

Article history:
Received: 19 May, 2024
Revised: 07 July, 2024
Accepted: 08 July, 2024
Online: 18 August, 2024

Keywords:
Internet of Things
IoT
LPG
Sensing approach
Business Intelligence
Liquefied Petroleum Gas

ABSTRACT

Gas leakage caused by various causes poses significant risks to public safety. To address this problem, an intelligent model is proposed for the accurate monitoring of Liquefied Petroleum Gas (LPG) distribution based on the integration of Internet of Things (IoT) and Business Intelligence (BI) technologies. Through the use of sensors and actuators, it seeks to mitigate risks and prevent accidents by enabling automated control of devices and infrastructures. The PRISMA methodology was used to perform a systematic review and obtain general characteristics of the components. Then, the proposed model was evaluated according to Y. 4908 which addresses IoT network interoperability, usability and security, the evaluation with 30 IT professionals who examined the BI model. The results obtained by the professionals were encouraging and favorable. The proposal, which enables remote LPG monitoring, establishes service through a website, mobile app or SMS when it detects fluctuations in humidity, temperature and gas indicators, shuts off the flow of LPG and notifies immediately. The research led to the development of a model that combines an IoT component with a four-tier BI, demonstrating its effectiveness and acceptance in the professional arena. At the overall medium level, 49% strongly agree, 38% agree, 12% neither agree and 1% disagree. It is concluded that the model has an overall average level of approval of 87%.

1. Introduction

The growing demand for liquefied petroleum gas (LPG) and natural gas leads to an increase in production, so equipment and components become essential to maintain operability. It is necessary to improve company safety conditions by optimizing control and detection techniques for all machinery to ensure its correct operation for the safety of companies and people. Energy demand for heating, manufacturing, household appliances, and vehicle fuel comes mostly from liquefied petroleum gas; gas replaces gasoline or coal because of its environmental advantages [1]. On the other hand, gas dissipation occurs due to various factors, including gas leakage along the distribution line, compromised regulating mechanisms, faulty service joints, and variations in inlet gas pressure, among other contributing elements: gas leaks or leaks cause human hazards, economic losses and an increase in greenhouse gases. LPG distribution has no detection or control of gas leaks; it only generates manual reports or complaints from users, which creates large amounts of carbon emissions into the atmosphere [2].

It is known that GAS detection is essential across diverse domains, including medicine, industry, environmental studies, and surveillance of confined environments; in the automotive sector, it uses gas sensors in engines to minimize contamination and maximize power, in the treatment and monitoring of gases in hospitals, in mines gas monitoring ensures air quality and avoids harm to people due to lack of oxygen. LPG leaks cause fires, poisoning, and explosions upon combustion; costs are also related to the aftermath of gas leaks, such as the destruction of infrastructure, medical expenses, and loss of income [3]. LPG is highly flammable and used in households, vehicles, businesses, and other areas; there are varieties of LPG with propane and butane content; it has an explosive characteristic that poses an imminent risk of fire at any location or during transportation. Negligence and ignorance cause accidents; some implementations use gas sensors and are outdoor solutions for environmental monitoring, chemical processes, fire detection, and air quality. LPG leak accidents cause loss of human and material lives; accidents are indeed few but very dangerous; there are houses or citadels with LPG connections, although they

[*]Corresponding Authors: Amalia Rodriguez Espinoza de los Monteros & Maximo Giovani Tandazo Espinoza, Emails: arodriguez@ups.edu.ec; maximo.tandazo@samara.com.ec

https://dx.doi.org/10.25046/aj090409

are very few, and there are no post-gas leak detection tasks. It is common knowledge that gas leaks are a serious problem, regardless of the sector related to gas, such as companies, homes, factories, kitchens, and restaurants [4].

As the use of LPG increases, so does the number of accidents caused by LPG explosions; accidents caused by gas leaks increase property damage and fatalities. Some problems are poor quality piping, poorly made pipe joints, substandard gas cylinders, worn regulators, obsolete valves, and improper handling of gas cylinders. Other problems are that the leak detection system has certain drawbacks, such as a long response time, a lack of prevention mechanisms, and the installation of faulty or untested leak detection devices. [5]. In 2022, LPG consumption in Ecuador was 14.4 million barrels, of which 12.7 million were imported and 1.7 million were produced locally [6]. At the national level, LPG consumption is distributed as follows: domestic is 91%, industrial 7%, agricultural 1%, and automotive 1% [7]. It is well known that even the slightest LPG leakage can cause serious damage to people, homes, buildings, or businesses; the safety of people is a major concern. To prevent gas leaks, it is essential to continuously monitor the LPG supply; Internet of Things (IoT) technology can perform sophisticated control and monitoring of the supply [8]. IoT for natural gas distribution has shortcomings such as centralized resource management, lack of data flow between stations, insecurity of transaction information or ledgers, and lack of contract consensus [9]. Poor quality rubber tubing or regulators that do not shut off when unused are two major problems that cause gas leaks [5]. The proposed system has a solenoid valve and an exhaust fan to remove the leaked gas. As mentioned in [1], the system focuses on gas preservation, safety precaution, and accident prevention.

It focuses on a model that can be used in the LPG distribution industry by contributing to continuous monitoring and controlling gas leaks remotely. The model exhibits fast detection and resolution, which ensures safety, reduced leakage, and reduced emissions [2]. The system's flexibility in accessing devices wirelessly from anywhere in the world means that one of its benefits is lower computational cost and energy consumption [10]. Emphasis is placed on the IEEE 802.15.4 standard for wireless communication between routers and sensors, among which the use of a middleware that manages data for the smart devices involved is mentioned [3]. Proposes an IoT-based security system that cuts the power supply in smart homes when it detects an LPG leak [4]. The system sends an alarm SMS to the relevant authority if it detects a gas leak; in addition to sending an alarm to the homeowner, the user can ask the web server about the presence of gas [8]. The tool detects leaks with classification in three LPG concentration categories and sends messages via Telegram or buzzers installed in the device [11]. It uses a Wi-Fi network to monitor the gas weight regularly, and a microcontroller to make gas reserves using IoT [12]. Using machine learning to predict data about potential risks on the specific location of pollutants such as CH4, CO, and CO2 in the air and notifies the user [13]. Mobile app for updating the daily rate of gas consumption in households; it also records the real-time weight of the cylinder and books a new one through the proposed mobile app. Turns off the gas regulator when it detects a leak and activates the buzzer [14]. The breakthrough is presented by the voice command used by the customer to turn the gas regulator on and off. The initiative, an intelligent cart, further helps prevent gas weight fraud conducted by agencies [15]. It combines a buzzer alarm, IoT, and Wi-Fi to make a network that receives and transmits data to detect gas leakage in residential and gas distribution locations [16]. It alerts the user via a mobile app about gas leakage and shuts off the cylinder knob when it exceeds a set limit. The app is complemented by a web interface to monitor and place orders with the supplier [17].

The use of devices such as sensors and actuators decreases risks and prevents accidents; IoT accesses automatic control of machines and infrastructure; objects are fed with data generated by sensors and result in intelligent decision-making; IoT is part of air conditioning, heating, and ventilation environments that integrate with gas sensors, windows, alarms, gas valves to minimize domestic and industrial accidents [3]. The technologies are used for gas management in gas metering, gas leakage monitoring, pressure monitoring of gas pipeline networks, and detection of abnormal events [9]. In addition, IoT plays a fundamental and essential role in instrument design, providing all the necessary details and specifications to protect companies or individuals against accidents that flammable gas leaks can cause; also, IoT devices, through the use of sensors, can detect, store and process data from the environment, and the use of sensors has advantages such as reliability, ease of use, flexibility, accuracy, and cost-effectiveness [4].

The objective is to design a general and intelligent tracking-monitoring model based on IoT and BI technology for LPG distribution. The specific objectives are a) to design a general architecture for the control and monitoring of LPG distribution, storage, and use based on IoT and BI technologies; b) to evaluate the IoT network and BI model to determine interoperability, ease of use, and security through Y.4908 Standard and c) to conduct and analyze the survey of IT specialists. This paper proposes a model for an intelligent system in the detection, control, and monitoring of LPG using the Internet of Things (IoT) and Business Intelligence (BI). IoT technology is used to capture gas quantities or gas leaks and send notifications faster; monitoring is remote, which helps in a system to detect LPG leaks in places such as restaurants, homes, or industries and take early prevention measures in case of leakage during transportation or use. BI determines indices (e.g., humidity, temperature, gas level) to monitor and model a data warehouse related to LPG management.

2. Materials

LPG has been produced since 1914. It is a mixture of hydrocarbon gases used in homes and industries. The gas is toxic, colorless, and odorless. It is a fuel for heating, cooking, and vehicles [3]. LPG is available in much of the world to meet domestic requirements and industry needs. It is a mixture of propane and butane, which are very flammable. LPG is heavier than air, and if a person inhales it, then it causes asphyxiation [8]. LPG has ethanol added as a characteristic odor to prevent harmful events and identify leaks at any time or space; it replaces other types of fuel systems in commercial, domestic, and industrial areas; it is an explosive gas containing between 1.8% and 9.5% of the volume of gas and air [4].

It is a network system for connecting things via the Internet based on protocols used in data-sensing equipment; it is an ”intelligent emerging technology” that maintains the interconnection between devices with computing capabilities that are interconnected to the Internet. It is a communication paradigm comprising sensing devices, modules, and microcontrollers that pass digitized data [2]. IoT consists of interconnected networks and people sharing data inside or outside the environment found in the home, industry, sports, agriculture, transportation, medicine, and environment [10], and allows establishing communication among all to achieve common goals for the people's interest.

Business intelligence allows optimization and improvement of the quality of business decision-making based on data. BI is compared to information compendiums, reporting tools, executive information applications, and corporate consulting because it uses "business intelligence"; the company's information is processed and visualized better with the help of BI tools. Within BI, a Data Warehouse (DW) is used, which is the logical design of the data repository and is independent of the database; this DW allows the integration of multiple computer application environments and assists the processing of information to generate a single base that consolidates data for analysis purposes [18].

In [2], the author propose to use IoT to detect gas leaks, monitor leakage data, and control those leaks from remote customers with an on-off via mobile applications and the Internet; they use a gas sensor, valve, microcontroller, a database, online update, and relay. In case of gas leakage, the components are activated, alarms are turned on, the leakage status is sent to the gas supplier's smartphone, and the supplier can shut off the gas valve remotely. In other cases, the gas supplies are shut off automatically if the leakage exceeds the threshold. The paper by [10] proposes a device to detect the leakage of butane or LPG; if the sensor value increases, then an alarm, sound, and lights are activated; also, the sensor reading is displayed; the gas smoke is revealed by a pattern which is convenient for detection.

The prototype of [3] is a multi-gas smart sensor to prevent accidents in innovative environments on carbon monoxide or LPG; the sensor is connected to an IoT environment, and it uses the communication standard to interconnect the data with other smart devices and achieve to make the right decisions.

In [4], the author designed and implemented a prototype gas sensor. If there is a leakage, the sensor sends a signal to the microcontroller. The microcontroller receives and processes the call, and then the relay module shuts off the current supply to avoid sparks or short circuits generated by turning off switches and preventing explosion accidents. The prototype detects changes in LPG concentration at a short distance and displays the amount of particles per million (PPM) as an indicator.

In the research, they developed an electronic system to monitor LPG, natural gas, butane, humidity, temperature, and heat indicators through a web server; if there is a gas leak or if a parameter exceeds its threshold value, then an alarm is triggered, and a text message is sent to the authority. The system has sensors, a controller, a web server, and a display showing the sensor value [8].

In [1], they implemented an automated method to detect gas leaks through sensors connected to an alert and command system; the peripheral is effective, simple, low cost, portable, ultralight, and reliable; this helps in the sanitary sector because when there is a gas leak pollution, and waste is minimized, and increases the economy.

It uses IoT technology that gets faster responses on LPG systems; this IoT passes alerts on gas leaks to avoid accidents; it uses a microcontroller device, a gas detector, the model activates preventive measures by buzzer, has a GSM module, a solenoid valve, a LED light for indication, an air extractor, gas sensor, and a humidity-temperature sensor [5]. The IoT-based system of [11] is developed in four phases: data collection, hardware design, program coding, and system testing; this generates a tool to detect LPG leaks. When detecting anomalies, alert messages are sent by telegrams or buzzers; the system categorizes the concentration of LPG into two categories.

The [19] system has a first segment that monitors and a second segment that regulates the smoke and gas output through a controller; if the smoke sensor detects a leak, then a signal is emitted to the controller to turn on the exhaust fan then, it also stops the gas flow through the solenoid valve. The display shows the sensor value and a switch turns on the exhaust fan and the valve.

The article by [9] uses Artificial Intelligence (AI), Blockchain, and IoT technologies for a gas architecture in an intelligent city; AI algorithm uses the prediction model for gas production after detecting the change in gas delivery capacity; Blockchain is used to establish a secure transaction scheme, and maintain the buy-sell contract; simulations show the prediction on gas production in real-time and selection of dynamic sales transactional plan. The research of [20] explores BI in the domain of industries in gas refining by identifying and classifying deployment factors in Iran.

The importance and relevance of LPG as an energy source used industrially and domestically stands out by pointing out its use's characteristics, risks, and safety measures. The potential for efficient management of LPG is observed with the integration of technologies such as IoT and BI, as it is used in the early detection of gas leaks and preventive measures to avoid fatal accidents. The IoT contributes to the interconnection of devices and the generation of data that engages with the power of BI analysis to facilitate decision-making.

Numerous studies reviewed aim to channel alerts concerning elevated CO, CO2, or other relevant indicators to users, regardless of their geographical location. It highlights a global perspective of considerable interest in LPG management at the international level. Indeed, such a perspective illustrates how these technologies are deployed in various scenarios, from homes to industries and smart cities.

This wide range of applications underscores the versatility and heterogeneity of the technological approaches, including artificial intelligence and Blockchain. In this area of research, such technological advances play a critical role in amplifying safety, optimizing efficiency levels, and raising quality standards in LPG administration in various application areas.

3. Methodology

The methodology in this article explains the details of achieving and developing the specific objectives. Analyze scientific articles to learn about other IoT and BI models through a systematic literature review. The systematic review attached to the PRISMA statement of [21] is used, which has three phases, see Figure 1:

1. Identification of studies: it is based on a search in one or several library databases; IEEE, ACM Digital Library, and Web of Science are considered; peer-reviewed articles are used.

Mendeley software is used to eliminate duplicate articles. The filtering keywords are IoT, Business Intelligence, and LPG.;

2. Eligibility criteria: It is necessary to define the types of studies that could be considered for this study. The inclusion criteria for eligibility are as follows: The article is scientific. It is Written in English. The Content is on IoT or Business Intelligence in LPG management. The exclusion criteria are as follows: The article is younger than 2018. It is in a language other than English. It is an Abstract article. It is a Paid article.

   The following research questions (PI) are defined:

   - PI1: In which scenarios is IoT used (e.g., industry, household, housing estates, distribution, supplier)?
   - PI2: What devices are used in LPG monitoring (e.g., sensors, microcontrollers, buzzers)?
   - PI3: What other components are used (e.g., server, web application, mobile application)?
   - PI4: Are other technologies used for LPG monitoring (e.g., Blockchain, AI, Big Data)?
   - PI5: What is monitored (e.g., gas leakage, transport, pipelines)?
   - PI6: What is the result of the research (e.g., design, implementation)?
   - PI7: What gases are monitored (e.g., LPG, natural gas, butane, carbon monoxide, nitrogen dioxide, sulfur dioxide)?
   - PI8: What data are detected (e.g., humidity, temperature, heat indicator)?
   - PI9: What protocols are used (e.g. IEEE, 6LoWPAN, MQTT)?
   - PI10: What indicators are displayed (e.g., humidity, temperature, gas level)?
   - PI11: What software tools are used in Business Intelligence?
   - PI12: What general data do the data warehouses have?

3. Data collection and synthesis: Data covering the identified articles are extracted, the research questions are answered, and data analysis is performed in quantitative form and described for explanation. The study uses a quantitative approach.

Identification of Studies
Eligibility Criteria
Data collection and synthesis

Figure 1: Systematic review.

Based on IoT and BI technologies, a comprehensive architecture is designed for LPG distribution, storage, use, control, and monitoring. The architecture covers LPG plant storage, pipeline or transport distribution, and use in homes or businesses. The architecture design is based on analytical research that studies the feasibility of a measure through empirical evidence. The qualitative approach describes all the architecture details, such as participants, components, software, hardware, networks, and indicators. The architecture graph is presented. The Y. 4908 and a survey of IT specialists are used to evaluate the IoT network and BI model for interoperability, usability, and security. The Y.4908 standard evaluates the IoT network's interoperability, usability, and security; see Table 1.

Table 1: Evaluation factors.

	Capacity	**Yes**	**No**
Network Interoperability	Interconnected		
	Device List		
	Systems List		
Data Interoperability	Interconnect		
	Device List		
	Systems List		
Service Interoperability	Integrate		
	Device List		
	Systems List		

Source: Authors.

The survey technique is used with a specific group of at least 30 professionals in Information Technology, Information Systems, or Computer Science. Some of the questions are: a) The BI model presents the origins of the data, b) The Data Warehouse model is presented, c) The ETL is presented, d) The names of the indicators are presented, e) The names of the reports are presented. f) The software named is appropriate. and g) The DW contains dimensions and facts.

4. Results

4.1. Analysis of scientific articles using a systematic review of the literature.

The articles selected through the PRISMA methodology are considered. Figure 2 presents the flowchart on the themes identified, reviewed, and set. After the initial search in IEEE, ACM Digital Library, and Web of Science library databases with the keywords ”IoT” or ”Business Intelligence,” 148 articles were identified. After eliminating duplicates, illegible, and removed for other reasons, these were reduced to 113 pieces. A further 11 articles were excluded due to their title and abstract needing to fall into the eligibility criteria; this led to 102 articles being sought for retrieval. The articles were then reviewed for IoT and Business Intelligence descriptions or models. Exclusion criteria are applied sequentially according to age, article type, or payment; otherwise, the evaluation is continued. Other exclusion items are articles that mention the term IoT or BI but do not apply it; articles that mention IoT or BI in general and not as results; articles that do not demonstrate the influence of IoT or BI; and articles that are only concepts in total are 35. In the end, 40 papers were selected for analysis, as shown in Figure 2; all these articles are in the Reference section, and the complete list is in Table 2.

Table 2: Classified and selected papers.

Year	Papers	No
2019	[3, 12, 14, 20, 22, 23, 24, 25, 26]	9
2020	[8, 27, 28, 29, 30]	5
2021	[2, 4, 9, 10, 11, 17, 19, 31, 32, 33, 34, 35]	12
2022	[36, 37, 38, 39, 40]	5
2023	[1, 5, 15, 16, 41, 42, 43, 44, 45]	9

Source: Carried out by authors.

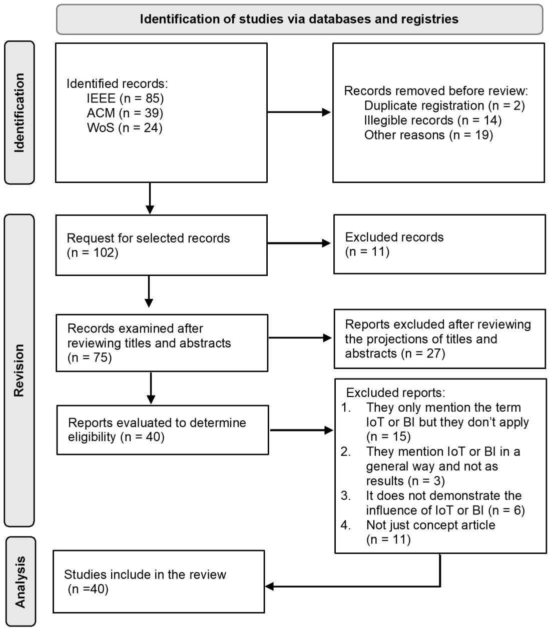

Identification of studies via databases and registries
Identification
Identified records:
IEEE (n = 85)
ACM (n = 39)
WoS (n = 24)
Records removed before review:
Duplicate registration (n = 2)
Illegible records (n = 14)
Other reasons (n = 19)
Revision
Request for selected records (n = 102)
Excluded records (n = 11)
Records examined after reviewing titles and abstracts (n = 75)
Reports excluded after reviewing the projections of titles and abstracts (n = 27)
Reports evaluated to determine eligibility (n = 40)
Excluded reports:
1. They only mention the term IoT or BI but they don't apply (n = 15)
2. They mention IoT or BI in a general way and not as results (n = 3)
3. It does not demonstrate the influence of IoT or BI (n = 6)
4. Not just concept article (n = 11)
Analysis
Studies include in the review (n =40)

Figure 2: Synthesis of article reviews using PRISMA.

The selection process for the peer-reviewed articles published in English between 2019 and 2020 was meticulous. The increased academic interest in this area can be attributed to the burgeoning number of urban platforms worldwide. As part of the eligibility criteria, each article described at least one urban crowdsourcing platform, with some even going beyond and mentioning multiple platforms. This comprehensive approach ensures that the selected articles hold significant value for our research.

Among the 40 meticulously selected articles, only 23 present models or architectures on IoT or Business Intelligence applied in LPG control; others present technology or alternatives for gas tank control. Each of the 40 articles is thoroughly reviewed to ensure it contains a property that falls into the following groups: IoT scenarios, Monitoring devices, Other components, Other technologies, Control, Research results, Gases they monitor, Data they detect, Protocols used, Indicators they show, Business Intelligence software, Data from Data Warehouse. This rigorous review process guarantees the validity and reliability of our research findings.

Each item is reviewed, and the properties of each item are as follows: item, year of production, item title, country of origin, and number of references. In addition, the IoT Scenarios segment has properties in the industry, home, distribution, and supplier sectors. The monitoring devices segment has the following properties: Sensors, microcontrollers, buzzers, valves, relays, Fans, displays, and LEDs. The segment has other components with properties: Server, web application, mobile application, database, GSM, and GPS. The segment includes other technologies with properties such as blockchain, artificial intelligence, and business intelligence. The Control segment has properties: Gas Leakage, Transportation, Pipelines, and Weight. The Research Result segment has Design, Implementation, and Prototype properties. The segment Gases Monitoring has properties: LPG, Natural Gas, Butane, Carbon Monoxide, Nitrogen Dioxide, Alcohol, Oxygen, Propane, Hydrogen, and Methane. The segment data they detect has properties such as humidity, temperature, smoke, fire, and gas—the segment Protocols used with properties: IEEE, 6LoWPAN, MQTT, TCP. The segment Indicators show the properties: Humidity, Temperature, Gas level, Smoke, Heat, and Cylinder status. Business Intelligence Software has properties such as BLYNK, Excel, and PowerBI. The Data Warehouse Data segment has weight, voltage, and level properties. Table 1 shows the 40 items analyzed and used to answer the research questions (PI).

4.2. *Answer to the research questions.*

PI01: In which scenarios is IoT used?
According to the analysis of the 40 papers, 58% of the proposals are for industry, 45% are for homes, 5% are for management in gas distribution, and 8% are suggestions for gas suppliers. Conversely, 12% specify industry and household, i.e., five things. See Figure 3 Scenarios.

PI02: What devices are used in LPG monitoring?
According to the analysis of the 40 papers, 88% use sensors, 83% use microcontroller cards, 48% use buzzers to alert, 55% use valves to close the tank, 55% of the documents use a relay to activate the valves; 23% of the items use a fan to dissipate the gas; 55% use a display to show the status or quantity; 40% use a LED light to alert of gas leakage. On the other hand, 83% use sensors and microcontrollers, i.e., 33 items. Another 33% use LED with relay and valve as prevention and warning, i.e., 13 papers. In addition, 48% that use buzzers to alert are among the 83% that use microcontrollers and

sensors, i.e., 19 items. The buzzer, fan, and LED are linked to the microcontroller. See Figure 4 Devices.

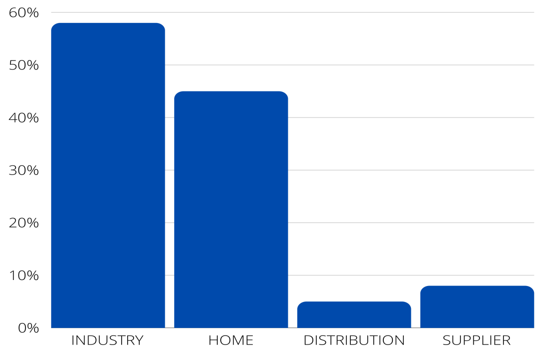


Figure 3: Types of scenarios for using IoT.

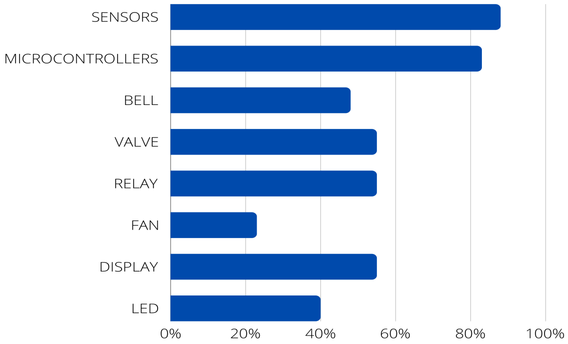


Figure 4: Physical objects for LPG control.

PI03: What other components are used? According to the analysis of the 40 papers, 48% use or have an application server, 30% have a web application, 58% use a mobile application, 33% use a database, 40% use GSM to send text messages, 8% use GPS for gas leak positioning. The tendency to use mobile applications remains high, as does using GSM to send simple text messages because they are very efficient for communication. GPS communication is rarely used because of the monthly costs providers charge for this service. On the other hand, databases are used by web applications or mobile applications; GSM modules may or may not be integrated into the microcontroller card. Web applications are linked to a server. See Figure 5 Other components.

PI04: Are other technologies used for LPG monitoring? According to the analysis of the 40 papers, 3% use Blockchain to maintain information security in encrypted form and maintain traceability/tracking in gas distribution; 8% use Artificial Intelligence for possible predictions in gas leakage; 13% use Business Intelligence to generate reporting or control indicators. On the other hand, the same article that proposes to use Blockchain also uses Artificial Intelligence. There is no overlap in using Business Intelligence with other technologies. See Figure 6 Other technologies.

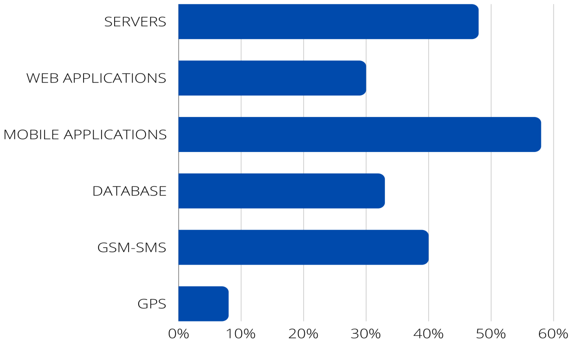


Figure 5: Components that interact on IoT platforms.

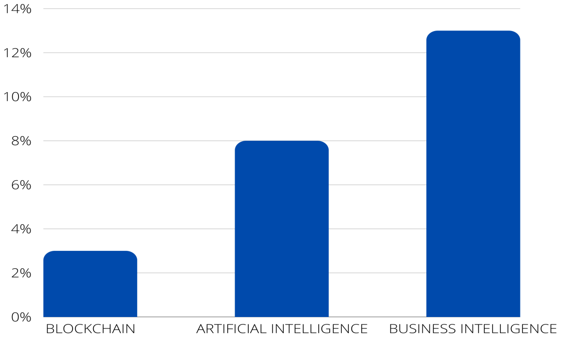


Figure 6: Other technologies that augment IoT applications.

PI05: What is controlled? According to the analysis of the 40 papers, 73% propose controlling gas leakage, 3% offer governing gas transportation, 15% propose controlling gas pipelines, and 15% suggest controlling LPG tank weight. However, only some of the articles submit gas leakage control. The papers that work on Business Intelligence only obtain data from files generated by sensors, and others are only theoretical articles. On the other hand, 13% of control pipelines are within the 73% gas leakage control, i.e., five pieces. In addition, only one paper proposes a rule during the gas transport process. See Figure 7 Control.

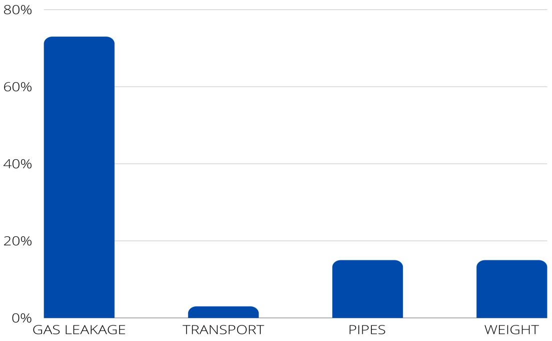


Figure 7: Real-time monitoring of gas usage.

PI06: Research results?

According to the analysis of the 40 papers, 83% present as a result of the network design or device design; 23% performed the implementation and use of their plans; 58% achieved a prototype to demonstrate their theories on device design in gas leakage control or LPG cylinder weight. On the other hand, two papers should have presented the method and implementation. The rest of the articles that were designed were also implemented. All the documents with prototypes did give their design, and all the prototypes presented photos of their work. See Figure 8 Results.

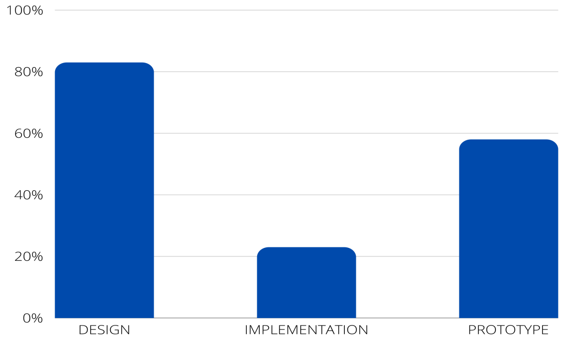


Figure 8: Results of the proposals in the articles.

PI07: What gases are monitored?

According to the analysis of the 40 papers, 85% monitor LPG from home or industry, 18% monitor Natural Gas, 25% monitor Butane, 20% monitor Carbon Monoxide, 10% monitor Nitrogen Dioxide, 10% monitor Alcohol, 10% monitor Oxygen; 18% monitor Propane; 23% monitor Hydrogen; 23% monitor Methane. On the other hand, 8% of natural gas monitoring is within LPG monitoring, i.e., three items. The other four items only monitor Natural Gas. All the articles monitoring Butane are within LPG monitoring, i.e., ten articles. The 18% that watch Carbon Monoxide is within the LPG monitoring, i.e., seven items. Items monitoring Nitrogen Dioxide, Alcohol, and Oxygen are within LPG monitoring, i.e., four items each. The things that observe Propane, Hydrogen, and Methane are within LPG monitoring. This happens because there are sensors that can detect few or many gases; the sensors used in the items are MQ2, MQ3, MQ4, MQ5, and MQ6. If the number is higher, then more gases can be detected. See Figure 9.

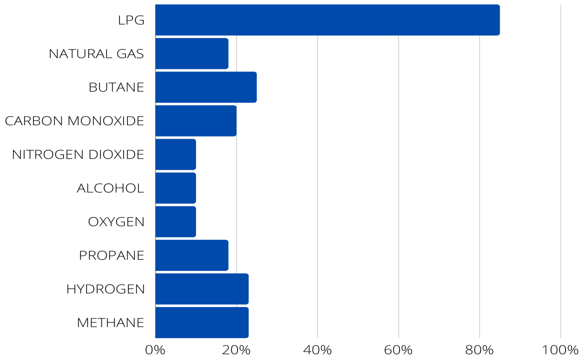


Figure 9: Real-time monitoring on various types of gases.

PI08: What data is detected?

According to the analysis of the 40 papers, 15% present designs that detect humidity, and 15% detect temperature; 10% of the designs detect smoke; 8% of the methods detect fire; 88% of the designs detect gas, either by gas leakage in cylinder or pipe leakage. On the other hand, the techniques that detect humidity and temperature are among the 88% that detect gas. The designs that detect smoke or fire are within 88% that detect gas. See Figure 10. Data detected.

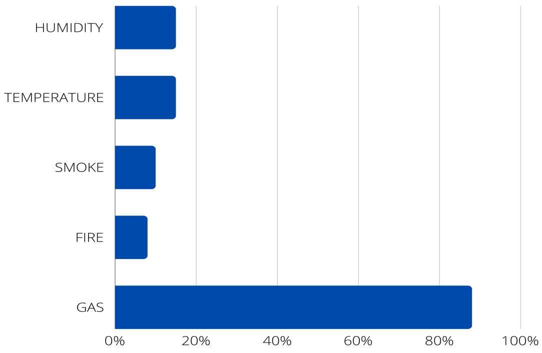


Figure 10: Real-time detected data on gas usage.

PI09: What protocols are used?

According to the analysis of the 40 papers, 43% name or use the IEEE protocol in their implementations or prototypes; 5% name or use the 6LowPan protocol in their designs; 8% name or use the MQTT protocol in their designs; 25% name or use the TCP protocol. On the other hand, 10% of the articles used any of these protocols in implementation, i.e., four papers. 38% of the documents used protocols in prototypes, i.e., 15 pieces. The 25% of the TCP articles are within the 43% using IEEE. See Figure 11 Protocols.

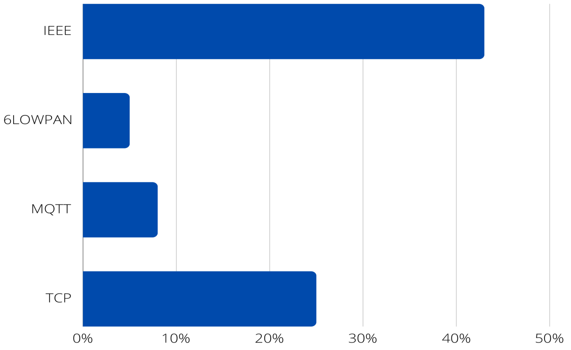


Figure 11: Protocols used in gas tracking.

PI10: What indicators are shown?

According to the analysis of the 40 papers, 13% show the humidity, 10% indicate the temperature, 73% offer the gas level, 8% show the smoke level, 3% show the heat level, and 25% show the bottle status. On the other hand, humidity, temperature, and gas levels are shown on the display in 10% of the items, i.e., in 4 pieces. In 8% of the items, i.e., three things, gas, and smoke levels are displayed on the screen. Only one report shows humidity, temperature, gas level, smoke, and heat as indicators. Only one article shows humidity, temperature, gas level, and smoke as indicators. Only one item presents humidity, temperature, and gas level as indicators. See Figure 12 Indicators.

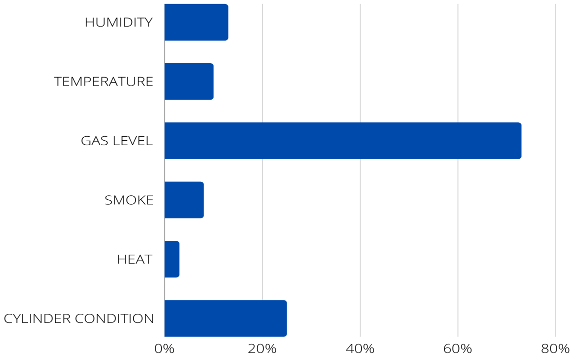


Figure 12: Real-time indicators used in gas tracking.

PI11: What software tools are used in Business Intelligence? According to the analysis of the 40 papers, 15% of the 40 articles use Blynk, 5% use Microsoft Excel, and 5% use Microsoft Power BI. On the other hand, only one piece uses Excel and Power BI; the other articles use these tools independently. See Figure 13 BI software. Blynk is used to connect IoT devices, assist in the visualization of sensor data, execute remote control with mobile web applications, perform firmware updates, and offer a secure cloud, user and access management, and alerts, among others. In addition, this platform promotes smart home hardware manufacturers [46]. Microsoft Power BI generates simple data sets with many data sources or origins, is also simple for aggregation to the Power-BI data connectivity hub, and generates a centralized, single, effective, and accessible source of information for data from multiple devices [47].

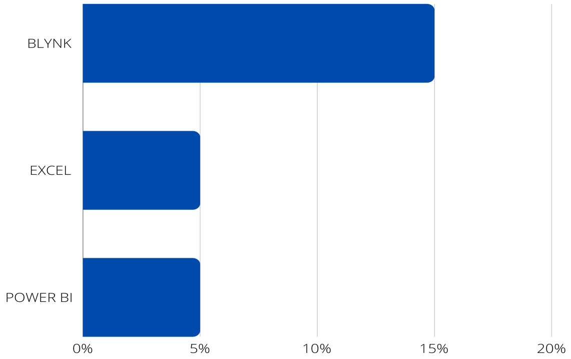


Figure 13: Tools in Business Intelligence projects.

PI12: What general data does the Data Warehouse have? According to the analysis of the 40 papers, 8% show the weight in the DW, 8% show the voltage in the DW, and 10% indicate the gas level in the DW. On the other hand, 8% of the articles present the weight and voltage together, i.e., in 3 pieces. The 10% of the items showing the gas level do not have any DW data in common with the other 8%. See Figure 14 for the DW data.

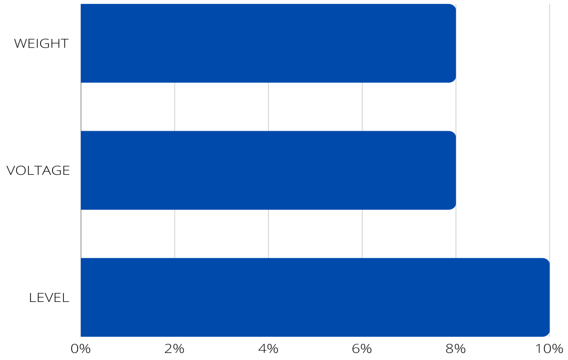


Figure 14: Data managed by DWs.

4.3. Design of a general architecture for LPG management based on IoT and BI.

In the event of a gas leak, the Gas Sensor detects the strength of the leak and sends the data to the microcontroller. This microcontroller takes that data takes it and sends it to a Firebase database in real-time via the Wi-Fi router and the Internet. The Firebase database sends the data to the smartphone via the Internet. In addition, GSM communication can be added to send text messages and an audible alarm on-site. The leak volume is displayed on the LCD screen and the smartphone. There is a predefined threshold of 500 PPM; if the leak volume is less than 500 PPM, the relay and solenoid valve are turned on, letting the gas pass through the pipeline. If the leak volume is more significant than 500 PPM or if pressing the RED toggle button on the smartphone, then a control signal is sent from the smartphone to the microcontroller via the Firebase database; this action turns off the relay and solenoid valve to not let gas pass through the pipeline.

Components:

1. MQ-5 gas sensor. Detects LPG and natural gas with excellent accuracy. Obtains the presence of gas with a concentration from 2000 PPM (Parts Per Million) up to 10000 PPM and operates with 5 volts of power.

2. MQ-6 gas sensor. It detects the presence of LPG. It is an analog sensor based on resistance. It obtains the presence of gas with a concentration from 200 PPM to 10000 PPM.

3. Temperature and humidity sensor. The DHT11 digital sensor is a low-cost sensor that measures air temperature and humidity. It can measure temperature from 0 to 500 °C with an accuracy of ±2 °C and humidity from 20 to 80% with an accuracy of 5%. It consumes power from 3 to 5 volts and draws a current of up to 2.5 milliamps while reading data.

4. LCD. The 16cm x 2cm liquid crystal display is connected to the NodeMCU via I2C communication protocol. The LCDs the data obtained by the sensors, such as humidity, temperature, and gas status, in real-time on-site.

5. NodeMCU DEVKIT 1.0. NodeMCU is open-source firmware for the IoT platform. This hardware is a microcontroller unit with a wifi chip. It is an excellent low-cost option for sending

data to a web server, LCD, GSM, and relay. This control unit takes the data obtained by the sensors. After analyzing the sensor data, this microcontroller executes the appropriate actions.

6. Audible alarm. The buzzer is added to notice nearby people. If the sensor detects the presence of gas in the air, then the NodeMCU activates the audible alarm.
7. GSM modem (SIM800L). This hardware connects to the NodeMCU to send and receive text messages (SMS). The modem incorporates a SIM card as necessitates a subscription with a designated mobile operator, specifically one denoted by the identifier. Upon detection by the sensor of either gas presence or value exceeding predetermined thresholds. If the sensor detects the presence of gas or out-of-range value, then the microcontroller sends an automatic notification to a cell phone number about the gas leak. In addition, it is possible to query the status of the gas leak by SMS remotely.
8. Relay. It is a device that operates the solenoid valve.
9. Solenoid valve: This device controls gas leakage; it turns on or off through the relay module according to the signal from the microcontroller.
10. Wifi router. It is a wifi router device for Internet connection.
11. Smartphone: This is a control unit. It can access mobile applications on the solenoid valve and remotely turn it on or off.
12. Google Firebase is a platform for storing and processing leakage data. This database sends the data from the microcontroller to the mobile applications in real time.
13. Arduino IDE and C++ programming. The microcontrollers are programmed in Arduino IDE and C++ programming language.

The utilization of MQ-5 and MQ-6 gas sensors presents a judicious approach to detecting combustible gases, including LPG, and their concentration within the ambient environment. Concurrently, temperature and humidity assessment is paramount in elucidating the heat index. Integration of these sensors with the Node MCU microcontroller facilitates an exploration of potential accident scenarios. The sensors transmit pertinent data to the microcontroller upon gas leakage detection and quantification of its volume in PPM units. Subsequently, the microcontroller undertakes a meticulous verification process, scrutinizing the received data against predefined thresholds. Should the measured leakage value surpass the threshold of 500 PPM, an automated shutdown mechanism for the solenoid valve is activated. See Figure 15.

The web application presents the important index values (gas percentage, heat index, humidity, smoke presence, and temperature). In the web application, you can understand the gas leakage situation or the normal state, and it is unnecessary to understand the sensor values. If there is a gas leak, an alarm is generated on-site, and an SMS text message is sent to minimize the possibility of an accident. The GSM module allows a message to be sent for the presence of gas or other out-of-range sensor values, and it is possible to query the current sensor values. The smartphone receives the data from the server via the mobile application. In addition, the on/off interface is displayed. If the leak volume exceeds 500 PPM, you must press the shutdown button (RED) on the smartphone to control the gas leak. The command data is transmitted to the microcontroller via the Internet database. The solenoid valve can be opened via the GREEN button on the smartphone. See Figure 16.

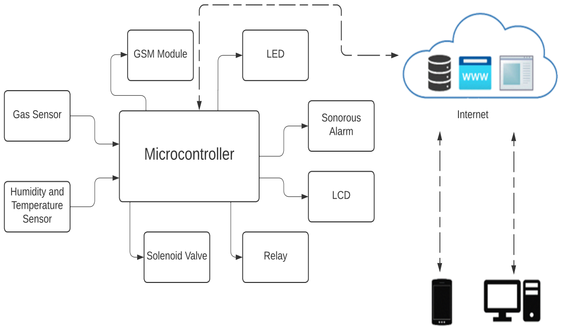


Figure 15: IoT device at control points.

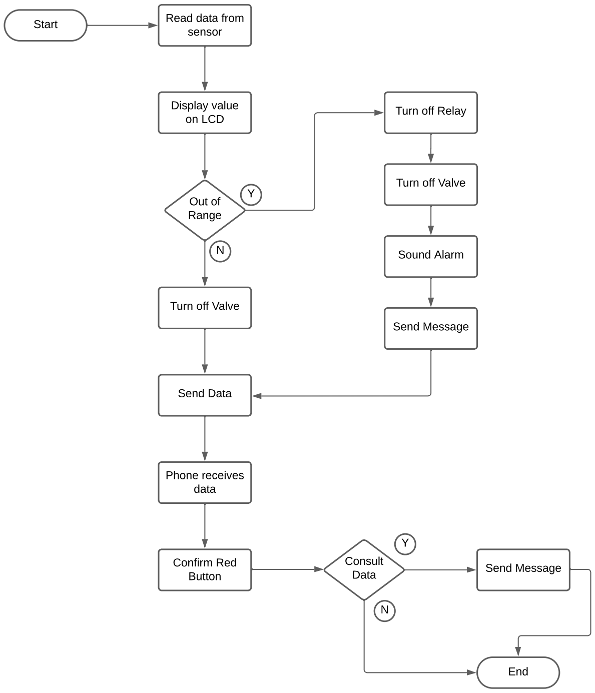


Figure 16: Data flow in gas monitoring and control.

The Business Intelligence part [48] takes the source data from the database in the cloud; in this database are the measurements of all the sensors in the LPG distribution network, i.e., humidity, temperature, and particles per million in each fraction of time. Other data that exist are transport, time, and quantity. It is recommended to use Microsoft Power BI Desktop (PowerBI, 2023) because it is very intuitive.

The BI model has four levels; Figure 17 represents the model based on IoT and BI.

1. Data Source: This is the Firebase in the cloud. The database contains the Sensors table with column identification, series, sensor name, location, start date, and status. The Measurements table has columns such as sensor, humidity level, temperature level, PPM level, date, and time.

2. ETL: There is the ETL process (Extraction-Transformation-Load); here, the Power BI tool performs the validation, cleaning, transformation, and aggregation of the data and then performs the load to the Data Mart. In this case, the source data belongs to a single database; the data is homogeneous in the extraction; the extraction is performed every hour or according to the Power BI configuration; in the data cleansing, unnecessary data is discarded. Data is considered valid because it is in a database; data such as sensor series and start date are discarded in data cleaning. The cleaned data is loaded into the Data Warehouse, and the data belonging to the Facts table is loaded into the Power BI tool.

3. Storage: There is the datamart, the data warehouse, and the cube; remember that the source database comprises two-dimensional tables or straightforward data. The Power BI tool obtains this multidimensional data on the sensors. A multifaceted analysis allows thinking, reducing confusion, avoiding lousy perspectives, and seeing from another angle and other facets.

4. Visualization: This BI results in view contains dashboard sorts; the previous steps could be performed in the Power-BIPower BI.

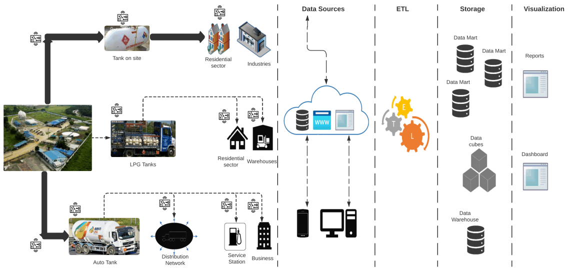


Figure 17: A general architecture for monitoring and control in gas distribution

The Data Warehouse gets the data from the IoT network; from this IoT, it receives the humidity, temperature, and PPM data, which are stored in the database in the cloud. In the database are other data such as city names, province names, sensors names, microcontrollers lists, gas supplies lists, threshold parameters by gas type, daily sensor activities, customer list supply lists, and customer services lists. The lists become the dimensions of the DW; the facts are the technical and daily indicators. The technical indicators help to track the LPG dispatch. The daily hands keep the history of the sensors to perform averaging and presentation of carvings on the dashboard. Data model in star type for facts and dimensions is shown in Figure 18.

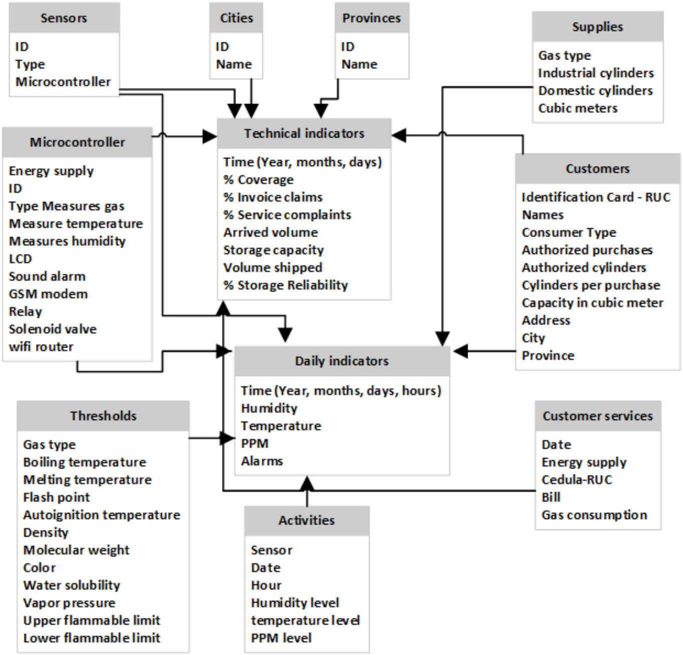


Figure 18: Data model in star type for facts and dimensions

The reports that the dashboard can present are a list of alarms by city, a list of thresholds, a list of technical indicators, a list of active microcontrollers, and a list of sensors. Microsoft Power-BI is recommended for implementation, and the information can be exported to a spreadsheet.

4.4. *Evaluate the IoT network and BI model using the Y.4908 Standard and the IT specialist survey.*

Standard Y.4908 (ITU-T, 2020) is used to evaluate the IoT network's interoperability, usability, and security; this task was performed by three professionals who participated in the same survey. All three agreed on the answers in the table of evaluation factors about the IoT network. They answered Yes on: List of Network interoperability devices, List of Network interoperability systems, List of Data interoperability devices, List of Data interoperability systems, and List of Services interoperability devices. They answered No on the Service Interoperability Systems List. Dividing 100% for six answers gives 16.66% for each answer. The three professionals answered five questions in the affirmative, i.e., a percentage of 83.33% was obtained. This means the IoT network has an outstanding approval because it exceeds 80%. See Table 3.

The survey was administered to 30 Information Technology, Systems, or Computer Science professionals. Ten questions with Likert scale responses were asked for the study, which is "1. Strongly agree", "2. Agree", "3. Neither agree nor disagree", "4. Disagree" and "5. Strongly disagree", used in [49].

If the percentages of "1. Strongly agree" and "2. Agree" are added together, it is assumed that both groups of people agree on the answer. Survey analysis: In question 1, 87% of the professionals agree that the model presents the data origins of IoT sensors, and 13% disagree. In question 2, 80% of the professionals agree that the article presents the DW model, and 20% disagree. In question 3, 70% of the professionals agree that the model does give the ETL, and 30% disagree. In question 4, 93% of the professionals agree that the research does present the indicators, and 7% have no opinion. In question 5, 94% of the professionals agree that the study shows only the reports' names and 7% do not know. In question 6, 87% of the professionals agree that using Microsoft Power-BI is appropriate, and 13% do not agree. In question 7, 87% of the professionals agree on the dimensions and facts presented in the model, and 13% do not agree. In question 8, 94% of the professionals agree that the indicators presented or named are appropriate, and 6% do not agree. In question 9, 80% of the professionals agree that the model promotes decision-making, and 20% do not agree. In question 10, 94% of the professionals agree that the model presented is transparent, and 6% do not agree. At the overall average level, 49% completely agree with the answers, 38% agree, 12% neither agree nor disagree, and 1% disagree. In other words, the model has an overall average approval rating of 87%.

Table 3: Evaluation factors.

	Capacity	Yes	No
Network Interoperability	Interconnected		
	Device List	X	
	Systems List	X	
Data Interoperability	Interconnect		
	Device List	X	
	Systems List	X	
Service Interoperability	Integrate		
	Device List	X	
	Systems List		X

Source: Authors.

5. Discussion

The proposed model has been designed to detect Liquefied Petroleum Gas (LPG) leaks and automatically activate a supply cut to prevent potential accidents. However, its capabilities could be strengthened by integrating IoT devices for accurate detection and more efficient management.

This proposal simplifies the process of monitoring and detecting LPG leaks through the adoption of IoT technology. The devised model notifies users about gas leak incidents through multiple communication channels: on-site audible alarms, text messages, an up-to-date web app, historical data, and a web app. Observation and management can be carried out remotely and effectively, even allowing the control of IoT devices through a web server or mobile device. Although a financial estimate is omitted, IoT devices' effectiveness and low cost in the research area are highlighted.

The eventual implementation of this model could be highly beneficial to the authorities in monitoring gas leak levels in critical areas, which makes it possible to issue immediate alerts to the public and prevent accidents, thus contributing to public safety. The gas sensors can detect various gases, fuels, and alterations in oxygen consumption, making it easy to take early leak containment measures. In the same way, it is emphasized that the implementation of the system could be effective both in locations near and distant from the base station, thus demonstrating the versatility and ease of use of the model in various environments. This contributes to the improvement in efficiency and response capacity in emergencies.

From the performance demonstrated in the research, the remarkable effectiveness of the proposed design in detecting LPG leaks is inferred. This is shown by integrating knowledge extracted from the literature, which gives the model considerable robustness. It is pertinent to highlight that this model benefits the industry and the commercial and residential sectors from gas distribution by offering an effective tool to prevent accidents and protect people's integrity.

Detailed analysis of the 40 selected articles reveals characteristic patterns in the application and design of similar systems. Notably, 58% of the articles focused on industrial applications, while 88% used sensors as an integral part of the submitted design. In addition, 58% of the studies implemented a mobile application to improve the system's accessibility and usability. The findings suggest a trend toward the integration of mobile technologies.

A significant portion of articles (approximately 73%) proposed specific measures for controlling gas leaks, and the majority (approximately 83%) presented results related to designing infrastructures or devices for this purpose. This data underscores the importance of actively addressing the detection and management of gas leaks in domestic and industrial settings. It is also highlighted that most studies (85%) have focused on LPG monitoring in residential or industrial contexts, suggesting a practical and relevant approach to gas supply security. A significant proportion of the articles (88%) focused on gas detection in cylinders or pipes, reflecting the importance of proactively identifying and mitigating leaks.

Finally, the IEEE protocol was mentioned or used in 43% of the articles reviewed, indicating a trend toward standardization and interoperability in this field. In addition, a small percentage of the studies (15%) used Blynk software for business intelligence, indicating a growing interest in data visualization and analysis to guide decision-making in this area.

The present research provides significant value in detecting LPG leaks in various contexts, such as industry, businesses, and private households. It stands out for its importance in accident prevention and protecting lives.

6. Conclusions

The proposed model that combines IoT with BI offers a solution for the real-time detection of LPG leaks during the distribution of this

Table 4: Results of the survey to professionals.

No	Questions	1	2	3	4	5
1	A BI model presents data sources	44	43	10	3	0
2	The Data Warehouse model is presented	33	47	20	0	0
3	ETL is presented	37	33	30	0	0
4	The names of the indicators are presented	53	40	7	0	0
5	The terms of the reports are presented	57	37	6	0	0
6	The named software is appropriate	34	33	13	0	0
7	The DW contains dimensions and facts	47	40	13	0	0
8	The hands are suitable for this case	57	37	3	3	0
9	The model promotes a culture of data-driven decision-making	57	30	7	3	3
10	The model is clear and specific	57	37	6	0	0
	Overall average	50	38	12	1	0

Source: Authors.

element to industries, businesses, or homes, helping to avoid fatal accidents. This approach seeks to manage gas distribution effectively and promotes the automation of facilities to safeguard the safety of individuals and the community in their environment while reducing carbon emissions and optimizing energy consumption. The construction of the architecture was based on the exhaustive analysis of the literature review achieved through the PRISMA Declaration. Standard Y. 4908 was used to assess interoperability, ease of use, and security. It should be noted that Information Technology professionals completed this last element. The sensors used to detect LPG and natural gas, in addition to temperature and humidity sensors, can respond to the variation of low temperatures, which helps achieve a better process in the detection of gas. The collection of data in the cloud, generated by the IoT network, such as transportation, time, and quantity, translates into dynamic two-dimensional visualizations through BI technology that simplifies and accelerates the tracking process. In addition, additional variables such as location, start date, and state are considered to reduce confusion and erroneous perspectives. Distribution companies, in their quest to improve competencies and meet the needs of their customers, benefit from daily indicator lists and a track record of performing average analyses and presentations on the dashboard, facilitating the monitoring of LPG dispatch.

Despite the high score obtained in network and data interoperability, the model has limitations regarding service interoperability. Regarding adequately presenting data sources, configuring a data warehouse with dimensions and facts, and identifying appropriate indicators received a generally favorable evaluation from the professionals who analyzed the model. This research proposed a design that makes it possible to monitor the distribution of LPG using a web server, mobile application, or remote warning of IoT sensors. The model automatically closes the LPG passage when it detects a leak. The article does not present prices of IoT devices; however, it emphasizes that the named components are low-cost in the market. The research led to the development of a model combining an IoT component with a four-level BI, proving its effectiveness and acceptance in the professional field. At the overall medium level, 49% strongly agree, 38% agree, 12% neither agree nor 17 disagree, and 1% disagree. It is concluded that the model has an overall average level of approval of 87%.

References

[1] G. Senthil, P. Suganthi, R. Prabha, M. Madhumathi, S. Prabhu, S. Sridevi, "An Enhanced Smart Intelligent Detecting and Alerting System for Industrial Gas Leakage using IoT in Sensor Network," in 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), 397–401, 2023, doi:10.1109/ICSSIT55814.2023.10060907.

[2] H. Paul, M. K. Saifullah, M. M. Kabir, "A Smart Natural Gas Leakage Detection and Control System for Gas Distribution Companies of Bangladesh using IoT," in 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 109–114, 2021, doi:10.1109/ICREST51555.2021.9331226.

[3] J. B. A. Gomes, J. J. P. C. Rodrigues, J. Al-Muhtadi, N. Arunkumar, R. A. L. Rabêlo, V. Furtado, "An IoT-Based Smart Solution for Preventing Domestic CO and LPG Gas Accidents," in 2018 IEEE 10th Latin-American Conference on Communications (LATINCOM), 1–6, 2018, doi:10.1109/LATINCOM.2018.8613241.

[4] B. B. Sharma, P. Vaidya, N. Kumar, C.-C. Chen, R. Sharma, R. P. Dwivedi, G. Gupta, "Arduino based LPG Leakage Detection and Prevention System," in 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), 161–166, 2021.

[5] S. Ahmed, M. J. Rahman, M. A. Razzak, "Design and Development of an IoT-Based LPG Gas Leakage Detector for Households and Industries," in 2023 IEEE World AI IoT Congress (AIIoT), 0762–0767, 2023, doi:10.1109/AIIoT58121.2023.10174377.

[6] PetroEcuador, "PetroEcuador E.P.-GLP," 2023.

[7] BCE, "Banco Central Ec GLP," 2023.

[8] N. Mahfuz, S. Karmokar, M. I. H. Rana, "A Smart Approach of LPG Monitoring and Detection System Using IoT," in 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 1–4, 2020, doi:10.1109/ICCCNT49239.2020.9225293.

[9] W. Xiao, C. Liu, H. Wang, M. Zhou, M. S. Hossain, M. Alrashoud, G. Muhammad, "Blockchain for Secure-GaS: Blockchain-Powered Secure Natural Gas IoT System With AI-Enabled Gas Prediction and Transaction in Smart City," IEEE Internet of Things Journal, **8**(8), 6305–6312, 2021, doi:10.1109/JIOT.2020.3028773.

[10] S. A. Yadav, S. Sharma, L. Das, S. Gupta, S. Vashisht, "An Effective IoT Empowered Real-time Gas Detection System for Wireless Sensor Networks," in 2021 International Conference on Innovative Practices in Technology and Management (ICIPTM), 44–49, 2021, doi:10.1109/ICIPTM52218.2021.9388365.

[11] M. Kholil, I. Ismanto, R. Akhsani, "Development Of LPG Leak Detection System Using Instant Messaging Infrastructure Based On Internet Of Things," in 2021 International Conference on Electrical and Information Technology (IEIT), 147–150, 2021, doi:10.1109/IEIT53149.2021.9587414.

[12] V. Suma, R. R. Shekar, K. A. Akshay, "Gas Leakage Detection Based on IOT," in 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), 1312–1315, 2019, doi:10.1109/ICECA.2019.8822055.

[13] P. R. Meris, E. Dimaunahan, J. C. Dela Cruz, N. A. Fadchar, M. C. Manuel, J. C. C. Bonaobra, F. J. I. Ranosa, J. L. D. Mangaoang, P. C. Reyes, "IOT Based – Automated Indoor Air Quality and LPG Leak Detection Control System using Support Vector Machine," in 2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC), 231–235, 2020, doi:10.1109/ICSGRC49013.2020.9232472.

[14] A. K. Srivastava, S. Thakur, A. Kumar, A. Raj, "IoT Based LPG Cylinder Monitoring System," in 2019 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS), 268–271, 2019, doi:10.1109/iSES47678.2019.00066.

[15] S. Chawla, H. Chawla, "IoT-Based Digital LPG Gas Cylinder Trolley to Prevent Hazards with Voice-Controlled Features," in 2023 6th International Conference on Information Systems and Computer Networks (ISCON), 1–6, 2023, doi:10.1109/ISCON57294.2023.10112147.

[16] K. Chakradhar, R. Deshmukh, P. P. Singh, B. Hazela, R. Taluja, "LPG Cylinder Leakage Monitoring by IoT," in 2023 International Conference on Inventive Computation Technologies (ICICT), 1386–1389, 2023, doi:10.1109/ICICT57646.2023.10134191.

[17] M. H. B. M. Yaya, R. K. Patchmuthu, A. T. Wan, "LPG Gas Usage and Leakage Detection Using IoT in Brunei," in 2021 International Conference on Green Energy, Computing and Sustainable Technology (GECOST), 1–5, 2021, doi:10.1109/GECOST52368.2021.9538647.

[18] P. P. Ramadhani, S. Hadi, R. Rosadi, "Implementation of Data Warehouse in Making Business Intelligence Dashboard Development Using PostgreSQL Database and Kimball Lifecycle Method," in 2021 International Conference on Artificial Intelligence and Big Data Analytics, 88–92, 2021, doi:10.1109/ICAIBDA53487.2021.9689697.

[19] K. Gavaskar, D. Malathi, G. Ravivarma, A. Arulmurugan, "Development of LPG Leakage Detection Alert and Auto Exhaust System using IoT," in 2021 7th International Conference on Electrical Energy Systems (ICEES), 558–563, 2021, doi:10.1109/ICEES51510.2021.9383633.

[20] R. Eshgarf, M. Deldardil, "Identifying Effective Factors on Deploying Business Intelligence in the Gas Refinery Industries of Iran Based on DEMATEL Approach: A Study of Parsian Gas Refinery Co." Management and Administrative Sciences Review, **3**, 523–531, 2014.

[21] A. Fornaroli, D. Gatica-Perez, "Urban Crowdsourcing Platforms across the World: A Systematic Review," Digit. Gov.: Res. Pract., **4**(3), 2023, doi:10.1145/3603256.

[22] S. Shrestha, V. P. K. Anne, R. Chaitanya, "IoT Based Smart Gas Management System," in 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 550–555, 2019, doi:10.1109/ICOEI.2019.8862639.

[23] R. K. Kodali, R. Greeshma, K. P. Nimmanapalli, Y. K. Y. Borra, "IOT Based Industrial Plant Safety Gas Leakage Detection System," in 2018 4th International Conference on Computing Communication and Automation (ICCCA), 1–5, 2018, doi:10.1109/CCAA.2018.8777463.

[24] G. V. Da Silva Medeiros, M. R. d. Santos, A. S. B. Lopes, E. C. Barbalho Neto, "Smartgas: a smart platform for cooking gas monitoring," in 2017 IEEE First Summer School on Smart Cities (S3C), 97–102, 2017, doi:10.1109/S3C.2017.8501387.

[25] N. Denić, Z. Nešić, M. Radojičić, J. V. Vasović, "Some considerations on business intelligence application in business improvement," in 2014 22nd Telecommunications Forum Telfor (TELFOR), 1142–1145, 2014, doi:10.1109/TELFOR.2014.7034609.

[26] U. u. R. Zia, M. Zulfiqar, U. Azram, M. Haris, M. A. Khan, M. O. Zahoor, "Use of Macro/Micro Models and Business Intelligence tools for Energy Assessment and Scenario based Modeling," in 2019 4th International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), 1–7, 2019, doi:10.1109/ICEEST48626.2019.8981691.

[27] P. R. Meris, E. Dimaunahan, J. C. Dela Cruz, N. A. Fadchar, M. C. Manuel, J. C. C. Bonaobra, F. J. I. Ranosa, J. L. D. Mangaoang, P. C. Reyes, "IOT Based – Automated Indoor Air Quality and LPG Leak Detection Control System using Support Vector Machine," in 2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC), 231–235, 2020, doi:10.1109/ICSGRC49013.2020.9232472.

[28] S. Z. Yahaya, M. N. Mohd Zailani, Z. H. Che Soh, K. Ahmad, "IoT Based System for Monitoring and Control of Gas Leaking," in 2020 1st International Conference on Information Technology, Advanced Mechanical and Electrical Engineering (ICITAMEE), 122–127, 2020, doi:10.1109/ICITAMEE50454.2020.9398384.

[29] F. Aman, T. P. Thiran, K. Huda Yusof, N. M. Sapari, "IoT Gas Leakage Detection, Alert, and Gas Concentration Reduction System," in 2022 IEEE 12th Symposium on Computer Applications & Industrial Electronics (ISCAIE), 55–60, 2022, doi:10.1109/ISCAIE54458.2022.9794559.

[30] M. R. Islam, A. Matin, M. S. Siddiquee, F. M. S. Hasnain, M. H. Rahman, T. Hasan, "A Novel Smart Gas Stove with Gas Leakage Detection and Multistage Prevention System Using IoT LoRa Technology," in 2020 IEEE Electric Power and Energy Conference (EPEC), 1–5, 2020, doi:10.1109/EPEC48502.2020.9320109.

[31] M. Kumaran, J. Pradeep, R. Hounandan, B. Prahatheesh, "Smart LPG Cylinder Monitoring and Explosion Management System," in 2021 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE), 1–7, 2021, doi:10.1109/ATEE52255.2021.9425101.

[32] F. A. Hambali, R. Fitriana, E. Joelianto, "Integration System of IoT Gas Sensor using Simple Network Management Protocol and Open Platform Communication," in 2021 IEEE 7th Information Technology International Seminar (ITIS), 1–6, 2021, doi:10.1109/ITIS53497.2021.9791564.

[33] L. Jinfeng, C. Chen, C. Xiaowei, W. Wei, "Management of Indoor Gas Safety based on the NB-IoT Gas Meter," in 2021 33rd Chinese Control and Decision Conference (CCDC), 2776–2780, 2021, doi:10.1109/CCDC52312.2021.9602560.

[34] S. I. Nahid, M. M. Khan, "Toxic Gas Sensor and Temperature Monitoring in Industries using Internet of Things (IoT)," in 2021 24th International Conference on Computer and Information Technology (ICCIT), 1–6, 2021, doi:10.1109/ICCIT54785.2021.9689802.

[35] K. Şahinbaş, B. Yılmaz, Business Intelligence Application in the Natural Gas Industry: A Company Case, 141–154, Springer International Publishing, Cham, 2021, doi:10.1007/978-3-030-76783-9_11.

[36] H. Saad, S. A. Siddiqui, N. F. Naim, N. Othman, "Development of LPG Leakage Simulation System Integrated with the Internet of Things (IoT)," in 2022 IEEE 18th International Colloquium on Signal Processing & Applications (CSPA), 161–166, 2022, doi:10.1109/CSPA55076.2022.9781880.

[37] R. B, G. K, M. D, N. R, G. V, S. R, "Smart Detection System for LPG Gas Leakage using IoT," in 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), 421–430, 2022, doi:10.1109/ICCMC53470.2022.9753894.

[38] B. Gokulavasan, K. Shrina, U. Sangeetha Sruthi, R. Sneka Darshini, D. Srivaishnavi, "Smart Gas Booking System and Leakage Detection Using IOT," in 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), volume 1, 1914–1917, 2022, doi:10.1109/ICACCS54159.2022.9785144.

[39] S. Reddy, R. P. K N, "LPG Gas Detection and Monitoring Using IoT," in 2022 International Interdisciplinary Humanitarian Conference for Sustainability (IIHC), 693–697, 2022, doi:10.1109/IIHC55949.2022.10060195.

[40] L. Khajavizadeh, M. Andersson, "MOSFET-based gas sensors for process industry IoT applications," in 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), 1–5, 2022, doi:10.1109/ICECCME55909.2022.9988741.

[41] R. B, M. D, G. K, S. N, S. R, K. T, "IoT based Automatic Electricity Cut off using LPG Gas Leakage Detection System," in 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), 1–8, 2023, doi:10.1109/ICEARS56392.2023.10085217.

[42] S. Lorthong, U. Janjarassuk, N. Jayranaiwachira, "LPG Leakage Risk Predictions from an IoT-Based Detection System Using Machine Learning," in 2023 9th International Conference on Engineering, Applied Sciences, and Technology (ICEAST), 14–17, 2023, doi:10.1109/ICEAST58324.2023.10157528.

[43] T. Kalavathi Devi, N. S. Kumar, G. Chandrasekaran, P. Sakthivel, N. Priyadarshi, M. S. Bhaskar, N. Kumar, "IoT Based Remote Monitoring of Gas Leakage in Power Plants," in 2023 IEEE 3rd International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET), 1–5, 2023, doi:10.1109/TEMSMET56707.2023.10150136.

[44] M. Reza, "Business Intelligence on Agile Natural Gas Supply Chain," Journak of Business Data Science Research, 2023.

[45] D. Gautam, S. Bhatia, N. Goel, B. Mallikaijuna, G. H S, B. Bhushan Naib, "Development of IoT Enabled Framework for LPG Gas Leakage Detection and Weight Monitoring System," in 2023 International Conference on Device Intelligence, Computing and Communication Technologies, (DICCT), 182–187, 2023, doi:10.1109/DICCT56244.2023.10110294.

[46] Blynk, "Blynk," 2023. Avaliable Online: https://blynk.io/

[47] Microsoft, "Power BI," 2023, Avaliable Online: https://powerbi.microsoft.com/es-es/

[48] K. Ralph, R. Margy, The Data Warehouse Toolkit, In John Wiley & Sons, Inc., Canada, 2018.

[49] S. Nishisato, Data Analysis and Likert Scale, 19–36, Springer Nature Singapore, Singapore, 2023, doi:10.1007/978-981-99-2295-6_2.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 4, 93-101 (2024)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Digitalization Review for American SMEs

Dharmender Salian, Steven Brown, Raed Sbeit

School of Computer and Information Sciences, University of the Cumberlands, Williamsburg, 40769, USA

ARTICLE INFO

Article history:
Received: 30 April, 2024
Revised: 18 July, 2024
Accepted: 31 July, 2024
Online: 16 August, 2024

Keywords:
Industry 5.0
Big Data
Digitalization

ABSTRACT

SME big data maturity models will be reviewed in this study to identify systematic publications related to the subject. For SMEs to remain competitive, digitalization is essential. Due to limited resources, SMEs need to be more proactive in digitalization. Still, the benefits, such as operational efficiency, cost reduction, quality improvement, and innovative culture, make digitalization attractive and valuable to customers. In recent years, there has been an increase in the use of big data techniques in operations. The paper discusses big data applications in SMEs through the lens of a big data maturity model. This paper met two objectives. First, this paper summarizes the most commonly used maturity models in the existing literature. Second, existing Big Data maturity models have limitations. Moreover, this paper outlines key considerations for selecting a Big Data maturity model to support data-driven decisions. Based on the Big Data maturity dimensions, further work aims to develop a new Big Data maturity model.

1. Introduction

This paper is an extension of work initially presented at the IEEE Conference [1]. The heterogeneous data sets in big data contain different types of information. Organizations worldwide have been focusing on big data for the last decade. Real-time decision-making relies on statistics, econometrics, math, simulations, optimizations, and other techniques for collecting and analyzing high-speed data from multiple sources. [2]. Big data must be understood and adapted to specific uses and requirements in most organizations [3]. Big data allows managers to monitor organizational processes, assets, production units, and supply chains [4]. Organizations can identify their competitors with the help of current and new web data [5]. Analyzing big data can also help producers discover what their customers are doing, what they are complaining about, and what they are asking.

Information management and strategic use are increasingly important to organizations and industries [6]. Organizations can gain a competitive edge by collecting big data. Costs and time can add up when acquiring organization-specific data [7]. Many factors go into collecting and storing data, analyzing the data correctly, and even knowing what questions to ask and which data to collect. [7]. Big data requires new skills that traditional analysts still need to gain [6]. The use of Big Data can fundamentally transform businesses through radical disruption, minor incremental improvements, or a complete reimagining of the business model. Using big data's volume, variety, and velocity can lead to new insights and better decision-making [8]. Digital technology generates much data from various sources, such as manufacturing, businesses, education, and entertainment. [9]. Due to its popularity, many fields have incorporated machine learning, cognitive science, and the semantic web [9]. Data security, privacy, and ethical concerns hinder big data's potential.

Strategic planning and execution guide business decisions [10]. Strategic planning aims to get things done, so intelligence needs to be deployed [10]. A company's strategy must be positioned strategically with the right vision as online, social media, and big data continue to grow in importance [11]. Big data can help businesses grow internationally and improve their performance in critical areas like international market orientation [11].

In addition to managing inventory, controlling quality, and minimizing costs, hybrid technology systems assist companies with their operations. Technology can be valued by core use cases instead of expert systems, robot localization, and visual surveillance [8]. There are 61.7 million small and medium businesses (SMEs) in America, which account for 46.4% of the workforce [12]. Governments and large tech companies are funding big data applications in advanced economies. Big data is

*Corresponding Authors: Dharmender Salian University of the Cumberlands, Emails: dsalian0302@ucumberlands.edu

https://dx.doi.org/10.25046/aj090410

being reviewed in engineering and manufacturing since previous reviews ignored them.

SMEs need to catch up to larger firms in their adoption of digital technologies, and digital adoption by SMEs is still confined mainly to essential services, with adoption gaps increasing with the sophistication of technologies. SMEs can significantly benefit from digital systems, but their development could be more expensive, time-consuming, and skill-intensive. Since the COVID-19 crisis, SME digitalization has accelerated, but barriers remain.

SMEs comprise most companies and industries in most countries and regions, and their contribution to inclusive and resilient societies is critical. Due to the information technology revolution, SMEs' identity and position have changed dramatically over the past two decades. Due to this period, they have become more vulnerable to global competition, but they also have acquired competencies and opportunities to reach previously unimagined audiences.SME adoption of digital transformation poses many challenges for governments, including encouraging digital adoption by SMEs, supporting SME training and upskilling, and strengthening managerial skills.

A growing number of small businesses have identified the importance of digital Infrastructure as a significant driver of long-term financial growth and adaptability. To remain competitive in an evolving digital landscape, SMEs must keep these trends in mind. Investing in digitalization can enable SMEs to gain valuable insights from their data, increase efficiency, reduce costs, improve competitiveness, and create more significant opportunities for scalability and growth. It may involve a substantial upfront cost in the medium to long term. Digitalization presents both opportunities and challenges for SMEs in contemporary business.

Digital Transformation in automotive manufacturing means using digital tools to improve efficiency and reduce costs. Optimizing production planning and improving quality is made possible by data-driven analytics and artificial intelligence. With digital transformation, businesses can reduce production costs and manufacturing time as they face cutthroat competition. Automated robots and automation help companies analyze data in real time, improving productivity, while artificial intelligence, additive manufacturing, and the Internet of Things revolutionize product design. Electric cars are becoming more popular as eco-friendly vehicles increase demand, so automakers need new suppliers. Models of maturity (M.M.s) serve as tools for answering the following questions.

RQ1 In automotive manufacturing, what role does big data maturity play?

RQ2 What are the different stages of maturity described in the literature review?

RQ3 Big data maturity models have what characteristics and goals?

A systematic review of big data M.M.s determines the dimensions to assess. The extensive data maturity assessment can help organizations evaluate their capabilities, identify gaps, and create a roadmap for building more successful big data programs [13]. This study will fill the research gap and guide future studies considering these dimensions. Following is a breakdown of the rest of the paper: Sections 2 and 3 present a SWOT and PESTLE analysis of SMEs. Section 4 is about Big Data applications. Digital Strategy is in section 5. The sixth section discusses big data. Section 7 provides a literature review. Section 8 is about the requirements for a new SME maturity model. Section 8 describes AI MM for the aerospace industry. Section 9 discusses future challenges, while section 10 summarizes the study's results.

2. SWOT

Table 1: SWOT Analysis of SMEs

Strength	**Weakness**
• Adaptable and more flexible organization structure. • Experience, knowledge, accomplishments, and skills. • Constant product innovation • Facilities in developing nation • Electric vehicles increased their market shares. • Incentives and support from the government	• Bargaining power of consumers • Government regulations • Investments are expensive • R&D facilities are not good enough • Lack of skilled workers • A declining aftermarket
Opportunities	**Threats**
• Aid from the government. • Meeting customer needs and making the organization more efficient. • Production costs are lower. • Batteries and electric cars have led to new supply chains. • Commercial vehicles are becoming more multifunctional.	• Rising competition • Sluggish economy • Competition in the market. • The inflation rate is high • An economic downturn • Enterprising newcomers

3. Pestle Analysis

New products and projects are launched in an environment tracked by pestle analysis (see Table 2).

Table 2: SME PESTLE Analysis

Political	**Economical**
• U.S. auto sales are growing.	• Automobile industry hit by pandemics.
Social	**Technological**

• The auto industry needs skilled talent.	• The digital revolution is taking off.
Legal	**Environmental**
• The auto industry has regulations and restrictions.	• Manufacturing's environmental impact.

4. Big Data Applications in Manufacturing

Intelligent systems have transformed smart manufacturing through the digital revolution. New methods for developing and deploying big data could benefit the manufacturing industry. The vast majority of proposed solutions in manufacturing build a big data platform from scratch by installing and testing each tool individually, owing to the fast-changing nature of many of the tools. As big data and manufacturing knowledge converge, smart manufacturing becomes more actionable in real-time. Big data are collected and analyzed for timely information, but manufacturing industries may only know the best approach with domain expertise. Manufacturers can create actionable knowledge by combining manufacturing knowledge with timely information in big data.

Machines, products, and supply chains are all examples of big data gathered in the manufacturing industry. Big Data Manufacturing finds previously unavailable insights into efficiency, productivity, and quality. With industrial Internet of Things (IIoT) sensors, manufacturing execution systems (MES) (software applications that track and control production processes on the factory floor), and industrial control systems, manufacturers are gathering a lot of big data from different sources throughout their operations. (crucial for managing automated processes in manufacturing). Many production activities affect yield, so manufacturers must diagnose and correct process errors in detail. Analytics provides actionable insights and patterns by analyzing data using statistical models, machine learning, and artificial intelligence. The capabilities of conducting sophisticated statistical analyses are now available to global manufacturers across various industries and geographies. Companies are now aggregating and analyzing previously isolated data sets to uncover valuable insights, which requires long-term planning and investment. By identifying opportunities for new products and improvements in existing processes, data-driven insights can give manufacturers a competitive advantage.

SMEs need more funding and talent. Currently, they are reskilling their employees and adopting a new employment model that emphasizes selecting employees with various skill sets. To understand big data, SMEs must develop extensive data knowledge and education, workplace training, recognition, awareness, and promotion. Everyone must understand the benefits of comprehensive data adoption regardless of level or type of organization.

- Maintenance Regulation

Integration of Machine Learning (ML) and Big Data (B.D.) into IoT (Internet of Things).

- Quality Checks

Digitalization has improved the accuracy and efficiency of manufacturing quality assurance (Q.A.).

- Supply Chain Management

Traditional systems must change to keep up with the complexity of supply chain management (SCM). Analyzing big data requires tools, processing systems, and algorithms to interpret insights. Making better decisions requires better data integration; automation is only the beginning.

- Risk Evaluation

Big data can help organizations predict and calculate risks so they can take preventative measures. When assessing risks, one must consider trends with unique impacts, including human talent, production materials, and global economic developments. Additionally, losses and low margins can pose risks.

- Production Optimization

Data analytics can improve aging manufacturing systems. The use of big data ranges from forecasting to enterprise performance management.

- Sustainable Development

Big data can prevent resource waste. Preventing excessive production costs begins with monitoring supply chain management and machine management.

5. Digital Strategy

Digitization converts analog to digital. As the business model changes, a new revenue and value-producing opportunity arises. In today's workplace, industry, and region, digital skills are a must. Companies are implementing digital technologies, which means jobs are changing. Whether transforming business processes or shifting work roles, automation is critical to digitalization. Transparency and efficiency boost the organization's bottom line with digitalization.

As a result of technology, manufacturing companies are becoming more efficient, customer-centric, and competitive. Manufacturers can improve their products with the help of artificial intelligence, machine learning, and big data analytics. Manufacturers can improve their products with the help of artificial intelligence, machine learning, and big data analytics. In addition to building brand equity, increasing employee effectiveness, and increasing profits, industries and businesses must have strategic plans. Figure 1 shows Firm inversion (14).

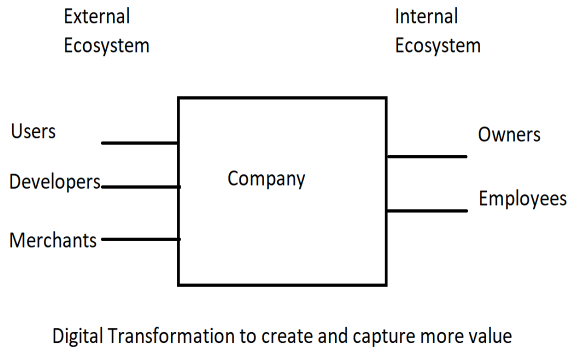


Figure 1: Inverted Firm

The inversion of a company is where the value creation is due to digital transformation. Creating value on their own is less crucial than orchestrating value with other firms, so users, developers, and merchants need to be able to partner with digital investments at scale to support firm inversion. Inverted firms rely on others' resources rather than controlling their own, and to enable this, managers need to switch from controlling to enabling. An external ecosystem can thrive if companies coerce partners into sharing ideas, investments, and effort. Leaders are responsible for encouraging and accelerating multiparty collaboration within and outside their organizations.

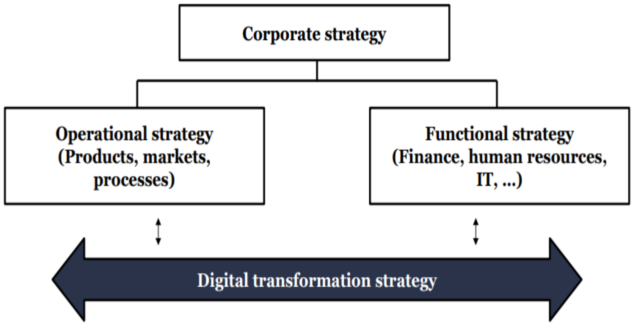

Corporate strategy
Operational strategy (Products, markets, processes)
Functional strategy (Finance, human resources, IT, ...)
Digital transformation strategy

Figure 2: Corporate strategy and digital transformation

Organizations, products, and processes are all being transformed by digital transformation. Digital transformation should align organizational and functional strategies (see Figure 2). Financials play a role in both driving and constraining transformation. Companies can finance transformations both internally and externally. The key to transforming business is to create new opportunities. Whatever the industry or firm, digital transformation strategies have some things in common: technology, value creation, structure, and finance (see Figure 3).

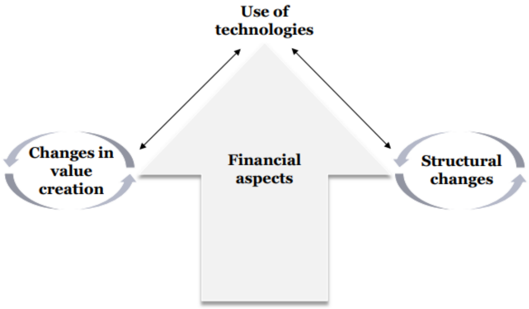


Figure 3: Balancing four transformational dimensions of digital transformation

Porter recommends three generic strategies for any business: differentiation, leadership, and focus. In the digital transformation, the inversion of a company that makes the most money reshapes how companies create value. Customized transformation measures can help achieve this. Companies can do this when they stop creating value independently and start orchestrating it with each other. Partnering with users, developers, and merchants at scale is the key to success. Network effects are not common in standalone companies, but they might occur if they look at their business frame in an ecosystem context. High market caps do not come from automation or transferring labor to capital but from coordinating external value creation. Shared value can help companies get multiple benefits.

Economies and markets are suffering from Uber's and Airbnb's systemic effects. Digital disruption - or radical digital innovation - and its broader systemic effects get much attention. From their perspective, companies heavily invested in old conditions tend to be disrupted by digital disruption. Digital processes and artifacts can radically change operations for established companies. There are a lot of digital innovations that have triggered and spurred the digital revolution, some of which might disrupt traditional industries. Embracing the digital transformation of manufacturing operations is critical for manufacturers seeking unprecedented efficiency and quality control.

6. Big Data and SME

Big data for businesses, the I.T. industry, and researchers are associated with challenges and opportunities. These days, SMEs need more resources to deal with big data. The high speed and volume of data exchange in industries have resulted in big data becoming one of the topics of discussion among managers and decision-makers these days because it provides the information needed to pinpoint inefficiencies in their operations and, as a result, resolve these issues. SMEs are more flexible and adaptable to technology and change than large firms, and managers of SMEs might be influenced by significant characteristics such as competition pressure, financial resources, and talent shortages when making decisions regarding extensive data adoption. This study's findings reinforce that technology and innovation are priorities for the growth of SMEs, and big data will contribute significantly to this growth. Using big data in SMEs gives organizations greater flexibility, efficiency, responsiveness, and ability to anticipate and meet customer needs

SMEs benefit from big data by offering real-time solutions to issues in every industry and fostering alliances. Openness helps to make decisions, and SMEs are chosen explicitly within the context. Because of their overall position in the economy, SMEs have the advantage and flexibility to adjust to changes faster toward efficiency, and a slight change can have a significant macro-level impact. Nevertheless, there are several contentious issues associated with Big Data, including storage, processing, and generating useful information from it, as well as privacy and security concerns. With the addition of machine-collected data (e.g., intelligent counters and sensors), the datasets can grow to enormous sizes, bringing about the possibility of improved decision-making and performance outcomes. The fourth industrial revolution and global digitization of consumers, customers, and providers have created abundant data that firms must learn to use [15]. Technology such as the Internet of Things (IoT), which has the potential to lead to many developments, is essential when it comes to exploring big data. In addition, manufacturers can use IoT to monitor machinery and track parts in real-time.

Furthermore, IoT connects everything. The internet allows us to connect things in our everyday lives, like cars, T.V.s, washing machines, pumps, shipping containers, and machinery. Cloud-based technology and mobile devices with connections to all these things mean SMEs must be bold in taking business-related risks for advancements.

Transparency makes data available to relevant users as promptly as possible, while open transparency improves data quality, consistency, reliability, availability, and accessibility. As a result, open, transparent data is valuable for improving product offerings, reducing time-to-market, and improving engineering processes. Data transparency is achieved by effectively managing people, data, tools, and intent collectively through proper Big Data management, and the most critical aspects of BDM are data governance, business intelligence/data warehouse management, and data quality management. BDM requires knowledge of new Big Data technologies and business practices. Processed and refined data becomes more valuable over time but also loses value over time. To overcome this challenge, BDM should be automated. Automated BDM processes make it easier to manage the data more efficiently and effectively while freeing up human resources simultaneously.

The West's major players in Big Data are Alphabet (Google), Amazing, Facebook, and Apple. [16]. Technology leadership, research, and the development of new processes can help develop innovative processes. Organizations must still utilize big data and adapt it to specific needs and requirements [17]. Businesses can increase performance and expose variability by creating and storing more digital transactional data. SMEs should focus on the growth opportunities that result from Big Data when evaluating its use and potential. Segmentation of customers can improve products and services, and Big Data can be used to analyze in-depth data for correlations, risks, and opportunities and to do predictive maintenance, demand forecasting, process optimization, inventory planning, and market segmentation, analysis, and forecasting. Researchers have a consensus that Big Data is transforming information sciences, manufacturing, retail, and healthcare [18]. A big data model for SMEs is possible with the leaders' support, mandate, and trust, and data from across all divisions will only be shared when management has given it a clear direction and a clear picture of the results. Lastly, Big Data can significantly improve revenue, cost, profit, and growth through improved productivity, efficiency, and new business creation.

7. Literature Review

An extensive literature review can help identify knowledge gaps, generate insights, and provide helpful advice. This paper reviews previous studies on big data maturity models and discusses developments in the future using an in-depth literature review.

Most organizations are still experimenting with big data. Big Data strategies and Big Data technology can benefit businesses by increasing value, decreasing costs, improving operational efficiency, and streamlining operations. Automation, employee productivity, and streamlining can all be achieved with big data. Simplified operations and reduced downtime are the results of Big Data.

The maturity model provides layered levels of accomplishment for assessing maturity in these areas and pinpointing areas for improvement for stakeholders to plan. Big data has different maturity models with different names, levels, and content. The following sections will summarize these models.

An organization's architecture is one way to describe its prominent data structure. The data organization model describes how different systems integrate. An organization's competitive advantage depends on data quality; thus, a maturity model is critical. [28]. Table 3 shows maturity models. Digital transformation for SMEs is a specialty of others, while some specialize in big data.

Table 3: Comparing big data dimensions overall

Author	[19]	[20]	[21]	[22]	[23]
Connectivity	X	X		X	X
Resilience	X	X		X	
Sustainability		X		X	
Expansive Growth	X	X			X
Strategy	X	X		X	X
Leadership		X	X	X	X
Customers	X			X	X
Culture		X			X
Production	X		X	X	X

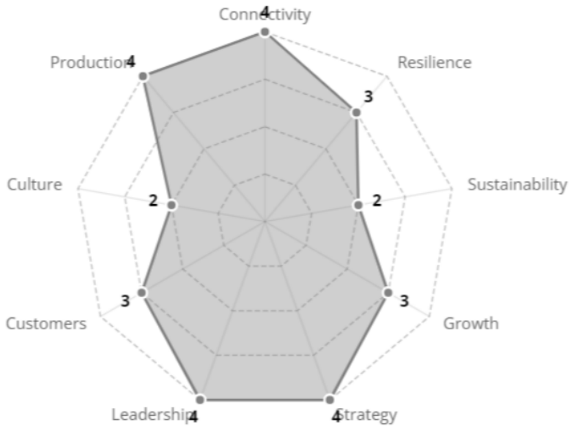


Figure 4: Dimensions

By identifying the shortcomings, the big data maturity model can be improved (see Figure 4). The maturity models discussed above only work for SMEs, are prevalent, and may result in inaccurate results. Regarding big data MM for automotive small and medium enterprises, the maturity assessment dimensions cover only some sizes and capabilities. Below are some critical dimensions of big data assessment, so a new big data MM should cover these dimensions.

8. Requirements for a new SME maturity model

Model construction based on an organization's future big data needs to identify strengths and areas for improvement and prioritize reaching higher levels of big data maturity. These maturity stages can serve the research gap:

1. Nascent: Beginners need strategic focus. Regarding environmental pursuits, organizations are still in the exploratory phase, looking at the benefits of big data.

2. Innovation: In the innovation stage, companies use big data techniques to reduce costs and increase revenue. A culture of innovation is being developed, focusing on environmental benefits.

3. Integration: During the Integration stage, organizations have relatively straightforward business strategies for using big data, and their processes, Infrastructure, and investments align with what big data requires. Big data-embracing culture and environment policies are in place in subunits pursuing big data innovations.

4. Mature: A big data-focused work culture is characteristic of organizations at the mature stage of big data maturity. The organization employs big data-enabled action plans to exhibit competitive superiority. The matrix below compares the four maturity stages with the selected big data dimensions.

Table 4: Most influencing parameters for each maturity stage

	STAGE					
		Parameter	Nascent	Innovation	Integration	Mature
DIMENSION	Leadership	Evaluate Objectives	Organizational priorities	Expectations and concerns	Integration planning	driving functional excellence
		Measure confidence				
	Expansive growth	Successful transformation	Setting clear, attainable goals	A collaborative and adaptive growth process	Identifying areas of concern and achieving integration goals.	Attained a stable and efficient operational state
		Grow healthily				
	Strategy	Governance	The goal is undefined, and no budget allocation	The focus is on cost leadership and efficiency	Aligned with the overall business strategy	Business development and competitiveness
		Cost leadership				
		security				
	Resiliency	Handle adversity	Risk initiation and command structure in place	The resilience of the market for essential supplies	System and Infrastructure established.	Professional and responsible. An insight into the market.
		insecurity				
	Sustainability	Sustainability business practices	Taking advantage of opportunities	Exploring benefits	Participating in stakeholder engagement	Environmentally friendly and have profit
	Culture	Culture shift	Getting the most out of big data	A culture that embraces big data	Embracing big data and evolving culturally	A proactive mindset and innovative decision-making
		Innovation mindset				
	Customer	Engaging customers	Engaging customers in a more meaningful way	Establishing a customer-centric approach	Customer-centricity and clearly defined accountability	Fulfill customer desires and needs
	Production	Empower employees	Developing but not prioritized	Big data professionals in charge need more enterprise-level centralization despite their expertise.	Decisions align with the larger ecosystem and boardroom decisions.	Automate and reduce inefficiencies
		Connectivity and Integration				

9. Big Data MM for the Automotive SME Market

Data maturity is the ability of an organization to use its data effectively. Companies can develop best practices for managing big data by assessing their data maturity. The maturity model assesses an organization's readiness, identifies its weaknesses, and identifies its capabilities. It is vital to fix those weaknesses and improve the assessment to advance. Organizations should assess their extensive data capabilities before moving on to maturity models since they need more standardized development methodologies. The goals of big data maturity models are:

- Analyze big data in critical areas of the organization with this tool.
- We are setting milestones during development.

Data maturity leads to increased revenue and lower costs. As part of Industry 5.0, big data maturity models should support sustainability, resilience, and human-centered design (see Figure 5). Productivity and efficiency are only part of the Industrial 5.0 vision [24]. Worker well-being is prioritized and emphasizes the industry's role in society. Instead of just assessing technology, industry 5.0 focuses on value rather than technology. Industrial output and profit alone are no longer enough to account for environmental and societal costs. Investing in automation, digitization, and artificial intelligence is essential, but not at the expense of sustainability and human beings. People and societies should benefit from technology, which is employed to make this ideal a reality. A sustainable approach to natural resources relies on repurposing, recycling, and reuse. Thus, it reduces waste and the negative impacts on the environment. Green solutions will be vital for companies (and society as a whole) in the next few decades. A resilient supply chain and production means making them more resilient to interruptions and ensuring that critical Infrastructure is available and supported during crises.

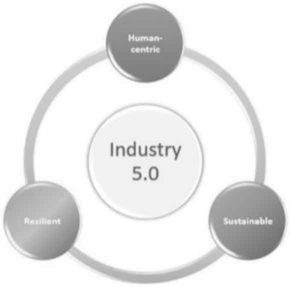


Figure 5: Industry 5.0

It is essential that the SME industry survives upheavals and provides assistance when needed. A sustainable approach can help SMEs recover, grow, and contribute to long-term survival.

Big data analytics makes it easier for businesses to manage large amounts of data. Businesses can use this emerging trend to manage complex datasets, democratize insights, and empower workers. Small companies are not using big data analytics because they are late adopters.

E.V. manufacturers can become more prosperous with a digital backbone. Contrary to established car companies, E.V.s are disrupting the market. The maturity model assessment must include electric vehicles (E.V.s) to measure the transition to zero-emission cars. Customers can use big data about charger networks to recharge their batteries en route. In electric vehicles, GPS navigation systems can incorporate charger information and directions to anticipate battery recharging needs.

In addition to 5G reducing data transmission latency, big data and IoT require fast data transfer rates, and factory automation relies heavily on real-time operations. Using big data can improve visibility, insights, and control for manufacturers. It can also improve production processes, reduce downtime, improve workplace safety, and predict maintenance issues. IoT and big data applications in industries like industrial automation get a fillip due to 6G networks. They will be able to deliver exceptionally high-performance connectivity. In 6G, machines and processes can exchange data instantly and with ultra-low latency. Augmented analytics will be a big trend in big data, with machine learning, A.I. helping, and advanced sensing allowing for low-latency, fast communications.

10. Future challenges and opportunities

There is constant refinement and development of new Big Data models, opening up new research opportunities; Big Data maturity models and the area of Big Data are constantly evolving. The Big Data domain will undoubtedly gain acceptance in the following years due to its potential for generating business value. Today, Big Data is still a relatively new domain, but it will undoubtedly prove to be a valuable area in the future. Big Data is becoming part of everyday business activities. Strategic Technological and critical business solutions will continue to evolve and improve. As Big Data evolves, new business approaches will emerge. As a result, future research should explore how Big Data contributes to business value creation. In addition, conducting a new benchmarking study for Big Data maturity models in the coming years would be worthwhile in assessing the new models and reevaluating the old models in light of evolutionary changes. Understanding the evolution of Big Data maturity modeling practices would be valuable.

Organizations operate, make decisions, and gain insights using big data, and as we look to the future, we see that it holds immense potential. With digital transformation, enterprises can improve their competitiveness. Big data technologies will increase The production of goods by 40% by 2035. Economic growth will average 1.7% across different industries. Intelligent manufacturing is complex for small and medium-sized businesses; they need not implement it better. Inventing must keep up, however. A study predicts that 75 billion IoT devices connect to the internet by 2025, and connecting all these devices to intelligence will be difficult. Organizations implement A.I. and IoT in production processes because of global competition. Intelligent manufacturing is the key to manufacturing's ability to compete globally. New revenue streams will arise as big data, machine learning, and IoT interconnect [25]. Big data opens up a whole new era of possibilities in the future. Several concrete recommendations have already been provided for developing and implementing the big data maturity model to ensure successful future integration and advancements of big data within an organization.

11. Conclusions

This study aimed to devise a benchmarking process to assess the business value created by maturity models and good practices for maturity modeling. Therefore, business organizations could compare maturity models to identify the best model for evaluating and improving their Big Data maturity. The purpose of Big Data MMs has expanded to include evaluation of the implementation of Big Data. Before implementing Big Data, organizations must adopt mature designs to maximize their success. In addition, this paper provides essential insights that can help relevant stakeholders select more effective models for their organizations.

Plant floors are experiencing technologies that were once considered science fiction. Machine learning algorithms enable manufacturing plants to go beyond tracking operations and use Big Data for decision-making. By doing so, they can control inefficiencies, evaluate alternative strategies, and develop new protocols without simultaneously backing up the supply chain. Manufacturing will likely suffer significant consequences due to an uncertain and turbulent environment, such as mandatory closures, logistical bottlenecks, supply problems, and volatile consumption patterns. Embracing advanced digital technologies, like big data, is critical to manufacturers' success. Scholars and practitioners are increasingly interested in big data due to the increasing amount of data collected and processed by firms. Digital transformation and big data are revolutionizing manufacturing, so organizations need a big data maturity model to evaluate their capabilities. SME maturity models need to be updated to meet the requirements of big data-focused SMEs. SME managers do not understand big data, like "unclarity regarding the benefits" and "insufficient understanding" of it.

References

[1] Salian, D.T., Sbeit, R., “Review of Digitalization Using Big Data Maturity Models: The Case of American Automotive SMEs,” *2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON),* 2023.

[2] Cakir, A., Akın, Ö., Deniz, H. F., Yılmaz, A., “Enabling real time big data solutions for manufacturing at scale,” *Journal of Big Data,* **9**(1):118, 2022, DOI: https://doi.org/10.1186/s40537-022-00672-6.

[3] “The big data bandwagon: How big data can benefit organizational practices,” *Strategic Direction,* **36**(10):13-14, 2020, DOI: https://doi.org/10.1108/SD-08-2020-0144.

[4] Khan, Z., Vorley, T., “Big data text analytics: an enabler of knowledge management,” *Journal of Knowledge Management,* **21**(1):18-34, 2017, DOI: 10.1108/JKM-06-2015-0238.

[5] Cappellesso, G., Thomé, K.M., “Technological innovation in food supply chains: systematic literature review,” *British Food Journal,* **121**(10):2413-2428, 2019, DOI: 10.1108/BFJ-03-2019-0160.

[6] “So, are the geeks inheriting the earth?: Big data, big decisions and best practice for information management,” *Strategic Direction,* **29**(9):12-15, 2013, DOI: https://doi.org/10.1108/SD-08-2013-0051.

[7] “The big data boost: Using proprietary big data to boost competitive advantage,” *Strategic Direction,* **37**(5):20-21**,** 2021, DOI: https://doi.org/10.1108/SD-03-2021-0027.

[8] Rosangela de Fátima, P. M., Tereza Cristina Melo de, Brito Carvalho., “Examining the nexus between the vs of big data and the sustainable challenges in the textile industry,” *Sustainability,* **14**(8):4638, 2022, DOI: https://doi.org/10.3390/su14084638.

[9] Dasari, S., Kaluri, R., “Big data analytics, processing models, taxonomy of tools, V’s, and challenges: State-of-art review and future implications,” *Wireless Communications & Mobile Computing (Online),* 2023, DOI: https://doi.org/10.1155/2023/3976302.

[10] “Maximizing the potential impact of competitive intelligence: Big data strategy lessons from Brazil’s private sector,” *Strategic Direction,* **36**(9):5-**7,** 2020, DOI: https://doi.org/10.1108/SD-06-2020-0126.

[11] “The next big thing?: How big data is shaping firms’ international strategies,” *Strategic Direction,* **35**(4):10-12**,** 2019, DOI: https://doi.org/10.1108/SD-01-2019-0004.

[12] Peinado-Asensi, I., Montés, N., García, E., “Industrial internet of things and big data techniques for the smart press shop 4.0 development in automotive industry,” *IOP Conference Series. Materials Science and Engineering,* **1284**(1):012012, 2023, DOI: https://doi.org/10.1088/1757-899X/1284/1/012012.

[13] Llopis-Albert, C., Rubio, F., Valero, F., “Impact of digital transformation on the automotive industry,” *Technological Forecasting and Social Change,* **162,** 2021, DOI: 10.1016/j.techfore.2020.120343.

[14] Alstyne, M.W.V., Parker, G.G., “Digital Transformation Changes How Companies Create Value,” *Harvard Business Review,* 2012.

[15] “Big data performance boost: Improving dynamic capabilities through big data analytics,” *Strategic Direction,* **39**(7):7-8, 2023, DOI: https://doi.org/10.1108/SD-06-2023-0073.

[16] “The big data bandwagon: How big data can benefit organizational practices,” *Strategic Direction,* **36**(10):13-14, 2020, DOI: https://doi.org/10.1108/SD-08-2020-0144.

[17] “The big data boost: Using proprietary big data to boost competitive advantage,” *Strategic Direction,* **37**(5):20-21, 2021, DOI: https://doi.org/10.1108/SD-03-2021-0027.

[18] Qasim, A. N., Nasir, N., Jamshed, S., Naz, S., Ali, M., Shahzad, A., “Big data management and environmental performance: Role of big data decision-making capabilities and decision-making quality,” *Journal of Enterprise Information Management,* **34**(4):1061-1096, 2021, DOI: https://doi.org/10.1108/JEIM-04-2020-0137.

[19] Hu, J., Gao, S., “Research and Application of Capability Maturity Model for Chinese Intelligent Manufacturing,” *Procedia CIRP,* 83:794-799, 2019, DOI: https://doi.org/10.1016/j.procir.2019.05.013.

[20] Bozic Yams, N., Richardson, V., Shubina, G., Albrecht, S., Gillblad, D., “Integrated A.I. and Innovation Management: The Beginning of a Beautiful Friendship,” *Technology,* 10:5-18, 2023, DOI: 10.22215/timreview/1399.

[21] Ansari, I., Barati, M., Sadeghi Moghadam, M. R., Ghobakhloo, M., “Readiness model for new technology exploitation,” *The International Journal of Quality & Reliability Management,* **40**(10):2519-2538, DOI: https://doi.org/10.1108/IJQRM-11-2022-0331.

[22] Chen, W., Liu, C., Xing, F., Peng, G., Yang, X., “Establishment of a maturity model to assess the development of industrial A.I. in smart manufacturing,” *Journal of Enterprise Information Management,* **35** (3):701-728, 2022, DOI: https://doi.org/10.1108/JEIM-10-2020-0397.

[23] Paschou, T., Rapaccini, M., Peters, C., Adrodegari, F., Saccani, N., “Developing a Maturity Model for Digital Servitization in Manufacturing Firms,” 2019.

[24] Valerio, P., “Industry 5.0 adds human-centric, sustainable, and resilient concepts to the industrial revolution,” *IOT Times,* 2021, URL: https://iot.eetimes.com/industry-5-0-adds-human-centric-sustainable-and-resilient-concepts-to-the-industrial revolution/.

[25] Friedenberg, M., “Our next great era: IDG enterprise's CEO spells out the future of big data, mobility and the impact of 50 billion connected devices by 2020,” *CIO,* **26** (11): 2013.

Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 4, 102-109 (2024)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Energy Management Policy and Strategies in ASEAN

Wai Yie Leong[1*], Yuan Zhi Leong [2], Wai San Leong [2]

[1]*INTI International University, Persiaran Perdana BBN Putra Nilai, 71800 Nilai, Negeri Sembilan, Malaysia*

[2]*Schneider Electric Singapore, 50 Kallang Avenue, Kallang, Singapore*

ARTICLE INFO

Article history:
Received: 30 June, 2024
Revised: 11 August, 2024
Accepted: 14 August, 2024
Online: 20 August, 2024

Keywords:
Energy Efficiency
ASEAN
Energy Infrastructure
Renewable Energy

ABSTRACT

This research analyses the challenges faced by ASEAN countries in managing its energy efficiencies and resources due to rapid economic growth, increasing energy demand, and diverse energy infrastructures across member states. This paper explores the energy management policies and strategies within the ASEAN region, focusing on the integration of energy efficiency measures, renewable energy initiatives, and cross-border energy trade. This paper analyse the region's progress towards its sustainable energy goals, the role of policy frameworks, and the impact of regional collaboration. Key challenges such as energy security, affordability, and environmental sustainability are examined, alongside opportunities for innovation in energy technologies and policy reforms. The findings highlight the importance of a cohesive energy management strategy that balances the diverse needs of ASEAN member states while advancing the region's transition towards a low-carbon future. This paper provides policy recommendations aimed at enhancing ASEAN's energy resilience and supporting its sustainable development goals.

1. Introduction

The ASEAN region's energy demand is growing significantly and rapidly as a result of urbanisation and economic advancement. The ASEAN region's reliance on fossil fuels, volatile geopolitics, and challenges associated with climate change make it vulnerable to energy supply vulnerabilities.

Consequently, member countries of ASEAN have been working together to develop and implement energy management policies that promote economic development, environmental sustainability, and energy security. The specifics of each ASEAN Member State's (AMS) energy efficiency (EE) and activities are shown in Figure 1. Remarkably, in the 2030s, Brunei, Singapore and Thailand, declared their intention to cut their Energy Intensity (EI) by 45%, 35%, and 30% [1]. Over the years, AMS has demonstrated a considerable reduction in energy intensity from 2005-2020 and the projected of energy Sumption to 2040 is shown in Figure 2.

In order to improve energy security, regional collaboration is emphasised in the ASEAN energy policy. Member nations cooperate to reduce supply disruptions and guarantee a steady supply of energy for their expanding economies by encouraging cross-border and international energy trade, international networkings, and energy resource sharing (Table I).

The ASEAN energy strategy encourages collaboration on energy-related projects, experience-sharing, financing access, and alliances with foreign organisations, development agencies, and other countries. The energy-related concerns is strengthened by this international cooperation.

The ASEAN energy strategy demonstrates a cooperative dedication to tackling the region's energy-related issues. The strategy lays the groundwork for member nations to collaborate on energy resources management and create a sustainable energy Future by fostering energy security, resource sustainability, and better economic growth. ASEAN Energy Statistics Leaflet (AESL) 2023 provides comprehensive visualised snapshots of the energy landscape in ASEAN. These include primary energy supply, final

*Corresponding Authors: Wai Yie Leong INTI International University, Emails: waiyie@gmail.com

https://dx.doi.org/10.25046/aj090411

energy consumption, electricity, renewable energy, energy-gender, and other energy-related indicators as shown in Figure 3.

2. Literature Review

The heterogeneous region of ASEAN has different energy needs, resources, and obstacles. In light of economic growth and urbanisation, there is an increasing need for energy, making it imperative to create and execute efficient energy management strategies in order to guarantee energy security, sustainability, and resilience. In order to better understand the literature and research on energy management policy in ASEAN, this review will focus on some of the major obstacles, frameworks for policy, and possible solutions.

Challenges in Energy Management: The vast array of problems posed by ASEAN's heterogeneous energy landscape is noteworthy. Coal, oil, and natural gas are examples of fossil fuels that continue to be major energy sources. These fuels raise difficulties with energy security and the environment. In addition, the region is vulnerable to price changes and geopolitical issues due to its reliance on imported fossil fuels. The necessity of developing sustainable energy sources and diversifying the energy mix is highlighted by this circumstance.

Energy Policies and Frameworks: The ASEAN Plan of Action for Energy Cooperation, or APAEC, is the cornerstone of the region's energy policy framework. In order to promote energy security, affordability, accessibility, and sustainability, APAEC was founded in 2016. It emphasises how important regional collaboration is to resolving energy-related problems, advancing energy trade within ASEAN, and encouraging energy technology knowledge transfer.

Development of Renewable Energy: As a result of its ability to improve energy security and lessen its negative effects on the environment, renewable energy policies have become more popular among ASEAN members. According to research [2], government initiatives on feed-in tariffs (FiT), incentives, and schemes could facilitate renewable energy technology on wind energy and solar power.

Initiatives for Energy Efficiency: ASEAN's energy management plans has been increasing energy efficiency. Studies [3] have demonstrated that energy efficiency initiatives aimed at families and businesses have resulted in significant energy savings as shown in Figure 4. These programmes include the creation of energy-efficient appliances, the implementation of best practices, and technological advancements.

Policy Coordination and Implementation: A number of studies highlight how crucial it is for ASEAN member nations to coordinate their policies to guarantee the successful application of energy management plans. Mechanisms to improve energy security and foster economic cooperation have been suggested, including cross-border energy commerce and harmonising energy norms. Policy coordination presents a number of issues that call for constant attention, especially when considering the disparities in national capacities and priorities.

The literature also suggests potential paths that may influence ASEAN's energy management laws. It has been proposed that the resilience and sustainability of energy systems can be improved by exploring energy technologies like energy storage systems and smart grids. The shift to low-carbon energy systems can also be facilitated by matching energy policies with global climate commitments like the Paris Agreement.

The analysis highlights the intricacy of ASEAN's energy management policies, involving issues with environmental sustainability, energy security, and policy coherence. It is encouraging to see how far the region has come in creating renewable energy sources and energy-saving techniques. However, to guarantee a safe, sustainable, and resilient energy future,

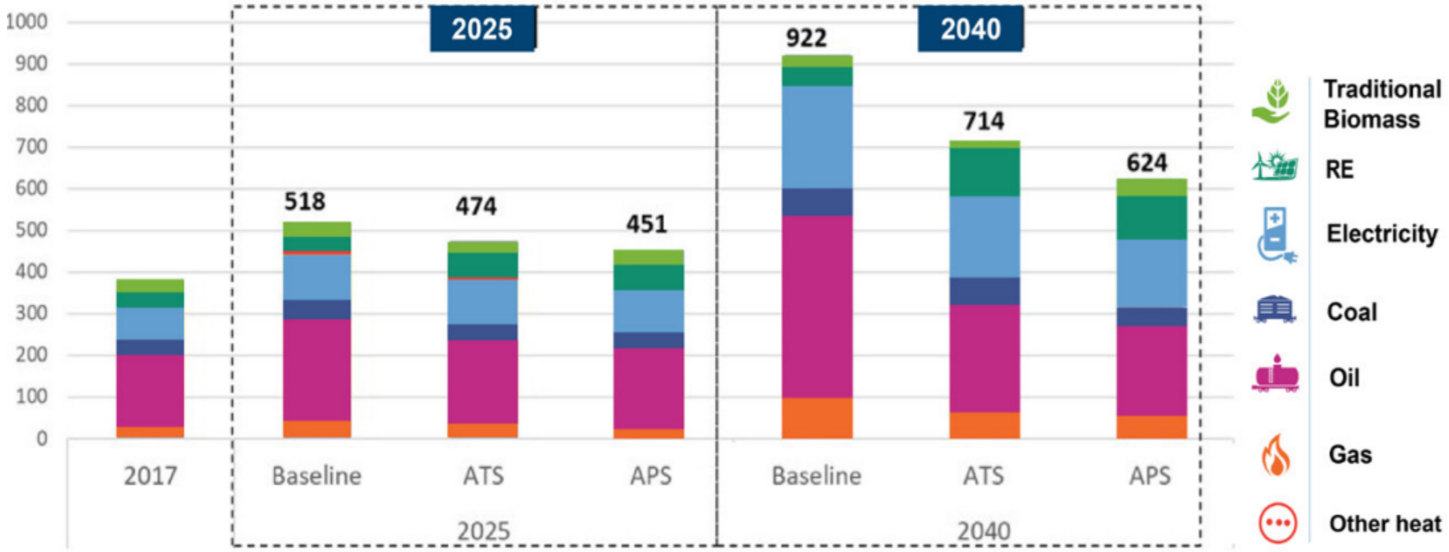


Figure 1. The projected ASEAN energy consumption based on 6th ASEAN Energy Outlook

cooperation must continue along with the development of novel tactics and changes to legislation [4].

3. Energy Efficiency Policy

The ASEAN region has a wealth of undeveloped renewable resources, but for now, fossil fuels control the majority of the energy systems in the area. ASEAN members aim to attain Net-Zero emissions by 2050 or later in order to combat climate change.

A. Singapore: The country has put in place a number of energy-saving initiatives:

- Energy Conservation Act mandates that major energy users increase their energy efficiency achievement and share on the energy consumption.

- Energy Efficiency National Partnership (EENP) initiative promotes the energy management strategies and energy savings objectives among organisations.

- Building designs and technology that are energy-efficient are promoted by the Green Mark certification programme.

B. Malaysia: The country has implemented energy efficiency policies such as the Energy Efficiency and Conservation Act, which attempts to increase energy efficiency in a number of industries.

- Projects and efforts pertaining to energy efficiency are supported by the Energy Efficiency and Conservation Fund.
- The programme called Malaysian Building Integrated Photovoltaic (MBIPV) promotes the integration of solar energy into buildings.

C. Thailand: The country's energy-saving initiatives include: Encouraging energy-efficient buildings and industry under the Energy Conservation Promotion Act.

- The plan encourages the use of energy-saving technologies and establishes goals for reducing energy intensity.
- Private investment in energy efficiency initiatives is encouraged by the Energy Performance Contracting (EPC) programme.

C. Indonesia: The country has implemented many energy efficiency programmes, such as the National Energy Policy, which endeavours to enhance energy efficiency while lowering energy intensity.

- The primary objective of the Energy Conservation Master Plan is to conserve energy in several sectors, including buildings, transportation, and industry.
- Building capacity and energy efficiency projects are supported by the Energy Efficiency and Conservation Programme.

D. Vietnam: The country has implemented many energy efficiency efforts, including:

- National Energy Efficiency Programme, which promotes energy-saving measures for public and industrial sectors.
- Energy Efficiency and Conservation Law creates energy labelling regulations and standards for energy-related equipment.
- Energy-efficient building designs are encouraged by the Green Building Certification programme.

E. Philippines: The country has enacted several energy-efficient laws, such as the Energy Efficiency and Conservation Act, to encourage energy-efficient technologies in buildings, industry, and transportation.

- The Energy Efficiency and Conservation Roadmap delineates objectives related to energy efficiency and provides a framework for accomplishing them.
- Programme implementation for energy efficiency is managed by the Energy Efficiency and Conservation Division of the Department of Energy.

G. Brunei: The energy-efficiency initiatives include:

- The National Energy White Paper lays out plans for enhancing energy-efficiency and advancing renewable energy.
- Targets for lowering energy use and advancing energy-efficient technologies are outlined in the Energy Efficiency Master Plan.

H.Vietnam, Cambodia, Myanmar, and Laos increase energy efficiency with assistance from partners and international organisations. Developing energy-efficient building rules, encouraging energy-efficient lighting, and spreading awareness of energy conservation are some of its endeavours.

It is imperative to acknowledge that these synopses offer a broad outline of energy efficiency regulations in every nation, as illustrated in Figure 3. These policies' efficacy is contingent upon a number of variables, including public participation, enforcement, and implementation. It is advised to consult government and professional energy related organisations for the information.

4. Strategies for Energy Management Policy in ASEAN

Diversification of Energy Sources: Changing your energy sources is one of the main tactics. Risks to energy security are increased by ASEAN's significant reliance on fossil fuels. These hazards can be reduced by encouraging the development and use of renewable energy sources, such as solar, wind, hydro, and geothermal. Figure illustrates energy and greenhouse gas emissions in ASEAN.

Improved Energy Economy: Enhancing energy efficiency is yet another essential tactic. Energy-efficient measures can be implemented by member states in a variety of sectors, such as buildings, transportation, and industries. This entails using cutting-edge technologies, encouraging energy-efficient behaviours, and upholding energy efficiency regulations.

Cross-Border Energy Trade: Energy security and dependability can be improved by facilitating cross-border energy trade and linkages among member states. By constructing infrastructure for the transmission of natural gas and electricity, this tactic enables excess energy in one nation to satisfy demand in another. These kinds of partnerships can reduce energy waste and maximise the use of resources.

Technology Innovation and Research: It is imperative to allocate resources towards technology innovation and research. Grid stability and energy management can be enhanced by developments in smart grids, energy saving and storage, and decentralised energy systems. To speed up technical advancements, ASEAN member states might encourage cooperation between academic institutions and business sectors.

Policy Coordination and Harmonisation: The prosperity of the area depends on member governments coordinating their energy policies. Fair competition can be encouraged and level playing fields can be created by harmonising norms, laws, and incentives. To enable cross-border trading of renewable energy, governance rules and key indicator targets for renewable energy can be aligned.

Building Human capability through Training Programmes and Educational Initiatives: Effective policy implementation depends on raising public awareness and fostering human capability. Increasing public knowledge of sustainable methods and energy saving can also encourage behavioural changes and advance energy management objectives.

The solutions presented in this paper provide a way forward for a sustainable, and safe energy, based on the energy issues that ASEAN is currently confronting. Through the adoption of strategies such as energy source diversification, efficiency enhancements, cross-border cooperation, technological innovation, and policy coordination, ASEAN can steer clear of obstacles to achieving its energy objectives and simultaneously support worldwide endeavours to tackle climate change and promote sustainable development [5].

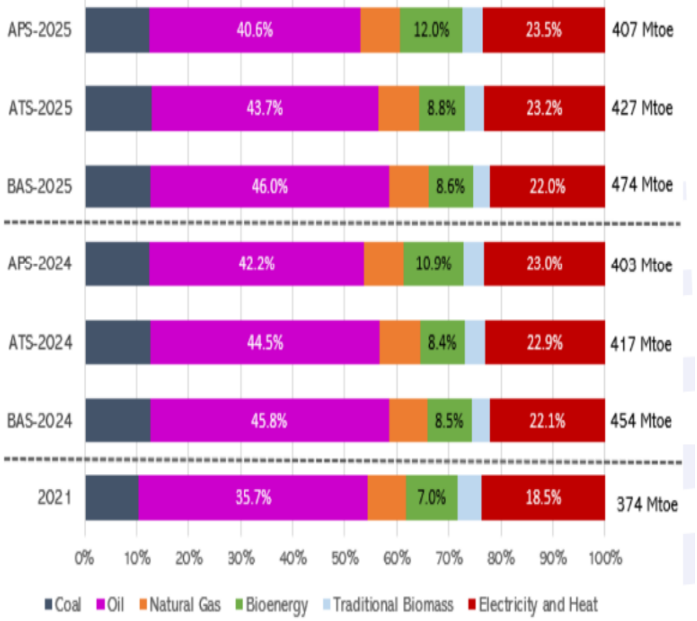


Figure 2: ASEAN energy demand 2024-2025 projection by fuel

Table I: Energy related pledges by committed by ASEAN countries

Pledge's Name	Targets	Countries
Global Renewable and Energy Efficiency pledge	Tripling the global's renewable energy generation capacity to at least 11,000 GW by 2030 and doubling the global average annual rate of energy efficiency improvement from 2% to over 4% per year until 2030	Brunei Darussalam, Malaysia, Singapore, Thailand
Global cooling pledge	Reduce cooling -related emission by a minimum of 68%to 2022 levels by 2050.	Brunei Darussalam, Cambodia, Singapore, Thailand, Vietnam
Declaration by Hydrogen and Derivatives	Mutual recognition of certification for renewable and hydrogen	Malaysia, Singapore

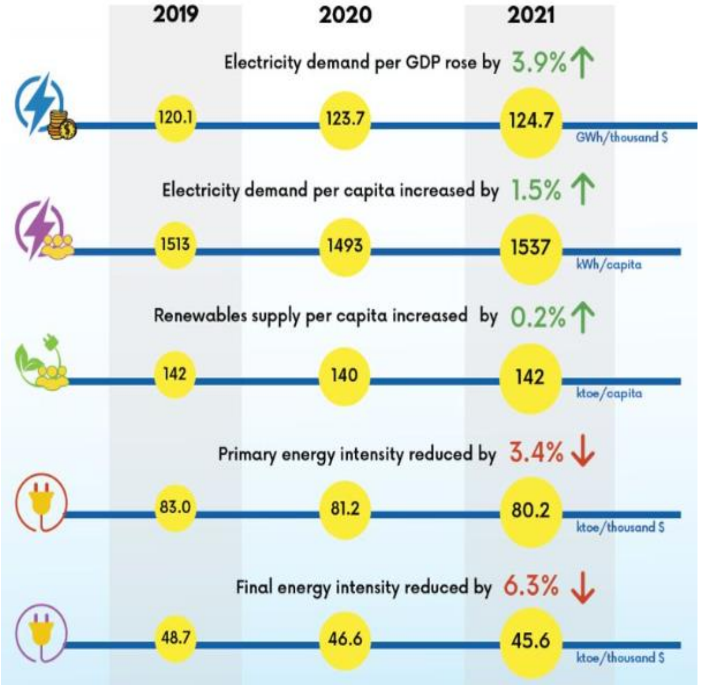


Figure 3: Primary energy supply, final energy consumption, electricity, renewable energy, energy-gender, and other energy-related indicators.

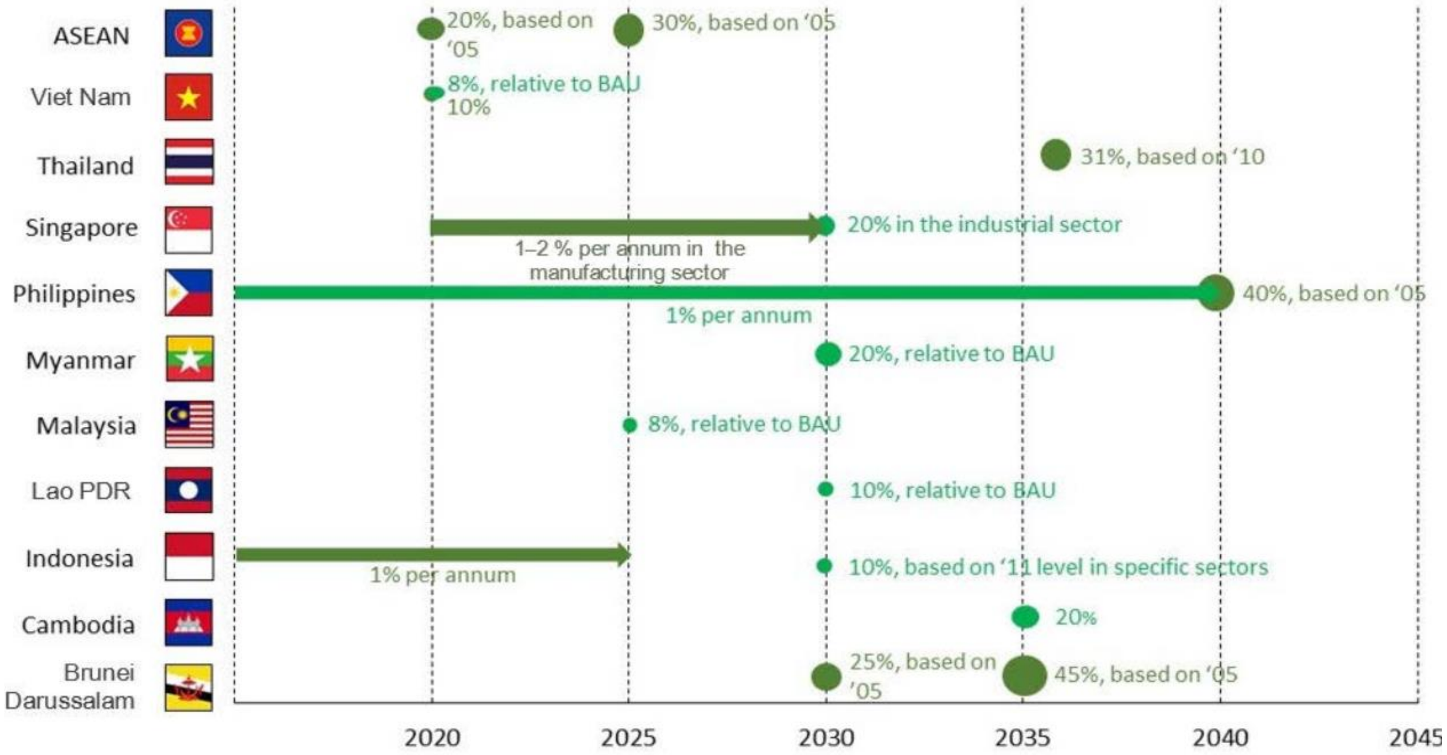


Figure 4: The ASEAN energy intensity (EI) and per capita total final energy consumption (TFC) [5]

5. ASEAN energy efficiency and policy Collaboration

Future prospects for efficient energy management in ASEAN are quite promising [6]. The region is witnessing economic expansion, urbanisation, and rising energy consumption. Consequently, there are multiple opportunities where energy efficiency and management initiatives might provide significant advantages (Table II):

A. The expansion of renewable energy: The ASEAN nations have an abundance of geothermal, hydro, wind, and solar energy resources. Increased utilisation of these resources offers a significant chance to improve energy security, lessen dependency on fuels, and reduce the climate change impacts [7, 8]. Infrastructure for renewable energy on wind turbines and solar energy, can draw investments, diversify the energy mix, and create jobs.

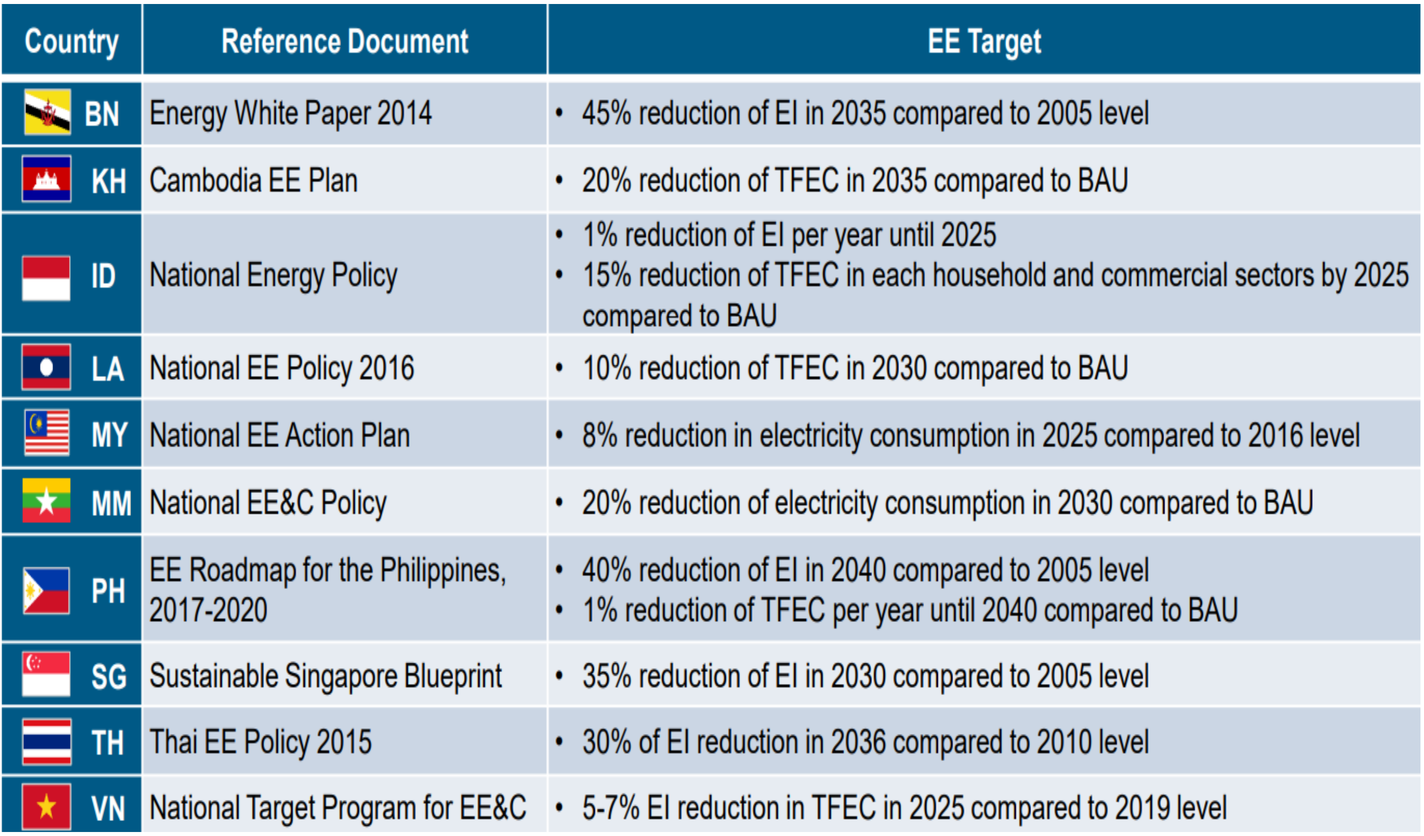
B. Smart Grid Implementation: ASEAN's energy management could undergo a radical change with the implementation of smart grid technologies. Demand response programmes, improved grid stability, optimal energy flow, and real-time monitoring and management of energy distribution are all made possible by smart grids [9]. Smart networks can decrease transmission and distribution losses and increase energy efficiency by incorporating renewable energy sources and strengthening system resilience.

C. Energy Efficiency in Buildings: An important portion of ASEAN's energy consumption comes from the building sector. Significant energy savings can be achieved by enforcing energy-efficient building rules, encouraging green building designs, and implementing technology like energy-efficient heating, cooling, and lighting systems [10]. Another big opportunity is to retrofit existing buildings with energy-efficient systems, as shown in Figure 4.

D. Industrial Energy Efficiency: ASEAN's energy-intensive sectors stand to gain from increased energy efficiency. Energy consumption and production costs can be decreased by using waste heat recovery, energy-efficient machinery, advanced manufacturing techniques, and process optimization [11]. Governments can use legislation and capacity-building initiatives to encourage industries to embrace best practices and technologies.

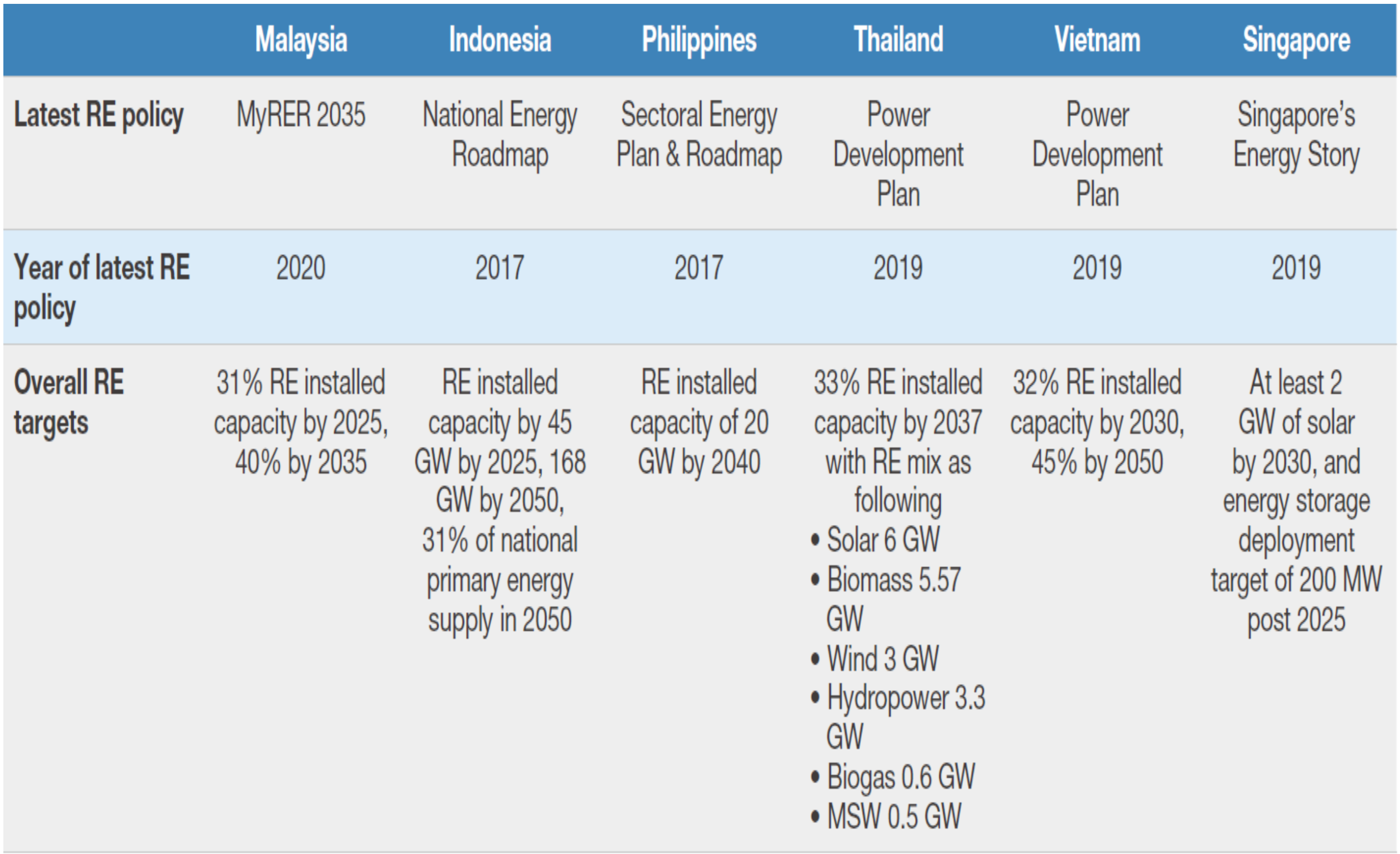
E. Electric Mobility and Transportation: The need for transportation grows along with urbanization [12]. One way to cut greenhouse gas emissions and lessen reliance on imported fossil fuels is to support public transit, establish infrastructure for EV charging, and promote electric vehicles (EVs). A sustainable transportation industry may be facilitated by the combination of renewable energy sources with electric mobility.

F. Energy Storage Solutions: The intermittency issues with renewable energy sources can be resolved by integrating energy storage technologies like batteries and pumped hydro storage [13]. In the end, energy storage contributes to a more dependable and resilient energy system by improving grid stability, facilitating the integration of renewable energy sources, and supporting demand-side management initiatives.



Country	Reference Document	EE Target
BN	Energy White Paper 2014	• 45% reduction of EI in 2035 compared to 2005 level
KH	Cambodia EE Plan	• 20% reduction of TFEC in 2035 compared to BAU
ID	National Energy Policy	• 1% reduction of EI per year until 2025 • 15% reduction of TFEC in each household and commercial sectors by 2025 compared to BAU
LA	National EE Policy 2016	• 10% reduction of TFEC in 2030 compared to BAU
MY	National EE Action Plan	• 8% reduction in electricity consumption in 2025 compared to 2016 level
MM	National EE&C Policy	• 20% reduction of electricity consumption in 2030 compared to BAU
PH	EE Roadmap for the Philippines, 2017-2020	• 40% reduction of EI in 2040 compared to 2005 level • 1% reduction of TFEC per year until 2040 compared to BAU
SG	Sustainable Singapore Blueprint	• 35% reduction of EI in 2030 compared to 2005 level
TH	Thai EE Policy 2015	• 30% of EI reduction in 2036 compared to 2010 level
VN	National Target Program for EE&C	• 5-7% EI reduction in TFEC in 2025 compared to 2019 level

Figure 5: ASEAN National Target on Energy Efficiency



	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 6: Overview of key ASEAN countries' renewable energy share targets [1]

G. Cross-Border Energy commerce: By connecting the ASEAN nations, cross-border energy commerce may improve energy security and maximise the use of energy resources [14]. Building cross-border transmission lines collaboratively can facilitate the pooling of excess energy and act as a safety net against supply interruptions as shown in Figure 6.

H. Energy and Digitalization Data Analytics: In energy management, digital technology and data analytics have the potential to revolutionise the field [15]. Adoption of sensors, Internet of Things (IoT) devices, and data analytics platforms can facilitate data-driven decision-making for increased energy efficiency, predictive repair of equipment, and real-time monitoring of energy consumption [16].

I. Green Finance and Investment: As green finance methods proliferate, funds for sustainable energy initiatives may be drawn to them. ASEAN nations may expedite the shift towards a sustainable and low-carbon energy future by endorsing investments in clean technology, energy-efficient infrastructure, and renewable energy.

J. International Collaboration and information Exchange: To speed up efforts to improve energy management and efficiency, ASEAN nations can take use of international partnerships and information exchange. Countries can adopt successful techniques and benefit from one other's experiences by exchanging best practices, lessons learned, and successful case studies [16, 17].

In general, there is a great deal of promise for improved energy security, job creation, economic growth, and environmental preservation in the ASEAN countries' future energy management and efficiency [18, 19]. Through deliberate utilisation of these opportunities, ASEAN nations can set the stage for a future in energy that is both sustainable and affluent [20, 21].

Table II: New and upcoming ASEAN energy policies based on 2023 regulations

Country	New Policy and Updates Announced in 2023
Brunei Darussalam	● Brunei Darussalam National Council on Climate Change (BNCCC) requires all greenhouse gas (GHG) emissions emitted by private and public sector facilities to be reported quarterly and annually. ● Brunei committed to cutting 20% of emissions compared to the business-as-usual scenario and moving towards net zero in 2050 through energy transition and forest conservation as stated under its 2030 Nationally Determined Contribution (NDC). ● The government plans to update their energy intensity reduction in 2024.
Cambodia	● Launched Power Development Master plan (PDP) 2022-2024, includes demand forecasts, generation expansion and a transmission and distribution plan. ● Increase renewable energy (RE) share and reduce fossil fuel energy share by 2040. ● Aims for a 21% coal power share of the total energy mix by 2030, down from an initially expected 40% in 2040. ● Cambodian government approved five new renewable projects that would generate 520 MW for the national power grid and aim to reduce CO2 emissions. ● Hydro and solar power generations spread throughout Cambodia, supporting the new PDP's targets of a capacity of 3,155 MW and 3,000 MW by 2040.
Indonesia	● Indonesia Minster of Energy and Mineral Resources (MEMR) issued Regulation Number 2 and the implementation of Carbon Capture and Storage (CCS) and Carbon Capture, Utilization and Storage (CCUS) in Upstream Oil and Gas Business Activities (MEMR Reg 2/2023). ● The MEMR Reg 2/2023 regulation covers technical, monetisation, operational, monitoring and measurement, reporting and verification (MRV) requirements, safety and environment and closure of CCS/CCUS activities. ● MMER Regulation Number 2/2024 to encourage rooftop solar by removing limits on capacity and increasing rooftop solar quota. ● Update Government Regulation Number 79 of 2014 concerning the National Energy Policy (NEP), targets and policies for energy and emissions in Indonesia for the period 2023-2060. ● Adjusted RE target from 23% to 17-19% by 2025.
Lao PDR	● In 2023, expanded RE generation as more clean emission technology is being implemented. ● National strategies on utilising hydrogen and ammonia for clean energy are being created.
Malaysia	● Launched policies in 2023 under National Energy Transition Roadmap (NETR). ● Low Carbon National Aspiration 2040 (LCNA 2040): set targets for energy transformation, reduce carbon emissions and lower coal power plants, increase RE power share, increase EE, adopts electric vehicles, increase the usage of public transport, increase carbon footprint tracking and sustainability reporting. ● Increase RE capacity from 40% in 2040 t o70% by 2050. More solar generation for government buildings and more RE trade with neighbouring countries.
Myanmar	● Ministry of Planning and Finance has exempted import taxes for solar generation technology. ● Incentives to increase energy investments in Myanmar.
Philippines	● Launched 2023 National Energy Efficiency and Conservation Plan (NEECP) and roadmap for the 2023-2050 period, aims for at least 30% emission reduction in the residential sector and 28% in utilities. ● Launched Fuel Conservation and Efficiency in Road Transport (FCERT) programs for higher fuel efficiency and electric vehicles (EVs)
Singapore	● Launched new emission standards for fossil-fuel- powered generation, to have at least 30% hydrogen ready to be used.

Thailand	● National Energy Policy Committee approved additional procurement of RE for 2022-2030. Increase the supply of RE, wind and solar generation in Thailand. ● Electricity Generating Authority of Thailand (EGAT) implemented a green tariff sandbox trial in 2023 for consumers to purchase RE easily. Full implementation in 2024. ● More EVs are being implemented on public transport. ● National Electric Vehicle Policy Committee extended the import fee exemption until the end of 2025, to attract domestic EV production in Thailand. ● Implemented Carbon Border Adjustment Mechanism (CBAM) certification in collaboration with the EU, to track and price carbon emissions for products to be able to be imported into the EU, effective from 2023 to 2025.
Vietnam	● Issued Directive No.20/CT-TTg to increase efforts for EE by reducing energy usage and using more energy efficient hardware. ● Vietnam’s government approved National Energy Master Plan (NEMP) 2021- 2030m to achieve energy security, reduce carbon emission, target to reach net zero by 2050. ● Aspired to export RE by 2030, for 5000 to 10000 MW. ● Green hydrogen production is expected to increase to 200,000 tonnes annually by 2030 and 20 million tonnes by 3million tonnes in 2050.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] "ASEAN Energy in 2024: Key Insights about ASEAN Energy Landscape and Predictions in 2024," *ASEAN Centre for Energy,* 2024. URL: aseanenergy.org. (Accessed on 9 January 2024).

[2] ASEANPOST, “5 Energy Companies to Look out for in ASEAN,” *ASEANPOST,* (Accessed on 9 January 2024). URL: https://theaseanpost.com/article/5-energy-companies-look-out-asean.

[3] "Energy: AEDS (ASEAN), IEA statistics (EU & Asia Pacific) Economics & Demographic: AEDS (ASEAN), APEC Statistics (Asia Pacific), WDI (EU)," (Accessed by July 2024).

[4] A.J.T.A.P. Gnanasagaran, “Renewable Energy Cooperation in ASEAN,” *The ASEAN Post,* 2019. URL: https://theaseanpost.com/article/renewable-energy-cooperation-asean. (Accessed on 1 January 2024).

[5] Y. Liu, R. Noor, “Energy Efficiency in ASEAN: Trends and Financing Schemes,” *ADBI Working Paper Series,* No. 1196, *Asian Development Bank Institute (ADBI),* Tokyo, 2020.

[6] W.Y. Leong, R. Kumar, “5G Intelligent Transportation Systems for Smart Cities,” *In Convergence of IoT, Blockchain, and Computational Intelligence in Smart Cities,* Edited by R. Kumar, V. Jain, W.Y. Leong, S. Teyarachakul, 1st Edition, *CRC Press,* 2023.

[7] W.Y. Leong, J.H. Chuah, B.T. Tee, *The Nine Pillars of Technologies for Industry 4.0, Institution of Engineering and Technology,* 2020.

[8] W.Y. Leong, *Human Machine Collaboration and Interaction for Smart Manufacturing: Automation, Robotics, Sensing, Artificial Intelligence, 5G, IoTs and Blockchain, Institution of Engineering and Technology,* Stevenage, United Kingdom, ISBN 1839534141, 2022.

[9] U. Mehrotra, W.Y. Leong, "NSEEAR: An Energy Adaptive Routing Protocol for Heterogeneous Wireless Sensor Networks," *2009 35th Annual Conference of IEEE Industrial Electronics,* Porto, Portugal, pp. 2647–2652, 2009, doi:10.1109/IECON.2009.5415255.

[10] W.Y. Leong, Y.Z. Leong, W.S. Leong, “Human-Machine Interaction in the Electric Vehicle Battery Industry,” *2024 10th International Conference on Applied System Innovation (ICASI), IEEE,* pp. 69–71, 2024.

[11] W.Y. Leong, Y.Z. Leong, W.S. Leong, "Green Building Initiatives in ASEAN Countries," *2023 Asia Meeting on Environment and Electrical Engineering (EEE-AM),* Hanoi, Vietnam, pp. 1–6, 2023.

[12] W.Y. Leong, *Medical Equipment Engineering: Design, Manufacture and Applications (Healthcare Technologies), Institution of Engineering and Technology,* 2023.

[13] W. Lee, W. Liu, P.H. Chong, B.L. Tay, W.Y. Leong, “Design of Applications on Ultra-Wideband Real-Time Locating System,” *2009 IEEE ASME International Conference on Advanced Intelligent Mechatronics,* 2009.

[14] W. Liu, E. Lupito, Y.L. Sum, B. Tay, W.Y. Leong, "Ultra Wideband Antenna for Real-Time Location System Application," *2009 35th Annual Conference of IEEE Industrial Electronics,* Porto, Portugal, pp. 2738–2742, 2009, doi:10.1109/IECON.2009.5415424.

[15] W.Y. Leong, J. Homer, "Implementing Nonlinear Algorithm in Multimicrophone Signal Processing," *2005 IEEE Workshop on Machine Learning for Signal Processing,* Mystic, CT, USA, pp. 33–39, 2005, doi:10.1109/MLSP.2005.1532870.

[16] W.Y. Leong, "Digital Technology for ASEAN Energy," *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT),* Kollam, India, pp. 1480–1486, 2023, doi:10.1109/ICCPCT58313.2023.10244806.

[17] W.Y. Leong, L.S. Heng, Y.Z. Leong, “Smart City Initiatives in Malaysia and Southeast Asia,” *Proceedings of International Conference on Renewable Power Generation,* Shanghai, China, pp. 1143–1149, 2023.

[18] W.Y. Leong, L.S. Heng, Y.Z. Leong, “Malaysia Renewable Energy Policy and Its Impact on Regional Countries,” *Proceedings of International Conference on Renewable Power Generation,* Shanghai, China, pp. 7–13, 2023.

[19] R. Kumar, V. Jain, W.Y. Leong, S. Teyarachakul, *Convergence of IoT, Blockchain, and Computational Intelligence in Smart Cities, CRC Press,* 2023.

[20] A. Arshad, M.A.M. Yajid, M. Daroonparvar, “Effect of Laser-Glazed Treatment on Thermal Cyclic Behavior of Plasma-Sprayed Lanthanum Zirconate/Yttria-Stabilized Zirconia Double Ceramic Layered on NiCoCrAlYTa-coated Inconel,” *Journal of Thermal Spray Technology,* 2023, doi:10.1007/s11666-023-01662-7.

[21] R. Kumar, A.K. Kapil, V. Athavale, W.Y. Leong, A. Touzene, “The Catalyst for Clean and Green Energy Using Blockchain Technology,” *In Modeling for Sustainable Development: A Multidisciplinary Approach,* Nova Science Publishers, Inc, pp. 23–39, 2023.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 9, No. 4, 110-118 (2024)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Assistive System for Collaborative Assembly Task using Augmented Reality

Woratida Sawangnamwong, Siam Charoenseang

Institute of Field Robotics, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand

ARTICLE INFO

Article history:
Received: 24 June, 2024
Revised: 13 August,2024
Accepted: 14 August, 2024
Online: 27 August, 2024

Keywords:
Augmented reality
STEM education
Education robot

ABSTRACT

Augmented reality (AR) technology has been increasingly used in developing teaching materials with the aim of sparking more interest in technology (T) and engineering (E) among students in STEM education. In the proposed system, AR is integrated with an educational robot controlled by a KidBright microcontroller board, developed by the Educational Technology research team (EDT) at the National Electronics and Computer Technology Center in Thailand. Moreover, the KidBright program has been implemented over 2,200 Thai schools. To maximize the benefits of the KidBright program, the Assistive System for Collaborative Assembly Task using Augmented Reality (ASCAT-AR) was created with the objective of enabling students to learn and collaborate in assembling robots. Students will work in pairs to assemble robots using the system and learn about mechanics, sensors, and 3D-printed parts. The students were divided into two groups: Group A read the manual and assembled the robot independently, while Group B used the ASCAT-AR system. In addition, AR applications offer smooth graphic rendering at 44-60 frames per second. Evaluation result showed that Group B students had a higher average success rate than average success rate of Group A students. The results showed that users of the ASCAT-AR system were more motivated in learning and obtained more knowledge about robot technology and programming.

1. Introduction

STEM education is a method of teaching that focuses on science, technology, engineering, and mathematics. Students receive training and preparation for the necessary 21st-century skills needed for success in the modern world [1]. Based on Bloom's Taxonomy [2], the related skills are classified into three categories: cognitive skills as shown in Table 1, such as critical thinking and problem-solving; social and emotional skills, such as communication and collaboration; and technological skills, such as the ability to use digital tools and platforms [3].

Previous works have revealed gaps in the effectiveness of STEM education in preparing students. One significant gap is the shortage of teachers specializing in STEM fields. A case study published in the Journal of Science Education and Technology highlights the challenges faced by many schools, particularly those in low-income neighborhoods, in providing STEM instruction due to difficulties in recruiting qualified teachers [4], [5]. Another challenge lies in the insufficient professional development opportunities for STEM teachers, as some educators lack the essential skills and knowledge needed to seamlessly incorporate STEM teaching and learning [6]. One such gap is the absence of standards and frameworks for developing and implementing MR teaching tools in STEM education. Although there is a growing interest in using MR for educational purposes, [7] a study published in the Journal of Science Education and Technology indicated a lack of sufficient guidelines and frameworks to assist educators in creating and utilizing MR resources effectively [8]. Table 2 show a comparison of the technologies.

Research published in the Journal of Science Education and Technology highlights the scarcity of mixed reality (MR) integration in STEM curricula. Rather than fully integrating mixed reality (MR) into the curriculum, some STEM instructors use it sporadically, which may reduce its effectiveness in teaching [9], [10]. Thailand has placed significant emphasis on STEM education and is committed to developing a more skilled and innovative workforce [11] The STEM education curriculum has been previously introduced and studied in Thailand[12], with initiatives focusing on curriculum development, digital media production, implementation in classrooms, and teacher training [13] The challenge is that students demonstrate low engagement in STEM disciplines. This lack of interest can hinder the effectiveness of STEM education and the development of a skilled future

[*] Corresponding Author: Siam Charoenseang, King Mongkut's University of Technology Thonburi, siam.cha@kmutt.ac.th

https://dx.doi.org/10.25046/aj090412

workforce [14]. Recently, robotics competitions have been organized in Thailand to motivate students' interest and creativity in robotics. In line with these efforts, schools have developed STEM education curricula that allow students to engage with technology. Although there is research on the design and implementation of MR learning games for robot assembly, there are still gaps in the teaching of technology and engineering subjects in STEM education. To enhance skills in robotics technology, this research project has developed an Assistive System for Collaborative Assembly Task using Augmented Reality (ASCAT-AR). This proposed system enables students to learn about the components and begin assembling a robot. The project implements augmented reality (AR) technology to captivate students' interests and relate their learning to future career opportunities.

Table 1: The Revised Taxonomy (2001) [4]

1.Remember	This level is about recalling information, such as facts, definitions, and concepts.
2.Understand	This level is about understanding the meaning of information, such as being able to explain it in your own words or apply it to new situations.
3.Apply	This level is about using information to solve problems or complete tasks.
4.Analyze	This level is about breaking down information into its component parts and understanding how they relate to each other.
5.Evaluate	This level is about making judgments about the value or worth of information.
6.Create	This level is about putting information together in new and original ways.

Table 2: XR Technology Comparison [9]

Features	AR	VR	MR
Definition	Limited interaction with virtual objects	Natural interaction with virtual objects	Natural interaction with both real and virtual objects
Hardware	Headset or smartphone	Headset required	Headset required
Application s	Navigation, wayfinding, product visualization , gaming	Gaming, entertainment , education, training	Gaming, entertainment, design, manufacturing , education, training

This research aims to develop an innovative educational tool to assist students in learning technology and engineering concepts within STEM education. AR technology is utilized within this system, and the educational robot is designed to be interfaced with Microsoft HoloLens 2 devices.

2. Proposed System

This research aims to create an innovative educational tool that can motivate students to understand STEM contents through robot assembly and control. Augmented reality (AR) technology is incorporated into this system to enhance the learning experience. The system overview, demonstrated in Figure 1, shows how users can use hand gestures to interact with a 3D model, manipulate its motion, and access information. Once the robot assembly is completed, the Microsoft HoloLens 2 provides a user interface for controlling the robot.

2.1. System Overview

The system overview depicted in Figure 1 illustrates that users can assemble an educational robot and utilize hand gesture recognition to rotate, move, zoom in, and zoom out a 3D prototype. The display is viewed through the Microsoft HoloLens 2. Once the educational robot is fully assembled, the system will show model a user interface and a window for simple programming, which is used to control the educational robot.

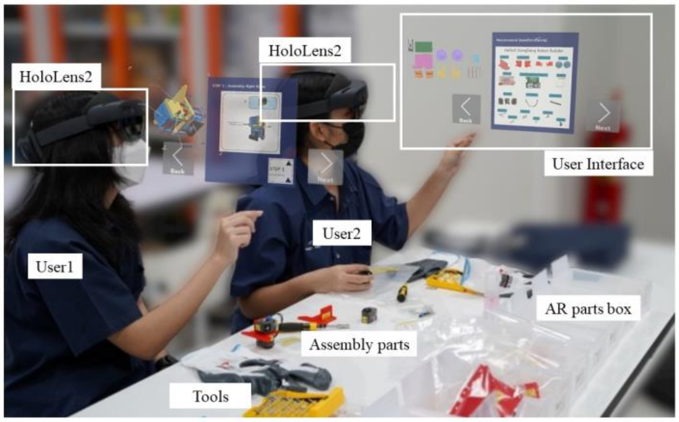


Figure 1: System Overview

2.2. System Configuration

The system operates through two primary processes: constructing the robot and controlling it. Figure 2 provides an overview of the involved steps. Initially, students are required to assemble the components of the educational robot. They can then write basic code to manage the robot using Microsoft HoloLens 2. Communication with the robot is conducted via a MQTT protocol. All instructions for building and controlling the robot are displayed directly on the Microsoft HoloLens 2.

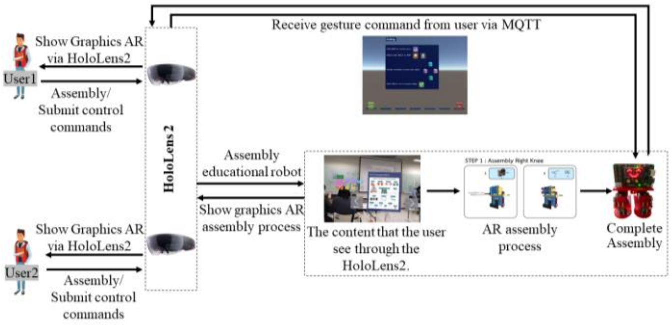
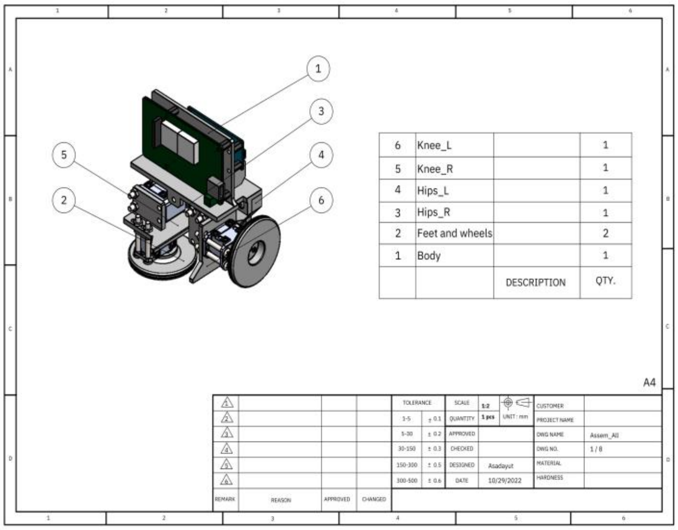

Receive gesture command from user via MQTT
Show Graphics AR via HoloLens2
User1
Assembly/ Submit control commands
HoloLens 2
Assembly educational robot
Show graphics AR assembly process
The content that the user see through the HoloLens2.
AR assembly process
Complete Assembly
Show Graphics AR via HoloLens2
User2
Assembly/ Submit control commands

Figure 2: System Configuration

2.3. Mechanical Design and Implementations

The Otto platform was effectively used in a robotics engineering class at MIT, demonstrating its suitability for educational purposes [15]. The education robot is based on the Otto DIY Ninja robot, which is an open-source robot designed to teach programming, mechanics, electronics, design, the internet of things, and artificial intelligence [16]. It was selected as the foundation for the project because of its well-designed, affordable nature that is suitable for Thai schools. Several modifications were made to the Otto DIY Ninja design to align it more closely with Thai curriculum and teaching style. The robot is a low-cost, user-friendly, and educationally rich tool that can be used to teach various STEM concepts to students in Thailand.

The education robot in this research was designed using the SolidWorks program for 3D modeling robot parts. The robot has circular feet for walking and wheels for fast movement. The wheel was designed to allow the robot to walk on its feet or move on wheels. The robot's head is designed to accommodate the KidBright microcontroller board, OpenCM9.04 for operating the DYNAMIXEL XL-320 and a battery pack. The robot's legs can be folded and are designed to support two DYNAMIXEL XL-320s on each side. The component parts of the robot that will be used for 3D printing are shown in Figure 3. The foot part of the robot is shown in Figure 4, and the robot's head with the control board is shown in Figure 5.



6	Knee_L		1
5	Knee_R		1
4	Hips_L		1
3	Hips_R		1
2	Feet and wheels		2
1	Body		1
		DESCRIPTION	QTY.

TOLERANCE		SCALE	1:2	CUSTOMER	
1-5	± 0.1	QUANTITY	1 pcs	PROJECT NAME	
5-30	± 0.2	APPROVED		DWG NAME	Assem_All
30-150	± 0.3	CHECKED		DWG NO.	1 / 8
150-300	± 0.5	DESIGNED	Asadayut	MATERIAL	
300-500	± 0.6	DATE	10/29/2022	HARDNESS	

UNIT : mm

REMARK REASON APPROVED CHANGED

Figure 3: Robot design

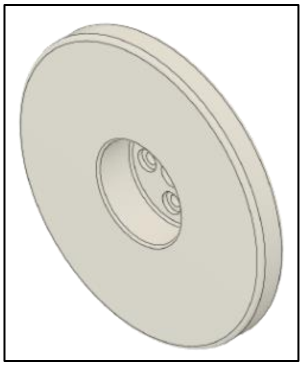


Figure 4: Wheel part

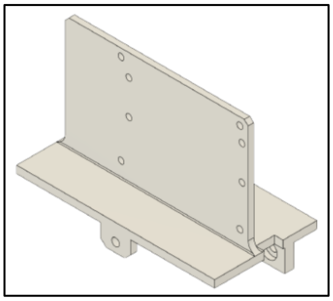


Figure 5: Head part

The assembled robot is set to the walk mode by default as shown in Figure 6. In this mode, the robot can move forward, backward, turning left, and turning right. Additionally, the robot can be switched to the wheel mode as shown in Figure 7, where it can move using its wheels. Even in wheel mode, the robot retains the ability to move forward, backward, turn left, and turn right.

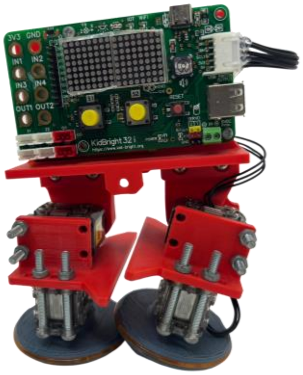


Figure 6: Walk mode configuration

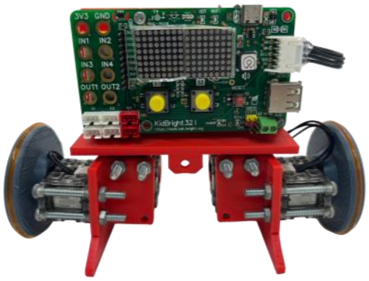


Figure 7: Wheel mode configuration

2.4. Electronic Design and Implementations

The KidBright microcontroller board was developed by the educational technology research team (EDT) at the National Electronics and Computer Technology Center (NECTEC), Thailand. It has been implemented over 2,200 schools across the country, promoting STEM learning on a wide scale. The board, based on the ESP32 microcontroller, enables device-to-device connectivity with internet of things feature and supports the integration of various external sensor modules via the I2C communication port. Figure 8 shows the connections of the KidBright board, the OpenCM9.04 controller, and the four DYNAMIXEL XL-320 servo motors. The KidBright microcontroller board offers a user-friendly interface, affordability, and IoT capabilities, making it an asset for educational robots, suitable for both in-class learning and remote-control applications.

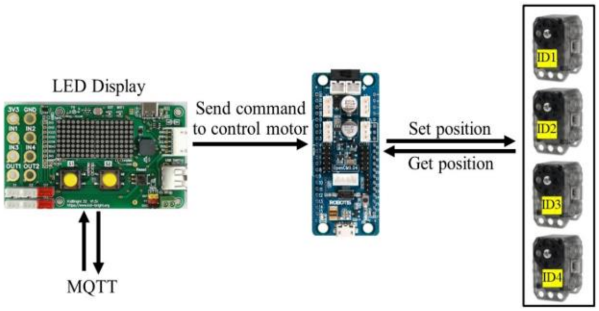


Figure 8: Microcontroller connection diagram

Figure 9 shows the OpenCM9.04 microcontroller and battery pack while Figure 10 displays the circuit battery charging design. The OpenCM9.04 is a microcontroller board used for controlling the DYNAMIXEL XL-320 motors. The battery pack supplies power to both the OpenCM9.04 and the DYNAMIXEL XL-320 motors. These components are essential for the education robot as they provide the necessary power for control the robot's movement.

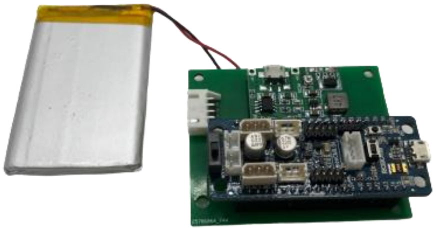


Figure 9: Battery charger and OpenCM 9.04 board

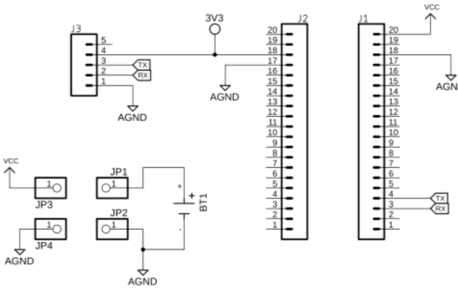


Figure 10: Battery charging and OpenCM 9.04 circuit design

2.5. Software Design and Implementations

2.5.1. Education Robot

The robot control program is divided into two parts. The KidBright program part: This part is used to as display graphics on screen and sends and receives data through the MQTT protocol. The KidBright board receives the robot's commands from the user. The OpenCM9.04 program part: This part is connected to the KidBright board using the I2C communication protocol. It is used to control the position of the DYNAMIXEL XL-320 motors to move to the received position by using the C++ programming language. The MQTT protocol is a lightweight messaging protocol that is well-suited for IoT applications. It is used to send and receive messages between devices over a network. The MQTT protocol is used in the education robot to send and receive commands between the KidBright board and the Microsoft HoloLens 2. The robot can be controlled by the user with two modes, which are wheel mode and walk mode as shown in Figure 11 and Figure 12, respectively.

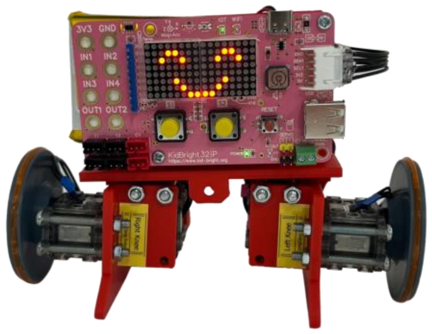


Figure 11: Wheel mode with display

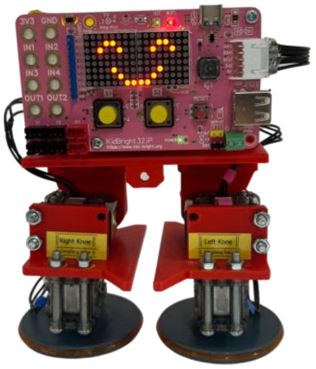


Figure 12: Walk mode with display

2.5.2. Robot Assembly Procedures

The assembly instructions for this research were modeled after LEGO's format, catering to a young audience. They were crafted to be intuitive, and accommodating users who may have no prior experience in robot assembly. These instructions utilize a 2D graphical format with distinct images and directional arrows to guide users through the assembly steps. Figure 13, the first page of the instructions, introduces the robot's parts along with a count of the pieces and components required for assembly. Subsequently, Figure 14 shows the detail of the initial step for the assembly, focusing on constructing the left knee, It is divided into two sub-steps. The guide follows a systematic structure, with each step clearly labeled and numbered. These instructions are a crucial component of the educational robot kit, facilitating the assembly process and enriching the user's understanding of the robot's various parts and components.

The 2D manual is implemented as a user interface (UI) on the display of Microsoft HoloLens 2. The UI on Microsoft HoloLens 2 is in a 3D format. 3D models can be scaled, rotated, or moved. These models can be animated to form an animation loop of the assembly process. Figure 15 shows Next and Back UI buttons for the next step or to go back to the previous step when the wrong assembly occurs. Assembly figure and description can be shown in Figure 16. The UI was developed using the Unity game engine.

The UI was designed to be easy to use and understand, even for users with no prior experience with 3D applications.

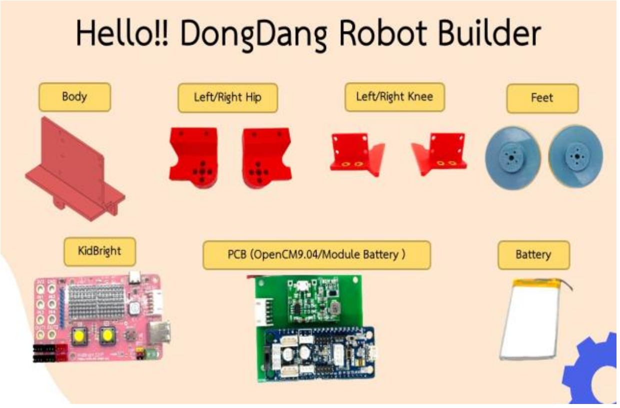

Hello!! DongDang Robot Builder
Body
Left/Right Hip
Left/Right Knee
Feet
KidBright
PCB (OpenCM9.04/Module Battery)
Battery

Figure 13: Introduction page of robot parts

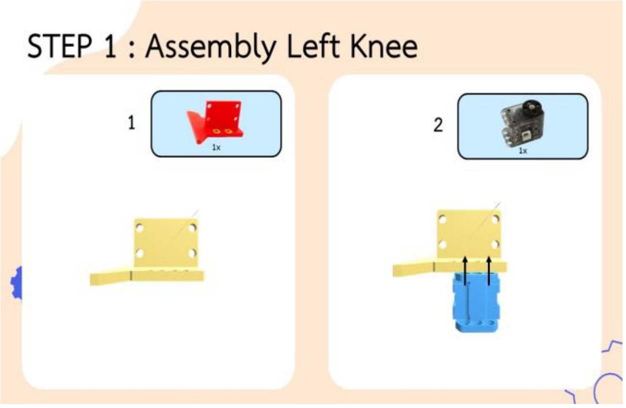


Figure 14: Assembly manual

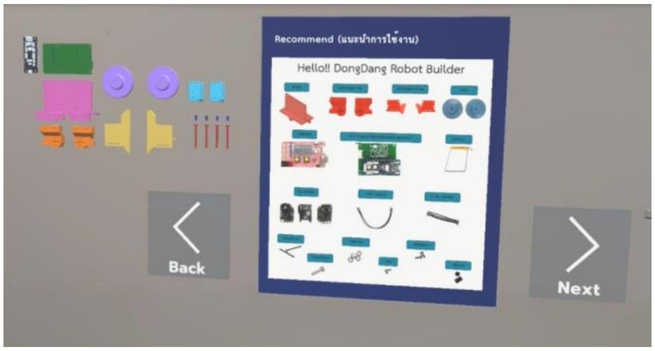


Figure 15: 3D introduction of robot parts via Microsoft HoloLens 2

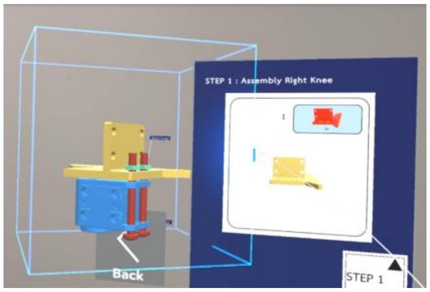


Figure 16: 3D guide for robot assembly via Microsoft HoloLens 2

3. System Implementations

3.1. Assembly System

Figure 17 shows the state diagram of ASCAT-AR system. First, the user launches the ASCAT-AR application. The application presents a window showing details of the modeled robot's components, as well as the models of necessary tools and equipment. Next, the user follows the on-screen guide to assemble the robot. This interactive guide will lead the user through the assembly process, from step one to step eight. AR highlights the exact location of each tool required for the current step, and enhances the ease of the assembly process. After completing the assembly, the application offers a choice of reassembling the robot or proceeding to program its controller. At this step, the user can verify each step of the assembly to ensure the completion. Once the assembly is confirmed to be completely correct, the application moves to a display window for robot control programming.

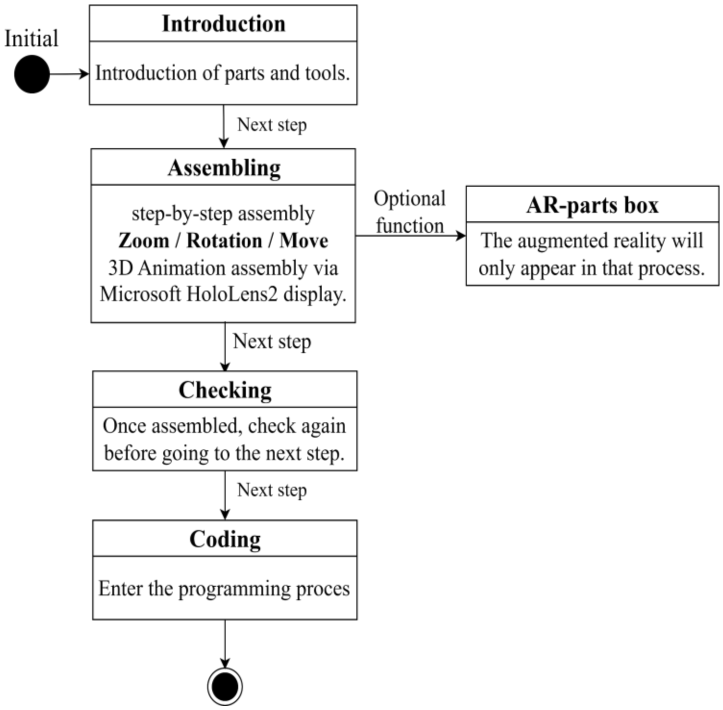

Initial
Introduction
Introduction of parts and tools.
Next step
Assembling
step-by-step assembly
Zoom / Rotation / Move
3D Animation assembly via
Microsoft HoloLens2 display.
Optional
function
AR-parts box
The augmented reality will
only appear in that process.
Next step
Checking
Once assembled, check again
before going to the next step.
Next step
Coding
Enter the programming proces

Figure 17: State diagram of proposed system

Figure 18 shows the sequences of assembly process. Each step must be completed before the next step can be proceeded. However, some robot modules can be assembled independently from the others.

3.2. Blockly System and Robot Control

Figure 19 shows the code using block or Blockly, an open-source visual programming language. Blockly allows users to write code using a visual block-based interface. Blockly is based on the prototype concept from Google’s Blockly [14] and is designed to be easy to use and intuitive. This makes Blockly an excellent tool for teaching programming concepts to children and other beginners. Blockly lets beginners build programs fast and easily with visual blocks. Blockly is also flexible and can be used to create many kinds of programs [17]. Overall, Blockly is a powerful and user-friendly visual programming language that has become a popular choice for both teaching and development.

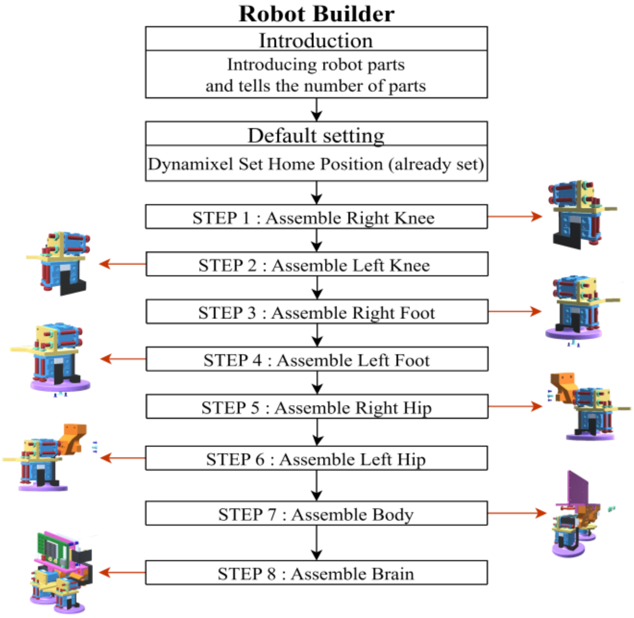

Robot Builder
Introduction
Introducing robot parts
and tells the number of parts
Default setting
Dynamixel Set Home Position (already set)
STEP 1 : Assemble Right Knee
STEP 2 : Assemble Left Knee
STEP 3 : Assemble Right Foot
STEP 4 : Assemble Left Foot
STEP 5 : Assemble Right Hip
STEP 6 : Assemble Left Hip
STEP 7 : Assemble Body
STEP 8 : Assemble Brain

Figure 18: Robot assembly steps

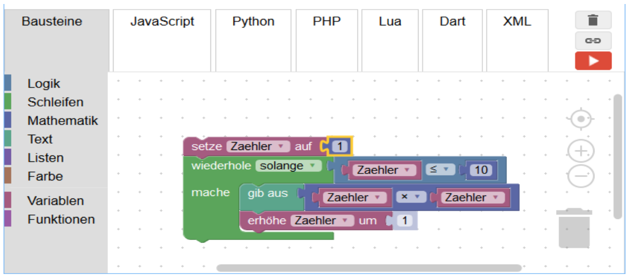
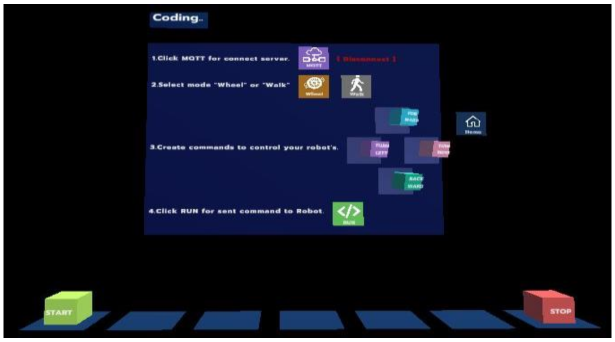


Figure 19: Google's Blockly demo [17]

The command set designed for ASCAT-AR system includes the following command buttons, which are MQTT connection button, robot transformation button, movement command buttons (front, back, left, and right), run the command button, as shown in Figure 20. These movement command buttons are used to control the movement of the robot. When the user presses a movement command button, it will create a command block in scene. When the command block appears, the user can pick it up and sort it into any available position. The programming window allows only 5 command blocks at one time as shown in Figure 21. Once the command is in the correct position, the user presses the run button to send the command to the robot. The instruction set is designed to be simple to use. This makes the instruction set a valuable tool for controlling the movement of the robot.


Coding..

1.Click MQTT for connect server.

2.Select mode "Wheel" or "Walk"

3.Create commands to control your robot's.

4.Click RUN for sent command to Robot.

Figure 20: Robot control interface

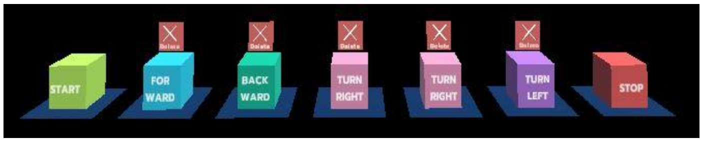


Figure 21: Samples of command blocks

The state diagram of robot control in Figure 22 describes the process of sending the robot command to the education robot. First, the user connects to the MQTT server by clicking the MQTT button on the screen. The user can select the working mode of the robot after the proposed application is connected with the MQTT server. There are two operational modes which are wheel mode and walk mode. In wheel mode, the robot moves like a car. In walk mode, the robot moves like a human walk. The user can select the Front, Back, Left, Right, and Turn commands to move the robot in the desired direction. Once the user completes the block coding, the Run button can be clicked to send the robot commands. As shown in Figure 23, the robot moves accordingly to the selected commands given by the user.

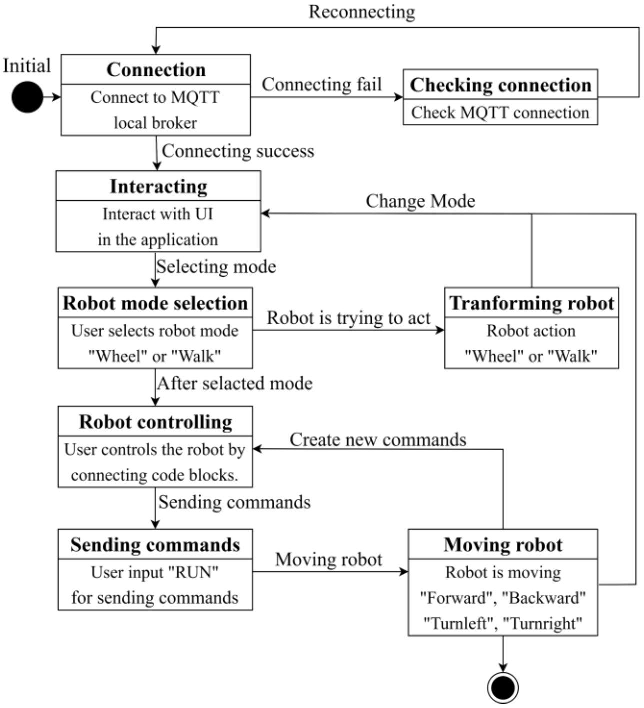


Figure 22: State diagram of Blockly-based robot control

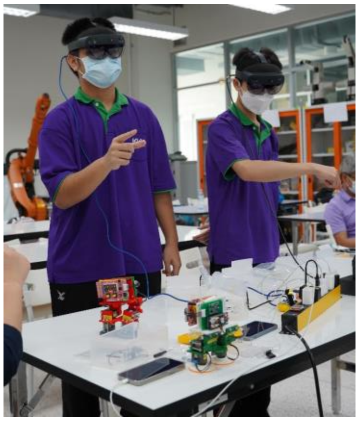


Figure 23: Users send robot commands to the real robot via HoloLens 2

4. Experimental Results

The experimental results of this research cover system performance, usability, and values for specific tasks. System performance shows rendering frame rate of the ASCAT-AR system program on Microsoft HoloLens 2. Usability tests cover the evaluation of user's satisfaction, ability to learn, and ease of learning. Values for specific tasks are evaluated from data records of the assembly times, success rates, and error rates.

4.1. Population and Requirement

The number of users in the system are 24 persons who are high school students in Thailand. The students were divided into two groups to compare between the uses of the ASCAT-AR system and the assembly manual.

4.2. Evaluation Tools

4.2.1. Robot assembly record form

This form aims to collect the amount of time that the students spent on robot assembly. It can record the time spent, the point of assembly failure, and the time it takes to fix the wrong position, and the number of errors.

4.2.2. Pre-questionnaire and post-questionnaire

The pre-questionnaire and post-questionnaire forms include general questions about user background, experience, and suggestions for improving the system, as well as questions about values for specific tasks and evaluation of satisfaction on proposed system.

4.2.3. Satisfaction questionnaire

The satisfaction form aims to collect feedback on user's satisfaction about the provided content and system performance of the ASCAT-AR system.

4.2.4. Comparative assessment form

This assessment form was applied for both experimental sets to compare motivation and system differences.

4.3. Procedure of Experiment

The experimental sets were established for Thai high school students aged 13-18 years who have some or no experience in engineering. The experimental sets provide the users about practical experience in robot assembly and the use of Microsoft HoloLens 2 headset.

Volunteers were divided into two groups to avoid the memorization of operation flow and provided AR contents. The ASCAT-AR system was tested with 12 students while the other 12 students used the assembly manual. There were pre- and post-questionnaires to collect the students' knowledge and interest about the robot and the opinions on the experiment.

In Figure 24, students, who have no experience about robot assembly, are divided into two groups. The experiment consists of the following steps:

Step 1: Students need to complete a preliminary questionnaire.

Step 2: Both groups were asked to assemble the robot with different tools. The aim of this exercise is to provide experience about engineering tools and AR technology for students. The expected learning outcome is that both groups of students can complete the robot assembly and control the robot's movement.

Step 3: After completion of robot assembly and robot control, each group of students was asked to answer a post questionnaire testing.

Step 4: Students, who worked with ASCAT-AR system, need to complete the satisfaction questionnaire.

During experiments were conducted, researchers also observed the students' behaviors and reactions and interviewed the students regarding their experiences on using the ASCAT-AR system and the instruction manual.

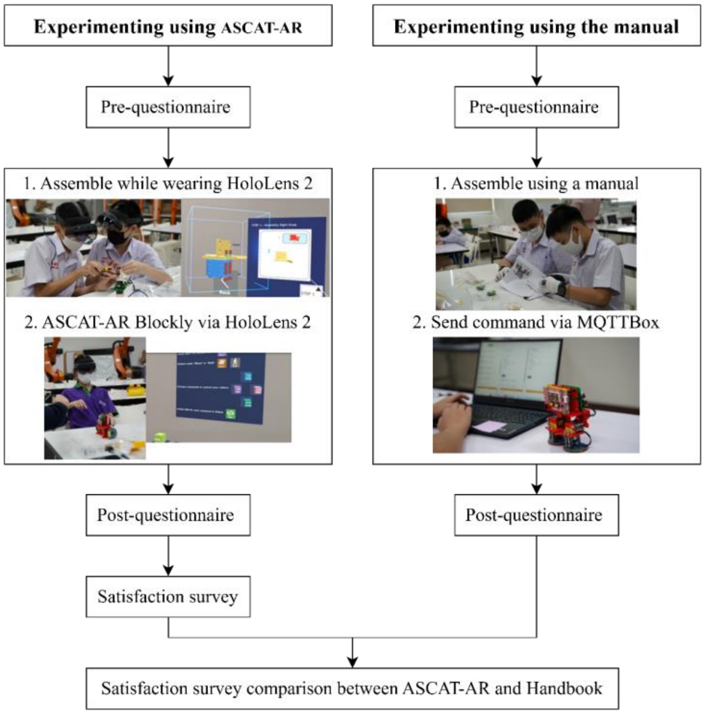

Experimenting using ASCAT-AR
Experimenting using the manual
Pre-questionnaire
Pre-questionnaire
1. Assemble while wearing HoloLens 2
1. Assemble using a manual
2. ASCAT-AR Blockly via HoloLens 2
2. Send command via MQTTBox
Post-questionnaire
Post-questionnaire
Satisfaction survey
Satisfaction survey comparison between ASCAT-AR and Handbook

Figure 24: Experiment process flow

4.4. Experimental Results

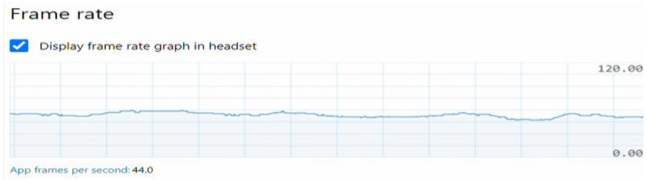
The study was designed to evaluate the effectiveness of the ASCAT-AR system in enhancing student learning in STEM education, particularly in robot assembly task. The participants were 24 high school students aged 13-18 from Thailand, who had no prior experience with robotics. The study investigated two learning tools, which are the ASCAT-AR system utilizing augmented reality guidance via the Microsoft HoloLens 2 and the traditional 2D printed manual.

Data were collected using a combination of observation forms, questionnaires, and system logs. During the assembly phase, completion times, error rates, and task success rates were recorded for each participant. Pre- and post-experiment questionnaires were conducted to assess all participants' knowledge about educational robot assembly and movement control, and their satisfaction on the learning process.

The collected data were analyzed using statistical methods to compare the performance of the two groups. Assembly times were averaged, and error rates were calculated as percentages of incorrect assembly steps.

4.4.1. System Performance

AR contents are rendered on the Microsoft HoloLens 2 devices with a reasonable frame rate. The system delivers a smooth and responsive graphics to enhance the user's experience. Figure 25 shows that the application can render at frame rates of 44-60 FPS with a resolution of 1440x936 pixels per eye.



Frame rate

☑ Display frame rate graph in headset

App frames per second: 44.0

Figure 25: Frame rates of ASCAT-AR system

4.4.2. Usability

The comparative evaluations from learners, who used the ASCAT-AR system and used instruction manual to assemble and control robot, were conducted to assess system benefits, ease of use, ease of learning, and user satisfaction. Each aspect was rated on a 5-point Likert's scale (1=lowest, 5=highest) by 24 participants.

Table 3: Results of satisfaction survey comparing between the uses of ASCAT-AR system and instruction manual

Learnability	**ASCAT-AR**	**Manual**
I can learn about more parts of the robot.	3.67	4.08
I can perform cooperative task in assembling robots.	4.42	4.75
I want to complete the final robot assembly.	4.83	4.75
I think it took a long time after using this learning tool.	3.58	3.5
Easiness of control		
I find it difficult, and I want to quit.	2.25	2.67
Satisfaction		
I am interested and would like to learn more.	4.17	4.58
How much motivation does the system provide for me to build a robot?	4.17	3.92

Table 3 shows the results of the satisfaction survey comparing the uses of ASCAT-AR system and instruction manual. Scores of learnability shows that the use of ASCAT-AR system can motivate and help the users to assemble robots more successfully than the use of instruction manual. Scores of easiness of control identifies that the use of ASCAT-AR system slightly more difficult than the use of instruction manual since users were just using Microsoft HoloLens 2 for the first time. Finally, users think that the ASCAT-AR system can motivate and help them to assemble and control robot.

4.4.3. Values for Specific Task

To collect the feedback from the experiments, the users were scheduled to perform the task within 2 hours for both of the uses of ASCAT-AR system and instruction manual. Table 4 shows the minimum, maximum, and average of user's robot assembly time for both cases.

Table 4: Time spent on task completion

Groups	Min time	Max time	Average time
ASCAT-AR	0:49:47	1:53:56	1:12:24
Manual	0:59:39	1:40:35	1:18:12

In Figure 26, task completion rate (TSR) is a performance chart used to measure usability and show the percentage of task completion in each step and the system effectiveness of helping users to achieve their goals. The success of the task can be influenced by factors such as user interaction with the interface, overall user experiences, and user motivation. Students using the ASCAT-AR system had a constant percentage success rate from assembly step 1 to final assembly step 8, ranging from 75% to 100%. The results show that students, who used ASCAT-AR system, had a better success rate during step 1 and step 2 than the ones who used the instruction manual with success rate of 33% in task 1.

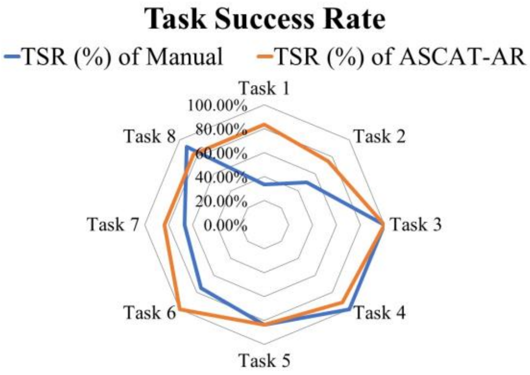


Figure 26: Robot assembly success rate chart

Comparing error rates, students using the ASCAT-AR system assisted robot assembly system had fewer errors than those assembling robots manually. The robot assembly assisted system had a prototype model that could be rotated 360 degrees for a detailed and accurate view of the assembly position, unlike the 2D figures in the manual.

5. Conclusions and Discussions

This research focuses on the development of an ASCAT-AR system using AR technology. Augmented Reality has been developed and applied in various industries, such as education, medicine, tourism, industry, entertainment, and etc [18]. AR technology has the potential to improve the learning experience. It can help the learner to better understand and remember educational content because it allows users to be engaged in learning, which may lead to generate creative teaching methods and better educational outcomes [19]. Currently, efforts and focus on promoting STEM curricula in Thailand have been found, and it has also been found that some teachers have limited knowledge of STEM education and lack of capability of how to integrate STEM into their teaching practices [20]. The proposed system provides AR content that is utilized to comply with the STEM curriculum integrating multiple various subjects. The goal is to equip students with knowledge and understanding of technology and engineering, enhance their collaborative skills, and foster their interests in applying this competency for future careers. The system specifically helps the development of collaborative skills by engaging students in a collaborative robot assembly task. The system's core hardware consists of the Microsoft HoloLens 2 device and an educational robot. The ASCAT-AR system displays 3D model data and animations to help students explore structures and components in augmented reality. The system is designed with a user-friendly interface that makes it accessible to non-engineering students. The application and the robot can interact in real time, enhancing learning and creating a robot control experience. The study found that students who used ASCAT-AR system were more motivated to complete the robot assembly than the ones who used the manual. Times spent on task completion of two student groups are slightly different. These results may depend on the learning ability of each student. In addition, visualization of 3D models from the ASCAT-AR system assists the user to know the positions of robot parts and sequence of robot assembly more clearly. This increases the success rate of robot assembly.

The evaluation covered all three aspects: system performance, usability, and values for specific tasks. The evaluation results demonstrated that the ASCAT-AR system can increase the user's interests in learning to gain more knowledge and skill about robot technology and programming. Additionally, the robot was found to be easy to control using the ASCAT-AR system. The proposed system not only aids in teaching the users how to program and control the robot but also enhances their understanding of technological concepts and fostering their creativity and collaboration skills.

Furthermore, the improved 3D object detection can improve the current system performance in order to help the user to assemble the robot more easily. The system can be applied to other learning subjects to give users a visualization of the concept.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We would like to acknowledge the funding from the Program Management Unit for Human Resources & Institutional Development, Research and Innovation (PMU-B)'s AI for All 1-2 projects and the Thailand Science Research and Innovation's Fundamental Fund for research financial supports, and the Human-Computer Interface Lab and Institute of Field Robotics for supporting the facilities and hardware used in this research.

References

[1] B.E. Penprase, "STEM Education for the 21st Century," *Springer International Publishing,* 2020, doi:10.1007/978-3-030-41633-1.

[2] P. Armstrong, "Bloom's Taxonomy," *Vanderbilt University Center for Teaching,* 2010.

[3] B. Bai, H. Song, "21st Century Skills Development through Inquiry-Based Learning: From Theory to Practice," *Asia Pacific Journal of Education,* **38**(4):584–586, 2018, doi:10.1080/02188791.2018.1452348.

[4] A.L.W., K.D.R., A.P.W., C.K.A., R. Mayer, P.P.R., J. Raths, W.M.C., "A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives," 2001.

[5] World Economic Forum, "Schools of the Future: Defining New Models of Education for the Fourth Industrial Revolution," *World Economic Forum Reports,* January 2020.

[6] S. Wachira, L. Deborah, "HANUSCIN and CHATREE FAIKHAMTA Perceptions of In-service Teachers toward Teaching STEM in Thailand," *Asia-Pacific Forum on Science Learning and Teaching,* **18**(2):1, 2017.

[7] J. Garzón, J. Pavón, S. Baldiris, "Systematic Review and Meta-analysis of Augmented Reality in Educational Settings," *Virtual Reality,* **23**(4):447–459, 2019, doi:10.1007/s10055-019-00379-9.

[8] C.E. Hughes, C.B. Stapleton, D.E. Hughes, E.M. Smith, "Mixed Reality in Education, Entertainment, and Training," *IEEE Computer Graphics and Applications,* **25**(6):24–30, 2005, doi:10.1109/MCG.2005.139.

[9] O.B., Gleb; I., "VR vs AR vs MR: Differences and Real-Life Applications," *RubyGarage,* 2020.

[10] T.L. Hutner, V. Sampson, L. Chu, C.L. Baze, R.H. Crawford, "A Case Study of Science Teachers' Goal Conflicts Arising when Integrating Engineering into Science Classes," *Science Education,* **106**(1):88–118, 2022, doi:10.1002/sce.21690.

[11] A. Koolnapadol, P. Nokkaew, P. Tuksino, "The Study of STEM Education Management Connecting the Context of Science Teachers in the School of Extension for Educational Opportunities in the Central Region of Thailand," *International Journal of Science and Innovative Technology,* **2**(June):121–130, 2019.

[12] F.N. Promboon, K.K. Finley, "The Evolution and Current Status of STEM Education in Thailand: Policy Directions and Recommendations," *STEM Education in the Nation: Policies, Practices, and Trends,* Springer Singapore, pp. 423–459, 2018, doi:10.1007/978-981-10-7857-6_17.

[13] S. Sutaphan, C. Yuenyong, "STEM Education Teaching Approach: Inquiry from the Context Based," *Journal of Physics: Conference Series,* **1340**(1):012003, 2019, doi:10.1088/1742-6596/1340/1/012003.

[14] Suriyabutr, J. Williams, "Integrated STEM Education in Thai Secondary Schools: Challenges and Addressing of Challenges," *Journal of Physics: Conference Series,* **1957**(1):012025, 2021, doi:10.1088/1742-6596/1957/1/012025.

[15] E.B. Olson, "Otto: A Low-Cost Robotics Platform for Research and Education," 2001.

[16] O. DIY, "Build Your Own Robot Like a Ninja," *Brno, Czech,* 2021.

[17] G.D. Group, "Try Blockly," *Google,* 2021.

[18] F. Eishita, K. Stanley, "The Impact on Player Experience in Augmented Reality Outdoor Games of Different Noise Models," *Entertainment Computing,* **27**:2018, doi: 10.1016/j.entcom.2018.04.006.

[19] M. Brizar, D. Kažović, "Potential Implementation of Augmented Reality Technology in Education," *2023 46th MIPRO ICT and Electronics Convention (MIPRO),* pp. 608–612, 2023, doi:10.23919/MIPRO57284.2023.10159865.

[20] S. Pitiporntapin, P. Chantara, W. Srikoom, P. Nuangchalerm, L.M. Hines, "Enhancing Thai In-service Teachers' Perceptions of STEM Education with Tablet-based Professional Development," *Asian Social Science,* **14**(10):13, 2018, doi:10.5539/ass.v14n10p13.