



Balkan Journal of Electrical & Computer Engineering

An International Peer Reviewed, Referred, Indexed and Open Access Journal

www.bajece.com

Vol : 13

No : 2

Year : 2025

ISSN : 2147 - 284X



It is abstracted and indexed in, Index Google Scholarship, the PSCR, Cross ref, DOAJ, Research Bible, Indian Open Access Journals (OAJ), Institutional Repositories (IR), J-Gate (Informatics India), Ulrich's, International Society of Universal Research in Sciences, DRJI, EyeSource, Cosmos Impact Factor, Cite Factor, SIS Scientific Indexing Service, IJIF, iijFactor. ULAKBİM-TR Dizin.

General Publication Director & Editor-in-Chief
Musa Yilmaz, University of California Riverside, US

Vice Editor
Hamidreza Nazaripouya, Oklahoma State University, US

Scientific Committee
Abhishek Shukla (India)
Abraham Lomi (Indonesia)
Aleksandar Georgiev (Bulgaria)
Arunas Lipnickas (Lithuania)
Audrius Senulis (Lithuania)
Belle R. Upadhyaya (USA)
Brijender Kahanwal (India)
Chandar Kumar Chanda (India)
Daniela Dzhonova-Atanasova (Bulgaria)
Deris Stiawan (Indonesia)
Emel Onal (Turkey)
Emine Ayaz (Turkey)
Enver Hatimi (Kosovo)
Ferhat Sahin (USA)
Gursel Alici (Australia)
Hakan Temeltaş (Turkey)
Ibrahim Akduman (Turkey)
Jan Izykowski (Poland)
Javier Bilbao Landatxe (Spain)
Jelena Dikun (Lithuania)
Karol Kyslan (Slovakia)
Kunihiko Nabeshima (Japan)
Lambros Ekonomou (Greece)
Lazhar Rahmani (Algerie)
Marcel Istrate (Romania)
Marija Eidukeviciute (Lithuania)
Milena Lazarova (Bulgaria)
Muhammad Hadi (Australia)
Muhamed Turkanović (Slovenia)
Mourad Houabes (Algerie)
Murari Mohan Saha (Sweden)
Nick Papanikolaou (Greece)
Okyay Kaynak (Turkey)
Osman Nuri Ucan (Turkey)
Ozgur E. Mustecaplioglu (Turkey)
Padmanaban Sanjeevikumar (India)
Ramazan Caglar (Turkey)
Rumen Popov (Bulgaria)
Tarek Bouktir (Algeria)
Sead Berberovic (Croatia)
Seta Bogosyan (USA)
Savvas G. Vassiliadis (Greece)
Suwarno (Indonesia)
Tulay Adali (USA)
Yogeshwarsing Calleecharan (Mauritius)
YangQuan Chen (USA)
Youcef Soufi (Algeria)

Aim & Scope

The journal publishes original papers in the extensive field of Electrical-Electronics and Computer engineering. It accepts contributions which are fundamental for the development of electrical engineering, computer engineering and its applications, including overlaps to physics. Manuscripts on both theoretical and experimental work are welcome. Review articles and letters to the editors are also included.

Application areas include (but are not limited to): Electrical & Electronics Engineering, Computer Engineering, Software Engineering, Biomedical Engineering, Electrical Power Engineering, Control Engineering, Signal and Image Processing, Communications & Networking, Sensors, Actuators, Remote Sensing, Consumer Electronics, Fiber-Optics, Radar and Sonar Systems, Artificial Intelligence and its applications, Expert Systems, Medical Imaging, Biomedical Analysis and its applications, Computer Vision, Pattern Recognition, Robotics, Industrial Automation.



ISSN: 2147- 284X
Vol: 13
No : 2
Year: June 2025

CONTENTS

Research Article

Eren Gündüzvar, Abdulsamet Kayık, Mehmet Ali Altuncu; Alzheimer's Disease Diagnosis in MRI Images Using Transfer Learning Methods: Evaluation of Different Model Performances,119–127

Research Article

İsmail Kırbaş, Ahmet Çifci; Leveraging SHAP for Interpretable Diabetes Prediction: A Study of Machine Learning Models on the Pima Indians Diabetes Dataset,128–139

Research Article

Süleyman Dal, Necmettin Sezgin; Heart Attack Classification with a Machine Learning Approach Based on the Random Forest Algorithm,140–147

Research Article

Şilan Fidan Vural, Nida Kumbasar; Comparison of VT-based and CNN-based Models on Teeth Segmentation,148–156

Research Article

Cemanur Aydinlal, Gülşah Yıldız Altıntaş; Breast Cancer Detectability and Tumor Differentiation based on Microwave Dielectric Property Changes with Reverse Time Migration,157–163

Research Article

Aykut Satici; Control Through Contact using Mixture of Deep Neural-Net Experts,164–173

Research Article

Ahmet Hamdi Özkurt, Emrah Aydemir, Yasin Sönmez; Large Language Models vs. Human Interpretation: Which is More Accurate in Text Classification?.174–182

Research Article

Hadjer Brioua, Havvanur Siyambaş, Durmuş Özkan Şahin; Phishing E-mail Detection with Machine Learning and Deep Learning: Improving Classification Performance with Proposed New Features,183–193

Research Article

Emre İrtem, Nesli Erdoğan; Fingerprint Generation for DNN Training: A Case Study in Fingerprint Classification,194–202

Research Article

Emrah Aslan, Yıldırım Özüpak; Performance Comparison of Deep Learning Models in Brain Tumor Classification,203–209

Research Article

Vedat Yılmaz; A Bibliometric Analysis on Cybersecurity Using VOSviewer: An Evaluation for Public Security,210–218

Research Article

Abdulkadir Gozuoglu; Intelligent Modular Energy Hub: Advanced Optimization of Second-Life Lithium-Based Batteries for Sustainable Power Utilization,219–229

Research Article

Serdar Özyön, Hasan Temurtaş, Burhanettin Durmuş, Celal Yaşar; Application of Average Differential Evolution Algorithm to Lossy Fixed Head Short-Term Hydrothermal Coordination Problem,230–242

BALKAN JOURNAL OF ELECTRICAL & COMPUTER ENGINEERING
(An International Peer Reviewed, Indexed and Open Access Journal)

Contact: Batman University Department of Electrical-Electronics Eng.

Bati Raman Campus Batman-Turkey

Web: <https://dergipark.org.tr/en/pub/bajece>

<https://www.bajece.com> **e-mail:** bajece@hotmail.com

Alzheimer's Disease Diagnosis in MRI Images Using Transfer Learning Methods: Evaluation of Different Model Performances

Eren Gunduzvar, Abdulsamet Kayik and Mehmet Ali Altuncu


Abstract— The most common type of dementia in older adults is Alzheimer's disease. Currently, there is no known cure for this illness. The progression of the disease can lead to loss of cognitive and physical abilities. In addition, the process of caring for patients causes both economic and psychological difficulties for their relatives. Therefore, early detection of Alzheimer's disease is vital. With an early diagnosis, patients' quality of life can be improved, and the progression of the disease can be slowed. Many clinical methods are used in the diagnosis of Alzheimer's disease. One of the most preferred of these methods is Magnetic Resonance Imaging (MRI). This study compares the performance of transfer learning-based Convolutional Neural Network (CNN) models, including VGG-19, ResNet-50, DenseNet-201, and InceptionV3. These models are used to classify Alzheimer's disease into four stages: Non Demented, Very Mild Demented, Mild Demented, and Moderate Demented. To compare the performance of the models, accuracy, precision, sensitivity, F1-score, and area under the curve (AUC) metrics were measured. A publicly available dataset of MRI images was used in the study. SMOTE (Synthetic Minority Over-sampling Technique) was applied to overcome the class imbalance in the dataset. Experimental results show that transfer learning-based CNN models are effective in classifying Alzheimer's stages. In particular, the DenseNet-201 model outperformed the other models with an accuracy of 96.52%.

Index Terms— Alzheimer, CNN architectures, MRI imaging, Transfer learning.


I. INTRODUCTION

DEMENTIA, A type of disease caused by impairment in cognitive functions, leads to problems in skills such as memory formation, speech, thinking, judgment and behavior.


Eren Gündüzvar is with the Department of Computer Engineering University of Kocaeli, Kocaeli, Turkey, (e-mail: gunduzvareren@gmail.com).

 <https://orcid.org/0009-0009-2375-9080>

Abdulsamet Kayik is with the Department of Computer Engineering University of Kocaeli, Kocaeli, Turkey, (e-mail: sametkayik@gmail.com).

 <https://orcid.org/0009-0002-3212-8618>

Mehmet Ali Altuncu is with the Department of Computer Engineering University of Kocaeli, Kocaeli, Turkey, (e-mail: mehmetali.altuncu@kocaeli.edu.tr).

 <https://orcid.org/0000-0002-2948-3937>

Manuscript received Aug 20, 2023; accepted Feb, 17, 2025.

DOI: [10.17694/bajece.1535631](https://doi.org/10.17694/bajece.1535631)

According to the World Alzheimer Report 2023, the number of dementia cases, which reached 55 million in 2020, is estimated to reach 139 million in the 2050s as societies age. According to the same report, the cost of treating dementia patients, which was USD 1.3 trillion per year in 2019, is projected to increase to USD 2.8 trillion in 2030 [1].

In Alzheimer's disease (AD), symptoms develop gradually but progressively worsen, particularly affecting those over the age of 65. Since there is no definitive cure, it has become one of the most important diseases for the elderly. AD patients are in need of care, and caring for them is both costly and psychologically burdensome for families. Therefore, early diagnosis is of great importance to prevent the disease from progressing [2, 7]. Moreover, early diagnosis can prolong the survival of AD patients by an average of 3 years [3].

Depending on the level of brain damage and the patient's condition, AD is usually classified into four stages. These stages are defined to reflect the progression of the disease and the severity of cognitive impairments. The first is Mild Cognitive Impairment (MCI), when symptoms begin to appear, and the second is Mild Alzheimer's, when memory loss and cognitive impairments become more pronounced. In the third stage, Moderate Alzheimer's, the person may have difficulty recognizing himself or herself and the people around them, and may struggle to perform complex tasks. In the last stage, Severe Disability, abilities for daily living are severely impaired, and patients need full-time care to meet their basic needs [8].

In clinical practice, various imaging techniques are employed to identify the stages of Alzheimer's disease; these techniques help in understanding the effects, structure, and function of the disease. These methods include Computed Tomography (CT), Positron Emission Tomography (PET), MRI, and ultrasonography [4]. Particularly in the early stages of Alzheimer's disease, MRI is the most commonly used imaging modality because it can capture small structural changes in different brain regions more clearly [5].

Detecting the first stage of Alzheimer's disease, MCI, enables the implementation of necessary measures to prevent the disease from progressing to other stages. While clinical assessments and expert evaluations are necessary to prevent the progression of Alzheimer's disease, symptoms often need to become quite pronounced for experts to accurately identify the stages. Therefore, the presence of automated assistive detection systems in conjunction with expert evaluations is of vital

importance [6]. A number of sophisticated and successful machine learning techniques have been developed in recent research to classify Alzheimer's disease stages, such as from MRI scans.

This study assessed how well CNN-based transfer learning models performed when used to categorize MRI images into the four stages of Alzheimer's disease. The examined models underwent experiments, and the output of each model was contrasted with each other. The rest of this essay is structured as follows: In Section II, the methods, and studies that are currently being used to identify AD stages are thoroughly reviewed. Section IV provides a detailed presentation of the experimental outcomes, whereas Section III outlines the suggested technique. The paper is concluded in Section V, also addressing future research directions.

II. RELATED WORKS

Various methods have been proposed for the classification of AD stages. This section reviews studies that used MRI. A model for multi-class AD diagnosis based on Linear Discriminant Analysis (LDA) was proposed by Lin et al. [7]. The MR images were initially adjusted based on the patient's age. Using the least absolute shrinkage and selection operator (LASSO) approach, features were selected in the second stage, and Principal Component Analysis (PCA) was used to reduce dimensionality. Finally, a decision tree based on an Extreme Learning Machine (ELM) was employed to perform multi-class categorization. The AD Neuroimaging Initiative (ADNI) dataset was used for the studies, and the proposed model outperformed an approach that relied only on raw characteristics.

Acharya et al. [8] proposed transfer learning models for predicting AD stages. In this study, the proposed modified AlexNet architecture was compared with traditional CNN, VGG-16, and ResNet-50 models, and it was reported to achieve higher accuracy, F-score, Recall, and Precision on the Kaggle MRI dataset [9].

The stages of Alzheimer's disease were also categorized by Shamrat et al. [10] using transfer learning models. The ADNI dataset's image quality was first enhanced using a histogram equalization-based approach, and the dataset's class imbalance was eliminated using data augmentation approaches. After training five CNN models (VGG-16, MobileNetV2, AlexNet, ResNet-50, and InceptionV3), the InceptionV3 model achieved the highest accuracy. In the final phase of the study, the InceptionV3 model, which had the best accuracy, was modified, resulting in an improvement in accuracy beyond that of the classical approach.

A modified version of the VGG-16 model was presented by Mehmood et al. [11] to categorize AD patients into four stages. The present study used the Open Access Series of Imaging Studies (OASIS) dataset and applied data augmentation techniques. The suggested model, which has two modified VGG-16 layers operating in parallel, includes three batch normalization layers, three Gaussian noise layers, five max-pooling layers, and 14 convolutional layers. This approach

resulted in a significant improvement in the accuracy of multi-class AD classification.

Using a portion of the ADNI dataset, Nawaz et al. [12] presented a unique CNN architecture for the classification of AD phases. The suggested model outperformed the conventional AlexNet and VGG-16 models in terms of accuracy, when tested on a dataset that was produced from a subset of the ADNI images.

Fu'adah et al. [13] attempted to detect AD stages using a CNN model based on AlexNet architecture. The dataset used in this study also comprised MRI scans. The study compared different learning rates using Adam optimization, and the best accuracy and loss parameters were obtained with a learning rate of 0.0001.

A CNN model was proposed by Ajagbe et al. [14] to categorize AD stages from MRI images. The accuracy area under curve (AUC), F1 score, precision, recall and computation time of the suggested model were evaluated using the Kaggle MRI dataset and compared with the results obtained with the VGG-16 and VGG-19 models. The proposed model demonstrated better performance in terms of the F1 score, recall, and computational time.

Rao et al. [15] applied machine learning techniques to classify the AD stages. The proposed method generally comprises data preprocessing, feature extraction and selection, and classification stages. Correlation Matrices and Exhaustive Feature Selection were employed for feature selection, and the Support Vector Machine (SVM) and Multilayer Perceptron (MLP) methods were used for classification. The proposed model was compared with the classical AlexNet method and yielded better results in terms of accuracy, F1 score, precision, and recall metrics.

Ramzan et al. [16] proposed a hybrid method that combines residual neural networks (RNN) and transfer learning (ResNet-18) to classify the six stages of AD. The proposed model was tested on the ADNI dataset and was reported to achieve better results than some previous studies.

Nawaz et al. [17] also proposed a hybrid method. In this study, the AlexNet model was used for feature extraction, and K-Nearest Neighbors (KNN), SVM, and Random Forest (RF) methods were used for classification. The model's performance was evaluated on the OASIS dataset, and the highest accuracy was achieved using the SVM method during classification.

Savaş [18] compared the performance of 29 different pretrained models for detecting AD stages using MRI images from the ADNI dataset. The performances of the models were evaluated using accuracy, precision, sensitivity, and specificity metrics, and the EfficientNet versions demonstrated the best performance in terms of these parameters.

Degadwala et al. [19] proposed another CNN-based model to better capture specific features in AD dataset images and enhance classification performance. The performance of the proposed model was compared with those of the AlexNet, VGG-16, and ResNet-50 models, and it was reported to have better performance than all of these methods.

In terms of accuracy and loss parameters, Mirchandani et al. [20] evaluated the performance of three CNN-based algorithms for multi-class classification of AD stages (four stages): AlexNet, Faster R-CNN, and YOLOv4. In this study, data augmentation techniques were applied to address the class imbalance in the Kaggle MRI dataset. The results demonstrate that AlexNet and YOLOv4 achieved the highest accuracy.

A hybrid CNN approach based on the ResNet-50 architecture was proposed by Yildirim and Çinar [21] for the classification of four distinct AD phases. The study used the Kaggle MRI dataset [3], and the performance of the proposed model was compared to that of the classical AlexNet, ResNet-50, DenseNet-201 and VGG-16 methods. It was reported that the developed hybrid model improved accuracy by 3% compared to traditional CNN architectures.

Esam and Mohammed [33] also proposed an original CNN architecture for the classification of four different AD stages. In the pre-processing stage of the study, images were converted to 150-x-150 pixels and data augmentation was applied using the SMOTE technique. The accuracy of the proposed model was compared with pre-trained CNN models. As a result of the evaluations, the original CNN model exhibited superior performance.

Arafa et al. [34] proposed a method consisting of data pre-processing, data augmentation, cross-validation and classification stages for the detection of AD. In the classification stage, they first performed AD detection with an original CNN architecture. In the second method, they tried to improve the performance of the model by testing the pre-trained CNN model VGG-16 with different optimizers. The results show that the original CNN architecture provides higher classification accuracy than the optimized VGG-16 model.

Singh and Kumar [35] compared the performance of seven pre-trained CNN models for the detection of six different stages of AD. In order to improve the images used in the study, methods such as image reorientation, shadow correction, and segmentation were applied. In the study, the performance of the models was evaluated with metrics such as accuracy, AUC, loss values, F1 score, and it was stated that the best performance was obtained with the EfficientNet0 model.

Khalid et al. [36] developed three different feature extraction methods for the detection of AD stages and compared their performance. In the first method, features of MRI images were separately obtained using the GoogLeNet and DenseNet-121 models were used. In the second method, features obtained using the same models were combined, and the dimensions of the features were reduced using the Principal Component Analysis (PCA) algorithm. In the third method, the features obtained with the GoogLeNet and DenseNet-121 models were combined with the hand-crafted features obtained with the Discrete Wavelet Transform (DWT), Local Binary Pattern (LBP) and Gray Level Equivalence Matrix (GLCM) methods. The features obtained with the three different methods were classified, using the Feed Forward Neural Network (FFNN). It was stated that the highest accuracy was obtained from FFNN classification by combining the features obtained with the

DenseNet-121 model with the hand-crafted features.

III. METHODOLOGY

The proposed method for classifying the AD stages is illustrated in Fig. 1. The SMOTE technique was initially applied to address class imbalance in a publicly available dataset. The dataset was then divided into 80% training data and 20% test data. During the classification phase, four transfer learning methods—VGG-19, ResNet-50, InceptionV3, and DenseNet-201—were employed. These models were used to classify AD stages into four categories.

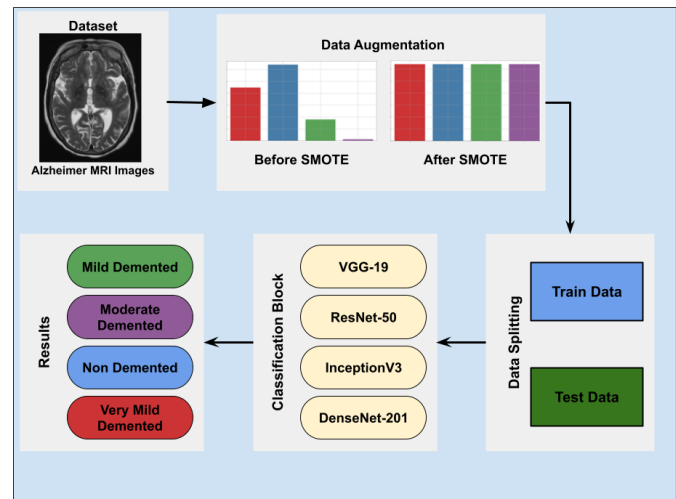


Fig.1. Architecture of the proposed system

A. VGG-19

The University of Oxford's Visual Geometry Group (VGG) created the VGG-19 model, a CNN architecture trained on millions of images from the ImageNet collection (Fig. 2). Three fully connected layers and sixteen convolutional layers comprise VGG-19. Each of the five blocks that comprised the convolutional layers contained 3×3 filters. The max-pooling layer lowers the dimensionality and computing cost of the model, which follows each convolutional block. The nonlinear activation function is the rectified linear unit (ReLU). The last layer, which is a fully connected layer, performs classification using the softmax activation function [22, 23].

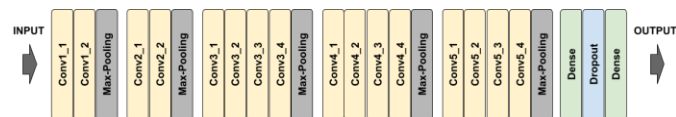


Fig.2. The architecture of the VGG-19 model [24]

B. ResNet-50

He et al. created the Residual Network (ResNet) architecture in 2015 to address the vanishing gradient problem, which occurs when deep networks have more layers [25]. The existence of structures known as "residual blocks" which appear every few layers is the primary way that ResNet differs from conventional CNN systems. With the help of skip connections, these blocks directly add input to the output layer, which improves the learning efficiency of the network. This method solves problems like disappearing gradients and makes training

deeper networks easier. Depending on the number of layers, there are variations in the ResNet design with 50, 101, and 152 layers [26, 27]. These versions are frequently used in various applications, such as location, object identification, and image categorization. ResNet-50 is a model that was trained using millions of images from the ImageNet database, just like VGG-19. The 50 layers of the ResNet-50 architecture, which include pooling, activation functions, and convolutional layers, are shown in Fig. 3.

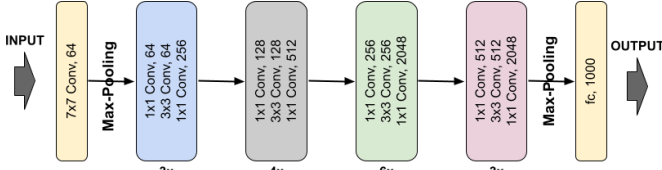


Fig.3. The architecture of the ResNet-50 model [28]

C. DenseNet-201

A variation of the DenseNet architecture called DenseNet-201 has dense blocks that comprise several convolutional layers connected directly to one another. 1x1 convolutions make up the Transition layers between Dense blocks, which lessen feature maps and maximize computational load [29]. Fig. 4 depicts the DenseNet-201 model's overall network structure.

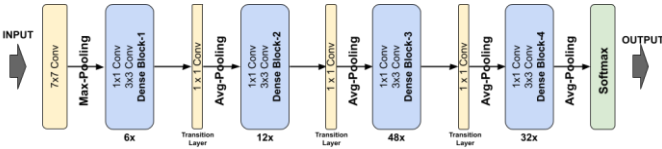


Fig.4. The architecture of the DenseNet-201 model [30]

D. InceptionV3

InceptionV3, an evolution of GoogLeNet (InceptionV1), offers significant improvements in terms of both performance and computational efficiency. This architecture features "Inception Blocks," which enable the model to learn features at various scales by simultaneously performing convolutions and max-pooling operations of different sizes. Additionally, similar to the 1x1 convolutions used in the DenseNet-201 model, InceptionV3 employs 1x1 convolutions. These convolutions reduce the size of the feature maps, thereby reducing the computational load of the model and facilitating a faster learning process [31].

IV. EXPERIMENTAL RESULTS

In this study, the performance of popular transfer learning models, including VGG-19, ResNet-50, DenseNet-201, and InceptionV3, was analyzed for the classification of AD stages. The experiments were conducted using the TensorFlow framework, which is commonly used in deep learning applications. The dataset was divided into 80% training and 20% testing, for model training and evaluation. The training and testing data were randomly selected using Python's scikit-learn library. To ensure a fair comparison of model performance, the same data were used during both training and testing phases for each model.

A. Dataset Description

This study utilized the open-source Alzheimer's Dataset (4 Classes of Images) [9], which is available on Kaggle. The dataset contains 6,400 MRI images in JPEG format, each with dimensions of 176x208 pixels. The dataset is organized into four classes and the distribution of each class is given in Table I.

TABLE I
DISTRIBUTION OF CLASSES IN THE KAGGLE MRI DATASET

Class Name	Number of samples
Mild Demented	896
Moderate Demented	64
Non Demented	3200
Very Mild Demented	2240

As shown in Table I, the distribution among classes is imbalanced. To address this imbalance, the SMOTE technique [32] was applied to equalize the number of images in each class to 3200.

B. Evaluation metrics

Four distinct criteria were used to assess the performance of the proposed transfer learning models: accuracy, precision, sensitivity, and F1 score. Four basic parameters were used to calculate these metrics: False Positive (FP), True Negative (TN), True Positive (TP), and False Negative (FN). Equations (1-4) include the formulas required to compute the metrics. In addition, the loss value, which measures the difference between the model's predictions and actual values and the AUC (Area Under the Curve) metric indicating how well the model differentiates between classes at various thresholds, are reported for each transfer learning model individually.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity} \quad (4)$$

The hyperparameters used during the training of VGG-19, ResNet-50, DenseNet-201, and InceptionV3 models are detailed in Table II. The Adam algorithm was preferred in model optimization; learning rate was fixed at 0.001; mini-batch size was determined to be 32; and the number of training iterations (epochs) was configured as 100.

TABLE II
HYPERPARAMETERS USED IN MODEL TRAINING

Hyperparameter Name	Value
Optimizer	Adam
Learning Rate	0.001
Batch Size	32
Maksimum Epoch	100

C. Results obtained with the VGG-19 model

The accuracy and loss function values of the VGG-19 model during the training process are shown in Figure 5. Upon examining the accuracy graph history in Figure 5, the training accuracy value increased steadily as the number of epochs increased, while validation accuracy fluctuated but stabilized towards the end of training. This shows that the model achieved convergence without overfitting. When the history of loss graph is examined, the training loss decreased over time and reached low levels, which showed that the model's errors are gradually decreasing.

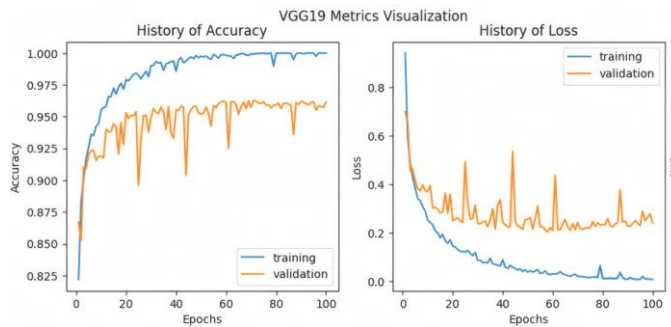


Fig.5. Accuracy and loss functions during VGG-19 training

The confusion matrix obtained using the test data after training the VGG-19 model is shown in Fig. 6.

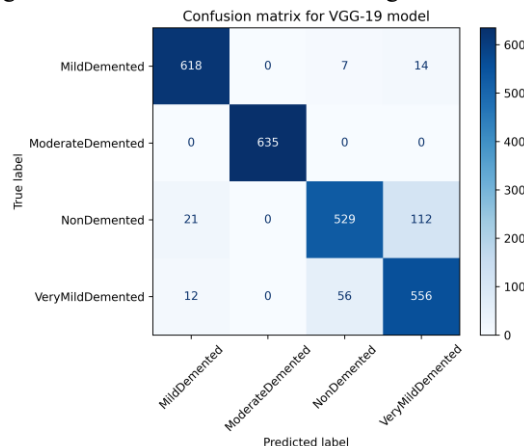


Fig.6. Confusion matrix for the VGG-19 model

With the VGG-19 model, 91.32% overall accuracy was obtained. Other performance metrics of the model are presented in Table III. According to the table, the 'Moderate Demented' class achieved the highest performance with 100% Precision, Sensitivity, and F1-Score. On the other hand, the 'Very Mild Demented' class showed the lowest performance with Precision (81.52%), Sensitivity (89.1%), and F1-Score (85.15%). This shows that the model has greater difficulty in distinguishing the class.

TABLE III
EVALUATION METRICS FOR VGG-19 MODELS

Class	Precision (%)	Sensitivity (%)	F1-Score (%)
Mild Demented	94.93	96.71	95.81
Moderate Demented	100	100	100
Non Demented	89.36	79.91	84.37
Very Mild Demented	81.52	89.1	85.15
Weighted average	91.48	91.33	91.29

The ROC curve for the VGG-19 model is shown in Figure 7. When the ROC curve is examined, it is seen that the lowest performance among the four classes is in the 'Very Mild Demented' class with 0.97. However, the AUC value for the 'Moderate Demented' and 'Mild Demented' classes was calculated as 1.0.

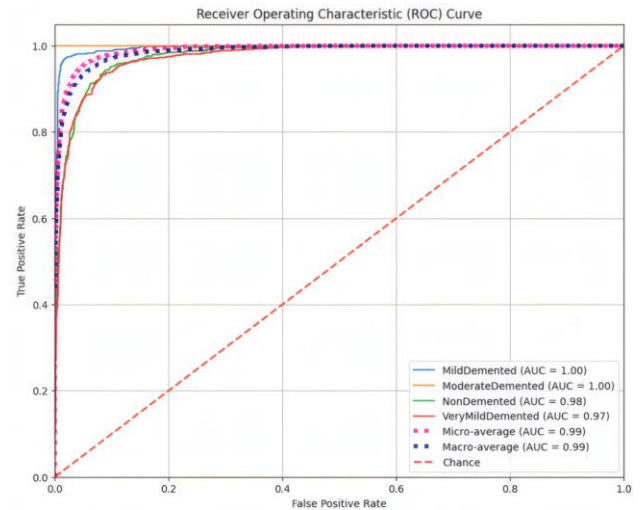


Fig.7. ROC curve of the VGG-19 model

D. Results obtained with the ResNet-50 model

The accuracy and loss function values of the ResNet-50 model during the training process are shown in Figure 8. When the 'history of accuracy' graph in Figure 8 is examined, it shows that the model generalized successfully on the validation set and that there was no overfitting. The history of loss graph shows that the validation loss is lower than the training loss.

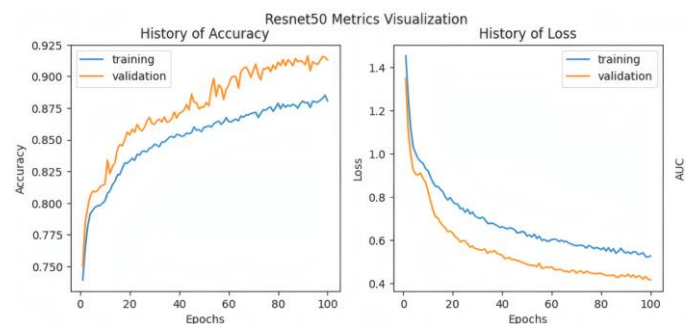


Fig.8. Accuracy and loss functions during ResNet-50 training

The confusion matrix obtained using the test data after training the ResNet-50 model is shown in Fig. 9.

With the ResNet-50 model, 80.62% overall accuracy was obtained. Other performance metrics of the model are presented in Table IV. According to Table IV, the 'Moderate Demented' class showed the highest performance with 99.84% Precision, 100% Sensitivity, and 99.92% F1-Score. On the other hand, the 'Very Mild Demented' class exhibited the lowest performance with Precision (61.35%), Sensitivity (62.82%) and F1-Score (62.07%). This shows that the model does not adequately distinguish the 'Very Mild Demented' class. In addition, although the Precision (71.18%) is low for the 'Non Demented'

class, it was observed that the false negative rate was low because the Sensitivity value was high at 97.52%.

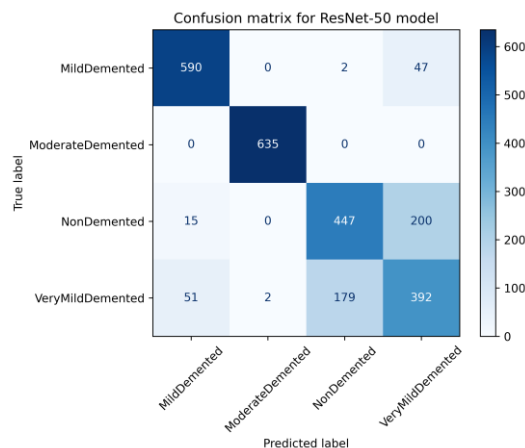


Fig.9. Confusion matrix for the ResNet-50 model

TABLE IV
EVALUATION METRICS FOR THE RESNET-50 MODEL

Class	Precision (%)	Sensitivity (%)	F1-Score (%)
Mild Demented	89.94	92.33	91.12
Moderate Demented	99.84	100	99.92
Non Demented	71.18	97.52	69.3
Very Mild Demented	61.35	62.82	62.07
Weighted average	80.54	80.62	80.56

The ROC curve for the ResNet-50 model is presented in Figure 10. When the ROC curve is examined, it is seen that the lowest performance among the four classes is in the 'Very Mild Demented' class with an AUC value of 88%. The AUC for the 'Moderate Demented' class is calculated as 100%. The results show that the ResNet-50 model needs improvements, especially in the 'VeryMildDemented' and 'NonDemented' classes.

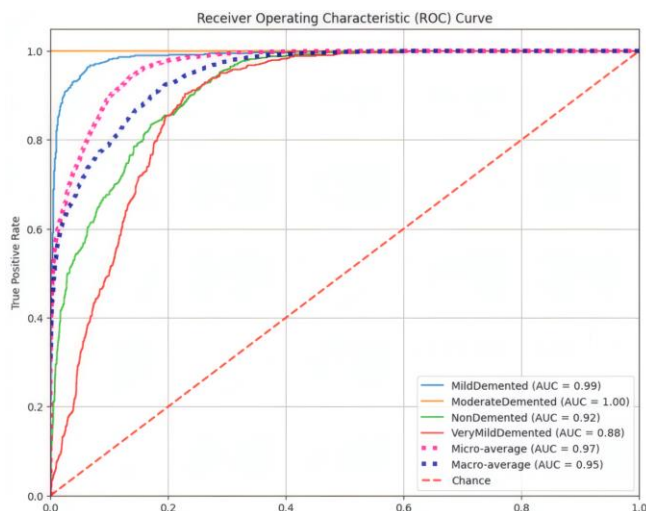


Fig.10. ROC curve of ResNet-50

E. Results Obtained with the DenseNet-201 model

The accuracy and loss function values of the DenseNet-201 model during the training process are shown in Figure 11. When the 'history of accuracy' graph in Figure 11 is examined, the difference between the model's training and validation accuracy

is small, and stable convergence is achieved. The "history of loss" graph shows that the fluctuations in the validation loss are small and the model progresses stably during the training process.

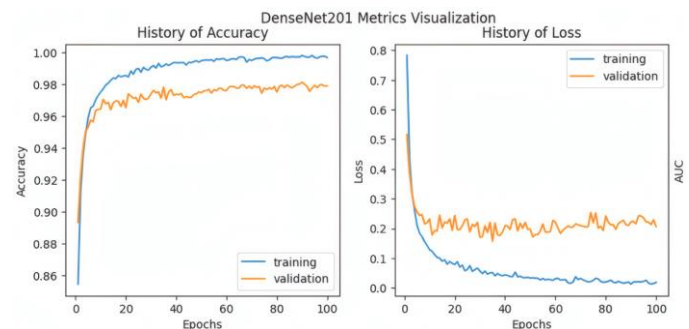


Fig.11. Accuracy and loss functions during the DenseNet-201 training

The confusion matrix obtained using the test data after training the DenseNet-201 model is shown in Fig. 12.

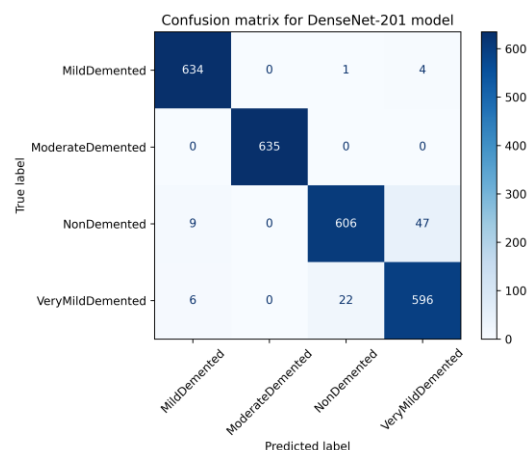


Fig.12. Confusion matrix for the DenseNet-201 model

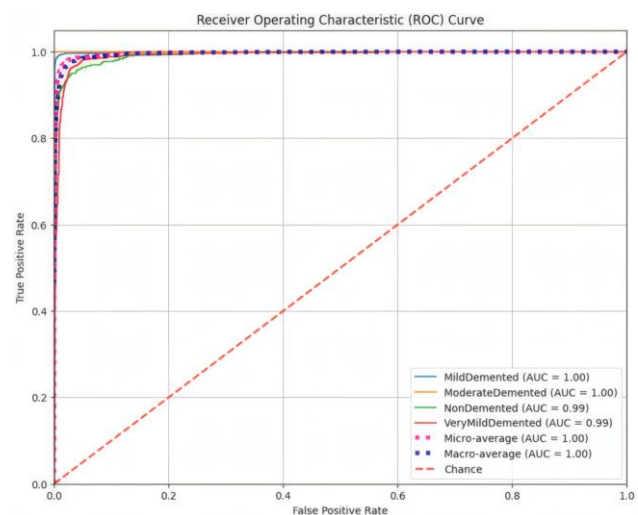


Fig.13. ROC curve of the DenseNet-201 model

With the DenseNet-201 model, 96.52% overall accuracy was achieved. Other performance criteria of the model are presented in Table V. When the table is examined, although there are low scores especially in the 'Non Demented' and 'Very Mild

Demented' classes, the overall model performance is highly balanced.

TABLE V
EVALUATION METRICS FOR THE DENSENET-201 MODEL

Class	Precision (%)	Sensitivity (%)	F1-Score (%)
Mild Demented	97.69	99.22	98.45
Moderate Demented	100	100	100
Non Demented	96.34	91.54	93.88
Very Mild Demented	92.12	95.51	93.78
Weighted average	96.56	96.52	96.51

The ROC curve for the DenseNet-201 model is shown in Figure 13. When the ROC curve is examined, it is seen that it can distinguish all classes quite well.

F. Results Obtained with the InceptionV3 model

The accuracy and loss function values of the InceptionV3 model during the training process are shown in Figure 11. As in the DenseNet-201 model, the difference between the training and validation accuracy of the InceptionV3 model is minimal, which reveals that the learning process of the model is completed efficiently and progresses without overfitting.

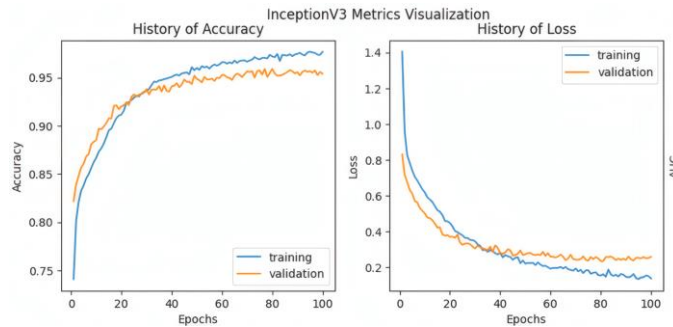


Fig.14. Accuracy and loss functions during the InceptionV3 training

The confusion matrix obtained using the test data after training the InceptionV3 model is shown in Fig. 15.

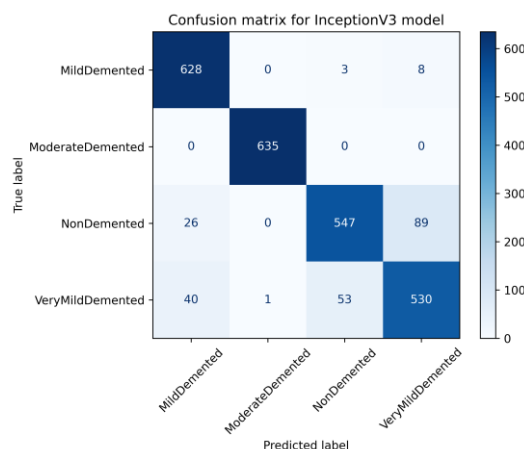


Fig.15. Confusion matrix for the InceptionV3 model

With the InceptionV3 model, 91.41% overall accuracy was achieved. Other performance criteria of the model are presented in Table VI. Upon examining the table, it is observed that

although there are differences in the 'Non Demented' and 'Very Mild Demented' classes compared to other classes, the overall model performance is balanced with an F1-score of 91.32%.

TABLE VI
EVALUATION METRICS FOR INCEPTIONV3 MODELS

Class	Precision (%)	Sensitivity (%)	F1-Score (%)
Mild Demented	90.49	98.28	94.22
Moderate Demented	99.84	1.0	99.92
Non Demented	90.71	82.63	86.48
Very Mild Demented	84.53	84.94	84.73
Weighted average	91.41	91.41	91.32

The ROC curve for the InceptionV3 model is shown in Figure 16. When the ROC curve is examined, the AUC was 0.97 and above in all classes. These values show that the model is a strong classifier in distinguishing between all classes.

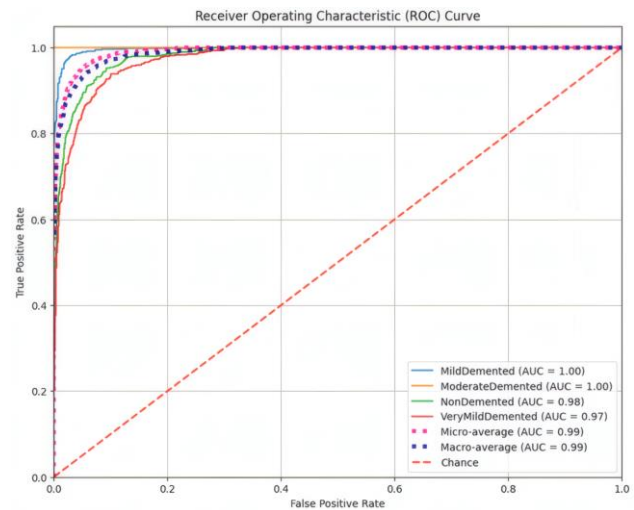


Fig.16. ROC curve of the InceptionV3 model

V. CONCLUSION

In this study, the performances of transfer learning-based CNN models are compared to classify four different stages of Alzheimer's disease via MRI images. In the experiments performed on VGG-19, ResNet-50, DenseNet-201, and InceptionV3 models, DenseNet-201 performed better than other models with Accuracy: 96.52%, Precision: 96.56%, Sensitivity: 96.52%, and F1-Score: 96.51%. However, the ResNet-50 model showed lower performance than DenseNet-201 with accuracy rates of 80.62%, InceptionV3 with 91.41%, and VGG-19 with 91.32%. The results show that transfer learning models can be used in the detection of AD stages. However, the errors observed in the early stages (Very Mild Demented) also suggest that the classification performance needs to be improved. In addition, the SMOTE technique was used in the study to eliminate the imbalance between the classes in the dataset. However, while SMOTE produces synthetic data, it can negatively affect the generalization performance of the model, as it cannot fully reflect the variations in real MRI images.

To eliminate these problems, it is aimed to use ensemble models created by combining different CNN architectures and to train these models on larger data sets. In addition, more

realistic data augmentation can be achieved by GAN-based synthetic data generation instead of the SMOTE technique. It is anticipated that these approaches will provide more reliable and consistent results in the classification of the stages of Alzheimer's disease.

REFERENCES

- [1] World Alzheimer Report 2023. Available: <https://www.alzint.org/u/World-Alzheimer-Report-2023.pdf>. Accessed 14 August 2024.
- [2] F. Karakaya, C. Gurkan, A. Budak, and H. Karataş, "Classification and Segmentation of Alzheimer Disease in MRI Modality using the Deep Convolutional Neural Networks," *Avrupa Bilim ve Teknoloji Dergisi*, no. 40, pp. 99-105, 2022.
- [3] M. Leela, K. Helenprabha, and L. Sharmila, "Prediction and classification of Alzheimer disease categories using integrated deep transfer learning approach," *Measurement: Sensors*, vol. 27, no. 100749, 2023.
- [4] H. S. Suresha and S. S. Parthasarathy, "Alzheimer disease detection based on deep neural network with rectified Adam optimization technique using MRI analysis," in *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAEECC)*, 2020, pp. 1-6.
- [5] A. W. Salehi, P. Baglat, B. B. Sharma, G. Gupta, and A. Upadhyay, "A CNN model: earlier diagnosis and classification of Alzheimer disease using MRI," in *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, 2020, pp. 156-161.
- [6] N. Raza, A. Naseer, M. Tamoor, and K. Zafar, "Alzheimer disease classification through transfer learning approach," *Diagnostics*, vol. 13, no. 4, p. 801, 2023.
- [7] W. Lin, Q. Gao, M. Du, W. Chen, and T. Tong, "Multiclass diagnosis of stages of Alzheimer's disease using linear discriminant analysis scoring for multimodal data," *Computers in Biology and Medicine*, vol. 134, no. 104478, 2021.
- [8] H. Acharya, R. Mehta, and D. K. Singh, "Alzheimer disease classification using transfer learning," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021, pp. 1503-1508.
- [9] Alzheimer's Dataset (4 class of Images). Available: <https://www.kaggle.com/datasets/tourist55/alzheimers-dataset-4-class-of-images>. Accessed 14 August 2024.
- [10] F. J. M. Shamrat, S. Akter, S. Azam, A. Karim, P. Ghosh, Z. Tasnim, and K. Ahmed, "AlzheimerNet: An effective deep learning based proposition for alzheimer's disease stages classification from functional brain changes in magnetic resonance images," *IEEE Access*, vol. 11, pp. 16376-16395, 2023.
- [11] A. Mehmood, M. Maqsood, M. Bashir, and Y. Shuyuan, "A deep Siamese convolution neural network for multi-class classification of Alzheimer disease," *Brain Sciences*, vol. 10, no. 2, p. 84, 2020.
- [12] A. Nawaz, S. M. Anwar, R. Liaqat, J. Iqbal, U. Bagci, and M. Majid, "Deep convolutional neural network based classification of Alzheimer's disease using MRI data," in *2020 IEEE 23rd International Multipoint Conference (INMIC)*, 2020, pp. 1-6.
- [13] Y. N. Fu'adah, I. Wijayanto, N. K. C. Pratiwi, F. F. Taliningsih, S. Rizal, and M. A. Pramudito, "Automated classification of Alzheimer's disease based on MRI image processing using convolutional neural network (CNN) with AlexNet architecture," in *Journal of Physics: Conference Series*, 2021, vol. 1844, no. 1, p. 012020.
- [14] S. A. Ajagbe, K. A. Amuda, M. A. Oladipupo, F. A. Oluwaseyi, and K. I. Okesola, "Multi-classification of Alzheimer disease on magnetic resonance images (MRI) using deep convolutional neural network (DCNN) approaches," *International Journal of Advanced Computer Research*, vol. 11, no. 53, pp. 51, 2021.
- [15] K. N. Rao, B. R. Gandhi, M. V. Rao, S. Javvadi, S. S. Vellela, and S. K. Basha, "Prediction and classification of Alzheimer's disease using machine learning techniques in 3D MR images," in *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, 2023, pp. 85-90.
- [16] F. Ramzan, M. U. G. Khan, A. Rehmat, S. Iqbal, T. Saba, A. Rehman, and Z. Mehmood, "A deep learning approach for automated diagnosis and multi-class classification of Alzheimer's disease stages using resting-state fMRI and residual neural networks," *Journal of Medical Systems*, vol. 44, pp. 1-16, 2020.
- [17] H. Nawaz, M. Maqsood, S. Afzal, F. Aadil, I. Mehmood, and S. Rho, "A deep feature-based real-time system for Alzheimer disease stage detection," *Multimedia Tools and Applications*, vol. 80, pp. 35789-35807, 2021.
- [18] S. Savaş, "Detecting the stages of Alzheimer's disease with pre-trained deep learning architectures," *Arabian Journal for Science and Engineering*, vol. 47, no. 2, pp. 2201-2218, 2022.
- [19] S. Degadwala, D. Vyas, A. Jadeja, and D. D. Pandya, "Enhancing Alzheimer Stage Classification of MRI Images through Transfer Learning," in *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2023, pp. 733-737.
- [20] R. Mirchandani, C. Yoon, S. Prakash, A. Khaire, A. Naran, A. Nair, and S. Ganti, "Comparing the Architecture and Performance of AlexNet Faster R-CNN and YOLOv4 in the Multiclass Classification of Alzheimer Brain MRI Scans." Available: https://ai-4-all.org/wp-content/uploads/2021/04/Comparing_the_Architecture_and_Performance_of_AlexNet_Faster_R_CNN_and_YOLOv4_in_the_Multiclass_Classification_of_Alzheimer_Brain_MRIScans_Final.pdf. Accessed 14 August 2024.
- [21] M. Yildirim and A. Cinar, "Classification of Alzheimer's Disease MRI Images with CNN Based Hybrid Method," *Ingénierie des Systèmes d'Inf.*, vol. 25, no. 4, pp. 413-418, 2020.
- [22] A. Khattar and S. M. K. Quadri, "Generalization of convolutional network to domain adaptation network for classification of disaster images on twitter," *Multimedia Tools and Applications*, vol. 81, no. 21, pp. 30437-30464, 2022.
- [23] V. Sudha and T. R. Ganeshbabu, "A Convolutional Neural Network Classifier VGG-19 Architecture for Lesion Detection and Grading in Diabetic Retinopathy Based on Deep Learning," *Computers, Materials & Continua*, vol. 66, no. 1, 2021.
- [24] J. Jaworek-Korjakowska, P. Kleczek, and M. Gorgon, "Melanoma thickness prediction based on convolutional neural network with VGG-19 model transfer learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0-0.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
- [26] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," *Applied Sciences*, vol. 12, no. 18, p. 8972, 2022.
- [27] Z. Qin, Q. Zeng, Y. Zong, and F. Xu, "Image inpainting based on deep learning: A review," *Displays*, vol. 69, p. 102028, 2021.
- [28] S. Jahromi, M. N., P. Buch-Cardona, E. Avots, K. Nasrollahi, S. Escalera, T. B. Moeslund, and G. Anbarjafari, "Privacy-constrained biometric system for non-cooperative users," *Entropy*, vol. 21, no. 11, p. 1033, 2019.
- [29] A. P. Syahputra, A. C. Siregar, and R. W. S. Insani, "Comparison of CNN models with transfer learning in the classification of insect pests," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 17, no. 1, pp. 103-114, 2023.
- [30] M. Bakr, S. Abdel-Gaber, M. Nasr, and M. Hazman, "DenseNet based model for plant diseases diagnosis," *European Journal of Electrical Engineering and Computer Science*, vol. 6, no. 5, pp. 1-9, 2022.
- [31] S. Singh and R. Kumar, "Breast cancer detection from histopathology images with deep inception and residual blocks," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 5849-5865, 2022.
- [32] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.
- [33] S. Esam and A. Mohammed, "Alzheimer's disease classification for MRI images using Convolutional Neural Networks," in *2024 6th International Conference on Computing and Informatics*, Mar. 2024, pp. 1-5.
- [34] D. A. Arafa, H. E. D. Moustafa, H. A. Ali, A. M. Ali-Eldin, and S. F. Saraya, "A deep learning framework for early diagnosis of Alzheimer's disease on MRI images," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 3767-3799, 2024.
- [35] A. Singh and R. Kumar, "Brain MRI image analysis for Alzheimer's disease (AD) prediction using deep learning approaches," *SN Computer Science*, vol. 5, no. 1, p. 160, 2024.
- [36] A. Khalid, E. M. Senan, K. Al-Wagih, M. M. Ali Al-Azzam, and Z. M. Alkhraisha, "Automatic analysis of MRI images for early prediction of Alzheimer's disease stages based on hybrid features of CNN and handcrafted features," *Diagnostics*, vol. 13, no. 9, p. 1654, 2023.

BIOGRAPHIES



Eren GÜNDÜZVAR was born in Yalova, Turkey in 2001. He completed his B.S. in Computer Engineering at Kocaeli University in 2023. During his studies, he worked on projects and conducted research in areas such as Data Science, Full Stack Development, Mobile

Development, Image Processing, and Artificial Intelligence. In 2022, he gained practical experience through internships at Döktaş and the Embedded Systems Laboratory at Kocaeli University, where he focused primarily on Full Stack Development. As of 2024, he is continuing his professional career as a FullStack Developer at DefineX Consulting, Technology, and Labs. His interests include web development, data science, image processing, and artificial intelligence.



Abdulsamet KAYIK was born in 2000 in Istanbul, Turkey. He completed his A.S. in Computer Programing at Marmara University in 2021 and began his studies in Computer Engineering at Kocaeli University in the same year. Throughout his academic career, he engaged with various aspects of software development.

In 2023, he gained practical experience through web development internships at Innova and Falla, where he developed frontend and backend development skills. His interests include web development, image processing, and artificial intelligence.



Mehmet Ali ALTUNCU received his B.S. in Computer Engineering from the University of Sakarya in 2006, his M.S. in the same field from Kocaeli University in 2015, and his Ph. D. from Kocaeli University in 2021.

From 2013 to 2022, he worked as a Research Assistant at the Embedded Systems Laboratory at Kocaeli University. Since 2022, he has been serving as an Assistant Professor in the Department of Computer Engineering at Kocaeli University. His research interests include embedded systems, computer network security, pattern recognition, image processing, and machine learning.

Leveraging SHAP for Interpretable Diabetes Prediction: A Study of Machine Learning Models on the Pima Indians Diabetes Dataset

Ismail Kırbas and Ahmet Cıfci


Abstract—This paper investigates the application of machine learning (ML) models for predicting diabetes using the Pima Indians Diabetes Database, with a focus on enhancing model interpretability through the use of SHapley Additive exPlanations (SHAP). The study evaluates eight ML models, including Adaptive Boosting (AdaBoost), k-Nearest Neighbors (k-NN), Logistic Regression (LR), Multi-layer Perceptron (MLP), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGBoost), utilizing both test/train split and 10-fold cross-validation methods. The RF model demonstrated superior performance, achieving an accuracy of 82% and an F1-score of 0.83 in the test/train split, and an accuracy of 83% and an F1-score of 0.84 in the 10-fold cross-validation. SHAP analysis was employed to identify the most influential predictors, revealing that glucose, BMI, pregnancies, and insulin levels are the key factors in diabetes prediction, aligning with established clinical markers. Additionally, the use of the Synthetic Minority Over-sampling TEchnique (SMOTE) for class balancing and data scaling contributes to robust model performance. The study emphasizes the necessity for interpretable ML in healthcare, proposing SHAP as a valuable tool for bridging predictive accuracy and clinical transparency in diabetes diagnostics.

Index Terms—Diabetes Prediction, Explainable Artificial Intelligence, Machine Learning Models, Model Interpretability, SHapley Additive exPlanation.


I. INTRODUCTION

DIABETES IS a chronic and increasingly prevalent condition with profound implications for public health worldwide [1-3]. According to recent estimates, the global prevalence of diabetes among adults continues to rise, creating a significant burden on healthcare systems and underscoring the urgent need for effective preventive and diagnostic tools [4].

Ismail Kırbas, is with Department of Computer Engineering Burdur Mehmet Akif Ersoy University, Burdur, Türkiye, (e-mail: ismailkirbas@mehmetakif.edu.tr).

 <https://orcid.org/0000-0002-1206-8294>

Ahmet Cıfci, is with Department of Electrical-Electronics Engineering Burdur Mehmet Akif Ersoy University, Burdur, Türkiye, (e-mail: acifci@mehmetakif.edu.tr).

 <https://orcid.org/0000-0001-7679-9945>

Manuscript received Nov 1, 2024; accepted Dec 27, 2025.
DOI: [10.17694/bajece.1577929](https://doi.org/10.17694/bajece.1577929)

Early prediction of diabetes can enable timely interventions, reducing the likelihood of severe complications, improving patient outcomes, and potentially decreasing healthcare costs. ML models have shown promise in predicting diabetes by identifying patterns within clinical and demographic data, facilitating early detection [5-9]. However, despite advances in predictive accuracy, traditional ML models often lack interpretability, which is a critical limitation in clinical settings where transparency is paramount.

The interpretability of a model is especially crucial in healthcare, as it provides clinicians with insights into the decision-making process, enhances trust in model predictions, and supports more informed and individualized patient care. Conventional ML models, such as neural networks and ensemble methods, typically operate as “black boxes,” yielding high predictive accuracy but offering limited understanding of how predictions are derived [10, 11]. This opacity creates challenges in clinical applications, as healthcare providers require an explanation of model decisions to comply with ethical standards, support diagnostic conclusions, and facilitate shared decision-making with patients. The field of XAI aims to address these challenges by developing methods that enhance the transparency and interpretability of ML models, making them more suitable for sensitive applications such as diabetes prediction [12-14].

One promising XAI method is SHAP [15], which assigns importance values to each feature in a model’s prediction process, helping to elucidate the contribution of specific patient characteristics to the overall prediction. SHAP is based on Shapley values, a concept from cooperative game theory [16], and provides consistent, theoretically grounded explanations that allow clinicians to understand which features most strongly influence the likelihood of diabetes in individual cases. By incorporating SHAP, healthcare providers can make more confident decisions, potentially identifying high-risk patients based on meaningful patterns in data [17].

The utilization of ML models for diabetes prediction has garnered significant attention in research, driven by the rising global prevalence of diabetes and the pressing need for early diagnostic solutions. A substantial body of studies has concentrated on evaluating and comparing various ML algorithms, with a particular emphasis on the Pima Indians Diabetes Database [18]. Verma and Khatoon [19] compared LR, SVM, k-NN, and RF models, identifying RF as the best

performer with an accuracy of 80.08%. Similarly, Xie [20] demonstrated that LR slightly outperformed RF and SVM, achieving a prediction accuracy of 79.13%. Chang et al. [21] emphasized the importance of interpretable models in the context of the Internet of Medical Things (IoMT), exploring NB, RF, and J48 Decision Tree (DT) models. RF was particularly effective for datasets with more features, while NB excelled in simpler configurations. Sahoo et al. [22] conducted a comparative analysis of supervised classification algorithms, including LR, SVM, and RF. Their study highlighted LR and DT classifiers as achieving the highest accuracy, demonstrating their suitability for diabetes prediction tasks. Similarly, You and Kang [23] identified glucose, BMI, and age as the most significant predictors using correlation analysis and employed SVM and DT models, achieving an accuracy of 70%. Ashour et al. [24] evaluated feedforward neural networks (FNN) and convolutional neural networks (CNN), with the former achieving the highest accuracy of 82%. Akyol and Şen [25] examined ensemble learning methods such as AdaBoost, Gradient Boosted Trees, and RF, reporting that AdaBoost combined with stability selection achieved the best accuracy of 73.88%. Reza et al. [26] achieved the highest accuracy of 79.33% using the RF model for the Pima Indian Diabetes Dataset, highlighting its effectiveness in diabetes classification. Pyne and Chakraborty [27] implemented an artificial neural network (ANN) without feature extraction, achieving a classification accuracy of 80.79%.

Efficient preprocessing plays a critical role in enhancing model performance. Jain et al. [28] analyzed imputation techniques such as Multivariate Imputation by Chained Equations (MICE), k-NN, and mean/mode replacement, finding that k-NN based imputation improved the predictive accuracy of RF models. Karatsiolis and Schizas [29] proposed a region-based SVM approach that integrated clustering and kernel selection, achieving an accuracy of 82.2%.

While previous studies have primarily focused on predictive accuracy and have implemented only a limited selection of ML classifiers, they often lack the integration of XAI methods necessary for clinical application. This study addresses this gap by systematically integrating SHAP into the classification

process, thereby enhancing interpretability without sacrificing accuracy. Our objective is to improve transparency in diabetes prediction models, making them more useful for healthcare providers and ultimately contributing to better patient care through informed decision-making.

In this study, we employ a variety of ML models, including AdaBoost, k-NN, LR, MLP, NB, RF, SVM, and XGBoost. These models were selected for their diverse characteristics, which range from linear to non-linear and from probabilistic to tree-based approaches. The diversity of model types allows for a comprehensive evaluation of classification performance and interpretability when XAI methods are applied. Furthermore, we utilize the Pima Indians Diabetes Database [18], a widely referenced dataset in diabetes prediction research, which includes relevant clinical and demographic variables.

II. MATERIALS AND METHOD

This section provides a comprehensive explanation of the dataset, the application of SHAP for model interpretability, the performance evaluation metrics utilized to assess the models, and the ML algorithms implemented in this study.

Fig. 1 outlines the process for predicting diabetes using ML models. It begins with data collection, where the Pima Indians Diabetes Database is utilized as the dataset. This is followed by data normalization, ensuring all features are scaled to a consistent range for unbiased model training. Next, class balancing is performed, employing techniques like SMOTE to address any class imbalances in the dataset, such as unequal representation of diabetic and non-diabetic cases. In the test-train splitting step, the data is divided into training and testing sets (80% training, 20% testing) to facilitate model evaluation. For enhanced reliability, 10-fold cross validation is applied, splitting the training data into ten subsets to train and validate the model iteratively. During model training, ML algorithms are used to learn patterns from the training data. The trained model's effectiveness is then assessed through performance evaluation, using metrics like accuracy, precision, recall, and F1-score. Finally, SHAP analysis is conducted to interpret the model's decision-making process, identifying the relative importance of features in predicting diabetes.

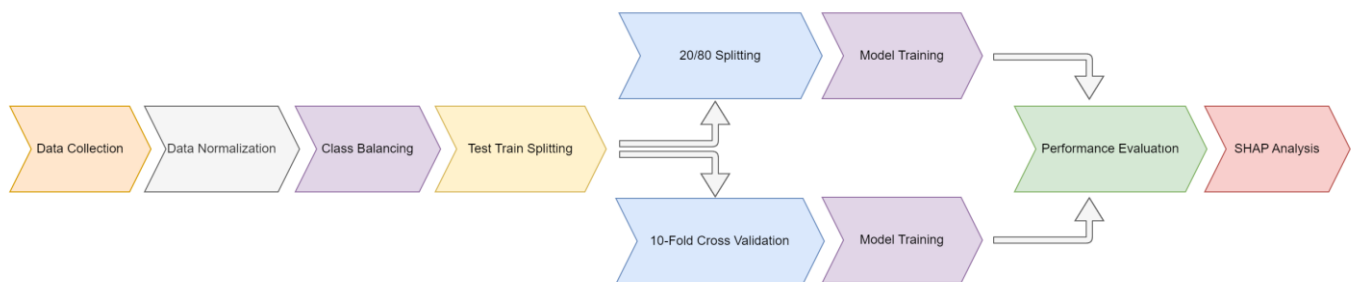


Fig.1. Workflow of the proposed system

A. Dataset

The dataset employed in this study is the Pima Indians Diabetes Database [18], sourced from the National Institute of

Diabetes and Digestive and Kidney Diseases repository. This dataset is frequently utilized in diabetes research due to its comprehensive inclusion of health-related attributes that are predictive of diabetes onset. It is widely available for research and has become a standard benchmark in diabetes prediction modeling studies.

The Pima Indians Diabetes Database comprises 768 instances, each representing an individual of Pima Indian heritage who is 21 years or older. The dataset includes a total of eight predictive features, each capturing an essential physiological or clinical variable. These features include:

1. Number of pregnancies
2. Plasma glucose concentration (measured two hours post-oral glucose tolerance test)
3. Diastolic blood pressure (mm Hg)
4. Triceps skinfold thickness (mm)
5. Serum insulin level ($\mu\text{U/mL}$)
6. Body mass index (BMI) (weight in kg/height in m^2)
7. Diabetes pedigree function (a score representing the genetic predisposition to diabetes)
8. Age (years)

In addition to the eight features, there is a binary target variable, representing whether a subject is diagnosed with diabetes (1) or not (0). The collection of these variables provides a multidimensional view of the factors associated with diabetes risk, facilitating the development of predictive models in epidemiological studies.

The data was initially collected as part of a longitudinal study aimed at understanding the prevalence of diabetes and its related risk factors within the Pima Indian population, which has a historically high prevalence of Type 2 diabetes. Variables were measured through clinical tests and self-reported metrics under controlled conditions, ensuring consistency and reliability in the dataset. This database remains a valuable resource for diabetes research, especially in exploring predictive analytics and the relationship between physiological markers and diabetes onset.

The dataset contains no null or missing values. However, based on domain knowledge [21], certain features—blood pressure, BMI, glucose, insulin, and skin thickness—have inconsistent values. Specifically, zero values for these features are inaccurate as they fall outside the normal range (see Table 1).

Table 1 presents the descriptive statistics of eight features used in the classification task. The descriptive statistics for the eight features in the diabetes dataset provide a comprehensive view of each variable's central tendency, variability, and range. The age feature has a mean of 33.24 years, with a mode of 22, suggesting a concentration of younger individuals. The median age of 29, combined with a dispersion of 0.35, indicates a relatively balanced distribution, covering a broad age range from 21 to 81. Blood pressure shows a mean of 69.11 mm Hg, with central measures closely aligned (mode of 70 and median of 72), suggesting symmetry in the distribution. However, the minimum value of 0 may indicate missing or erroneous data, given that blood pressure values typically exceed zero, and the maximum value of 122 reflects a wide range, with a dispersion of 0.28. For BMI, the mean and mode both stand at approximately 32 kg/m^2 , suggesting a balanced distribution, yet

the minimum of 0 and maximum of 67.1 kg/m^2 indicate considerable variability, which might signal the need for data cleaning or further investigation, given its dispersion of 0.25. The diabetes pedigree function, with a mean of 0.47, mode of 0.25, and median of 0.37, reveals significant variability, indicated by a high dispersion of 0.70 and a range from 0.08 to 2.42. This variability likely reflects diverse genetic or familial risks for diabetes across the dataset. Glucose levels have a mean of 120.89 mg/dL and display moderate dispersion (0.26), with values spanning from 0 to 199 mg/dL. A zero glucose value may point to missing data, as this measure is generally non-zero. Insulin levels, with a mean of 79.80 $\mu\text{U/mL}$, show a high dispersion of 1.44, and values range widely from 0 to 846, suggesting individual differences or data collection issues given the unusually high mode of 0. Pregnancies exhibit a mean of 3.85 and a mode of 1, showing a right-skewed distribution with a dispersion of 0.88 and a range from 0 to 17 pregnancies, aligning with typical reproductive variability. Lastly, Skin thickness has a mean of 20.54 mm and shows a right-skewed distribution with zeros appearing as the mode, likely indicating missing or incomplete data, and a median of 23. The range extends from 0 to 99 mm, with a relatively high dispersion of 0.78.

1) Data scaling

In ML, scaling is a crucial preprocessing step that standardizes the range of independent features to ensure consistent and effective model performance. Data often contains features with varying ranges and units, which can introduce bias during the training process, particularly when models rely on distance calculations, such as in k-NN and SVM [30]. By normalizing or standardizing features, scaling brings data into a uniform range, reducing the influence of features with larger numerical values and ensuring that each feature contributes equitably to the model. This step is especially beneficial in gradient-based algorithms, where unscaled data may lead to suboptimal convergence or slower learning as the model becomes prone to oscillating toward larger-scale features. Beyond enhancing model efficiency, scaling offers several key advantages, such as improved algorithm accuracy and increased computational speed. Models trained on scaled data demonstrate better generalization and tend to avoid overfitting by reducing variance related to magnitude differences between features [31]. As ML solutions are increasingly applied to varied datasets across domains, implementing scaling is vital for achieving reliable, reproducible, and high-performance models.

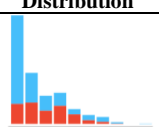

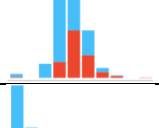
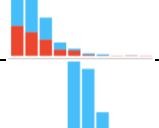

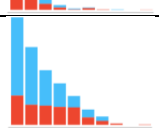
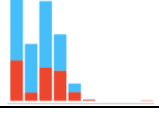
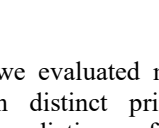
2) Class balancing

In this study, the Pima Indian Diabetes dataset, characterized by two classes—diabetic and non-diabetic individuals—exhibits a marked class imbalance, with a significantly higher number of non-diabetic cases relative to diabetic ones. Class imbalance is a critical issue in ML and statistical modeling, as it can lead to biased models that disproportionately favor the majority class, consequently compromising the predictive performance, especially for the minority class [32]. To mitigate this imbalance and enhance model efficacy, we employed SMOTE [33]. SMOTE is a sophisticated resampling method that generates synthetic examples in the minority class by interpolating between existing samples, thus balancing the dataset without simply duplicating minority instances. By

equalizing the sample size across classes, SMOTE promotes a more representative learning process, enabling the model to better capture patterns pertinent to both diabetic and non-diabetic individuals. This process not only improves classification accuracy but also ensures that the model's performance is more robust and reliable, particularly in real-

world applications where balanced prediction accuracy across classes is crucial.

TABLE I
DESCRIPTIVE STATISTICAL VALUE OF THE DATASET

Feature	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.
Age		33.24	22	29	0.35	21	81
Blood pressure		69.11	70	72	0.28	0	122
Body mass index (BMI)		31.99	32	32	0.25	0	67.1
Diabetes pedigree function		0.47	0.25	0.37	0.70	0.08	2.42
Glucose		120.89	99	117	0.26	0	199
Insulin		79.80	0	30.50	1.44	0	846
Pregnancies		3.85	1	3	0.88	0	17
Skin thickness		20.54	0	23	0.78	0	99

B. Machine Learning Models

To predict diabetes diagnoses, we evaluated multiple ML classification models, each with distinct principles and methodologies tailored to enhance predictive performance and interpretability. Below, we detail the models considered for this study.

Adaptive Boosting (AdaBoost) is a ML algorithm that combines multiple weak learners to create a strong learner [34]. It works by iteratively weighting the training data, giving more weight to misclassified instances in each iteration. This process results in a final model that is more accurate and robust than any individual weak learner.

k-Nearest Neighbors (k-NN) works by measuring the distances between a test data instance and all instances in the training dataset. It then identifies the k nearest training instances to classify the test instance [35]. The model is advantageous in scenarios with well-defined clusters and is

non-parametric, requiring minimal assumptions, making it adaptable for varying diabetes-related datasets.

Logistic Regression (LR) is based on a statistical model that predicts binary outcomes by estimating probabilities through a logistic function [36]. In the context of diabetes prediction, LR is favored for its simplicity and interpretability, particularly in assessing linear relationships between predictors and the likelihood of disease presence.

Multi-Layer Perceptron (MLP) is a neural network model that consists of multiple layers of interconnected nodes, where each node represents a neuron [37]. MLPs are characterized by their ability to capture non-linear relationships in data through backpropagation and activation functions. This model is advantageous for its flexibility in learning complex patterns, which is beneficial when diagnosing diabetes based on various patient features.

Naive Bayes (NB) assumes independence among predictor features and calculates the probability of class membership using Bayes' theorem [38]. Despite the simplicity of this

independence assumption, NB often performs well in medical contexts where conditional probabilities are informative, thus providing a quick and computationally efficient option for diabetes prediction.

Random Forest (RF) enhances the decision tree method by constructing an ensemble of multiple trees, each trained on different data subsets and feature splits [39]. This model improves generalization and reduces overfitting, making it robust for the variability present in medical data, such as diverse patient demographics and health indicators relevant to diabetes.

Support Vector Machine (SVM) attempts to find an optimal hyperplane that maximizes the margin between classes [40], effectively separating diabetic and non-diabetic instances in high-dimensional spaces. SVM is particularly suited for datasets where feature dimensions are high, offering strong performance with appropriate kernel selection, especially for complex, non-linear decision boundaries.

eXtreme Gradient Boosting (XGBoost) optimizes the Gradient Boosting algorithm with techniques such as regularization, parallelization, and efficient handling of sparse data [41]. It has demonstrated success in improving both speed and accuracy, which is beneficial in a diabetes diagnosis setting where quick, reliable predictions are essential for timely patient interventions.

Each of these models was selected for its unique properties, strengths, and applicability to the problem of diabetes diagnosis, providing a comprehensive approach to exploring the predictive capacity of different ML techniques.

C. SHapley Additive exPlanation (SHAP)

SHAP is a model-agnostic interpretability approach rooted in cooperative game theory [16]. SHAP aims to enhance model transparency by assigning importance scores to features based on their contribution to prediction, enabling researchers to gain insights into the influence of each feature. By calculating Shapley values, SHAP helps to decompose the model output in a way that considers all possible feature interactions, making it a robust tool for feature interpretability in complex ML models [15]. This subsection provides a detailed methodology on two SHAP visualizations: SHAP Feature Importance (meanSHAP) and SHAP Summary Plot (Beeswarm Plot), each of which plays a distinct role in elucidating model behavior.

1) SHAP feature importance (meanSHAP)

SHAP feature importance, commonly expressed as meanSHAP, quantifies the average effect of each feature on the model output by calculating the mean of the absolute SHAP values for each feature across all instances in the dataset. Mathematically, for a given feature f , the mean SHAP value is computed as [42]:

$$\text{meanSHAP}(f) = \frac{1}{N} \sum_{i=1}^N |\phi_{f,i}| \quad (1)$$

where N represents the number of instances, and $\phi_{f,i}$ denotes the SHAP value for feature f in instance i . This aggregation of absolute SHAP values provides a singular, intuitive metric that ranks features by their average importance, capturing both the magnitude and frequency of their impact on model predictions.

The meanSHAP metric serves as a foundational interpretability measure, offering a clear, quantitative assessment of feature relevance. By focusing on absolute values, meanSHAP accounts for both positive and negative contributions of each feature, facilitating a comprehensive view of feature importance. Unlike traditional importance measures that may overlook feature interactions or nonlinear effects, meanSHAP is based on Shapley values, which incorporate the complete range of feature interdependencies, thus offering a reliable and interpretable importance ranking that aligns closely with model behavior [15, 43].

2) SHAP summary plot (beeswarm plot)

The SHAP summary plot, also referred to as the beeswarm plot, visually represents the distribution of SHAP values across all instances for each feature. In this plot, each feature is displayed along the vertical axis, while the horizontal axis represents the range of SHAP values, indicating the magnitude and direction of feature impact on model predictions. Each point on the plot corresponds to the SHAP value for an individual instance, with the points color-coded to represent the feature values, typically on a blue-to-red gradient, where red denotes higher feature values and blue denotes lower ones. The beeswarm plot is particularly useful in examining how each feature affects the model output, providing insights into the distribution of feature impact. For instance, a wide horizontal spread of points suggests that a feature has variable importance across instances, while clustering around zero indicates minimal influence. Additionally, by examining the color gradients, researchers can infer the relationship between feature values and their corresponding SHAP values, revealing patterns such as whether higher feature values lead to increased or decreased predictions [15, 44].

D. Performance Evaluation Metrics

Evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix are employed to assess the models. These statistical measures are derived from ground truth values, namely True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The calculations for accuracy, precision, recall, and F1-score can be found in Eqs. (2), (3), (4), and (5), respectively.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

These variables are calculated using the confusion matrix, which is a tabular representation showing the values of the actual outcome classes and the predicted outcome classes on the testing dataset as shown in Fig. 2.

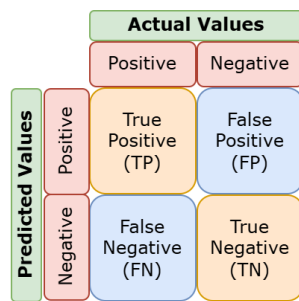


Fig. 2. Confusion matrix

III. RESULTS

This section provides a comprehensive analysis of the study's findings, including the evaluation of eight ML models.

TABLE II
CONFUSION MATRICES OF ML MODELS USING TEST/TRAIN SPLIT

AdaBoost	Actual	Predicted		NB	Actual	Predicted		
		Diabetic	Non-Diabetic			Diabetic	Non-Diabetic	
		Diabetic	70			29	Diabetic	79
		Non-Diabetic	21	80		Non-Diabetic	26	75
k-NN	Actual	Predicted		RF	Actual	Predicted		
		Diabetic	Non-Diabetic			Diabetic	Non-Diabetic	
		Diabetic	75			27	Diabetic	76
		Non-Diabetic	12	89		Non-Diabetic	13	88
LR	Actual	Predicted		SVM	Actual	Predicted		
		Diabetic	Non-Diabetic			Diabetic	Non-Diabetic	
		Diabetic	72			27	Diabetic	71
		Non-Diabetic	23	78		Non-Diabetic	12	89
MLP	Actual	Predicted		XGBoost	Actual	Predicted		
		Diabetic	Non-Diabetic			Diabetic	Non-Diabetic	
		Diabetic	68			31	Diabetic	71
		Non-Diabetic	14	87		Non-Diabetic	17	84

In Table 2, the AdaBoost algorithm demonstrates a balanced performance in terms of sensitivity and specificity, producing 70 true positives (TP) and 80 true negatives (TN), with 21 false positives (FP) and 29 false negatives (FN). The NB model performs well in identifying diabetic individuals with 79 true positives, though it shows a slight decline in specificity with 75 true negatives. The k-NN algorithm exhibits high specificity, achieving 89 true negatives and only 12 false positives, but relatively lower sensitivity with 27 false negatives. The RF model delivers balanced performance, achieving 76 true positives and 88 true negatives. LR provides an acceptable

balance between sensitivity and specificity, with 72 true positives and 23 false positives, but yields 27 false negatives, indicating limitations in identifying diabetic cases. The SVM model shows a comparable performance to LR but stands out with higher specificity, achieving 89 true negatives. The MLP model, with 68 true positives and 87 true negatives, demonstrates slightly lower sensitivity and yields 31 false negatives. Finally, the XGBoost model strikes a balance between sensitivity and specificity, achieving 71 true positives and 84 true negatives.

Additionally, the application of XAI method, specifically SHAP, is detailed to elucidate the models' internal mechanisms and decision-making processes, enhancing interpretability and transparency. To evaluate model performance, the dataset was divided into training (80%) and testing (20%) sets and applied a 10-fold cross-validation.

Table 2 presents confusion matrices for several ML models tested to predict diabetes in women of Pima Indian heritage. Each confusion matrix displays the performance of a specific model in classifying individuals as either having diabetes (positive) or not having diabetes (negative). The rows of the matrix represent the true class (actual diabetes status), while the columns represent the predicted class assigned by the model.

Table 3 provides a performance comparison of eight ML models based on accuracy, precision, recall, and F1-score using two approaches: the test/train split and 10-fold cross-validation.

For the test/train split (20/80), the RF model stands out with the highest performance metrics: an accuracy of 0.82, precision of 0.82, recall of 0.82, and F1-score of 0.83. This indicates that RF is particularly adept at capturing complex, non-linear relationships in the dataset, resulting in balanced and robust predictions. Following RF, the k-NN model demonstrates competitive performance, achieving an accuracy of 0.81, precision of 0.81, recall of 0.80, and F1-score of 0.82. Similarly, SVM also shows strong results with an accuracy of 0.80, precision of 0.81, recall of 0.80, and F1-score of 0.82, indicating its effectiveness in handling high-dimensional data and distinguishing between diabetic and non-diabetic cases. In contrast, models such as AdaBoost, LR, and NB exhibit moderate performance, with accuracy and F1-scores ranging between 0.75 and 0.77. While these models offer interpretability and computational efficiency, their lower sensitivity suggests potential limitations in identifying diabetic cases.

TABLE III
COMPARATIVE PERFORMANCE ANALYSIS OF ML MODELS FOR
TEST/TRAIN AND CROSS-VALIDATION

Data Splitting	Model	Accuracy	Precision	Recall	F1-score
Test/Train (20/80)	AdaBoost	0.75	0.75	0.75	0.76
	k-NN	0.81	0.81	0.80	0.82
	LR	0.75	0.75	0.75	0.76
	MLP	0.78	0.78	0.77	0.79
	NB	0.77	0.77	0.77	0.77
	RF	0.82	0.82	0.82	0.83
	SVM	0.80	0.81	0.80	0.82
	XGBoost	0.78	0.78	0.77	0.79
Cross-validation (10-fold)	AdaBoost	0.79	0.79	0.79	0.79
	k-NN	0.81	0.82	0.81	0.82
	LR	0.76	0.76	0.76	0.75
	MLP	0.80	0.80	0.80	0.81
	NB	0.74	0.74	0.74	0.72
	RF	0.83	0.83	0.83	0.84
	SVM	0.79	0.79	0.79	0.80
	XGBoost	0.81	0.81	0.81	0.81

Cross-validation results further reinforce RF's superior performance, with an accuracy of 0.83, precision of 0.83, recall of 0.83, and F1-score of 0.84. The consistency of RF's performance across both evaluation methods underscores its reliability and robustness in diabetes prediction. The k-NN model also performs exceptionally well, with an accuracy of 0.81, precision of 0.82, recall of 0.81, and F1-score of 0.82. Its strong results are indicative of its effectiveness in leveraging the local relationships among data points, particularly after the dataset has been balanced and scaled. The XGBoost model achieves an accuracy of 0.81 and maintains precision, recall, and F1-scores at 0.81 as well. The MLP model achieves comparable results, with accuracy, precision, and recall scores of 0.80, and an F1-score of 0.81. The slight improvement in F1-score compared to the test/train split suggests that MLP benefits from the diversified training subsets in cross-validation, allowing it to better generalize its predictions. The SVM model

also performs well, achieving an accuracy and precision of 0.79, with recall and F1-scores matching at 0.79 and 0.80, respectively. SVM's performance underscores its ability to construct decision boundaries effectively, especially in high-dimensional spaces, though its metrics are slightly lower compared to RF and k-NN. The AdaBoost model demonstrates moderate performance during cross-validation, with accuracy, precision, recall, and F1-scores all at 0.79. While its performance is slightly better than that in the test/train split, it still lags behind ensemble methods like RF and XGBoost in capturing the dataset's complexity. The LR model maintains consistent but relatively lower results compared to more advanced models, achieving an accuracy and precision of 0.76, recall of 0.76, and an F1-score of 0.75. Its simplicity and interpretability remain its key advantages, though its limited ability to handle non-linear relationships constrains its performance. The NB model exhibits the weakest performance among all models, with accuracy, precision, and recall scores at 0.74, and an F1-score of 0.72. Its simplistic assumption of feature independence may not align well with the real-world interactions within the dataset, leading to suboptimal results in distinguishing diabetic and non-diabetic cases.

Figs. 3 and 4 visually compare the performance of various ML models under two distinct evaluation methods: the test/train split and 10-fold cross-validation.

Fig. 3 provides a comparative visual analysis, showing that models such as RF, k-NN, and SVM consistently achieve superior performance across metrics, demonstrating their robustness and suitability for diabetes prediction. Simpler models, like LR and NB, exhibit moderate performance, which, while computationally efficient, reveal limitations in handling the dataset's complexity.

Fig. 4, on the other hand, presents the performance results obtained through 10-fold cross-validation. The results further validate the consistency and robustness of RF and k-NN models, which maintain high accuracy and F1-scores, underscoring their reliability in diabetes prediction tasks. Additionally, cross-validation highlights improvements in performance for models such as AdaBoost, MLP, and XGBoost.

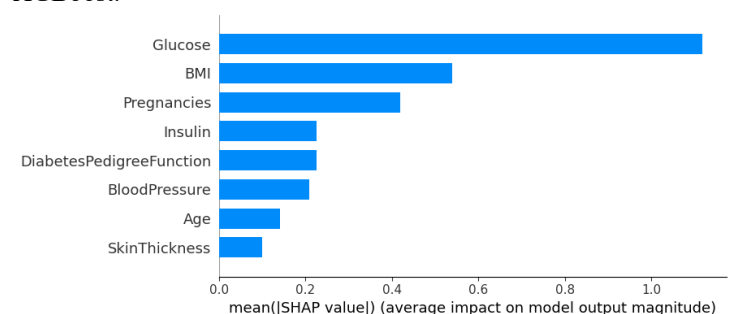


Fig. 5. Importance ranking of the model prediction features

The ranking indicates that glucose level is the most critical feature, suggesting that higher glucose levels strongly correlate with diabetes likelihood. This aligns with clinical expectations, as elevated glucose is a primary indicator of diabetes. BMI ranks as the second most important feature, highlighting the strong association between obesity and diabetes risk. This finding underscores the importance of body weight relative to

height in assessing metabolic health. Pregnancy is also identified as a critical factor, likely due to the increased metabolic stress during pregnancy, which can elevate the risk of developing diabetes. Insulin levels demonstrate a considerable impact on the model's predictions, capturing the intricate relationship between insulin metabolism and diabetes. Less influential features include diabetes pedigree function, blood pressure, age, and skin thickness, although they still contribute to the predictive model. The relatively lower importance of these features may indicate that while they provide valuable context, they are not as directly correlated with diabetes onset as glucose levels and BMI.

Fig. 6. is a visual representation that shows the importance of various features in predicting diabetes, using SHAP values. In this plot, each feature is displayed on the y-axis, ordered by its significance in the prediction, with the SHAP value distribution across instances plotted horizontally on the x-axis. The SHAP values indicate the magnitude and direction (positive or negative impact) of each feature on the model's prediction outcome for diabetes risk. Each dot represents an individual instance, with the color gradient (from blue to red)

corresponding to the feature value's magnitude, where red signifies higher values and blue lower ones.

Glucose often appears as the most significant predictor in diabetes-related models. Higher glucose levels, represented by red-colored points on the positive side of the SHAP values, usually increase the prediction probability for diabetes, reflecting the well-established clinical link between elevated blood glucose and diabetes risk. BMI, which indicates body weight relative to height, typically ranks high in importance. Higher BMI values (marked in red) are often associated with a higher probability of diabetes due to the strong association between obesity and diabetes risk. Pregnancies, which represent the number of times a patient has been pregnant, play an important role. Higher values (red) generally increase the SHAP values, indicating a higher likelihood of diabetes, potentially due to physiological changes associated with multiple pregnancies. Insulin demonstrates a moderate level of importance. Higher insulin levels (red) are positively associated with increased diabetes prediction scores, reflecting the body's compensatory response to insulin resistance. Lower values (blue), however, have a varying impact, suggesting a more complex relationship.

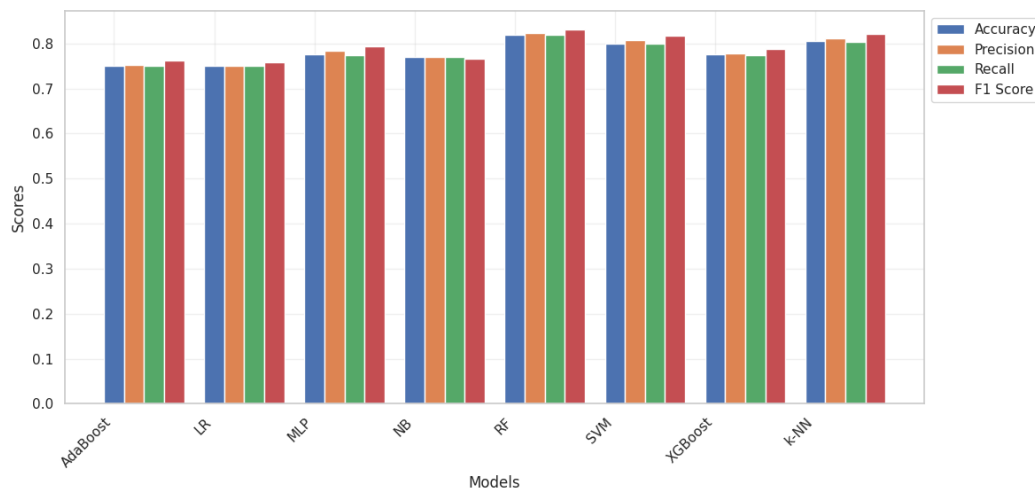


Fig. 3. Visualization of the performance metrics results for ML models using test/train split

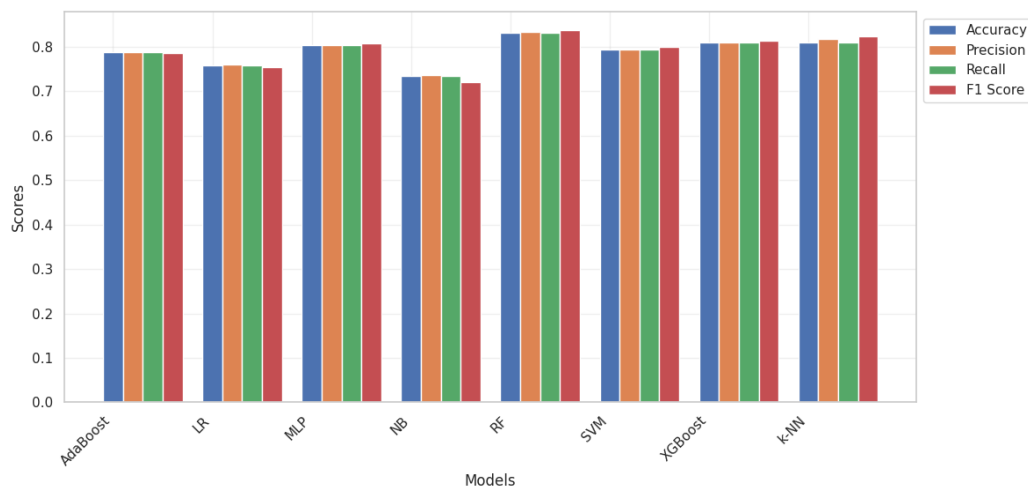


Fig. 4. Visualization of the performance metrics results for ML models using 10-fold cross-validation

RF was selected for SHAP analysis due to its superior predictive performance. Fig. 5 illustrates the significance of

various features in predicting diabetes using SHAP values. In this plot, features are ranked based on their average SHAP values, highlighting their relative impact on the model's predictions.

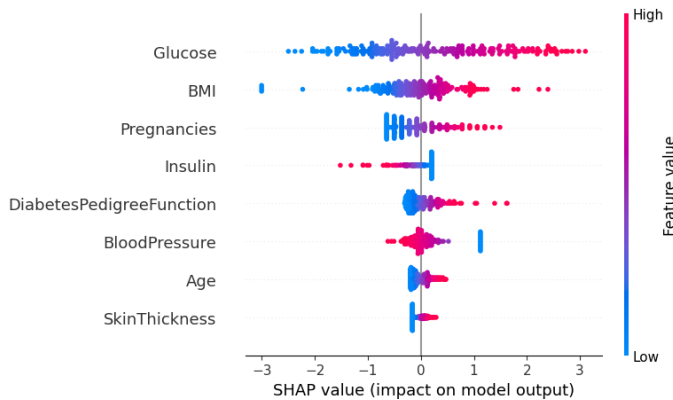


Fig. 6. SHAP summary plot

Diabetes pedigree function, blood pressure and age also contribute to the model, albeit with less impact compared to glucose and BMI. Diabetes pedigree function quantifies the genetic predisposition to diabetes. High values (red), indicating a stronger family history, consistently increase the SHAP values, reinforcing the heritability of diabetes risk. Low values (blue) contribute negatively, reducing the prediction probability. Older individuals (red dots for age) tend to show a higher likelihood of diabetes due to age-related declines in

metabolic health, while lower ages (blue) reduce the risk. Similarly, higher blood pressure values (red) are modestly associated with increased predictions, consistent with the link between hypertension and diabetes risk. Skin thickness, the least impactful feature, shows a nuanced pattern.

IV. DISCUSSIONS

This study reveals critical insights into the application of ML models for predicting diabetes, particularly highlighting the efficacy of ensemble methods and the importance of model interpretability. The most significant finding is the superior performance of the RF model, which achieved an accuracy of 82% and an F1-score of 0.83 in the test/train split evaluation, and an even higher accuracy of 83% and an F1-score of 0.84 in the 10-fold cross-validation (Table 3, Fig. 3, and Fig. 4). These results underscore RF's ability to capture complex, non-linear relationships within the Pima Indians Diabetes Database. Furthermore, the SHAP analysis identified glucose, BMI, pregnancies, and insulin as the most influential predictors, aligning with established clinical markers of diabetes (Fig. 5 and Fig. 6). The prominence of glucose levels, as indicated by the highest mean SHAP value, reinforces its well-documented role as a primary indicator of diabetes risk, while the importance of BMI reflects the known association between obesity and metabolic disorders.

TABLE IV
COMPARISON OF DIABETES PREDICTION STUDIES

Study	Models	Best Model	Accuracy
Verma and Khatoon [19]	LR, SVM, k-NN, RF	RF	80.08%
Xie [20]	k-NN, LR, SVM, RF	LR	79.13%
Chang et al. [21]	NB, RF, J48 DT	RF	79.57%
Sahoo et al. [22]	NB, LR, DT, RF, SVM, XGBoost	LR	74.03%
You and Kang [23]	SVM, DT	SVM	70.40%
Ashour et al. [24]	FNN, CNN	FNN	82%
Akyol and Şen [25]	AdaBoost, Gradient Boosted Trees, RF	AdaBoost	73.88%
Reza et al. [26]	Stacking Ensemble (Classical + Deep)	Stacking Ensemble (Deep NN)	75.03% (train/test), 77.10% (5-fold cross-validation)
Pyne and Chakraborty [27]	ANN	ANN	80.79%
Jain et al. [28]	DT, RF, SVM, NB	RF	79.08%
Karatsiolis and Schizas [29]	Modified SVM with RBF and Polynomial Kernel	Modified SVM	82.2%
This Study	AdaBoost, k-NN, LR, MLP, NB, RF, SVM, XGBoost	RF	82% (Train/Test), 83% (10-fold cross-validation)

The findings of this study contribute to a growing body of literature that evaluates ML models for diabetes prediction, as summarized in Table 4. Our results demonstrate that the RF model, particularly when combined with SHAP analysis, outperforms previously reported as applied to the same dataset.

For instance, Akyol and Şen [25] reported an accuracy of 73.88% using AdaBoost, while Verma and Khatoon [19] achieved 80.08% accuracy with RF. Our study's RF model surpasses these benchmarks, achieving 83% accuracy in the 10-fold cross-validation. This improvement can be attributed to our

use of the SMOTE for class balancing and the integration of SHAP values for enhanced interpretability. However, our results are comparable to those of Xie [20], who reported 79.13% accuracy with LR, and Jain et al. [28], who achieved 79.08% accuracy with Random Forest. These variations highlight the influence of different preprocessing techniques and model configurations on predictive performance.

A key strength of this study lies in its integration of XAI methods, specifically SHAP, to enhance model interpretability. While many previous studies have focused on predictive accuracy, the clinical applicability of ML models hinges on their transparency and the ability to provide actionable insights. By incorporating SHAP values, we provide a clear, quantitative assessment of feature importance, bridging the gap between predictive accuracy and clinical utility. This approach not only elucidates the model's decision-making process but also builds trust among healthcare providers by offering a deeper understanding of the factors driving predictions. Additionally, the use of SMOTE to address class imbalance ensures that the models are trained on a representative dataset, thereby enhancing their robustness and reliability in real-world scenarios.

Despite the strengths, this study has certain limitations. The reliance on the Pima Indians Diabetes Database, while a widely used benchmark, may introduce biases related to the specific population studied, potentially limiting the generalizability of the findings to other ethnic groups. Additionally, the study identified inconsistencies in the dataset, particularly in insulin, skin thickness, and blood pressure values, which could affect model performance. Although SMOTE was employed to mitigate class imbalance, the inherent limitations of the dataset cannot be entirely overcome. Furthermore, while the SHAP analysis enhances interpretability, it is essential to acknowledge that model interpretability is a complex and evolving field, and the explanations provided by SHAP, while valuable, may not fully capture the intricate decision-making processes of the models.

This study makes a significant contribution to the field of diabetes prediction by demonstrating the effectiveness of advanced ML models, particularly RF and k-NN, and by enhancing model interpretability through SHAP analysis. The findings underscore the importance of integrating XAI methods in healthcare applications to foster trust and facilitate clinical adoption. Future research should focus on validating these models with more diverse datasets and refining feature engineering to address the identified inconsistencies. Additionally, exploring the integration of other XAI techniques and investigating the longitudinal performance of these models in real-world clinical settings could further enhance their applicability. The insights gained from this study pave the way for developing more transparent, reliable, and clinically relevant predictive tools for diabetes, ultimately contributing to improved patient outcomes and more effective healthcare strategies. The study's findings open new research avenues, particularly in the development of personalized medicine approaches, where individual risk factors can be evaluated with greater precision and transparency.

V. CONCLUSION

This study demonstrates the successful integration of ML models and XAI techniques to enhance the predictive accuracy and interpretability of diabetes diagnosis using the Pima Indians Diabetes Database. The RF model emerged as the most effective classifier, achieving an accuracy of 83% and an F1-score of 0.84 in 10-fold cross-validation, underscoring its capability to model complex, non-linear relationships within the dataset. The incorporation of SHAP values provided critical insights into the contributions of various predictors, with glucose, BMI, pregnancies, and insulin identified as the most influential features. These findings align with established clinical markers of diabetes, affirming the validity of the model's decision-making process. This study, therefore, not only bridges the gap between predictive accuracy and clinical transparency but also provides a methodological framework for leveraging XAI to enhance the interpretability of ML models in healthcare. The incorporation of the SMOTE for class balancing further contributed to the robustness of the models, ensuring their reliability across diverse datasets and real-world scenarios.

The contributions of this research are multifold, extending the frontier of knowledge in both data mining and artificial intelligence applications within the healthcare domain. The integration of SHAP values into the diabetes prediction process is demonstrated to enhance transparency and trustworthiness in AI systems, facilitating their adoption in clinical practice. However, this study acknowledges its limitations, including the reliance on a single dataset, which may constrain the generalizability of the findings to other populations and clinical settings. Additionally, while SHAP analysis enhances interpretability, the inherent complexities of ML models mean that complete transparency remains an elusive goal. Future research should endeavor to validate these models with more diverse datasets and explore the integration of additional XAI techniques to further enhance model interpretability. A speculative, yet promising, direction could involve the development of longitudinal studies that track model performance and interpretability over time, providing insights into the dynamic nature of diabetes risk factors.

REFERENCES

- [1] P. David, S. Singh, R. Ankar, "A comprehensive overview of skin complications in diabetes and their prevention," *Cureus*, vol. 15, no. 5, p. e38961, 2023.
- [2] A. F. Walker et al., "Interventions to address global inequity in diabetes: international progress," *Lancet*, vol. 402, no. 10397, 2023, pp. 250-264.
- [3] M. Zakir et al., "Cardiovascular complications of diabetes: From microvascular to macrovascular pathways," *Cureus*, vol. 15, no. 9, p. e45835, 2023.
- [4] A. Avogaro, M. Rigato, E. di Brino, D. Bianco, I. Gianotto, G. Brusaporco, "The socio-environmental determinants of diabetes and their consequences," *Acta Diabetol.*, vol. 61, no. 10, 2024, pp. 1205-1210.
- [5] S. Gowthami, R. Venkata Siva Reddy, M. R. Ahmed, "Exploring the effectiveness of machine learning algorithms for early detection of Type-2 Diabetes Mellitus," *Measur. Sens.*, vol. 31, no. 100983, p. 100983, 2024.
- [6] A. A. L. Ahmad, A. A. Mohamed, "Artificial intelligence and machine learning techniques in the diagnosis of type I diabetes: Case studies," in

- Studies in Computational Intelligence, Singapore: Springer Nature Singapore, 2024, pp. 289-302.
- [7] T. Althobaiti, S. Althobaiti, M. M. Selim, "An optimized diabetes mellitus detection model for improved prediction of accuracy and clinical decision-making," *Alex. Eng. J.*, vol. 94, 2024, pp. 311-324.
- [8] R. F. Albadri, S. M. Awad, A. S. Hameed, T. H. Mandeel, R. A. Jabbar, "A diabetes prediction model using hybrid machine learning algorithm," *Math. Model. Eng. Probl.*, vol. 11, no. 8, 2024, pp. 2119-2126.
- [9] S. Buyrukoğlu, A. Akbaş, "Machine learning based early prediction of type 2 diabetes: A new hybrid feature selection approach using Correlation Matrix with Heatmap and SFS," *Balkan Journal of Electrical and Computer Engineering*, vol. 10, no. 2, 2022, pp. 110-117.
- [10] A. Adadi, M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, 2018, pp. 52138-52160.
- [11] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo, "Benchmarking and survey of explanation methods for black box models," *Data Min. Knowl. Discov.*, vol. 37, no. 5, 2023, pp. 1719-1778.
- [12] A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, 2020, pp. 82-115.
- [13] W. Ding, M. Abdel-Basset, H. Hawash, A. M. Ali, "Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey," *Inf. Sci. (N.Y.)*, vol. 615, 2022, pp. 238-292.
- [14] V. Hassija et al., "Interpreting black-box models: A review on explainable Artificial Intelligence," *Cognit. Comput.*, vol. 16, no. 1, 2024, pp. 45-74.
- [15] S. Lundberg, S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et al., Eds. Curran Associates, Inc., 2017.
- [16] L. S. Shapley, "Stochastic games," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 39, no. 10, 1953, pp. 1095-1100.
- [17] K. Aliyeva, N. Mehdiyev, "Uncertainty-aware multi-criteria decision analysis for evaluation of explainable artificial intelligence methods: A use case from the healthcare domain," *Information Sciences*, vol. 657, no. 119987, p. 119987, 2024.
- [18] Kaggle Dataset, "Pima Indian Diabetes Database," 2017. [Online]. Available: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>
- [19] P. Verma, A. Khatoun, "Data mining applications in healthcare: A comparative analysis of classification techniques for diabetes diagnosis using the PIMA Indian diabetes dataset," in *2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM)*, 2024.
- [20] L. Xie, "Pima Indian diabetes database and machine learning models for diabetes prediction," *Highlights in Science, Engineering and Technology*, vol. 88, 2024, pp. 97-103.
- [21] V. Chang, J. Bailey, Q. A. Xu, Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput. Appl.*, vol. 35, no. 22, 2022, pp. 1-17.
- [22] S. Sahoo, T. Mitra, A. K. Mohanty, B. J. R. Sahoo, and S. Rath, "Diabetes prediction: A study of various classification based data mining techniques," *International Journal of Computer Science and Informatics*, vol. 4, no. 3, 2022, pp. 1-13.
- [23] S. You, M. Kang, "A Study on Methods to Prevent Pima Indians Diabetes using SVM," *Korean Journal of Artificial Intelligence*, vol. 8, no. 2, 2020, pp. 7-10.
- [24] A. F. Ashour, M. M. Fouda, Z. M. Fadlullah, M. I. Ibrahim, "Optimized neural networks for diabetes classification using Pima Indians diabetes database," in *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*, 2024.
- [25] K. Akyol, B. Şen, "Diabetes mellitus data classification by cascading of feature selection methods and ensemble learning algorithms," *Int. J. Mod. Educ. Comput. Sci.*, vol. 6, 2018, pp. 10-16.
- [26] M. S. Reza, R. Amin, R. Yasmin, W. Kulsum, S. Ruhi, "Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data," *Heliyon*, vol. 10, p. e24536, 2024.
- [27] A. Pyne, B. Chakraborty, "Artificial Neural Network based approach to Diabetes Prediction using Pima Indians Diabetes Dataset," in *2023 International Conference on Control, Automation and Diagnosis (ICCAD)*, Rome, Italy, 2023.
- [28] A. V. Jain, S. Shukla, N. Khare, "Analysis of various data imputation techniques for diabetes classification on PIMA dataset," in *2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 2024, pp. 1-6.
- [29] S. Karatsiolis, C. N. Schizas, "Region based Support Vector Machine algorithm for medical diagnosis on Pima Indian Diabetes dataset," in *2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE)*, 2012.
- [30] M. Bilal, G. Ali, M. W. Iqbal, M. Anwar, M. S. A. Malik, R. A. Kadir, "Auto-Prep: Efficient and Automated Data Preprocessing Pipeline," *IEEE Access*, vol. 10, 2022, pp. 107764-107784.
- [31] L. B. V. de Amorim, G. D. C. Cavalcanti, R. M. O. Cruz, "The choice of scaling technique matters for classification performance," *Appl. Soft Comput.*, vol. 133, no. 109924, p. 109924, 2023.
- [32] A. D. Amiruddin, F. M. Muharam, M. H. Ismail, N. P. Tan, M. F. Ismail, "Synthetic Minority Over-sampling TEchnique (SMOTE) and Logistic Model Tree (LMT)-Adaptive Boosting algorithms for classifying imbalanced datasets of nutrient and chlorophyll sufficiency levels of oil palm (*Elaeis guineensis*) using spectroradiometers and unmanned aerial vehicles," *Comput. Electron. Agric.*, vol. 193, no. 106646, p. 106646, 2022.
- [33] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, 2002, pp. 321-357.
- [34] T. Yılmaz, "Microwave spectroscopy based classification of rat hepatic tissues: On the significance of dataset," *Balkan Journal of Electrical and Computer Engineering*, vol. 8, no. 4, 2020, pp. 307-313.
- [35] T. Tülgar, A. Haydar, İ. Erşan, "A distributed K Nearest Neighbor classifier for Big Data," *Balkan Journal of Electrical and Computer Engineering*, vol. 6, no. 2, 2018, pp. 105-111.
- [36] T. Pala, A. Y. Camurcu, "Design of decision support system in the metastatic colorectal cancer data set and its application," *Balkan Journal of Electrical and Computer Engineering*, vol. 4, no. 1, 2016, pp. 12-16.
- [37] C. Greco, P. Pace, S. Basagni, G. Fortino, "Jamming detection at the edge of drone networks using Multi-layer Perceptrons and Decision Trees," *Appl. Soft Comput.*, vol. 111, no. 107806, p. 107806, 2021.
- [38] İ. Kırbaş, A. Çifci, "Machine Learning-Based Rice Grain Classification Through Numerical Feature Extraction from Rice Image Data," in *9th International Zeugma Conference on Scientific Research*. Gaziantep, Türkiye, 2023.
- [39] A. Çifci, M. İlkuçar, "Analysis of window sizes in prediction of daily COVID-19 cases using machine learning models," *International Journal of Mechatronics, Electrical and Computer Technology (IJMEC)*, vol. 12, no. 45, 2022, pp. 5208-5217.
- [40] G. Bilgin, A. Çifci, "Eritematöz skuamöz hastalıkların teşhisinde makine öğrenme algoritmaları performanslarının değerlendirilmesi," *Journal of Intelligent Systems: Theory and Applications*, vol. 4, no. 2, 2021, pp. 195-202.
- [41] C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif. Intell. Rev.*, vol. 54, no. 3, 2021, pp. 1937-1967.
- [42] C. Molnar, *Interpretable machine learning: a guide for making black box models interpretable*. Morisville, North Carolina: Lulu, 2019.
- [43] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, 2020, pp. 56-67.
- [44] S. M. Lundberg, G. G. Erion, S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv [cs.LG]*, 2018.

BIOGRAPHIES

İsmail Kırbaş completed his Ph.D. in Electronics and Computer Education at Sakarya University and currently serves as the Dean of the Faculty of Computer and Informatics at Burdur Mehmet Akif Ersoy University, where he also holds the position of Director at the HAYTEK Collaborative Application Research Center. Dr. Kırbaş's research focuses on wireless body area network protocols, artificial intelligence, the Internet of Things,



and machine learning. With a strong academic background, including a bachelor's and master's degree from Kocaeli University, he has led a variety of innovative projects throughout his career.

Dr. Kırbaş's work significantly contributes to digital transformation across sectors such as agriculture and healthcare, utilizing artificial intelligence and data mining techniques. Actively involved as both a researcher and project leader in numerous national and international projects, Dr. Kırbaş has a robust portfolio of scientific publications and patents, underscoring his impactful presence in the field.



Ahmet Çifci received the B.S., M.S. and Ph.D. degrees in Electrical and Electronics Engineering from Sakarya University, Türkiye, in 2007, 2011 and 2015, respectively.

From 2009 to 2016, he was a Lecturer with the Vocational School of Burdur within Burdur Mehmet Akif Ersoy University.

Since 2016, he has been an Assistant

Professor with the Department of Electrical-Electronics Engineering, Faculty of Engineering and Architecture, Burdur Mehmet Akif Ersoy University, Burdur, Türkiye. Throughout his career, Dr. Çifci has held various administrative roles, including Senate Membership, Department Head of Electrical-Electronics Engineering, Vice-Director of the Institute of Science, and Vice-Dean. He currently serves as the Vice-Dean of the Faculty of Engineering and Architecture. Additionally, he is the Deputy Director at the Digital Technologies in Livestock Sector Joint Application and Research Center at Burdur Mehmet Akif Ersoy University. He has authored and co-authored numerous scholarly papers in peer-reviewed journals, books, and national/international conferences. His academic and research interests include high voltage engineering, power systems, and artificial intelligence.

Dr. Çifci is a member of the Chamber of Electrical Engineers of Türkiye.

Heart Attack Classification with a Machine Learning Approach Based on the Random Forest Algorithm

Suleyman Dal and Necmettin Sezgin

Abstract— Heart attack diagnosis delays constitute a critical health problem that increases the risk of mortality. Timely and accurate identification of cardiac events is therefore essential to improve patient outcomes and reduce preventable deaths. This study aims to develop a random forest based classification model using the Heart Disease Classification dataset published on the Kaggle platform to support early diagnosis. This dataset consists of 1319 samples and 8 demographic, clinical and biochemical features for the diagnosis of heart disease. To evaluate the model's reliability and generalizability, a 10-fold cross-validation technique was employed. Through this method, each data instance contributed to both training and testing phases, enabling a more stable and robust performance assessment. This approach also reduced the risk of overfitting and ensured more representative evaluation metrics. The performance of the model was evaluated with ROC curve, training-validation curves, confusion matrix. In the evaluation process, especially in Fold 6, 100% accuracy, precision, recall and F1 score were obtained and it was revealed that the model showed superior performance in the classification task. In addition, as a result of the feature importance analysis, it was determined that troponin, potassium (kcm) and age variables came to the forefront in the decision process. This study aims to fill an important gap in the literature in terms of both strong classification performance and interpretability in the field of machine learning models for heart attack diagnosis.

Index Terms— Heart Attack Classification, Machine Learning, Random Forest Algorithm, Clinical Decision Support Systems


I. INTRODUCTION

THE HEART is a vital organ that systematically pumps blood to maintain the body's life functions. The heart is the most basic component of the cardiovascular system, together with arteries, veins and capillaries, which are involved in the


efficient transport of oxygen and nutrients to the tissues [1]. Heart diseases are among the common health problems worldwide with high mortality risk. Among these diseases, heart attacks are responsible for more than 80% of all cardiovascular disease (CVD)-related deaths [1, 2]. Risks that trigger the occurrence of CVD include factors such as high cholesterol and blood pressure, sedentary lifestyle, age, genetic predisposition, obesity, diabetes, stress, excessive alcohol and smoking [3]. Some risk factors can be limited by lifestyle interventions such as smoking cessation, body weight control, regular physical activity and stress management. Diagnostic and imaging techniques such as medical history, physical evaluation, electrocardiography, echocardiography, cardiac magnetic resonance imaging and various blood analyses are widely used in the diagnosis of heart diseases. In the treatment of these diseases, methods such as lifestyle modifications, pharmacological treatment methods, angioplasty, coronary artery bypass surgery and pacemakers are applied by specialist physicians [4, 5].

The risk of death can be significantly prevented by early diagnosis of heart diseases and effective treatment options [6, 7]. In this context, the integration of developing technology into health systems is of vital importance. The use of data analysis-based methods effectively supports the medical decision-making process of specialised physicians in common diseases with high mortality rates such as cardiovascular diseases. In this context, machine learning (ML) methods have been widely embraced by researchers. In recent years, the development of ML methods has become an important auxiliary method by being actively used in the health sector as in almost every field [8, 9]. With the effective analysis of large data sets in the field of health, it can make significant contributions to disease prediction and treatment processes. These methods enable the development of clinical decision support models, especially by performing beyond human intuition. In this context, ML methods strongly support medical professionals with high accuracy in critical processes such as early diagnosis of heart diseases, patient risk classification and treatment response prediction. In this respect, random forest, which is one of the ML algorithms, is an effective method widely preferred in the field of health, especially in disease classification and risk prediction studies. This algorithm, which works on the principle of multiple decision trees, produces successful results in

Süleyman Dal, is with the Energy Coordination Office, Rectorate of Batman University, Batman, Türkiye States, (e-mail: suleyman.dal@batman.edu).

 <https://orcid.org/0000-0002-4564-8076>

Necmettin Sezgin, is with the Department of Electrical and Electronics Engineering, Batman University, Batman, Türkiye, (e-mail: necmettin.sezgin@batman.edu).

 <https://orcid.org/0000-0002-4893-6014>

Manuscript received May 5, 2025; accepted May 15, 2025.

DOI: [10.17694/bajece.1691905](https://doi.org/10.17694/bajece.1691905)

classification problems with medical data thanks to its low error tolerance [10, 11].

In recent years, studies based on ML for heart attack diagnosis have been increasing and the performances of the models developed in this field have been extensively examined in the literature. In their study, Natarajan et al. use Firefly algorithm-assisted feature selection and ensemble learning methods such as Stacking and Voting to identify important attributes related to heart disease and improve prediction accuracy. In the applications on the Z-Alizadeh Sani dataset, the Stacking method performed successfully with an accuracy rate of 86.79% [12]. In another study, Jabbar et al. propose an effective classification method by combining the K-nearest neighbour (KNN) algorithm and genetic algorithm to improve the accuracy of heart disease diagnosis. Experimental results show that the proposed method significantly improves the classification accuracy in heart disease diagnosis [13]. In Enad and Mohammed's study, a comprehensive analysis is performed on the Cleveland dataset using quantum machine learning (QML) methods to support early and accurate diagnosis of heart diseases. In the study, quantum-based approaches (QNN, QSVM, Bagging-QSVM) were compared with traditional classifiers (SVM, ANN) after preprocessing and feature selection; in particular, the Bagging-QSVM model achieved the highest accuracy with 100% success in all key performance measures [2]. In another study, El-Sofany compared ten different machine learning algorithms on feature sets generated by three different feature selection methods (Chi-square, ANOVA and Mutual Information) aiming to accurately predict heart diseases at an early stage. The unbalanced data problem was overcome with the SMOTE method, and the XGBoost algorithm achieved the most successful results with the SF-2 feature set with superior performance values such as 97.57% accuracy, 96.61% sensitivity and 95.00% precision [1].

The main objective of this study is to develop a machine learning model that provides highly accurate results for early prediction of heart attack risk. In this context, using the Heart Disease Classification Dataset published on the Kaggle platform, a clinical decision support mechanism that can classify individuals' susceptibility levels to heart disease has been created with the Random Forest algorithm. The main objective of the model is to provide a reliable prediction system that can contribute to early diagnosis processes in clinical settings by learning meaningful patterns from patient data. In this context, it both increases the speed and accuracy of clinical decision-making processes and strengthens effective intervention opportunities by providing data-based decision support to specialist physicians in the early detection and management of high-risk individuals. The main contributions of the study can be listed as follows.

- A reliable and generalisable machine learning model with high classification performance has been developed for early diagnosis of heart attack risk using Random Forest algorithm.
- The performance of the model was evaluated through a systematic cross-validation process, resulting in a stable

structure that can be integrated into clinical decision support systems.

- This study contributes to data-driven clinical interpretation by identifying key biomedical variables that influence classification decisions.

II. MATERIAL AND METHODS

A. Material

The dataset used in this study is the Heart Disease Classification Dataset, which was created and published on the Kaggle platform for the analysis of heart attack risk factors, one of the most common causes of death worldwide [37]. This dataset includes demographic, clinical and biochemical parameters that may be associated with heart attack. In the dataset recorded from 1319 individuals, a total of eight input parameters (age, gender, heart rate, systolic blood pressure, diastolic blood pressure, blood glucose level, CK-MB isoenzyme and troponin level) contain an output as an indicator of heart attack. Here, the input variables indicate clinical data on the individual's heart health, while the output variables indicate whether the individual has had a heart attack or not in a binary classification (0 = no heart attack, 1 = heart attack). These variables offer applicability for both clinical studies and artificial intelligence-based prediction models, as the effects of factors such as gender, hypertension, hyperglycaemia and cardiac enzyme levels, which reflect the main causative factors in heart diseases, are tested on heart attacks.

B. Methods

1) Data Preparation and Preprocessing

The dataset is a comprehensive dataset containing features for predicting the risk of heart attack. The data used in the study was loaded from a file named Heart_Attack.csv and includes demographic information, clinical measurements and biochemical parameters of individuals. In this context, the dataset is structured to analyse and classify the factors that may contribute to the occurrence of heart attack. In this process, the dataset was systematically organised and missing or inconsistent data were removed.

Label Encoding: Label Encoding is a method that enables categorical variables to be represented in numerical format, making it possible to be processed by machine learning algorithms [14, 15]. In this direction, the class column (Positive--1, negative--0), which is used as the target variable within the scope of the classification problem in this study, was converted into numerical values. This process enables the model to make numerical distinction between categorical classes and enables the algorithm to carry out the classification process effectively.

Data Balancing (SMOTE - Synthetic Minority Over-sampling Technique): SMOTE is an oversampling technique that eliminates data imbalance between classes. In imbalanced data sets, when one class has fewer samples than the other, machine learning algorithms usually prioritise the class with the highest number of samples, leading to a decrease in the prediction success of the model in the minority class [16].

SMOTE increases the learning capability of the model by eliminating the data imbalance for the minority class through synthetic examples. In this way, the overall accuracy of the model improves by balancing the data distribution between classes [17, 18]. Equations 1 and 2 present the mathematical representation of the synthetic data generated by the SMOTE method [16, 18].

$$d(x_i, x_{neighbor}) = \sqrt{\sum_{k=1}^n (x_{i,hi} - x_{neighbor,k_i})^2} \quad (1)$$

$$x_{synthetic} = x_i + \lambda \times (x_{neighbor} - x_i) \quad (2)$$

In these equations, for each minority class sample, a new synthetic data point is generated using the distance information determined in the previous step. Here x_i represents an instance of the minority class. $x_{neighbor}$ represents the selected neighbours. The parameter λ is a randomly chosen coefficient between 0 and 1. $x_{synthetic}$ represents the new synthetic data sample generated. In this study, the distribution of the number of samples in the dataset before and after the SMOTE application is presented in Table 1. The SMOTE method eliminated the imbalance between classes by increasing the number of samples in the minority class.

TABLE I
NUMBER OF DATA BEFORE AND AFTER CORRECTION OF
UNBALANCED CLASS DISTRIBUTION USING SMOTE METHOD

Operation	Total sample count	Number of class	Class(0) sample count	Class(1) sample count
Before SMOTE	1319	2	509	810
After SMOTE	1620	2	810	810

Standard Scaler: This method is a preprocessing and scaling method that equates the data mean to zero and the standard deviation to one. This technique is frequently used in machine learning applications to prevent the model from being stuck on a single feature. In this way, the use of the standard scaler ensures that all attributes are treated with similar weights during model training, which significantly prevents the model from developing bias towards certain features. This scaling method, expressed mathematically in Equation 3, is critical for improving the accuracy of machine learning models, optimising the training process, and reducing the imbalances that can be introduced by large-scale variables [16].

$$X_{scaled} = \frac{X - \text{mean}(X)}{\text{std}(X)} \quad (3)$$

X_{scaled} represents scaled data with mean 0 and standard deviation 1, where X is the data value, $\text{mean}(X)$ is the mean of the dataset and $\text{std}(X)$ is the standard deviation of the dataset.

III. MODELLING AND EVALUATION

A. Random Forest Classifier

Random Forest [10] is a classification algorithm that utilises multiple decision trees in the training phase and stands out with high accuracy rates among supervised learning methods [19]. Although each tree used is effective in the prediction process independently of each other, the final decision is made according to the weighted preference of all trees [20]. In this way, it improves the generalisation performance of the model by reducing the high variance that each individual tree may show. Furthermore, since the feature selection is a random process, the correlation between the trees is minimised and the overlearning of the model is avoided [21]. As shown in Figure 1, the random forest algorithm can minimise errors by providing highly accurate analyses of complex data sets. In addition, it can work in harmony with effective methods to eliminate data imbalances between classes [22]. This method can be effectively applied in many fields such as biomedical diagnostic systems, financial risk analyses, and development of educational systems [16]. The random forest hyper parameters used in this study are presented in Table 2.

TABLE II
RANDOM FOREST CLASSIFIER HYPER PARAMETERS

Parameters	Value	Description
n_estimators	100	The total number of trees to be created in the model.
max_depth	10	Determines the maximum depth of each decision tree.
max_samples	0.8	The proportion of samples to be used for training each tree (with bootstrap).
max_features	'sqrt'	Determines the maximum number of features to be used per tree; square root is taken.
class_weight	'balanced'	Used to automatically balance class imbalance in the dataset.
bootstrap	True	Ensures creation of sub-sample datasets using bootstrap sampling.
random_state	42	Fixes randomness to ensure reproducibility of results.

Evaluation Metrics

Cross-Validation: In order to evaluate the overall performance of the model more reliably, a 10-fold stratified cross-validation method was applied on the dataset. In this method, the dataset is divided into 10 equal parts (folds) and each part is used once as test data and the remaining parts are used as training data. Thanks to the stratified fold technique, the distribution between classes in each layer is preserved similar to the original dataset, and the generalisation ability of the model is measured more accurately even in imbalanced data situations. This approach makes the accuracy assessment of the model more fair and stable, especially in areas such as health data where class imbalance is significant [23].

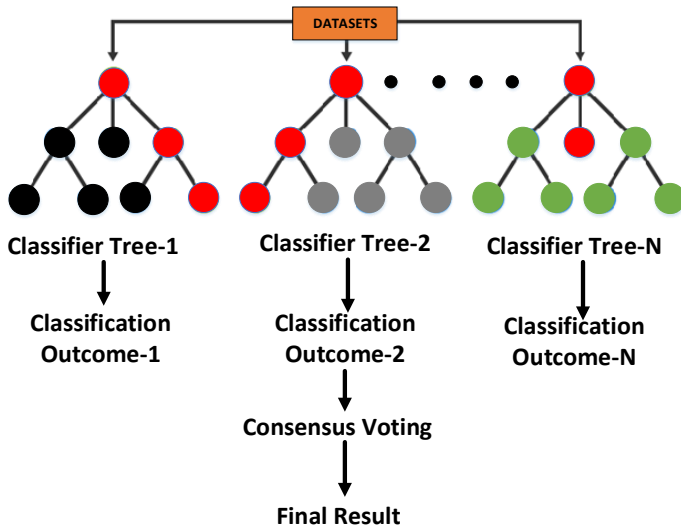


Fig.1. General operation of the random forest classification algorithm

Performance Metrics Used: The performance of the model was analysed with the following basic classification metrics:

Accuracy: Accuracy is defined as the ratio of the number of samples correctly classified by the model to the total number of samples. This metric is a fundamental measure of the overall success of the model. In this regard, it is calculated as the sum of true positive (TP) and true negative (TN) predictions divided by the sum of all predictions. By considering both positive and negative classes, it reflects the overall recognition ability of the model over all classes. The accuracy metric is expressed mathematically in equation 4 [24, 25].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Precision: Precision refers to the proportion of samples that the model predicts as positive that are actually positive. This metric gains importance when the number of false positives (FP) is high. Precision metric is expressed mathematically in equation 5 [24, 26].

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Recall: Sensitivity indicates how many true positive samples the model can accurately predict. In scenarios where false negatives (FN) need to be minimised, for example in disease detection, it is an important criterion in determining the success of the model. The Recall metric is expressed mathematically in equation 6 [24, 26].

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

F1 Score: The F1 score is the harmonic mean of precision and sensitivity and is used to balance both metrics. It is an ideal indicator to evaluate the overall performance of the model,

especially in cases where there are imbalances between classes. The F1 Score metric is expressed mathematically in the following equation [16, 27].

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

F1 score is a metric that is critical in evaluating the performance of classification models. This metric, which takes a value between 0 and 1, reflects how a model balances between precision and recall. Especially in cases where false positive and false negative results are important, the F1 score provides a more sensitive indicator of the overall reliability of the model [24].

Confusion Matrix : Confusion Matrix analyses the classification success by comparing the model's predicted classes with the actual class labels through four basic components: True Positive (TP), false positive (FP), true negative (TN) and false negative (FN). These components categorically reveal both the model's ability to classify correctly and its tendency to make errors. The Confusion Matrix metric is expressed mathematically in the following equation [28, 29].

$$\begin{aligned} Precision &= TP / (TP + FP) \\ Recall &= TP / (TP + FN) \end{aligned} \quad (8)$$

This 2×2 complexity matrix allows to observe not only the overall accuracy rate, but also how successful the model is for each class. In this respect, it is very valuable in evaluating the reliability of the model, especially in data sets with unbalanced class distribution or in applications where positive classes are critical (e.g. in clinical scenarios such as disease detection, readmission prediction). It also forms the basis for various performance metrics such as precision, recall and F1 score. Thus, not only the correct prediction rate of the model, but also the types of errors it makes and the possible effects of these errors in the application context can be analysed more systematically [24, 29].

IV. FEATURE IMPORTANCE

In machine learning applications, attribute importance ranking is a basic approach that reveals which attributes the model bases its predictions on the target variable [30]. This method both facilitates the understanding of decision-making processes and increases the transparency of the model by providing attributes related to the inner workings of the model [31]. Moreover, knowing the relative importance of attributes helps to eliminate unnecessary variables, especially in high-dimensional data structures, thus reducing the complexity of the model and minimising the risk of overfitting [32]. In this context, the main reasons for using attribute importance analysis are to determine which attributes contribute to the prediction process of the model, to highlight the critical variables required to improve the model performance and to strengthen the interpretability of the model. Especially in classification problems in the field of healthcare, understanding

which clinical indicators the decisions are based on is an important factor that enables expert physicians to use the model more confidently. In this way, artificial intelligence-based models not only make accurate predictions, but also clearly present the rationale for these predictions [33, 34].

In this study, attribute importance rating was performed using a model-based approach. Model-based attribute importance analysis is a method that aims to determine the contribution of each attribute to the decision process based on the internal structure of a trained machine learning model, derived directly from the prediction process [35, 36]. The multiple decision tree structure of the Random Forest algorithm allows the contribution of each attribute to the classification process to be statistically evaluated and a relative importance ranking is obtained according to these contributions. This model-based evaluation allows strong inferences to be made by revealing the attributes that interact with the data more effectively, especially in complex data structures where classical statistical methods are insufficient. This process also increases the accuracy and reliability levels of clinical decision support systems and strengthens both the performance and the acceptability of the model in the eyes of the user.

V. RESULT AND DISCUSSION

A. Result

In this study, the binary classification problem for heart attack diagnosis is addressed using the random forest algorithm. The dataset consists of 1319 samples and 8 demographic, clinical, and biochemical features, including age, gender, heart rate, systolic and diastolic blood pressure, blood glucose level, CK-MB isoenzyme, and troponin level. To ensure a robust and generalizable evaluation of model performance, a 10-fold cross-validation method was employed, allowing each data instance to contribute equally to model evaluation across different folds.

The experimental analyses show that the model works consistently and with high accuracy rates. The accuracy, precision, recall and F1 score values obtained in each layer are presented in detail in Table 3. The average accuracy value is 98.95%, the average precision is 99.38%, the average recall is 98.52% and the average F1 score is 98.94%. These high success rates indicate that the model has a strong discriminative ability between classes. In this context, especially Fold 6 stands out as the layer that best reflects the performance of the model.

TABLE III
RANDOM FOREST CLASSIFICATION PERFORMANCE (10-FOLD CROSS VALIDATION RESULTS)

Fold	Accuracy	Precision	Recall	F1 Score
1	0.9877	0.9877	0.9877	0.9877
2	0.9938	1.0000	0.9877	0.9938
3	0.9877	1.0000	0.9753	0.9875
4	0.9815	0.9875	0.9753	0.9814
5	0.9753	0.9753	0.9753	0.9753
6	1.0000	1.0000	1.0000	1.0000
7	0.9938	0.9878	1.0000	0.9939
8	0.9938	1.0000	0.9877	0.9938
9	0.9815	1.0000	0.9630	0.9811
10	1.0000	1.0000	1.0000	1.0000

The confusion matrix of Fold 6 presented in Figure 2 shows the classification performance of the random forest algorithm on the test data within the scope of cross-validation. In this particular fold, the model correctly classified all 162 samples with 100% accuracy. All true positive (TP = 81) and true negative (TN = 81) samples were predicted without error, and there were no false positive (FP) and false negative (FN) samples. This flawless classification within a single fold underscores the model's robustness under cross-validation settings. These findings suggest that the model may have the potential to maintain consistent performance in similar classification tasks. This result demonstrates that the model has high discriminative power for both classes and exhibits a strong generalisation performance without showing any signs of overfitting.

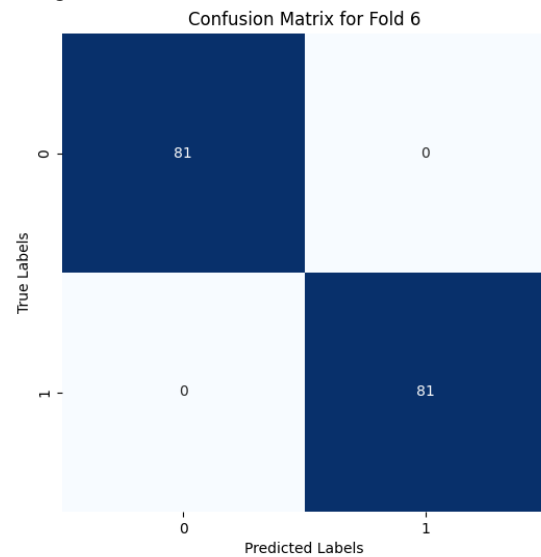


Fig.2. Confusion Matrix for Fold 6

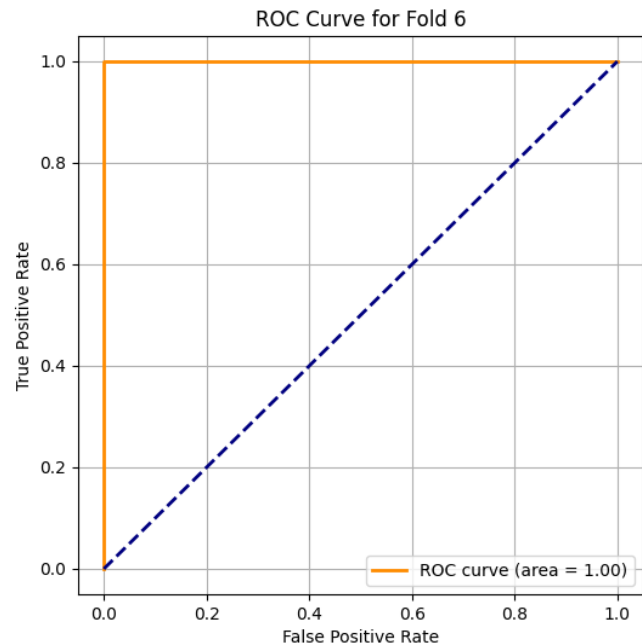


Fig.3. ROC Curve and AUC Value for Fold 6

The classification performance of the model is also supported by the ROC curve. As presented in Figure 3, the ROC curve of Fold 6 shows that the model provides 100% discrimination ability by maintaining both sensitivity and specificity at a high level. The AUC (Area Under Curve) value obtained was 1.00. This shows that the model can discriminate between classes with maximum accuracy.

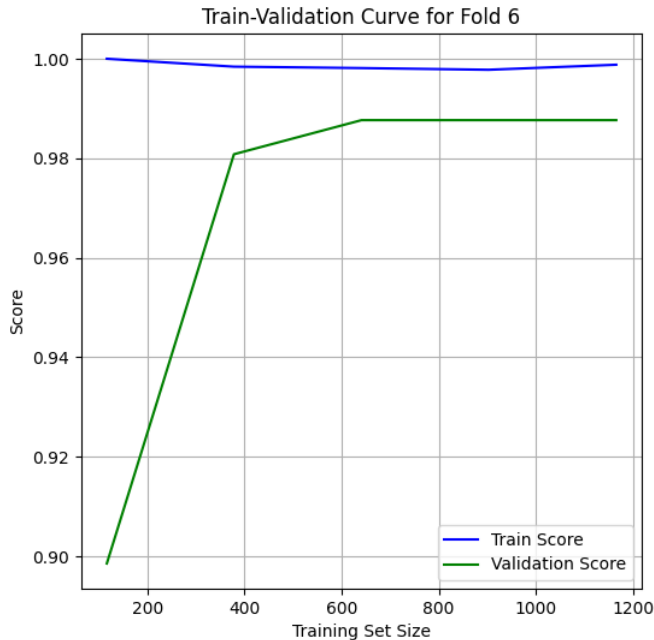


Fig.4. Training-Validation Curve for Fold 6

The training and validation curves generated to obtain insights into the learning process of the model are presented in Figure 4. When the graph is analysed, it is seen that as the training set grows, the validation success remains at a constant and high level (around 98.8%), while the training success is close to 100%. These results show that the model performs consistently throughout the learning process and gains a strong generalisation capability. Therefore, the model achieved high success not only on the training data but also on the validation data, providing a reliable and stable classification performance.

In this study, attribute importance rating is performed with a model-based approach. In model-based attribute importance analysis, the contribution of each attribute to the model is directly calculated by using the internal structure of the classification algorithm trained in the prediction process and the relative effect of the variables is revealed. Thanks to the multiple decision tree structure of the Random Forest algorithm, the contribution of each attribute to the classification process is statistically evaluated and the relative importance ranking is obtained according to these contributions. As presented in Table 4, according to the relative importance values calculated by the random forest algorithm, the troponin variable is the most dominant decision maker of the model with 58.13%. This variable is followed by kcm and age with 25.24% and 5.97%, respectively. The contributions of other attributes seem to be limited, indicating that certain variables play a dominant role in the decision process of the model. The

attribute importance distribution graph presented in Figure 5 visually supports this situation and clearly demonstrates the determinant effect of the troponin variable in the classification. These findings are consistent with the literature stating that troponin levels are one of the main biomarkers in determining the risk of heart attack and prove that the model has both an explainable and reliable structure.

TABLE IV
RELATIVE IMPORTANCE VALUES OF FEATURES ACCORDING TO RANDOM FOREST ALGORITHM

Feature	Importance
troponin	0.5813
kcm	0.2524
age	0.0597
pressurehigh	0.0255
glucose	0.0238
pressurelow	0.0213
impluse	0.0199
gender	0.0160

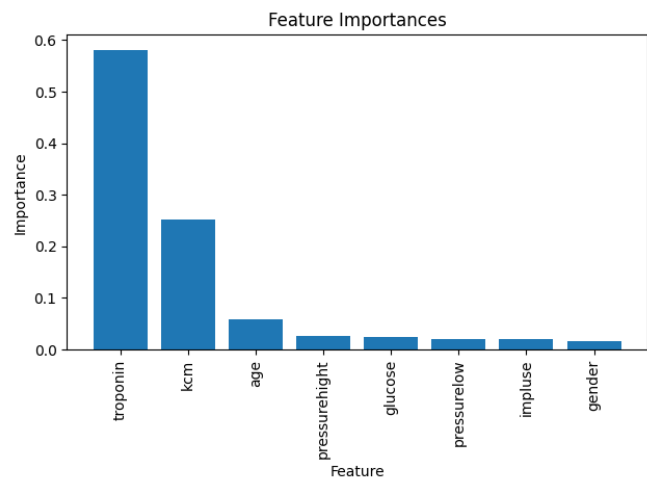


Fig.5. Importance distribution showing effects of attributes on the model

B. Discussion

In this study, a classification model using the random forest method developed for the diagnosis of heart attack has attracted attention with its high accuracy and stable performance results. The dataset consisting of 1319 individuals was evaluated using 10-fold cross-validation to assess the model's performance in a systematic and reliable manner. The metrics obtained show that the classification success is strong; especially the correct classification of all examples in the test data in Fold 6 clearly reveals the discriminative power of the model. These findings suggest that machine learning approaches can be an effective support tool for early diagnosis of cardiovascular diseases.

However, the study has some limitations. The dataset used consists of a relatively limited number of individuals and reflects the characteristics of a specific group, which may limit the generalisability of the model to different populations. Furthermore, the analysis is based on only eight basic clinical parameters. Therefore, it is important to re-evaluate the developed model with more diverse and comprehensive data sets in order to increase its generalisation capacity.

VI. CONCLUSION

This study aims to develop a classification model using the random forest algorithm in order to diagnose the risk of heart attack at an early stage. In order to evaluate the performance of the model, a dataset consisting of a total of 1319 individuals was used, and performance measures such as accuracy, precision, recall and F1 score were analysed through cross-validation-based evaluation. In the evaluation process, 10-fold cross-validation was applied and 100% classification success was achieved in Fold 6, demonstrating the strong discrimination capacity of the model. In addition, troponin and kcm parameters stood out as the most effective variables in the model-based feature importance analysis; these findings are consistent with clinical evaluations in the literature. Furthermore, a significant disadvantage is that the dataset contains only eight clinical parameters and represents a specific population. This may limit the generalisability of the model to different demographic groups. In future studies, retraining and evaluating the model with more diverse and comprehensive datasets containing more samples and attributes is considered. Such approaches can be extended not only to heart diseases, but also to early diagnosis of other diseases such as diabetes and cancer, and can contribute to decision support systems in the field of health.

REFERENCES

- [1] H. F. El-Sofany, "Predicting heart diseases using machine learning and different data classification techniques," *IEEE Access*, 2024.
- [2] H. G. Enad and M. A. Mohammed, "Cloud computing-based framework for heart disease classification using quantum machine learning approach," *Journal of Intelligent Systems*, vol. 33, no. 1, p. 20230261, 2024.
- [3] T. A. Gaziano, A. Bitton, S. Anand, S. Abrahams-Gessel, and A. Murphy, "Growing epidemic of coronary heart disease in low- and middle-income countries," *Current problems in cardiology*, vol. 35, no. 2, pp. 72-115, 2010.
- [4] C. Gupta, A. Saha, N. S. Reddy, and U. D. Acharya, "Cardiac Disease Prediction using Supervised Machine Learning Techniques," in *Journal of physics: conference series*, 2022, vol. 2161, no. 1: IOP Publishing, p. 012013.
- [5] A. K. Dubey, A. K. Sinhal, and R. Sharma, "Heart disease classification through crow intelligence optimization-based deep learning approach," *International Journal of Information Technology*, vol. 16, no. 3, pp. 1815-1830, 2024.
- [6] R. Rajkumar, K. Anandakumar, and A. Bharathi, "Coronary artery disease (CAD) prediction and classification-a survey," *Breast Cancer*, vol. 90, p. 94.35, 2006.
- [7] P. Rani *et al.*, "An extensive review of machine learning and deep learning techniques on heart disease classification and prediction," *Archives of Computational Methods in Engineering*, vol. 31, no. 6, pp. 3331-3349, 2024.
- [8] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN computer science*, vol. 2, no. 3, p. 160, 2021.
- [9] Ö. F. Ertuğrul, S. Dal, Y. Hazar, and E. Aldemir, "Determining relevant features in activity recognition via wearable sensors on the MYO Armband," *Arabian Journal for Science and Engineering*, vol. 45, pp. 10097-10113, 2020.
- [10] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, and H. Dickhaus, "Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier," *Computer methods and programs in biomedicine*, vol. 108, no. 1, pp. 10-19, 2012.
- [11] D. R. Edla, K. Mangalorekar, G. Dhavalikar, and S. Dodia, "Classification of EEG data for human mental state analysis using Random Forest Classifier," *Procedia computer science*, vol. 132, pp. 1523-1532, 2018.
- [12] K. Natarajan *et al.*, "Efficient heart disease classification through stacked ensemble with optimized firefly feature selection," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, p. 174, 2024.
- [13] B. Deekshatulu and P. Chandra, "Classification of heart disease using k-nearest neighbor and genetic algorithm," *Procedia technology*, vol. 10, pp. 85-94, 2013.
- [14] N. Kosaraju, S. R. Sankepally, and K. Mallikharjuna Rao, "Categorical data: Need, encoding, selection of encoding method and its emergence in machine learning models—a practical review study on heart disease prediction dataset using pearson correlation," in *Proceedings of International Conference on Data Science and Applications: ICDSA 2022, Volume 1*, 2023: Springer, pp. 369-382.
- [15] T. Amarbayasgalan, V.-H. Pham, N. Theera-Umpon, Y. Piao, and K. H. Ryu, "An efficient prediction method for coronary heart disease risk based on two deep neural networks trained on well-ordered training datasets," *IEEE Access*, vol. 9, pp. 135210-135223, 2021.
- [16] Y. Hazar and Ö. F. Ertuğrul, "Process management in diabetes treatment by blending technique," *Computers in Biology and Medicine*, vol. 190, p. 110034, 2025.
- [17] P. Soltanzadeh and M. Hashemzadeh, "RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem," *Information Sciences*, vol. 542, pp. 92-111, 2021.
- [18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321-357, 2002.
- [19] S. Hegelich, "Decision trees and random forests: Machine learning techniques to classify rare events," *European policy analysis*, vol. 2, no. 1, pp. 98-120, 2016.
- [20] G. A. B. Suryanegara and M. D. Purbolaksono, "Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 114-122, 2021.
- [21] S. Suparyati, E. Utami, and A. H. Muhammad, "Applying different resampling strategies in random forest algorithm to predict lumpy skin disease," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 4, pp. 555-562, 2022.
- [22] R. Oktafiani, A. Hermawan, and D. Avianto, "Max Depth Impact on Heart Disease Classification: Decision Tree and Random Forest," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 8, no. 1, pp. 160-168, 2024.
- [23] I. Tougui, A. Jilbab, and J. El Mhamdi, "Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications," *Healthcare informatics research*, vol. 27, no. 3, pp. 189-199, 2021.
- [24] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. "O'Reilly Media, Inc.", 2022.
- [25] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the effect of accuracy on trust in machine learning models," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019, pp. 1-12.
- [26] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233-240.
- [27] A. Humphrey *et al.*, "Machine-learning classification of astronomical sources: estimating F1-score in the absence of ground truth," *Monthly Notices of the Royal Astronomical Society: Letters*, vol. 517, no. 1, pp. L116-L120, 2022.
- [28] J. Liang, "Confusion matrix: Machine learning," *POGIL Activity Clearinghouse*, vol. 3, no. 4, 2022.

- [29] I. Markoulidakis and G. Markoulidakis, "Probabilistic Confusion Matrix: A Novel Method for Machine Learning Algorithm Generalized Performance Analysis," *Technologies*, vol. 12, no. 7, p. 113, 2024.
- [30] V. A. Huynh-Thu, Y. Saeys, L. Wehenkel, and P. Geurts, "Statistical interpretation of machine learning-based feature importance scores for biomarker discovery," *Bioinformatics*, vol. 28, no. 13, pp. 1766-1774, 2012.
- [31] F. Pan, T. Converse, D. Ahn, F. Salvetti, and G. Donato, "Feature selection for ranking using boosted trees," in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, pp. 2025-2028.
- [32] A. A. Megantara and T. Ahmad, "Feature importance ranking for increasing performance of intrusion detection system," in *2020 3rd International Conference on Computer and Informatics Engineering (IC2IE)*, 2020: IEEE, pp. 37-42.
- [33] M. A. Jamil and S. Khanam, "Influence of one-way ANOVA and Kruskal-Wallis based feature ranking on the performance of ML classifiers for bearing fault diagnosis," *Journal of Vibration Engineering & Technologies*, vol. 12, no. 3, pp. 3101-3132, 2024.
- [34] N. Silpa, V. M. Rao, M. V. Subbarao, R. R. Kurada, S. S. Reddy, and P. J. Uppalapati, "An enriched employee retention analysis system with a combination strategy of feature selection and machine learning techniques," in *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2023: IEEE, pp. 142-149.
- [35] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, vol. 8, pp. 1-21, 2007.
- [36] B. M. Greenwell, B. C. Boehmke, and A. J. McCarthy, "A simple and effective model-based variable importance measure," *arXiv preprint arXiv:1805.04755*, 2018.
- [37] Bharath011, Heart Disease Classification Dataset, Kaggle, 2022. [Çevrimiçi]. Erişim adresi: <https://www.kaggle.com/datasets/bharath011/heart-disease-classification-dataset>

BIOGRAPHIES



Süleyman Dal Dr. Lecturer. Assist. Süleyman Dal is an academic specialised in the field of electrical and electronics engineering. He completed his undergraduate education at Çukurova University, Department of Electrical and Electronics Engineering in 2018, completed his master's degree at Batman University, Institute of

Science and Technology in 2020, and completed his doctorate education at the Institute of Graduate Studies of the same university in 2021.

He is currently working as a Dr. Lecturer in the Energy Coordination Department within the Batman University Rectorate. His research interests include machine learning, signal processing and optimisation algorithms.



Necmettin Sezgin Prof. Dr. Necmettin Sezgin is an academic specialised in electronics. He received his bachelor's degree from Hacettepe University, his master's degree from Dicle University and his doctorate from İnönü University.

Since 2011, Prof. Sezgin has been working at Batman University, where he is currently a faculty member of the Department of Electrical and Electronics Engineering and Vice Rector of Batman University. His research interests include signal processing, biomedical signal analysis and electronic systems.


Comparison of VT-based and CNN-based Models on Teeth Segmentation

Silan Fidan Vural, Nida Kumbasar*


Abstract— Semantic segmentation is a crucial task in computer vision with a wide array of applications across various fields, especially in medical imaging. One of the most important applications of semantic segmentation is in the field of dentistry, where teeth segmentation plays a significant role in diagnosing and treating oral health issues. Accurate segmentation of teeth in dental images is vital for detecting abnormalities, planning treatments, and monitoring the progress of dental procedures. In this paper, a comprehensive comparative analysis is presented, focusing on the use of Convolutional Neural Network (CNN)-based and Vision Transformer (VT)-based models for image segmentation within the context of dentistry. The paper presents a comparison of eight different models, contributing to the literature on dental image segmentation and showcasing practical applications in clinical dental settings. The research presented in this study uses several state-of-the-art segmentation models, namely U-Net, LinkNet, and Swin U-Net, along with different backbones to perform teeth segmentation on publicly available two datasets: one representing adults and the other children. The experiments were conducted to determine which models and backbones provided the best segmentation performance for each dataset. The study also emphasizes that the segmentation modeling process should be handled separately since the alignment of child and adult teeth is different. The U-Net model with the ResNet101 backbone achieved the best performance on the adults dataset, while for the children dataset, the U-Net model with the same ResNet101 backbone also demonstrated superior results. The highest Dice scores obtained were 0.9543 for the adults dataset and 0.9019 for the children dataset, indicating the effectiveness of these models in accurately segmenting teeth. The findings from this research demonstrate the potential of deep learning techniques in improving the accuracy and efficiency of dental diagnosis and treatment planning. Codes used throughout the study will be publicly available at <https://github.com/FidanVural/Teeth-Segmentation-in-Panoramic-Radiography/tree/main>

Index Terms—Convolutional Neural Networks, Panoramic Radiography, Semantic Segmentation, Vision Transformer

Silan Fidan Vural, is with TUBITAK, Informatics and Information Security Research Center (BİLGEM), Gebze, Kocaeli, 41470, Turkey, (e-mail: fsilanvural@hotmail.com).

 <https://orcid.org/0009-0000-9488-3809>

Nida Kumbasar, is with TÜBİTAK, Informatics and Information Security Research Center (BİLGEM), (e-mail: nida.kumbasar@tubitak.gov.tr).

 <https://orcid.org/0000-0001-5497-4618>

Manuscript received Mar 23, 2024; accepted Feb 24, 2025

DOI: [10.17694/bajece.1457754](https://doi.org/10.17694/bajece.1457754)

I. INTRODUCTION

PANORAMIC RADIOGRAPHY (PR) is the most preferred 2 Dimension (2D) imaging technique in the dental field due to its relatively low radiation rate, fast results and low cost [1]. In PR teeth, gingiva, jaw bones, sinus cavities, temporomandibular joint and surrounding anatomical structures as well as intraoral diseases such as tooth decay, gum diseases, cysts, tumors, lesions, etc. can be analyzed in detail. PR is important for planning and follow-up in tooth extraction, dental implants, prosthetic applications and oral surgical procedures.

Deep Learning (DL) applications in the analysis of dental imaging techniques are becoming increasingly common. Caries detection [2], cyst and jaw tumor differentiation [3], root fracture detection [4], periodontal disease detection [5], tooth segmentation [6], jaw segmentation [7], wisdom teeth analysis [8], [9] dental biometric systems [10], [11] are some of the application areas. The combination of Artificial Intelligence (AI) tools with the dentist's vision improves the diagnostic treatment process by predicting diseases and outcomes for complex cases.

Teeth segmentation provides a visual reference for dentists to evaluate the condition of the teeth and closely follow the diagnosis and treatment process. In addition, teeth segmentation guides the diagnosis and treatment process of oral problems directly related to the teeth such as prosthesis - implant placement, tooth extraction, scaling, braces treatment, tooth decay detection. As various 3D anatomical structures overlap on 2D PR, distortions occur in the image. Moreover, the image quality varies from device to device and the low contrast of the image makes it difficult to perform teeth segmentation manually. Considering these factors, the correct teeth segmentation depends on the experience and availability of dentists [12]. Accurate and fully automated teeth segmentation is important to improve the performance of the clinical process.

This paper focuses its attention on the application of image segmentation in dentistry. The main motivation of the work is to present a comparative study on Vision Transformer (VT)-based and Convolutional Neural Network (CNN)-based tooth segmentation using two separate datasets which are adults and children with different patterns and sizes. Another motivation can be explained that performing these models with different backbone architectures.

The contributions in this paper can be outlined as follows:

- Inter-model and intra-model backbones comparison for tooth segmentation with PR is presented.
- The comparison of CNN-based models with VT-based models was performed on both children dental dataset with limited data and adults dental dataset with larger data volumes.
- Due to the relative scarcity of segmentation on children teeth datasets, experiments were conducted to contribute to the literature. To the best of our knowledge, there are no applications comparing DL and VT algorithms on a pediatric tooth dataset.

The rest of the study is organized as follows: Related works are explained in Section 2. Material and method are defined in Section 3. Experiments and experimental results are presented in Section 4. Section 5 which is discussion includes comparison with the literature. Finally, Section 6 outlines the study for teeth segmentation and mentions future work plans.

II. RELATED WORK

Semantic segmentation is one of the fundamental Computer Vision (CV) tasks used in a wide range of fields, from autonomous vehicle systems to medical images. It classifies each pixel in the image according to predefined categories. CNN-based segmentation models were quite popular in medical image segmentation task at first. U-Net [13] is one of the most used CNN-based models in medical image segmentation which has an encoder-decoder architecture with skip connections. There are other CNN-based models like Fully Convolutional Networks (FCN) [14] and LinkNet [15] has also demonstrated significant success in medical data. On the other hand, these models have been employed in teeth segmentation with considerable success. Furthermore, teeth segmentation plays a crucial role in assisting dentists in understanding the state of dental health. Even though CNN-based models have achieved substantial success, novel approaches based on transformers have emerged due to CNNs' lack global contextual understanding of images.

Transformers first gained prominence in the field of Natural Language Processing (NLP) [16], creating a profound impact. Subsequently, transformers entered the field of CV and initially achieved state-of-the-art (SOTA) performance in image classification [17]. After this paper [17], many new VT models emerged and vision transformers become quite popular in the field of CV such as image segmentation and object detection. Many models appeared like Swin U-Net [18], TransUNet [19] and PromptUNet [20] for medical image segmentation.

Many studies in the literature have demonstrated the potential of CNN-based or VT-based teeth segmentation approaches to assist clinicians in dental imaging.

Silva et al. [21] applied the Mask R-CNN technique for automatic teeth segmentation on PR. Koch et al. [22] and Sivagami et al. [23] proposed the use of the U-Net network for teeth segmentation, while Jader et al. [24] utilized Mask R-CNN for segmentation of teeth. Wirtz et al. [25] added a modeling process to the R-CNN mask and manually annotated individual tooth forms on PR. Lee et al. [26] performed teeth

segmentation automatically with a fine-tuned Mask R-CNN. Zhao et al. [27] integrated the U-Net architecture into the segmentation branch to improve the segmentation effect in the Mask R-CNN model and presented comparative results with U-Net and Mask R-CNN. Similarly, Silva et al. [28] comparatively analyzed teeth segmentation with Mask R-CNN, PANet, HTC, and ResNeSt. Sheng et al. [29] presented a comparative study of teeth segmentation with U-Net, LinkNet FPN, and Swin U-Net methods on PR images. Arora et al. [30] proposed a novel multimodal CNN architecture in which the encoder part consists of conventional CNN, atrous-CNN and separable CNN, and the decoder part consists of a single stream of deconvolutional layers for segmentation. Kanwal et al. [31] implemented a novel architecture for teeth segmentation on PR images that utilizes a dual-path transformer-based network integrated with a panoptic quality loss function. Dhar et al. [32] proposed a novel approach to teeth segmentation with PR by adding grid-based attention gates to the skip links of FUSegNet. Ghafoor et al. [33] proposed a new teeth segmentation model that combines an M-Net-like structure with swin transformers and teeth attention block. Zhang et al. [34] collected PR images used for different purposes in the literature for teeth segmentation. In addition, they prepared a child-specific PR dataset that was not previously available in the literature and shared it publicly available [34]. They performed teeth segmentation with U-Net, PSPNet, R2 U-Net and DeepLab V3+ using both adults and children's PR data separately and together. Brahmi et al. [35] employed Mask R-CNN for instance segmentation of teeth on PR. There is an increasing number of studies in the literature that create a model by grouping children's PRs separately from adults'. Asci et al. [36] utilized U-Net to perform dental caries segmentation in the PRs of children in primary dentition, mixed dentition and permanent dentition. Wathore et al. [37] introduced a new bilateral symmetry-based enhancement method specifically designed to improve tooth segmentation in PR and evaluated the effectiveness of the proposed method using U-Net, SE U-Net and TransUNet. Altan et al. [38] performed tooth segmentation using PR with Mask R-CNN on ResNet-50 backbones.

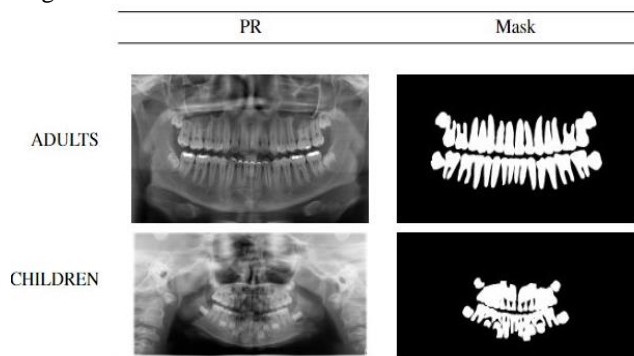


Fig.1. Examples of datasets

III. MATERIAL METHOD

A. Material

In this study, the publicly available "Dataset and Code" and "Children's Dental Caries Segmentation Dataset" mentioned in Zhang et al. [34], were used as two different datasets, namely

Adults Dataset and Children Dataset, respectively. The adults dataset consists of 1978 PR teeth images and masks with different image sizes, 1500 images for training, 202 images for validation and 276 images for testing. The children dataset consists of 193 image-mask pairs with non-standard image size obtained from pediatric patients between the ages of 2 and 13. Of these images, 148 were used for training, 15 for validation and 30 for testing. In the datasets, images have distinct sizes but we resized them to a fixed size. The examples of the datasets can be seen in Fig.1. Also, the distribution of the datasets was shown in Table 1.

TABLE I
LIST OF DATASETS USED IN THIS STUDY

	Train	Validation	Test	Total
Adults Dataset	1500	202	276	1978
Children Dataset	148	15	30	193

B. Method

Our experiments were conducted not only using well-known CNN-based architectures which are U-Net and LinkNet but also using VT architecture which is Swin U-Net.

1) U-Net

U-Net which is developed by Olaf Ronneberger et al. [13] is one of the most important and successful algorithms with an encoder-decoder architecture used in the field of medical image segmentation. U-Net can be represented as a function $f: X \rightarrow Y$, where X denotes the input image and Y represents the output segmentation mask. The encoder part can be mathematically defined as a function $E(x)$ maps the input image X to a latent feature space z .

z is calculated using Equation (1),

$$z = f(W * X + b) \quad (1)$$

where W and b are weights and biases, $*$ denotes the convolution operation, and f is the activation function which is ReLU. The encoder part of the network is responsible for extracting features and learning the representations of input image. The encoder network is usually nothing more than the classification architectures like VGGNet or ResNet. It can be used various networks for the encoder. On the other hand, the decoder part of the network is utilized for generating a segmentation belonging to the input image using encoder representation. In addition to this, there are lots of skip connections between the encoder and decoder networks. These skip connections, also known as shortcut connections, provide information transfer from the encoder to the decoder in order to obtain better segmentation results. Also, U-Net architecture is a modified and extended version of the FCN.

2) LinkNet

LinkNet [15] is also a DL architecture designed for image segmentation tasks, particularly semantic segmentation. The

LinkNet architecture, characterized by its encoder-decoder design, is quite similar to the U-Net. Input of each encoder layer is also passed to the output of its corresponding decoder to obtain better results in segmentation by preserving to spatial information. These processes are called skip connections. Moreover, layers in the LinkNet are not concatenated to each other through skip connections like in U-Net; instead, they are summed. The difference is visualized in Fig.2.

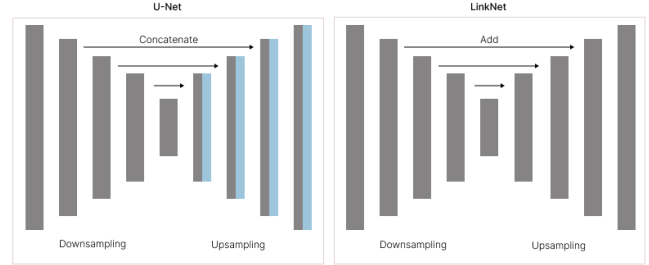


Fig.2. Difference between U-Net and LinkNet

3) Swin U-Net

Within the scope of this research, another architecture which is Swin U-Net was also employed. Swin U-Net model is U-Net shaped encoder-decoder architecture that consists of Swin transformers [18]. Firstly, the image is divided into non-overlapping patches. The number of patches, denoted as N , can be calculated using the formula in Equation (2),

$$N = \left(\frac{H \times W}{P^2} \right) \quad (2)$$

where H is the height of the image, W is the width of the image, and P is the patch size. After that, linear embedding is performed to change the channel size of the input. Swin transformer blocks are applied to these token patches. Each Swin transformer block consists of two successive swin transformer modules. While the first module consists of layer normalization (LN), window based multi-head self attention (W-MSA) and Multi Layer Perceptron (MLP), the second module composed of layer normalization, shifted window based multi-head self attention (SW-MSA), and MLP. A special mention can be made for the attention mechanism inside of the W-MSA and the SW-MSA. Self-attention formula is calculated using Equation (3),

$$A = \text{softmax} \left(\frac{Q \times K^T}{\sqrt{d_k}} \right) \times V \quad (3)$$

where Q, K, V are the query, key, and value matrices, and d_k is the dimension of the key vector. The self-attention mechanism allows a model to focus on different parts of the input data when making predictions. Also, residual (skip) connections are applied in each module. The patch expanding layer in the decoder part is utilized to upsample the feature maps. The linear projection layer is performed on these upsampled features in order to generate the pixel-level segmentation. Swin U-Net architecture is shown in Fig.3.

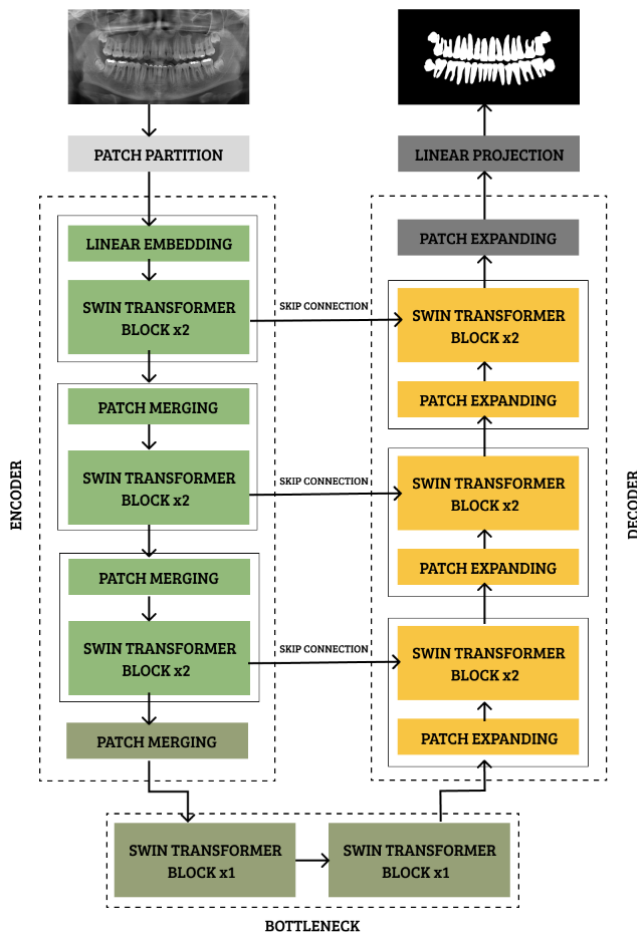


Fig. 3. Swin U-Net architecture

IV. RESULTS

A. Implementation Details

U-Net, LinkNet and Swin U-Net models trained based on Python 3.8.10 and Pytorch 1.12.1. For U-Net and LinkNet, the input image size is set as 256x256. On the other hand, the input image size for Swin U-Net is 224x224. Moreover, the learning rate, batch size and number of epochs were configured as 1e-4, 8, 100 for all models, respectively. Throughout the training period, Adam was used for optimizer and combination of binary cross-entropy and dice score as utilized for loss function. Our models were trained on four 16GB RAM GPUs, Tesla V100.

B. Evaluation Criteria

Dice Score (DS) and Intersection over Union (IoU) metrics express the performance of the predicted region in segmentation problems. DS and IoU are often preferred in segmentation problems as they provide sensitive measures of the overlap between the ground truth and the predicted region. DS is calculated two times the intersection between the predicted and ground truth image segmentations divided by the sum of pixels in both images. DS equation presented in Equation (4).

$$DS = \frac{2TP}{2TP + FP + FN} \quad (4)$$

Also, IoU is computed as the intersection of the areas covered by the predicted and ground truth segmentations over the union of these areas. These metrics provide us a quantitative measure of how well the predicted masks align with the ground truth masks. The equation can be seen in Equation (5).

$$IoU = \frac{TP}{TP + FP + FN} \quad (5)$$

C. Experimental Results

In this study, the experimental studies are presented in two separate sections as Adults Dataset and Children Dataset.

1) Adults Dataset

Table II shows the segmentation performance for adults dataset in terms of DS and IoU by standard deviation values. In the experiments conducted by varying a large number of parameters, all models performed well for the adults dataset. The most successful model is U-Net on ResNet101 backbone by 0.9543 DS and 0.9150 IoU. The model-based comparison shows that the highest performance belongs to U-Net ResNet101 by 0.9543 DS, LinkNet ResNet50 by 0.9542 DS and Swin U-Net T by 0.9529.

TABLE II
TEST RESULTS OF ADULTS DATASET

Adult Dataset		
Model	DS Mean± Std	IoU Mean± Std
U-Net ResNet34	0.9516 ± 0.0362	0.9099 ± 0.0631
U-Net ResNet50	0.9536 ± 0.0377	0.9137 ± 0.0659
U-Net ResNet101	0.9543 ± 0.0374	0.9150 ± 0.0654
LinkNet ResNet34	0.9491 ± 0.0378	0.9055 ± 0.065
LinkNet ResNet50	0.9542 ± 0.0350	0.9145 ± 0.0613
LinkNet ResNet101	0.9515 ± 0.0390	0.9100 ± 0.0677
Swin U-Net T	0.9529 ± 0.0359	0.9123 ± 0.0627
Swin U-Net S	0.9517 ± 0.0378	0.9104 ± 0.0658

Fig.4. shows a graphical comparison of the adults dataset.

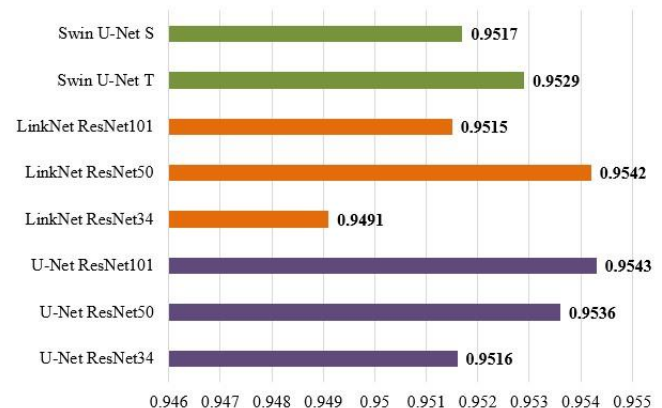


Fig.4. Comparison of DS on adults dataset

1) Children Dataset

Children dataset segmentation results are presented in Table III by DS, IoU and standard deviation values for eight different models. It has been observed that the highest performance in the children dataset, as in the adults dataset, belongs to U-Net ResNet101 by a DS of 0.9019 and an IoU value of 0.8217. In Fig.5. where the model-based comparison is presented, it is seen that Swin U-Net has the lowest performance and U-Net has the highest performance. The highest DS for Swin U-Net T, LinkNet ResNet50 and U-Net ResNet101 are 0.8189, 0.8914 and 0.9019 respectively.

TABLE III
TEST RESULTS OF CHILDREN DATASET

Children Dataset		
Model	DS Mean± Std	IoU Mean± Std
U-Net ResNet34	0.8996 ± 0.0127	0.8178 ± 0.0208
U-Net ResNet50	0.9013 ± 0.0118	0.8206 ± 0.0194
U-Net ResNet101	0.9019 ± 0.0123	0.8217 ± 0.0203
LinkNet ResNet34	0.8889 ± 0.0153	0.8004 ± 0.0243
LinkNet ResNet50	0.8914 ± 0.0156	0.8044 ± 0.0250
LinkNet ResNet101	0.8906 ± 0.0145	0.8032 ± 0.0232
Swin U-Net T	0.8189 ± 0.0389	0.6951 ± 0.0522
Swin U-Net S	0.8112 ± 0.0381	0.6841 ± 0.0514

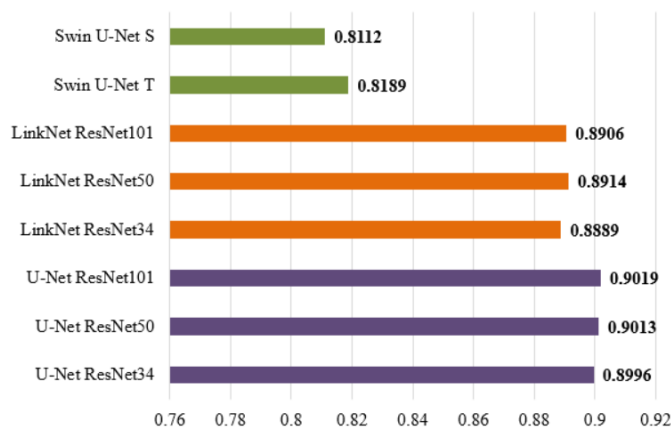


Fig.5. Comparison of DS on children dataset

Fig.6. shows the predicted segmentation results of ground truth and the models on one example each of adult and child PR images.

The experiments were carried out to identify the models and backbones that delivered the most accurate segmentation performance for each dataset. For the adults dataset, the U-Net model with the ResNet101 backbone yielded the best results, while the same U-Net model with the ResNet101 backbone also showed excellent performance on the children dataset. The highest Dice scores recorded were 0.9543 for the adults dataset and 0.9019 for the childrens dataset, highlighting the models' effectiveness in accurately segmenting teeth.

The high accuracy of segmentation in U-Net is ensured by the skip connections between layers so that attributes are not lost and details are preserved. U-Net can be particularly effective

when training with limited data because its structure helps to learn general features even in small data sets. This can improve the accuracy of segmentation in small-scale dental images, such as limited dental data. Although LinkNet has deeper networks, it is notable for its more efficient parameterization. This makes it a viable option for better performance on large datasets. In large-scale datasets such as PR, LinkNet processes the data more efficiently and enables fast segmentation. However, LinkNet has not been found to be as effective as U-Net in identifying more complex structures and details, as the structures used are not considered sufficient to capture less fine details. Since Swin is based on the U-Net transformer structure, it is expected to be able to learn local and global relationships better, especially with diversity in the dataset and large-scale images. In PRs, Swin-U-Net's strengths include the relationships between teeth, variability in jaw structure, and the different characteristics of each tooth, which require more contextual information. However, when Swin-U-Net works on smaller datasets, it has been observed that the large number of parameters degrades its performance.

The structure of U-Net does not have many parameters, as it works with an encoder-decoder architecture. However, components such as skip connections and upsampling layers can require high computational power during model training. U-Net generally has a medium level of computational complexity. LinkNet has a lower computational complexity compared to U-Net. This is an important advantage, especially when working on limited hardware. The fact that the model has fewer parameters allows for faster processing time. Swin-U-Net is the model with the highest computational complexity. Transformer-based structures require high computational power, especially for large data sets and high-resolution images. The large number of parameters and more complex computational processes make it necessary to run Swin-U-Net with higher hardware requirements.

V. DISCUSSIONS

PR is an important medical imaging tool for the diagnosis of oral diseases. In dental radiology, the automatic segmentation of the teeth structure with the help of PR is important as a first step to improve the performance of the diagnostic treatment process. Automatic analysis of PR images increases the efficiency of dentists in densely populated areas and speeds up the processes of patients waiting for treatment.

By the integration of highly computational hardware into machines, AI-based DL is increasingly preferred for problem solving and decision-making tasks. The advancement of technology has increased the rate of data accumulation, making it easier to access the data that DL algorithms need to be trained. CNN-based and VT-based DL models have become increasingly popular in medical and dental imaging.

Recently, teeth segmentation with PR has been widely used in the literature. Silva et al. [21] obtained an accuracy of 0.9208 with Mask R-CNN on 1500 PR data that they brought to the literature. On the same dataset, Koch et al. [22] achieved 0.936 DS with FCN based U-Net. Zhao et al. [1] validated their Two-Stage Attention Segmentation Network (TSASNet) designed

for teeth segmentation with Silva et al. [21] dataset and achieved 0.9272 DS. Hou et al. [39] used 1500 PRs collected by themselves to measure the performance of Teeth U-Net, which they designed with various additional modules between encoder and decoder and at the bottleneck, and obtained a DS of 0.9428. Arora et al. [30] achieved a precision of 0.9501 with a new encoder-decoder architecture based on multimodal feature extraction on 1500 PRs. Ghafoor et al. [33] validated their 540 PRs in their proposed M-Net-like structure with swin transformers teeth attention block cooperating model and obtained a DS of 0.9102. Zhang et al. [34] presented a preliminary study with 1978 and 193 PRs for adults and children, respectively. In the adult data, the DS values of U-Net, R2 U-Net, PSPNet, DeepLab V3+ are 0.9392, 0.9411, 0.9299, 0.9267 respectively.

For children data, the DS values of U-Net, R2 U-Net, PSPNet, DeepLabV3+ are 0.9120, 0.9027, 0.9083, 0.8961 respectively. Since this study utilizes the datasets presented by Zhang et al. [34] a detailed comparison is presented in Table VI. When Table VI is analyzed, as a result of the experiments, it is observed that the proposed model has an improvement of 1.40% with U-Net ResNet101 for the adults dataset. Unfortunately, the relatively small number of data in the children dataset did not result in a significant increase in segmentation performance results. Due to the low number of data, it was observed that the swin transformers based approach was inferior in the child dataset compared to the adult dataset. This study is important in terms of evaluating separate models for teeth segmentation for children and adults, comparing CNN and VT based segmentation on small and large datasets, and analyzing a model with respect to backbones of different depths. On the other hand, considering the dental segmentation achievements, the proposed study emphasizes that dental segmentation is suitable for practical application in clinics.

Although teeth segmentation on PR has great potential in clinical settings, some limitations and challenges can be encountered. Since PR usually presents a 2D projection of the teeth and surrounding tissues, lack of depth information and deformations can complicate segmentation processes. Another major challenge is the variety and quality of the data. PR images obtained in the clinical setting can have a wide range of quality depending on the different devices and acquisition techniques. This diversity can complicate the generalization ability of the model and cause preprocessing steps to become more complex to ensure accurate segmentation. Furthermore, the lack of a sufficient number of different patient samples in the datasets can reduce the generalization success of the model and affect the reliability of the results.

When considering teeth segmentation with supervised learning, the labeling part of the data is a major challenge that affects its applicability in a clinical setting. Correct labeling of PR images is a fundamental step in the training of segmentation models; however, this process is time-consuming and labor-intensive. Teeth need to be correctly labeled, the boundaries of each tooth identified, and associated with the appropriate anatomical structures. This is a specialized task and requires a meticulous examination of each image. In clinical settings, the input of dentists or radiologists is often required to perform this labeling process. However, these specialists may not have the time to perform manual labeling for each image. This increases the

workload and can make it difficult to efficiently implement tooth segmentation in a clinical setting. Furthermore, labeling errors can also negatively affect the accuracy of the model; mislabeled data can lead to incorrect learning of the model, reducing the reliability of the segmentation results. The clinical use of teeth segmentation can also bring real-time analysis requirements. Fast and accurate results are important, especially in busy clinical environments. However, some DL models can require high computational power and time, which can pose practical challenges.

VI. CONCLUSION

In this study, we conducted a detailed comparison between CNN-based architectures, such as U-Net and LinkNet, and transformer-based architectures, specifically Swin U-Net, using dental radiography datasets. These datasets consist of PR images from both adults and children, allowing us to evaluate the models' performance across diverse demographic groups. The evaluation results revealed interesting insights: while CNN-based models achieved superior performance on smaller, limited datasets, transformer-based Swin U-Net was affected by the limited data and did not perform as well on smaller datasets. This highlights the potential advantages of transformer-based models in handling more complex, large-scale datasets, which is a significant consideration in real-world clinical applications. Our findings suggest that teeth segmentation through DL models can provide substantial benefits for dental professionals by offering precise and automated segmentation of teeth in radiographic images. This segmentation can significantly assist in diagnosing dental conditions, planning treatment procedures, and monitoring the progress of treatments over time. By automating this process, dentists can save time, reduce human error, and focus more on patient care rather than manual image analysis. Moreover, the high DS achieved in both adults and children datasets emphasizes the potential of these models in diverse clinical settings, making them a versatile tool for dental practitioners. Furthermore, this research lays the groundwork for future studies in the field of dental image analysis. The results of this study, especially the performance comparison between CNN-based and transformer-based models, can be useful for researchers exploring the integration of DL techniques into dental applications.

The publicly available datasets used in this study also provide an excellent resource for future work in this area, fostering further innovation and improvements in the field of dental radiography analysis. Overall, we believe that the insights gained from this study will contribute significantly to the advancement of automated dental diagnosis and treatment planning.

Future studies are planned to collect larger datasets from individuals with different age groups and ethnic backgrounds. Thus, the performance of the models on various demographic characteristics will be evaluated in more detail and strategies to improve segmentation accuracy can be developed. Such an approach has the potential to provide more effective and generalizable solutions for different patient groups in medical imaging fields such as dental segmentation.



Fig.6 Visualization of the models on adults and children dataset

TABLE VI
COMPARISON OF THIS STUDY AND THE STUDY [34] RESULTS.

		Adults	Dataset	Children	Dataset
	Models	DS	IoU	DS	IoU
Literature [34]	U-Net	0.9392	0.8858	0.9120	0.8387
	R2 U-Net	0.9411	0.8892	0.9027	0.8247
	PSPNet	0.9299	0.8693	0.9083	0.8324
	DeepLab V3+	0.9267	0.8639	0.8961	0.8121
	U-Net ResNet34	0.9516	0.9099	0.8996	0.8178
Proposed Study	U-Net ResNet50	0.9536	0.9137	0.9013	0.8206
	U-Net ResNet101	0.9543	0.9150	0.9019	0.8217
	LinkNet ResNet34	0.9491	0.9055	0.8889	0.8004
	LinkNet ResNet50	0.9542	0.9145	0.8914	0.8044
	LinkNet ResNet101	0.9515	0.9100	0.8906	0.8032
	Swin U-Net T	0.9529	0.9123	0.8189	0.6951
	Swin U-Net S	0.9517	0.9104	0.8112	0.6841

ACKNOWLEDGMENT

This study was conducted in the TÜBİTAK-BİLGEM B3LAB. We would like to express our profound gratitude to Dr. Mehmet HAKLIDIR, Director of the TÜBİTAK BİLGEM Artificial Intelligence Institute for supplying of hardware throughout the work and supporting us to do the research.

REFERENCES

- [1] Y. Zhao *et al.*, "TSASNet: Tooth segmentation on dental panoramic X-ray images by Two-Stage Attention Segmentation Network," *Knowl Based Syst*, vol. 206, p. 106338, 2020.
- [2] A. Haghanifar, M. M. Majdabadi, S. Haghanifar, Y. Choi, and S.-B. Ko, "PaXNet: Tooth segmentation and dental caries detection in panoramic X-ray using ensemble transfer learning and capsule classifier," *Multimed Tools Appl*, vol. 82, no. 18, pp. 27659–27679, 2023.
- [3] Y. Arijji *et al.*, "Automatic detection and classification of radiolucent lesions in the mandible on panoramic radiographs using a deep learning object detection technique," *Oral Surg Oral Med Oral Pathol Oral Radiol*, vol. 128, no. 4, pp. 424–430, 2019.
- [4] M. Fukuda *et al.*, "Evaluation of an artificial intelligence system for detecting vertical root fracture on panoramic radiography," *Oral Radiol*, vol. 36, pp. 337–343, 2020.
- [5] B. C. Uzun Saylan *et al.*, "Assessing the effectiveness of artificial intelligence models for detecting alveolar bone loss in periodontal disease: a panoramic radiograph study," *Diagnostics*, vol. 13, no. 10, p. 1800, 2023.
- [6] L. Schneider *et al.*, "Federated vs local vs central deep learning of tooth segmentation on panoramic radiographs," *J Dent*, vol. 135, p. 104556, 2023.
- [7] S. Park *et al.*, "Deep learning-based automatic segmentation of mandible and maxilla in multi-center ct images," *Applied Sciences*, vol. 12, no. 3, p. 1358, 2022.
- [8] N. Kumbasar, M. T. Güller, Ö. Miloğlu, E. A. Oral, and I. Y. Ozbek, "Deep-learning based fusion of spatial relationship classification between mandibular third molar and inferior alveolar nerve using panoramic radiograph images," *Biomed Signal Process Control*, vol. 100, p. 107059, 2025.
- [9] M. T. Güller, N. Kumbasar, and Ö. Miloğlu, "Evaluation of the effectiveness of panoramic radiography in impacted mandibular third molars on deep learning models developed with findings obtained with cone beam computed tomography," *Oral Radiol*, pp. 1–16, 2024.
- [10] A. B. Oktay, Z. Akhtar, and A. Gurses, "Dental biometric systems: a comparative study of conventional descriptors and deep learning-based features," *Multimed Tools Appl*, vol. 81, no. 20, pp. 28183–28206, 2022.
- [11] Ö. Miloğlu, N. Kumbasar, Z. T. Tosun, M. T. Güller, and İbrahim Yücel Ozbek, "Gender Classification With Hand-Wrist Radiographs Using the Deep Learning Method," *Current Research in Dental Sciences*, vol. 35, no. 1, pp. 2–7, 2025.
- [12] C.-W. Wang *et al.*, "A benchmark for comparison of dental radiography analysis algorithms," *Med Image Anal*, vol. 31, pp. 63–76, 2016.
- [13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18, 2015, pp. 234–241.
- [14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [15] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE visual communications and image processing (VCIP)*, 2017, pp. 1–4.
- [16] A. Vaswani, "Attention is all you need," *Adv Neural Inf Process Syst*, 2017.
- [17] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [18] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.
- [19] J. Chen *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.
- [20] J. Wu, "Promptunet: Toward interactive medical image segmentation," *arXiv preprint arXiv:2305.10300*, vol. 2, 2023.
- [21] G. Silva, L. Oliveira, and M. Pithon, "Automatic segmenting teeth in X-ray images: Trends, a novel data set, benchmarking and future perspectives," *Expert Syst Appl*, vol. 107, pp. 15–31, 2018.
- [22] T. L. Koch, M. Perslev, C. Igel, and S. S. Brandt, "Accurate segmentation of dental panoramic radiographs with U-Nets," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019, pp. 15–19.
- [23] S. Sivagami, P. Chitra, G. S. R. Kailash, and S. R. Muralidharan, "Unet architecture based dental panoramic image segmentation," in *2020 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, 2020, pp. 187–191.
- [24] G. Jader, J. Fontineli, M. Ruiz, K. Abdalla, M. Pithon, and L. Oliveira, "Deep instance segmentation of teeth in panoramic X-ray images," in *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2018, pp. 400–407.
- [25] A. Wirtz, S. G. Mirashi, and S. Wesarg, "Automatic teeth segmentation in panoramic X-ray images using a coupled shape model in combination with a neural network," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part IV* 11, 2018, pp. 712–719.
- [26] J.-H. Lee, S.-S. Han, Y. H. Kim, C. Lee, and I. Kim, "Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs," *Oral Surg Oral Med Oral Pathol Oral Radiol*, vol. 129, no. 6, pp. 635–642, 2020.
- [27] S. Zhao, Q. Luo, and C. Liu, "Automatic tooth segmentation and classification in dental panoramic X-ray images," 2020.
- [28] B. Silva, L. Pinheiro, L. Oliveira, and M. Pithon, "A study on tooth segmentation and numbering using end-to-end deep neural networks," in *2020 33rd SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, 2020, pp. 164–171.
- [29] C. Sheng *et al.*, "Transformer-based deep learning network for tooth segmentation on panoramic radiographs," *J Syst Sci Complex*, vol. 36, no. 1, pp. 257–272, 2023.
- [30] S. Arora, S. K. Tripathy, R. Gupta, and R. Srivastava, "Exploiting multimodal CNN architecture for automated teeth segmentation on dental panoramic X-ray images," *Proc Inst Mech Eng H*, vol. 237, no. 3, pp. 395–405, 2023.
- [31] M. Kanwal, M. M. Ur Rehman, M. U. Farooq, and D.-K. Chae, "Mask-transformer-based networks for teeth segmentation in panoramic radiographs," *Bioengineering*, vol. 10, no. 7, p. 843, 2023.
- [32] M. K. Dhar, M. Deb, D. Madhab, and Z. Yu, "A Deep Learning Approach to Teeth Segmentation and Orientation from Panoramic X-rays," *arXiv preprint arXiv:2310.17176*, 2023.
- [33] A. Ghafoor, S.-Y. Moon, and B. Lee, "Multiclass Segmentation Using Teeth Attention Modules for Dental X-Ray Images," *IEEE Access*, vol. 11, pp. 123891–123903, 2023.
- [34] Y. Zhang *et al.*, "Children's dental panoramic radiographs dataset for caries segmentation and dental disease detection," *Sci Data*, vol. 10, no. 1, p. 380, 2023.
- [35] W. Brahmi and I. Jdey, "Automatic tooth instance segmentation and identification from panoramic X-Ray images using deep CNN," *Multimed Tools Appl*, vol. 83, no. 18, pp. 55565–55585, 2024.
- [36] E. Asci *et al.*, "A Deep Learning Approach to Automatic Tooth Caries Segmentation in Panoramic Radiographs of Children in Primary Dentition, Mixed Dentition, and Permanent Dentition," *Children*, vol. 11, no. 6, p. 690, 2024.
- [37] S. Wathore and S. Gorthi, "Bilateral symmetry-based augmentation method for improved tooth segmentation in panoramic X-rays," *Pattern Recognit Lett*, vol. 188, pp. 1–7, 2025.
- [38] G. Altan and A. Al Samar, "Tooth segmentation on dental panoramic X-rays using Mask R-CNN," in *Mining Biomedical Text, Images and Visual Features for Information Retrieval*, Elsevier, 2025, pp. 481–498.
- [39] S. Hou, T. Zhou, Y. Liu, P. Dang, H. Lu, and H. Shi, "Teeth U-Net: A segmentation model of dental panoramic X-ray images for context semantics and contrast enhancement," *Comput Biol Med*, vol. 152, p. 106296, 2023.

BIOGRAPHIES



Şilan Fidan Vural received the bachelor's degree in computer engineering from Yildiz Technical University, Turkey. Vural is a Machine Learning Engineer at Wiro AI, Turkey. Her field of study includes generative AI.



Nida Kumbasar received the B.Sc. degree in Computer Engineering Department from Ataturk University, Erzurum, Turkey, in 2015. She received the integrated Ph.D. degree in Electrical and Electronics Engineering Department from Ataturk University, Erzurum, Turkey, in 2024. She completed their PhD as a recipient of the YÖK 100/2000 PhD Scholarship

Program, a competitive funding initiative by the Council of Higher Education of Turkey (YÖK) to support doctoral research in priority fields. Throughout this period, she gained professional experience by working in various companies in the private sector within the field of computer science. She is currently a Senior Researcher at TÜBİTAK, Informatics and Information Security Research Center (BİLGEM), Kocaeli, Turkey. Her research interests include medical image processing, remote sensing, data valuation, signal processing, and deep learning.

Breast Cancer Detectability and Tumor Differentiation Based on Microwave Dielectric Property Changes with Reverse Time Migration

Cemanur Aydinalp and Gulsah Yildiz

Abstract—Breast cancer detection and treatment have advanced significantly with imaging technologies, but challenges remain in distinguishing the type and stage of tumors. Microwave imaging (MWI) offers a promising alternative due to its non-ionizing nature and its ability to exploit dielectric property (DP) contrast. This study investigates the effectiveness of MWI in detecting and characterizing tumors using a phantom for breast tissue and tumor-mimicking NaCl solutions with various DPs (0.1 M, 0.2 M, 0.4 M and 0.8 M). First, the Cole-Cole parameters of these materials were calculated using DP measurements obtained from the open-ended coaxial probe method in order to provide broadband frequency analysis. Furthermore, the developed MWI system was utilized to evaluate tumor detectability and differentiation based on these DP changes. The MWI experiment was performed with 12 Vivaldi antennas between 0.6-2.6 GHz, and the results were analyzed from two different positions. The results indicate that the MWI system can effectively distinguish tumors with different DPs from each other using quantitative differential imaging due to its sensitivity to variations. To this end, the inverse time migration (RTM) method was employed to compare reference-target pairs (RTP) to generate an image of a tissue-mimicking phantom with tumors. The results show a high correlation between RTP image contrast and the target-reference DP difference.

Index Terms— Dielectric property of tissue-mimicking materials, microwave imaging, breast cancer detection, open-ended coaxial probe, Cole-Cole parameters.

I. INTRODUCTION

BREAST cancer is one of the most common cancers worldwide [1]–[3]. The complexity of breast tissue has made the detection, diagnosis and treatment of breast cancer the subject of extensive and long-term research [4]–[7]. These studies comprise the measurement of tumor dielectric properties and tissue classification [4], [5], accurate tumor localization [6], [7], tumor stage detection [8], [9] and tumor treatment [10], [11]. Accurate localization and differentiation of the tumor into carcinogenic stages are crucial for the patient's treatment process and early diagnosis is essential in the breast cancer treatment plan [12]. Diagnostic methods include mammography, ultrasound and magnetic resonance imaging

(MRI). Mammography uses X-rays, a type of ionizing radiation, to detect tumors [8]. Although ultrasound and MRI do not emit harmful radiation, ultrasound provides low-resolution images and lacks consistency in archival quality due to differences in measurement positions by different practitioners [13]. In contrast, while MRI does not emit harmful waves, it is a costly option for early detection [14].

Microwave imaging (MWI) is an emerging technique with applications in medical imaging, earth observation and other fields [15]–[17]. MWI employs microwaves, a non-ionizing portion of the electromagnetic spectrum ranging from 0.3 to 30 GHz. This specific range provides the precise frequencies necessary to penetrate biological tissues and allows imaging at acceptable resolution without emitting high power levels into the body. MWI primarily utilizes the dielectric properties (DPs) contrast of materials to create images. Normal and abnormal tissues exhibit a wide range of DPs, which provide the contrast necessary for imaging [18]. This is particularly advantageous in dense breast tissue, where tumors have significantly different dielectric properties compared to healthy tissues. These differences increase the efficiency of MWI. In contrast, other imaging techniques are less sensitive in detecting differences between healthy and tumor regions in dense breast tissue [19].

Quantitative MWI techniques aim to provide an image where the contrast directly depends on the magnitude of physical parameters such as DPs. These methods are often iterative, computationally expensive, and time-consuming [20]. On the other hand, qualitative imaging techniques are faster and easier to implement, providing an indication of the magnitude correlation [21]. The linear sampling method (LSM) and the factorization method (FM) are the two most commonly used qualitative methods for MWI [21]. These methods have similar mathematical backgrounds and do not require prior knowledge. They can be formulated from the electric field [22] or scattering parameters [21], while the latter is more efficient to use directly from an experimental setup. Near-field orthogonality sampling method (NOSM), multiple signal classification (MUSIC) and reverse time migration (RTM) methods are some other qualitative methods used in MWI [23]–[25]. In [25], the electric field mathematical background for the RTM method is provided and extended to use the scattering parameters obtained from an experimental setup. A frequency range of 2-4 GHz with a step size of 0.1 GHz is used for single-slice imaging of a tissue-mimicking phantom. Furthermore, this approach is applied to detect breast cancer using differential imaging, which increases the contrast of the target region relative to the

© Cemanur Aydinalp and © Gulsah Yildiz are with the Department of Electronics and Communication Engineering, Faculty of Electrical and Electronics Engineering, Istanbul Technical University, Istanbul, 34469, TURKEY e-mail: aydinalp16@itu.edu.tr
Manuscript received Jul 30, 2024; accepted Jan 21, 2025.
DOI: [10.17694/bajece.1521841](https://doi.org/10.17694/bajece.1521841)

background by using the breast pair as the reference and the target [26]. In addition, DL-enhanced RTM has the potential to improve computational efficiency and result accuracy in medical microwave imaging [27].

The Cole-Cole method, a mathematical model, is frequently used in the literature to represent dielectric behavior over a wide frequency range using a few parameters [28]. This model provides essential parameters to establish the DP of materials based on the operation frequency. Particle swarm optimization and the generalized Newton-Raphson methods are employed to determine the Cole-Cole parameters based on DP measurements. Therefore, it is crucial to accurately obtain the DP of tissue-mimicking phantoms based on the research to ensure precision in biomedical applications.

In this study, the dielectric properties of tumors were altered using breast tissue-mimicking phantom, and the detectability of these changes was analyzed using MWI methods. The main contributions of this study are:

- The MWI experimental setup and the MWI algorithm to identify the normal and abnormal tissue characteristics are demonstrated.
- A detailed analysis of how tumors with varying DPs can be differentiated from each other using the MWI algorithm is given.
- The correlation between DP differences and the corresponding MWI results is presented.
- The Cole-Cole parameters for the DPs of breast tissue-mimicking phantom and tumor materials are provided.

These contributions collectively offer the potential to distinguish the type or stage of a tumor rather than merely identifying the presence of a tumor. Furthermore, during treatment, the DP of the tumor can undergo changes, and the results demonstrate that the MWI system is capable of detecting these alterations. The remainder of this paper is organized as follows: Section 2 details the formulation and preparation of the tissue-mimicking phantom, the experimental setups for DP measurements, and the MWI system. Additionally, this section describes the extraction of Cole-Cole parameters from DP characterization. Section 3 presents the results regarding the sensitivity of the MWI system to variations in DPs. The sensitivity analysis is performed based on a comparison of DP measurements and MWI methods. Conclusions are drawn in Section 4.

II. MATERIAL AND METHOD

A. Tissue-Mimicking Phantom

Tissue-mimicking phantoms are employed in medical fields to develop and assess emerging devices before conducting animal experiments and clinical tests. Therefore, several studies in the literature have been carried out to simplify the preparation of the phantom and enhance its long-term usability [18], [29], [30]. In this study, a breast fat-mimicking phantom shown in Fig. 1a was prepared using the formula obtained from [18]. The phantom preparation procedure was repeated twice to designate areas for the placement of NaCl solutions (0.1 M, 0.2 M, 0.4 M,

and 0.8 M). To summarize the phantom preparation, the following steps were followed:

- 17 g (dry mass) of calfskin gelatin (obtained from Vyse Gelatin Company, Schiller Park, IL, USA) was added to 95 ml of distilled water. The key procedure to facilitate gelatin dissolution in water involves sprinkling the gelatin over the water to prevent clumping.
 - A total of 400 ml of oil and the mixture of distilled water and gelatin were covered with plastic film and placed into a hot water jacket. The transition of the gelatin-water mixture to a transparent yellow color is a crucial part of this step.
 - When the oil and mixture cooled down to 50°C, the oil was gradually added while stirring with a magnetic stirrer.
 - Next, 0.56 ml of Fairy liquid surfactant (Procter and Gamble, Turkey) was introduced into the mixture while stirring. To prevent bubbling, the stirrer speed was reduced during the addition of the detergent.
 - Then, a formaldehyde solution (1.08 g) was incorporated into the mixture.
 - Finally, the liquid form of the phantom was poured into half of the breast model. After the first phantom solidified, a straw with a sealed bottom was placed over the solid phantom. A second liquid phantom, prepared using the same procedure, was poured into the remaining part of the breast model and allowed to solidify. The final breast phantom model consists of two layers, with an empty hole for the tumor in the second layer.
- Note that the diameter and height of the breast phantom model are 12 cm and 14 cm, respectively. The diameter and length of the straw placed in the phantom to represent the tumor are 1 cm and 10 cm.

B. Experimental Setup

The experimental setup consists of the dielectric property measurement and the imaging systems. First, N5230A PNA Series Network Analyzer (Santa Clara, CA, USA) and Speag DAK 3.5-mm-diameter open-ended coaxial probe (Zurich, Switzerland) were utilized to measure the dielectric property of the breast phantom and four different NaCl solutions (0.1 M, 0.2 M, 0.4 M, and 0.8 M). The dielectric property measurement setup is displayed in Fig. 1b. The frequency range of measurements was between 0.5-12 GHz with a resolution of 100 MHz, resulting in 116 frequency points.

Second, the imaging system is composed of 24 Vivaldi antennas connected to a 24-channel Rohde & Schwarz ZNBT8 network analyzer with RF cables (50Ω impedance). Two groups of 12 antennas were placed facing each other with a 14 cm gap in between, as in Fig. 1c. Antennas were placed between wooden insulators to prevent coupling.

C. Dielectric Property Characterization

In this section, the results obtained from the dielectric property measurement setup are examined for use in the imaging system. The dielectric property (real $-\epsilon'$ - and imaginary $-\epsilon''$ - parts) of NaCl solutions and breast phantom is shown in Fig. 2 and 3, respectively. Furthermore, the dielectric properties of phantom and NaCl solutions at 0.6 and 2.6 GHz are listed in Table I. The Cole-Cole parameters are calculated to allow broadband frequency analysis. In addition, these parameters can

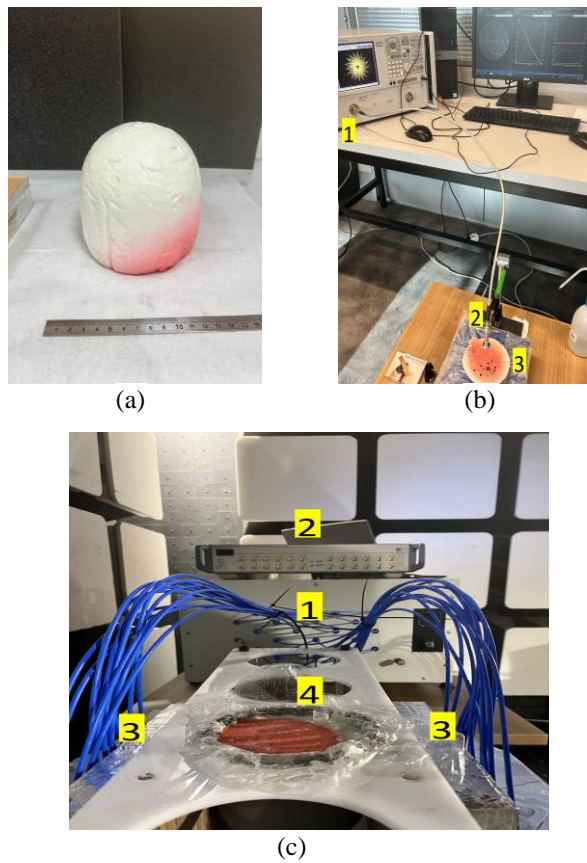


Fig.1. Breast phantom and experimental setup consisted of two measurement systems. (a) Breast phantom (colored to indicate two layers). (b) Dielectric property measurement system 1. N5230A PNA Series Network Analyzer, 2. Speag DAK 3.5-mm-diameter probe, 3. phantom. (c) Imaging system 1. 24-channel Rohde & Schwarz ZNB8 Network Analyzer, 2. Rohde & Schwarz ZN-Z154 calibration kit, 3. two sets of 12 antennas and 4. phantom.

be used in imaging systems with different frequency ranges without dielectric property measurements. To this end, the generalized Newton-Raphson method, an iterative method for solving nonlinear least squares problems, is used to calculate parameters of the one-pole Cole-Cole model [31], [32]:

$$\varepsilon(\omega) = \varepsilon_{\infty} + \frac{(\varepsilon_s - \varepsilon_{\infty})}{1 + (i\omega\tau)^{1-\alpha}} \quad (1)$$

$\varepsilon(\omega)$ is the complex dielectric permittivity as a function of angular frequency (ω). ε_{∞} and ε_s represent the high-frequency permittivity and the static (or DC) permittivity, respectively. τ stands for the relaxation time. α is the Cole-Cole parameter, typically between 0 and 1. Generalized Newton Raphson calculates the Cole-Cole parameters using the Euclidean Distance to evaluate the error rate between the calculated and measured dielectric properties, and the corresponding error calculation is given below:

$$e = \frac{1}{N} \sum_{i=1}^N \left[\left(\frac{\varepsilon'_{\omega_i} - \hat{\varepsilon}'_{\omega_i}}{\text{median}[\varepsilon'_{\omega_i}]} \right)^2 + \left(\frac{\varepsilon''_{\omega_i} - \hat{\varepsilon}''_{\omega_i}}{\text{median}[\varepsilon''_{\omega_i}]} \right)^2 \right] \quad (2)$$

ε' and ε'' are the real and imaginary parts of the measured dielectric property. $\hat{\varepsilon}'_{\omega_i}$ and $\hat{\varepsilon}''_{\omega_i}$ represent the real and

TABLE I
THE DIELECTRIC PROPERTY OF BREAST PHANTOM AND NaCl SOLUTIONS AT 0.6 AND 2.6 GHz.

Samples	ε'		ε''	
	0.6 GHz	2.6 GHz	0.6 GHz	2.6 GHz
Phantom	8.22	7.45	1.37	1.39
0.1 M	77.88	75.31	37.18	17.79
0.2 M	77.09	73.53	71.98	25.90
0.4 M	73.53	70.56	133.20	39.35
0.8 M	68.91	64.82	248.90	64.82

TABLE II
THE COLE-COLE PARAMETERS OF THE SAMPLES: BREAST PHANTOM AND NaCl SOLUTIONS WITH FOUR DIFFERENT MIXTURES.

Samples	Cole-Cole Parameters					Error (10^{-4})
	ε_s	ε_{∞}	τ (ps)	α	σ (S/m)	
Phantom	8.38	2.62	12.90	0.2710	0.028	9.33
0.1 M	77.03	9.45	8.9	0.0281	1.17	8.02
0.2 M	75.92	11.60	9.48	0.0473	2.35	8.91
0.4 M	73.04	6.04	8.13	0.0697	4.40	6.02
0.8 M	67.72	2.56	7.14	0.1257	8.19	6.37

imaginary parts of the data fitted to the Cole-Cole model. N is the number of frequency points. The algorithms continued to run until the error rate was below a threshold value of 0.001. The obtained Cole-Cole parameters of phantom and NaCl solutions are presented in Table II.

D. Imaging System Protocol

In the imaging system, a three-stage protocol was implemented to detect tumor location. First, the 24 cables connected to VNA were calibrated at 0.6-2.6 GHz using Rohde & Schwarz ZN-Z154 calibration kit. Thus, the reference plane was shifted from the network analyzer to the cable tip. Then, the cables are connected to the antennas. Note that the antennas are secured in a fixed position to prevent the bending of rigid RF cables. The phantom was then placed between the two antenna arrays. A dock (shown in Fig. 1c as white holder) was used to rotate the phantom from its center, with the tumor position closest to the fifth and ninth antennas throughout the experiments. Second, the measurements were initiated when the tumor location was not filled with NaCl solutions. After, NaCl solutions (0.1 M, 0.2 M, 0.4 M and 0.8 M) were sequentially injected into the empty tumor location with a syringe, 10 ml at a time, and the tumor location was drained before each injection. Thus, the measurement capabilities of the antennas were investigated for tumors with four different dielectric properties at two different locations in the imaging system. For the imaging results, a differential MWI approach was applied in which reference and target measurements were taken successively to detect the target tumor [7], [26]. The Inverse Time Migration (RTM) method was used as the imaging algorithm, and the inputs for the algorithm were the reflection and transmission coefficients obtained from the antennas [25]. Note that since the RTM method uses the absolute value of the difference of the scattering parameters,

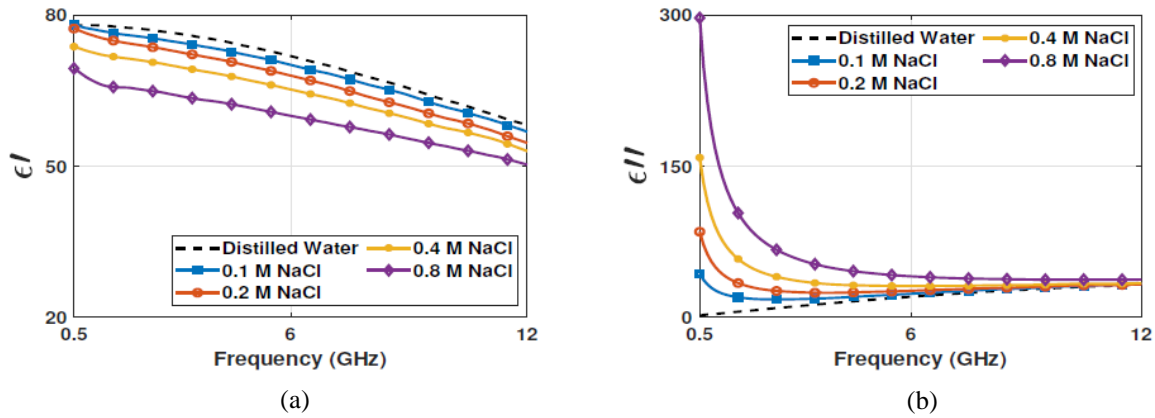


Fig. 2. Dielectric property of distilled water and NaCl solutions (0.1 M, 0.2 M, 0.4 M, and 0.8 M). (a) Real part and (b) imaginary of the dielectric properties at 0.5-12 GHz.

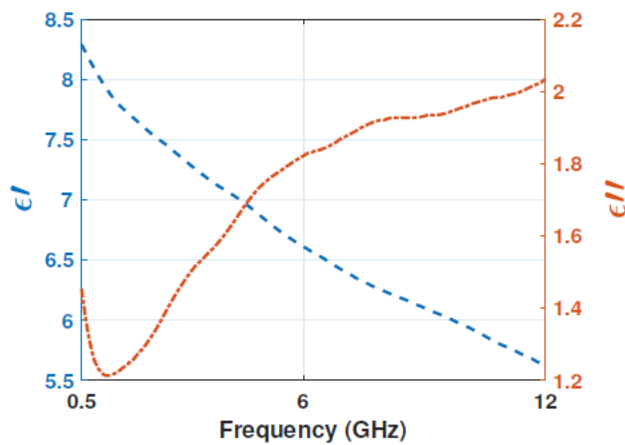


Fig. 3. Dielectric property (real part and imaginary) of breast phantom at 0.5 - 12 GHz.

the reference and target solutions can be used interchangeably. However, we will use these terms to facilitate clarity.

III. RESULTS

In this work, the sensitivity of the MWI system to changes in dielectric properties was investigated. To this end, four different NaCl solutions (0.1 M, 0.2 M, 0.4 M, and 0.8 M) were injected into the empty tumor location of the phantom. The percentage differences (PD) of the dielectric properties for the injected target NaCl solution with respect to the reference measurement at 0.6 and 2.6 GHz frequencies were calculated by the following equation:

$$PD = \left| \frac{\text{Target Value} - \text{Reference Value}}{\text{Reference Value}} \right| \times 100\%$$

The percentage differences of the dielectric properties are listed in Table III. For 0.6 GHz, in the case of 0.1 M reference solution, the percentage differences of 0.2 M, 0.4 M, and 0.8 M are 33, 63.28, and 82.05, respectively. When the reference solution was 0.2 M, the difference values 40.38 for 0.4 M and 68.58 for 0.8 M were calculated. Between 0.4 M reference and 0.8 M target solutions, the difference equals 44.83. In the case

of distilled water (DW) reference solution, the percentage differences of 0.1 M, 0.2 M, 0.4 M, and 0.8 M are 41.17, 66.69, 86.65, and 95.80, respectively. The corresponding difference values for 2.6 GHz are provided in the second row of Table III. Therefore, the percentage differences derived from the data collected at 0.6 GHz and 2.6 GHz indicate that the highest differences occur at 0.6 GHz. This trend suggests that lower frequencies are more effective for tumor differentiation.

The results, the images of reference-target pairs (RTP), shown in Fig. 4 and Fig. 5, are obtained from the qualitative differential MWI algorithm using reference-target pairs suggested in Table III. For example, to analyze the results (in Fig. 4a) for the 0.1 M-0.2 M difference, an experiment was performed for the fifth antenna position. First, the scattering parameters were collected when the tumor region was filled with the reference 0.1 M NaCl solution and the tumor was positioned as close to the fifth antenna as possible using the dock. Then, the region was drained, and the second set of measurements was collected when the tumor region was filled with 0.2 M NaCl solution. The collected scattering parameters were subtracted from each other as the differential imaging suggests, and the final image was retrieved by the RTM method, as displayed in Fig. 4a. The highest value of the image, which is 0.049, is close to the fifth antenna, as expected, in accordance with the tumor position. Keeping the reference as 0.1 M, two more measurements were collected when the tumor region was filled with 0.4 M and 0.8 M NaCl solutions, and the highest values of the retrieved images are at the same position as the first image (Fig. 4b and Fig. 4c). However, the highest values are 0.103 for 0.1 M-0.4 M image of RTP and 0.144 for 0.1 M-0.8 M image of RTP. For Fig. 4a-4c, the increase of the highest value in the images is in good agreement with the increase of the difference values for the two different frequency points given in Table III. Changing the reference solution to 0.2 M, two measurements were performed for targets 0.4 M and 0.8 M solutions, and the highest values of the resulting images of RTP are 0.058 and 0.115, respectively, as illustrated in Figs. 4d and 4e. Fig. 4f shows the image of RTP for 0.4 M-0.8 M pair with 0.063 as the highest value. For all the images in Fig. 4, the highest value positions indicate the tumor positions.

TABLE II
THE PERCENTAGE DIFFERENCES OF THE DIELECTRIC PROPERTIES FOR REFERENCE AND TARGET SOLUTIONS AT 0.6 AND 2.6 GHz.

Samples	0.1-0.2 M	0.1-0.4 M	0.1-0.8 M	0.2-0.4 M	0.2-0.8 M	0.4-0.8 M	DW-0.1 M	DW-0.2 M	DW-0.4 M	DW-0.8 M
PD (%)	33	63.28	82.05	40.38	68.58	44.83	41.17	66.69	86.65	95.80
	10.65	27.33	52.56	17.05	43.51	28.48	10.87	21.44	37.73	61.74

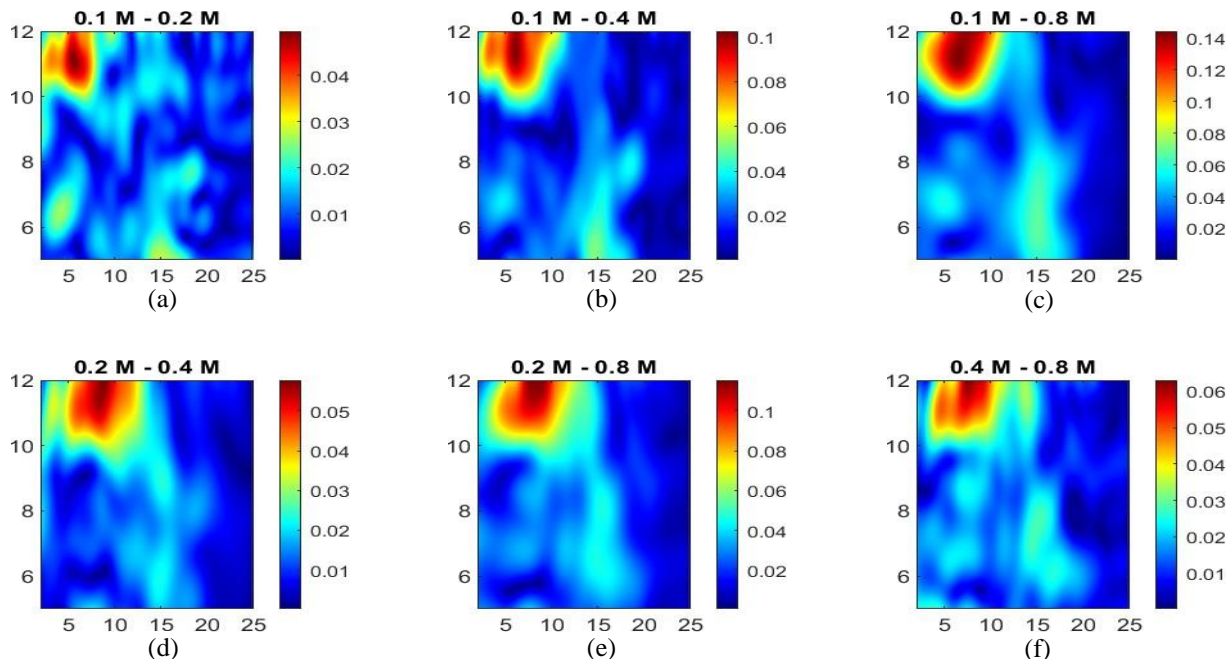


Fig. 4. Qualitative differential microwave imaging (MWI) results when the tumor is positioned near the fifth antenna. The reference-target pairs (RTP) are as follows: (a-c) 0.1 M solution against 0.2 M, 0.4 M, and 0.8 M, respectively; (d-e) 0.2 M solution as the reference for 0.4 M and 0.8 M targets; (f) 0.4 M solution as the reference for 0.8 M target.

Fig. 5 shows the images of RTP when the tumor region was positioned as close to the ninth antenna as possible. The highest values of the images are all at the same position, indicating the position of the tumor region. When the reference solution was 0.1 M, the highest values of the RTP images were 0.051, 0.111, and 0.158 for the targets 0.2 M, 0.4 M, and 0.8 M, respectively. Moreover, when 0.2 M is the reference, the highest values of the RTP images for 0.4 M and 0.8 M targets are 0.061 and 0.118, respectively. Last, the 0.4 M - 0.8 M RTP image has the highest value of 0.053. Note that the highest values obtained for the same reference-target pairs are comparable when the tumor position is in the fifth and ninth positions. Finally, distilled water was utilized as reference, and four NaCl solutions were measured as targets (Fig. 5g, 5h, 5i and 5j). Since the dielectric properties between distilled water and the targets exhibit significant differences, the corresponding highest values (0.88, 0.138, 0.199, and 0.245) are higher than those of the other pairs.

IV. CONCLUSION

Breast cancer remains one of the most prevalent cancers worldwide, driving comprehensive research into its detection, diagnosis, and treatment, including the dielectric properties measurement of tumors, tissue classification, accurate tumor localization, staging of tumors, and the development of

treatment strategies. This study aims to investigate whether the MWI system can distinguish differences in the DPs of tumor tissues. To this end, a phantom was prepared to mimic healthy breast tissue, and four different NaCl solutions (0.1 M, 0.2 M, 0.4 M, and 0.8 M) were utilized to represent tumors producing variation in DPs. The DPs of the phantom and NaCl solutions were measured with an open-ended coaxial probe, and the Cole-Cole parameters were retrieved from these measurements using the generalized Newton-Raphson method. Furthermore, the MWI system was utilized to collect data using 12 Vivaldi antennas operating between 0.6 and 2.6 GHz from a sample containing both healthy and tumor-mimicking materials. These data were used in the RTM imaging algorithm to demonstrate that MWI can accurately detect and differentiate tumors with different DPs. The results are analyzed for two different positions: close to the fifth and ninth antennas. As the DPs discrepancy between tumors increases, the RTP image shows higher values due to the characteristics of the imaging algorithm representing tumor location based on reference and target pairs. For instance, with a 0.1 M reference solution, the RTP images produced the highest values of 0.051, 0.111, and 0.158 for targets of 0.2 M, 0.4 M, and 0.8 M, respectively. Additionally, the ability of the MWI system to detect the changes in DPs would be an essential tool for monitoring the temperature-dependent DPs during the heat treatment processes. Therefore, the results highlight that the MWI system not only detects the presence of tumors but also distinguishes their type and stage

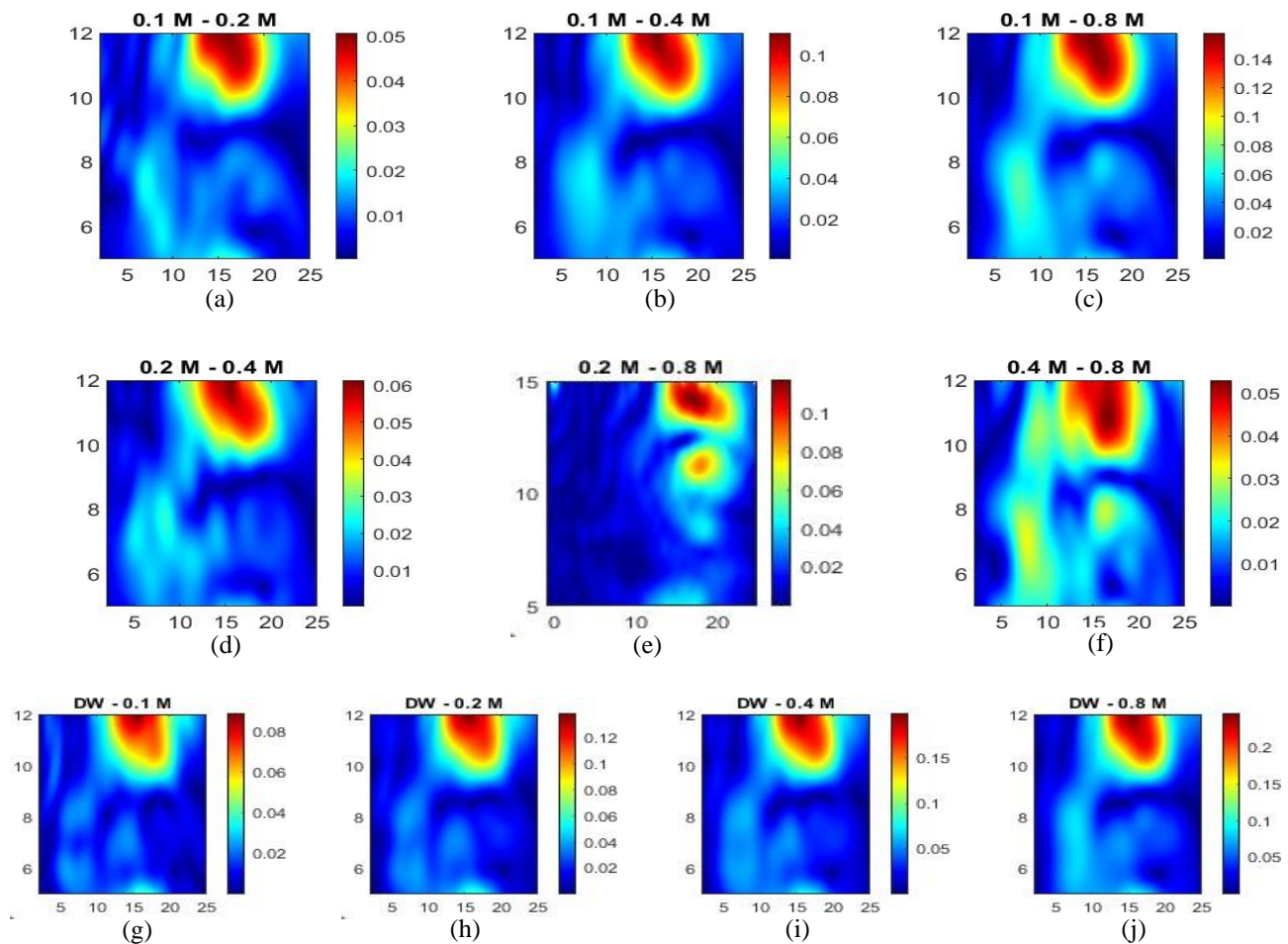


Fig. 5. Qualitative differential MWI results when the tumor is positioned near the ninth antenna. The RTP images are as follows: (a-c) 0.1 M solution against 0.2 M, 0.4 M, and 0.8 M, respectively; (d-e) 0.2 M solution as the reference for 0.4 M and 0.8 M targets; (f) 0.4 M solution as the reference for 0.8 M target; (g-j) distilled water (DW) as the reference for 0.1 M, 0.2 M, 0.4 M, and 0.8 M targets, respectively.

based on changes in DPs. Apart from these, providing Cole-Cole parameters in this study enhances both the reproducibility and sensitivity of MWI in biomedical applications. To conclude, MWI systems are a promising, non-invasive, and cost-effective method for breast cancer detection and monitoring. In future work, further adjustment is required to be implemented to utilize in clinical environment. To this end, extensive datasets can be generated to enable the training, validation, and testing of DL models in conjunction with the RTM algorithm. Therefore, with these advancements, MWI systems could significantly contribute to early diagnosis and more effective treatment strategies for breast cancer. Although numerous studies have been conducted on tumor detection, the importance of distinguishing between different tumor types has often been overlooked. This study emphasizes the significance of recognizing variations among tumor conditions, highlighting the need for accurate differentiation in medical imaging for improved diagnosis and treatment. However, several limitations should be noted. Firstly, the experiments were conducted using phantoms, which, although useful for simulating tumor conditions, do not fully replicate the complexity of human tissues. Clinical validation is essential to confirm the results and refine the system's performance in real-world applications. Additionally, the study focused on a limited

frequency range (0.6-2.6 GHz), and further investigation into the effect of different frequencies and more diverse tissue models could enhance the system's accuracy. Continued advancements in this technology could significantly improve early diagnosis and treatment.

ACKNOWLEDGMENT

The authors are with the Laboratory for Medical Device Research, Development and Application, Istanbul Technical University, Istanbul, 34469, TURKEY. Furthermore, this study was supported by the Research Fund of the Istanbul Technical University Project Number MAB-2024-45450 and by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant Number 123N484. The authors thank TUBITAK for their support.

REFERENCES

- [1] A. E. Giuliano, R. C. Jones, M. Brennan, and R. Statman, "Sentinel lymphadenectomy in breast cancer." *Journal of Clinical Oncology*, vol. 15, no. 6, pp. 2345-2350, 1997.
- [2] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209-249, 2021.

- [3] R. L. Siegel, N. S. Wagle, A. Cercek, R. A. Smith, and A. Jemal, "Colorectal cancer statistics, 2023," *CA: a cancer journal for clinicians*, vol. 73, no. 3, pp. 233–254, 2023.
- [4] C. Ss, R. K. Mishra, A. Swarup, and T. Jm, "Dielectric properties of normal & malignant human breast tissues at radiowave & microwave frequencies," *Indian journal of biochemistry & biophysics*, vol. 21 1, pp. 76–9, 1984. [Online]. Available: <https://api.semanticscholar.org/CorpusID:35827569>
- [5] E. Onemli, S. Joof, C. Aydinalp, N. Pastacı Özsoğacı, F. Ates, Alkan, N. Kepil, I. Rekik, I. Akduman, and T. Yilmaz, "Classification of rat mammary carcinoma with large scale in vivo microwave measurements," *Scientific reports*, vol. 12, no. 1, p. 349, 2022.
- [6] P. C. Götzsche and K. J. Jørgensen, "Screening for breast cancer with mammography," *Cochrane database of systematic reviews*, no. 6, 2013.
- [7] A. Janjic, M. Cayoren, I. Akduman, T. Yilmaz, E. Onemli, O. Bugdayci, and M. E. Aribal, "Safe: A novel microwave imaging system design for breast cancer screening and early detection—clinical evaluation," *Diagnostics*, vol. 11, no. 3, p. 533, 2021.
- [8] J. C. Lashof, I. C. Henderson, and S. J. Nass, "Mammography and beyond: developing technologies for the early detection of breast cancer," 2001.
- [9] J. N. Wolfe, "Breast patterns as an index of risk for developing breast cancer," *American Journal of Roentgenology*, vol. 126, no. 6, pp. 1130–1137, 1976.
- [10] N. I. of Health Consensus Development Panel et al., "Special report. treatment of primary breast cancer," *N Engl J Med*, vol. 301, p. 340, 1979.
- [11] G. Yildiz, H. Yasar, I. E. Uslu, Y. Demirel, M. N. Akinci, T. Yilmaz, and I. Akduman, "Antenna excitation optimization with deep learning for microwave breast cancer hyperthermia," *Sensors*, vol. 22, no. 17, p. 6343, 2022.
- [12] I. Barco, C. Chabrera, M. G. Font, N. Gimenez, M. Fraile, J. M. Lain, M. Piqueras, M. C. Vidal, M. Torras, S. Gonza'lez et al., "Comparison of screened and nonscreened breast cancer patients in relation to age: a 2-institution study," *Clinical Breast Cancer*, vol. 15, no. 6, pp. 482–489, 2015.
- [13] B. Ranger, P. J. Littrup, N. Duric, P. Chandiwalla-Mody, C. Li, S. Schmidt, and J. Lupinacci, "Breast ultrasound tomography versus MRI for clinical display of anatomy and tumor rendering: preliminary results," *American Journal of Roentgenology*, vol. 198, no. 1, pp. 233–239, 2012.
- [14] E. Aslan and Y. Ozupak, "Comparison of machine learning algorithms for automatic prediction of Alzheimer's disease," *Journal of the Chinese Medical Association*, pp. 10–1097, 2024.
- [15] S. Dey and A. O. Asok, "A review on microwave imaging for breast cancer detection," in *2024 IEEE Wireless Antenna and Microwave Symposium (WAMS)*, 2024, pp. 1–5.
- [16] S. Di Meo, A. Cannata, C. Blanco-Angulo, G. Matrone, A. Martinez-Lozano, J. Arias-Rodriguez, J. M. Sabater-Navarro, R. Gutierrez-Mazon, H. Garcia-Martinez, E. Avila-Navarro et al., "Multi-layer tissue-mimicking breast phantoms for microwave-based imaging systems," *IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology*, 2024.
- [17] D. Bhargava, P. Rattanadecho, and K. Jiamjiroch, "Microwave imaging for breast cancer detection-a comprehensive review," *Engineered Science*, vol. 30, p. 1116, 2024.
- [18] M. Lazebnik, E. L. Madsen, G. R. Frank, and S. C. Hagness, "Tissue-mimicking phantom materials for narrowband and ultrawideband microwave applications," *Physics in Medicine & Biology*, vol. 50, no. 18, p. 4245, 2005.
- [19] P. E. Freer, "Mammographic breast density: impact on breast cancer risk and implications for screening," *Radiographics*, vol. 35, no. 2, pp. 302–315, 2015.
- [20] A. Yago Ruiz, M. Cavagnaro, and L. Crocco, "An effective framework for deep-learning-enhanced quantitative microwave imaging and its potential for medical applications," *Sensors*, vol. 23, no. 2, p. 643, 2023.
- [21] M. N. Akinci, T. Caglayan, S. Ozgur, U. Alkasi, H. Ahmadzay, M. Abbak, M. Cayoren, and I. Akduman, "Qualitative microwave imaging with scattering parameters measurements," *IEEE Transactions on Microwave Theory and Techniques*, vol. 63, no. 9, pp. 2730–2740, 2015.
- [22] M. N. Akinci, M. Abbak, S. Özgür, M. Çayören, and I. Akduman, "Experimental comparison of qualitative inverse scattering methods," in *2014 IEEE Conference on Antenna Measurements & Applications (CAMA)*, 2014, pp. 1–4.
- [23] M. N. Akinci, M. Çayören, and I. Akduman, "Near-field orthogonality sampling method for microwave imaging: Theory and experimental verification," *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 8, pp. 2489–2501, 2016.
- [24] R. Fazli, M. Nakhkash, and A. A. Heidari, "Alleviating the practical restrictions for music algorithm in actual microwave imaging systems: Experimental assessment," *IEEE transactions on antennas and propagation*, vol. 62, no. 6, pp. 3108–3118, 2014.
- [25] E. Bilgin, M. Çayören, S. Joof, G. Cansiz, T. Yilmaz, and I. Akduman, "Single-slice microwave imaging of breast cancer by reverse time migration," *Medical Physics*, vol. 49, no. 10, pp. 6599–6608, 2022.
- [26] A. Abbosh, B. Mohammed, and K. Bialkowski, "Differential microwave imaging of the breast pair," *IEEE Antennas and Wireless Propagation Letters*, vol. 15, pp. 1434–1437, 2015.
- [27] M. Safak Kaplan, "Machine learning based augmentation of medical microwave imaging," Istanbul Technical University Graduate Program, 2022.
- [28] M. Lazebnik, M. Okoniewski, J. H. Booske, and S. C. Hagness, "Highly accurate Debye models for normal and malignant breast tissue dielectric properties at microwave frequencies," *IEEE microwave and wireless components letters*, vol. 17, no. 12, pp. 822–824, 2007.
- [29] C. Gabriel, "Tissue equivalent material for hand phantoms," *Physics in Medicine & Biology*, vol. 52, no. 14, p. 4205, 2007.
- [30] M. Y. Kanda, M. Ballen, S. Salins, C.-K. Chou, and Q. Balzano, "Formulation and characterization of tissue equivalent liquids used for R F dosimetry and dosimetry measurements," *IEEE Transactions on microwave theory and techniques*, vol. 52, no. 8, pp. 2046–2056, 2004.
- [31] B. Saçlı, C. Aydinalp, G. Cansız, S. Joof, T. Yilmaz, M. Çayören, B. Önal, and I. Akduman, "Microwave dielectric property-based classification of renal calculi: Application of a knn algorithm," *Computers in biology and medicine*, vol. 112, p. 103366, 2019.
- [32] U. B. Çalışkan, C. Aydinalp, and T. Y. Abdolsaheb, "Comparing two fitting algorithms to determine Cole-Cole parameters," in *2023 31st Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2023, pp. 1–4.

BIOGRAPHIES



Cemanur Aydinalp received the B.S. degree from the Department of Electronics Engineering, Ankara University, Ankara, Turkey, in 2011. The M.S. degree from the Department of Electrical and Computer Engineering, San Diego State University, San Diego, USA, in 2015. She completed her Ph.D. degree in the Department of Telecommunication

Engineering at Istanbul Technical University, Istanbul, Turkey. She is currently working as a research assistant at Istanbul Technical University. Her research interests include microwave dielectric spectroscopy, data analysis, optimization of open-ended coaxial probes, and application of supervised machine learning algorithms to engineering problems.



Gulsah Yildiz received the B.S. and M.Sc. degrees from the Department of Electrical and Electronics Engineering, İ.D. Bilkent University, Ankara, Turkey, in 2016 and 2018. She completed her Ph.D. degree in Telecommunication Engineering, Istanbul Technical University, Istanbul, Turkey. She is currently working as an assistant professor at Istanbul Technical University.

Her research interests are microwave imaging, microwave hyperthermia, the application of deep learning, and physics-induced neural network algorithms to engineering problems.

Control Through Contact using Mixture of Deep Neural-Net Experts

Aykut C. SATICI

Abstract—We provide a data-driven control design framework for hybrid systems, with a special emphasis on contact-rich robotic systems. These systems exhibit continuous state flows and discrete state transitions, which are governed by distinct equations of motion. Hence, it may be impossible to design a single policy that can control the system in all modes. Typically, hybrid systems are controlled by multi-modal policies, each manually triggered based on observed states. However, as the number of potential contacts increase, the number of policies can grow exponentially and the control-switching scheme becomes too complicated to parameterize. To address this issue, we design contact-aware data-driven controllers given by deep-net mixture of experts (MoE). This architecture automatically learns switching-control scheme that can achieve the desired overall performance of the system, and a gating network, which determines the region of validity of each expert, based on the observed states.

Index Terms—Robotics; nonlinear control; machine learning.

I. INTRODUCTION

A power grasp [1] involves firmly grasping the object on many sides, often with large contact forces, to immobilize the object relative to the gripper; immobilization simplifies object manipulation. This approach is widespread in factories and warehouses.

In household or agricultural settings, several tasks cannot be solved by power grasps. The reasons for this limitation include the object being fragile (fruit picking, contact with humans), constraints on its motion (transporting a bowl of soup that should not be tilted), or size (moving an exercise ball or large box). In effect, the gripper can no longer apply a net wrench (force-torque pair) in any direction on the object, and so the object is no longer guaranteed to move rigidly with the gripper. This scenario, known as *nonprehensile manipulation*, is critical for many tasks humans easily perform such as buttoning a shirt, making a salad, or spreading peanut butter on toast, but cannot be robustly performed by state-of-the-art robots.

Successful task execution under nonprehensile manipulation requires complex reasoning about how wrenches applied on the object interact with its natural dynamics. This reasoning is challenging to carry out due to the multiplicity of possible contact modes and the uncertain contact dynamics governing each mode. Recent advances automate the reasoning over multiple modes, resulting in high-quality plans for achieving a task using a sequence of contact modes [2]. However, to achieve computational tractability, these methods simplify the

robot-object dynamics by assuming that the robot-and-object are either always in equilibrium (quasi-static), or moving with constant velocity (quasi-dynamic), which constrain the versatility of nonprehensile manipulation. Moreover, the execution of this sequence typically relies on applying low-level controllers designed for single-mode tasks.

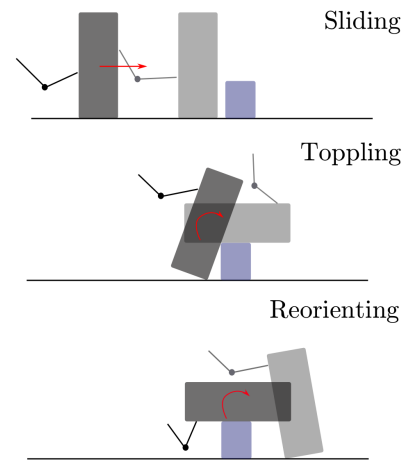



Fig. 1: Nonprehensile manipulation of a box via primitives

Another approach to nonprehensile manipulation is splitting tasks into *manipulation primitives* [3], [4], [5], such as rolling, sliding, throwing or toppling, where each primitive has a corresponding dynamics and is assigned its own controller. Consider the task of moving a box past an obstacle using a series of primitives such as sliding, toppling and reorienting as shown in Figure 1. Each primitive has a region of applicability in the state space, where the dynamics of that primitive describes the flow of the system [6]. Thus, manipulation planning involves identifying a successful primitive sequence, such as the order of primitives shown in Figure 1, and stabilizing the system under each primitive [7], [8]. This approach can be viewed as partitioning the state space and allocating a control law in each subdomain that results in a successful transition to the desired primitive until the goal state is reached[9]. However, the task of ordering the primitives and identifying the distinct controllers in each state partition is done manually [7], [9], and the control design problem is handled in a case-by-case basis.

Data-driven methods to nonprehensile object manipulation have been proposed in several recent works such as [10], which utilizes reinforcement learning ideas, [11], which leverages the recent advances on diffusion models. These approaches

 **Aykut C. Satici** is with the Mechanical and Biomedical Engineering Department, Boise State University, Boise, ID, 83725 USA e-mail: aykut-satici@boisestate.edu
Manuscript received Jul 13, 2024; accepted Jan 3, 2025.
DOI: 10.17694/bajece.1515854

are promising, the former suffering from typical reinforcement learning pitfalls, such as sample inefficiency and the latter typically requiring an imitation trajectory to learn from.

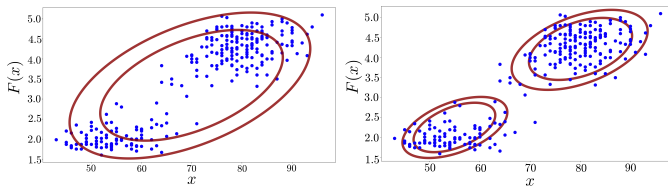
In this work, we present a data-driven approach for constructing dynamic motion plans and stabilizing control laws for complex locomotion and manipulation tasks that make and break contact. Our framework leverages the *mixture of experts* architecture from supervised learning to infer multi-modal controllers for contact-rich systems. This approach *automatically* learns the boundaries of the state partitions and allocates the appropriate expert controller to each partition in order to drive the system to the desired mode, and overall to the goal state. We demonstrate the efficacy of this technique on the swing-up task of the cartpole enclosed by wall barriers, both in simulation and real-world experiments.

II. BACKGROUND: MIXTURE OF EXPERTS

The mixture of experts (MoE) framework is a technique primarily used to learn an ensemble of regression models (experts) that best fit high variance or multi-modal datasets, such as the one shown in Figure 2a[12]. This technique provides a way to train several specialized expert models simultaneously, where each expert is well curated for a cluster of datasets as seen in Figure 2b. The MoE architecture uses a routing function called a *gating network* to allocate the appropriate local expert for each input data[13]. The objective is to learn the parameters of each local experts and the gating network to best fit the dataset.

Let $F(x; \theta)$ denote a collection of N_F expert models $F(x; \theta) := \{F_1(x; \theta_1), \dots, F_{N_F}(x; \theta_{N_F})\}$, whose parameters are given by the set $\theta = \{\theta_1, \dots, \theta_{N_F}\}$. The gating network is responsible for dividing the input space $\mathcal{X} \subset \mathbb{R}^m$ into *state partitions*, and assigning local expert models capable of providing specialized predictions for each partition. We represent the gating network with the discrete probability distribution $\mathbf{P}(x|\psi) := (P_1(x|\psi), \dots, P_{N_F}(x|\psi))$, where $P_i(x|\psi)$ denotes the probability of state x belonging to the state partition $\mathcal{X}^i \subset \mathcal{X}$ with the index $i \in \{1, \dots, N_F\}$. In the standard MoE framework[14], the prediction $u(x)$ of the MoE is given by

$$u(x) = \sum_{i=1}^{N_F} F_i(x; \theta_i) P_i(x|\psi), \quad (1)$$



(a) Multi-modal dataset fit with one model (b) Multi-modal dataset fit with MoE

Fig. 2: Comparison of multi-modal dataset fit with one regression model and MoE

which requires evaluating all the experts for each input x . We can reduce the computation cost of (1) by utilizing the output of the single best expert as determined by the gating network[15]

$$u(x) = \{F_a(x; \theta_a) \mid a = \underset{i}{\operatorname{argmax}} \{P_i(x|\psi)\}\}. \quad (2)$$

Model Structure: The expert models and the gating network can take several forms. Gaussian process (GP) models are commonly used in the MoE framework to infer a multi-modal probabilistic model from a small amount of data[13]. Despite the expressive power and tractability of GP experts, the inference procedure requires repeated matrix inversions that scale cubically with the size of the dataset[16]. In order to circumvent the large computational and memory overhead while also preserving the expressive power of GP experts, we leverage the universal approximation capabilities of neural networks for both the experts and the gating network. For regression problems that require the flexibility of nonlinear models, the experts can be given by deep neural-nets with point-estimate parameters, which can be extended to probabilistic models with the use of Bayesian neural networks, whose weights and biases are given by probability distributions[17]. Similarly, the gating network can be given by a neural network $\mathbf{P}(x|\psi) : \mathcal{X} \rightarrow \mathbb{R}^{N_F}$ with parameters ψ , and the output corresponds to the vector $[P_1(x|\psi), \dots, P_{N_F}(x|\psi)]$. In order to ensure that the probabilities $P_i(x|\psi)$ over all state partitions i sum to one, we use the SOFTMAX activation function[18] on the last layer of the gating network.

Training: Given the training dataset $\mathbb{D} = \{(x_1, y_1), \dots, (x_N, y_N)\}$ with N input state-label pairs, we can use gradient-based techniques to find the optimal parameters (ψ, θ) that best fit the dataset[15]. In such techniques, we construct the cost function we wish to minimize as

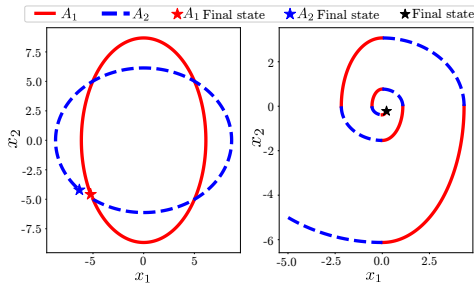
$$\mathbb{L}(\mathbb{D}) = \sum_{j=1}^N \sum_{i=1}^{N_F} \|F_i(x_j; \theta_i) - y_j\| P_i(x_j, \psi), \quad (3)$$

where $\|F_i(x_j; \theta_i) - y_j\|$ is the error in the prediction made by the expert i . Notice that the cost function (3) is minimum when the parameter θ_i has the lowest prediction error and the highest probability of getting selected by the gating network. So long as the complexity of the experts and the cost function allow for the pertinent gradients $\partial \mathbb{L} / \partial \psi, \partial \mathbb{L} / \partial \theta$ to be evaluated, we can invoke stochastic gradient descent (SGD) to update the decision parameters as follows:

$$\begin{aligned} \psi &\leftarrow \psi - \frac{\partial \mathbb{L}}{\partial \psi}, \\ \theta &\leftarrow \theta - \frac{\partial \mathbb{L}}{\partial \theta}. \end{aligned}$$

III. MOTIVATING APPLICATION: Switching Linear System

In the following discussion, we present an example to motivate and lay the foundation for the use of MoE in the control design problem. In particular, we propose a data-driven technique to automatically seek switching controllers



(a) Two marginally stable closed-loop systems (b) Asymptotically stable switching systems

Fig. 3: Stable switching between two marginally stable systems

for multi-modal systems. Suppose we have two linear systems of the form

$$\begin{aligned} \dot{x} &= A_1 x = \begin{bmatrix} 0 & -1 \\ 2 & 0 \end{bmatrix} x, \\ \dot{x} &= A_2 x = \begin{bmatrix} 0 & -2 \\ 1 & 0 \end{bmatrix} x, \end{aligned} \quad (4)$$

where each system is marginally stable as shown in Figure 3a. Although the individual systems are not asymptotically stable, it is possible to find a state-dependent switching rule that makes the resulting switched system stable[19] (Figure 3b). We aim to learn a gating network $\mathbf{P}(x|\psi)$ to automatically divide the state space into partitions and identify which of the two systems to execute in each state partition, with the goal of asymptotically stabilizing the origin.

Akin to the regression problem in Section II, the training dataset consists of *input state-label* pairs, where the labels are the performances of the trajectories generated under the current control law. In the case of the switching-control problem, we generate a trajectory and the corresponding performance metric (labels) as follows. Starting from some initial state $x(t=0)$, we sample a state partition index i from the categorical distribution, whose probabilities are provided by the gating network:

$$i \sim \text{Categorical}(\mathbf{P}(x(t)|\psi)).$$

Given the partition index i , the expert (control law) is given by a sample from the Bernoulli probability distribution

$$F_i(\theta_i) = \begin{cases} 0, & \theta_i > \frac{1}{2}, \\ 1, & \theta_i \leq \frac{1}{2}, \end{cases} \quad (5)$$

where $F_i = 0$ corresponds to the first dynamics $\dot{x} = A_1 x$ and $F_i = 1$ corresponds to $\dot{x} = A_2 x$. The parameter θ_i of the expert is to be learned, and it determines which of the two experts to execute in each partition. In order to ensure that the parameter θ_i of the expert serves as the probability of the Bernoulli distribution, we use the SIGMOID function[18] to limit θ_i between 0 and 1. The next state $x(t+\Delta t)$ in the trajectory is obtained from the following integration scheme:

$$x(t+\Delta t) = (1 - F_i)A_1 x(t) + F_i A_2 x(t).$$

We repeat this process to generate a trajectory for the time-

horizon T . The performance of the trajectory generated under the current parameters (ψ, θ) can be quantified by the metric ℓ as

$$\ell(x(t+\Delta t)) := \frac{1}{2} \|x(t+\Delta t)\|^2.$$

In Section IV-A, we generalize the performance metrics to be applicable to various dynamical systems and discuss how we can encode desired characteristics of the controller. From the performance metric ℓ , we can construct the cost function \mathbb{L} similar to the standard MoE framework in (3) as

$$\mathbb{L}(\{x(0), \dots, x(T)\}) = \sum_{t=0}^T \sum_{i=1}^{N_F} \ell_i(x_i(t+\Delta t)) P_i(x(t), \psi).$$

In the upcoming sections, we generalize the MoE control-search problem and provide techniques to efficiently learn the optimal decision parameters from appropriate cost functions.

IV. METHODS

Based on the motivating example provided in Section III, we present a generalized data-driven control design framework for hybrid dynamical systems. In this framework, the controller is given by deep-net mixture of experts $F(x;\theta)$, and the control switching scheme is governed by the gating network $\mathbf{P}(x|\psi)$. This technique allows us to observe the effects of mode changes from the closed-loop trajectories and learn a switching mechanism to best control the hybrid system across modes. The objective is to learn the parameters θ_i of each expert and the gating network ψ that can achieve the desired performance.

Let $\phi(x_0, u, T)$ denote a closed-loop trajectory generated from a hybrid dynamical model starting from initial state x_0 . We represent the dynamics of a hybrid system with the differential inclusions[20]

$$\begin{cases} \dot{x} \in f(x, u), & x \in C, \\ x^+ \in g(x, u), & x \in D, \end{cases} \quad (6)$$

where $x \in \mathbb{R}^m$ is the state vector, and $u \in \mathbb{R}^n$ is the input. The set-valued mappings $f: \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $g: \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}^m$ denote the flow and jump maps, respectively, where C and D are subsets of \mathbb{R}^m consisting of the feasible states under the flow and jump rules, respectively. The notation x^+ indicates the state resulted by the jump rule g .

Remark Another popular formulation for hybrid systems (e.g. mechanical systems that undergo contact) is through the utilization of measure differential inclusions (MDI)[21], which leads to a more succinct representation in terms of differential measures. This formulation represents equations (6) in the form

$$dx = f(x, u) dt + dR,$$

where dx , dt and dR represent various measures. The term dR is responsible for the changes in the system's continuous vector fields and state jumps whenever a contact is made or broken. We have used this formulation in Section V-B1 in order to model an example system on which we apply our data-driven controllers for verification. Readers who are interested in the theory of the MDI formulation are referred to the piece[21].

For every state x in a trajectory, the control law first samples state partition index i from a categorical distribution and evaluates the corresponding expert as

$$u(x; \psi, \theta) = \{F_i(x; \theta_i) \mid i \sim \text{Categorical}(\mathbf{P}(x|\psi))\}. \quad (7)$$

We use the metric $\ell : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R}$ to measure the performance of the sampled experts, which we discuss in depth in Section IV-A. The goal is to learn the decision parameters (ψ, θ) that minimize the metric ℓ for all initial states in the state space. We pose the search over the parameters of the experts and the gating network as the following optimization problem.

$$\begin{aligned} & \underset{\psi, \theta}{\text{minimize}} && \int_0^T \ell(x(t), u) dt, \\ & \text{subject to} && \begin{cases} \dot{x} \in f(x, u), & x \in C, \\ x^+ \in g(x, u), & x \in D, \end{cases} \\ & && u = \{F_i(x; \theta_i) \mid i \sim \text{Categorical}(\mathbf{P}(x|\psi))\}. \end{aligned} \quad (8)$$

In Section IV-C, we provide a procedure to solve the optimization problem (8) via stochastic gradient descent.

Remark Without prior knowledge injected to the gating network, the samples from the categorical distribution in (7) initially explore the performance of most, if not all, of the expert controllers. As the parameters converge to their optimal values, the samples from the categorical distribution correspond to the indices of the single best experts and the control law in (7) is equivalent to (2).

A. Performance Metrics

We present two viable choices for the performance metric ℓ .

1) **Accumulated cost**: is the total quadratic loss between the desired state x^* and the states generated under the current control law. We can also enforce control saturations for underactuated systems by incurring a cost on the control input as follows:

$$\ell(x, u) = \frac{1}{2}(x - x^*)^\top \mathcal{Q}(x - x^*) + \frac{1}{2}u^\top \mathcal{R}u, \quad (9)$$

where \mathcal{Q} and \mathcal{R} represent a positive definite and positive semi-definite gain matrices, respectively. This construction encourages trajectories to reach the desired equilibrium with minimum effort and shortest time. We modify the cost function \mathbb{L} presented in (3) to incorporate the quadratic loss $\ell(x, u)$ as follows:

$$\mathbb{L}(\phi) = \sum_{t=0}^T \sum_{i=1}^{N_F} \ell_i(x_i(t + \Delta t), F_i) P_i(x(t)|\psi). \quad (10)$$

Similar to the regression problem provided in Section II, the accumulated cost (10) is minimum when the metric ℓ_i achieved by expert i is low and the responsibility $P_i(x(t)|\psi)$ of the expert is high. Notice that the accumulated cost checks the performance of each expert at every state. When training for few experts, this cost function provides ample exploration, resulting in fast convergence to an optimal control strategy.

However, for numerous experts, the accumulated cost incurs large computational overhead.

2) **Minimum trajectory loss (MTL)**: is designed to minimize the computational complexities of the accumulated cost. Compared to (10), MTL may also better represent the desired behavior of some dynamical systems. Consider the classical control problem of swinging-up the simple pendulum to the upright equilibrium. For an underactuated pendulum, a successful controller needs to swing the pendulum clockwise and counterclockwise, passing through the downward equilibrium point multiple times until enough kinetic energy is built up to reach the upward equilibrium. Accumulated loss incurs high cost in such scenarios and the control search would get stuck in a local minimum. In such cases, a successful cost function encourages trajectories that *eventually* lead to the goal state. This is achieved by MTL, which is composed of the lowest cost incurred across the entire trajectory and the responsibilities of the experts that led to the minimum cost. The resulting cost function \mathbb{L} is given by

$$\begin{aligned} t_{\min} &= \inf_t \{ \ell(x(t), u) : x(t) \in \phi(x_0, u, T) \}, \\ \mathbb{L}(\phi) &= \frac{\ell(x(t_{\min}), u)}{C} \sum_{t=0}^{t_{\min}} P_i(x(t)|\psi), \end{aligned} \quad (11)$$

where $C > 0$ is a normalization factor. Unlike the accumulated cost, MTL does not particularly reward low effort or short time trajectories, but it equally rewards two trajectories as long as they both reach the goal state within the time horizon T .

B. State Sampling

We intend to find a solution to the optimization problem in (8) for all initial states x_0 in the state space. To do so, we compose the performance metric ℓ from a *batch of initial states* and update the parameters (ψ, θ) *iteratively* via stochastic gradient descent (SGD). To efficiently sample the initial states, we use a combination of greedy and explorative state sampling techniques. An example of greedy state sampling technique, commonly known as *Dataset Aggregation* (DAGGER), is a method adapted from imitation learning[22]. This technique collects states most visited under the current parameters (ψ, θ) and concentrates on refining the performance of the controller on these states. In detail, we first discretize the state space and uniformly sample several initial states. Starting from those initial states, we generate trajectories using the current parameters. In order to improve the controller at the states favored by the current policy, we draw N_d initial state samples from the states visited in the trajectories. This efficient state exposition is pivotal for a fast convergence to the optimal parameters.

The explorative state sampling technique exposes the training to the rewards of approaching and remaining close to x^* . It also uses random sampling to explore new control strategies and recover from locally optimal solutions. This method collects N_r initial states around the neighborhood of the desired equilibrium by drawing samples from the normal distribution $x_0 \sim \mathcal{N}(x^*, \Sigma)$, with mean x^* and the variance Σ is kept small. For each parameter update in SGD, we compute

Algorithm 1 Solution to the Optimization Problem (8)

```

1:  $\mathcal{D}_N \leftarrow \{x_0\}_{(N_D)}$   $\triangleright N_D$  initial state samples
2: while !(is converged) do
3:    $J \leftarrow 0$   $\triangleright$  Average cost function
4:   for  $x_0 \in \mathcal{D}_N$  do
5:      $\mathbb{L} = \text{Performance metric}(x_0, \psi, \theta)$   $\triangleright$  Section IV-A
6:      $J \leftarrow J + \mathbb{L}/N_D$ 
7:    $\theta \leftarrow \theta - \alpha \partial J / \partial \theta$   $\triangleright$  SGD step
8:    $\psi \leftarrow \psi - \alpha \partial J / \partial \psi$ 
9:    $\mathcal{D}_N \leftarrow \{x_0\}_{(N_D)}$   $\triangleright$  New initial state samples (Section IV-B)
10:   $i \leftarrow i + 1$ 
11: return  $\theta$ 

```

the performance metric as an expectation over $N_D = N_d + N_r$ samples as follows:

$$J(\phi, u) = \mathbb{E}_{x_0 \in \mathcal{D}_N} [\mathbb{L}(\phi(x_0, u, T))], \quad (12)$$

where \mathcal{D}_N is a *replay buffer* consisting of N_D initial state samples.

C. Training Mixture of Experts Controller

We solve the optimization problem in (8) following the procedure outlined in Algorithm (1). At the beginning of the training, we collect N_D initial states samples using the greedy and explorative state sampling techniques discussed in Section IV-B and save them in the replay buffer \mathcal{D}_N . For every initial state in the replay buffer, we generate a trajectory using the current decision parameters (ψ, θ) and assign the cost function \mathbb{L} . The average cost incurred by the current policy is given by J in (12), from which we compute the pertinent gradients $\partial J / \partial \psi, \partial J / \partial \theta$ via forward-mode auto-differentiation techniques[23]. We invoke a variant of stochastic gradient descent (SGD), known as ADAM[24] to efficiently update the parameters with adaptive learning rates α . The training is terminated when the average running cost J is below a small threshold for trajectories generated from various initial states.

V. CASE STUDIES

We demonstrate the efficacy of the MoE controller in simulation and real-world experiments. In the first case study, we learn a gating network that switches between two marginally stable closed-loop systems to result in a piecewise-asymptotically-stable system. Then, we find switching MoE controller to swing up the classical cartpole mechanism enclosed with wall barriers.

A. Switching Linear System

We find the stable switching scheme through the MoE framework discussed in Section III. We aim to learn the parameters ψ of the gating network $\mathbf{P}(x|\psi)$ and the expert parameters θ_i such that the switching system converges to the desired equilibrium $x^* = (0, 0)$. The gating network is a fully-connected neural net with one hidden layer (2 input states \rightarrow 6 neurons \rightarrow 4 outputs) and an ELU activation function[25].

We constrain the maximum number of state partitions to 4. Each state partition has a corresponding controller parameter $\theta_i \in \mathbb{R}$.

The response of the learned switching system is shown in Figure 4. Figure 4a shows the single best expert F_a given by (2) in each state partition, where purple corresponds to $F_a = 0$ or $\dot{x} = A_1x$ and yellow corresponds to $F_a = 1$ or $\dot{x} = A_2x$. The sample trajectory starts at $x_0 = [-5, -5]$ and successfully converges to the origin shown by the red star. The state partition index of the single best expert is shown in Figure 4b, and it depicts that the training uses only 3 out of the 4 state partitions available. The partitions in Figure 4b matches the analytical solution to the successful stable switching system given by[19]

$$\dot{x} = \begin{cases} A_1x, & x_1x_2 \leq 0, \\ A_2x, & x_1x_2 > 0, \end{cases}$$

where $x = [x_1, x_2]$.

The training progress is shown in Figure 5. The three rows in the figure depict the performance of the training after 0, 200 and 1400 parameter updates, respectively. The sample trajectory in Figure 5a shows that the initial parameters result in unstable switching between the two systems. After only 200 parameter updates, the training finds a stable switching mechanism, but it does not yet converge to the desired equilibrium x^* . In order to create an asymptotically stable system, the corners of each state partition must intersect at the origin, which the training finds successfully after 2000 parameter updates (Figure 4). This is thanks to the explorative state sampling technique, which samples states close to the desired equilibrium, assisting the training in finding the distinct boundaries of each partition at the origin.

B. Cartpole with Wall Contacts

In this section, we take the classical cartpole swing-up problem and introduce potential contacts from two barriers as shown Figure 6. The potential contacts serve as a way to convert the standard cartpole system into a multi-modal dynamics. The objective is to swing-up the pendulum on the

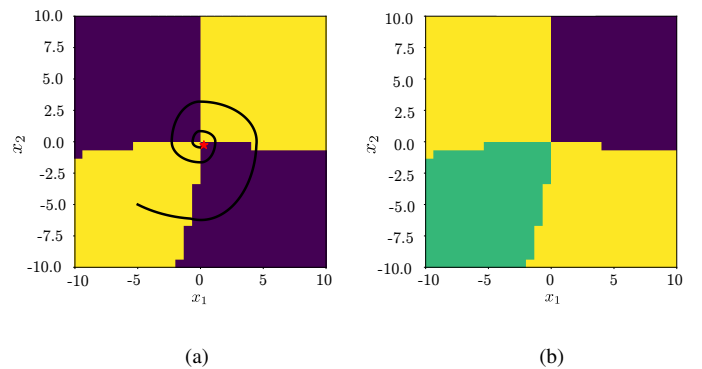


Fig. 4: Final stable switching system: (a) The single best expert F_a in each state partition, where purple corresponds to $F_a = 0$ or $\dot{x} = A_1x$ and yellow corresponds to $F_a = 1$ or $\dot{x} = A_2x$, (b) State partition index of the single best expert

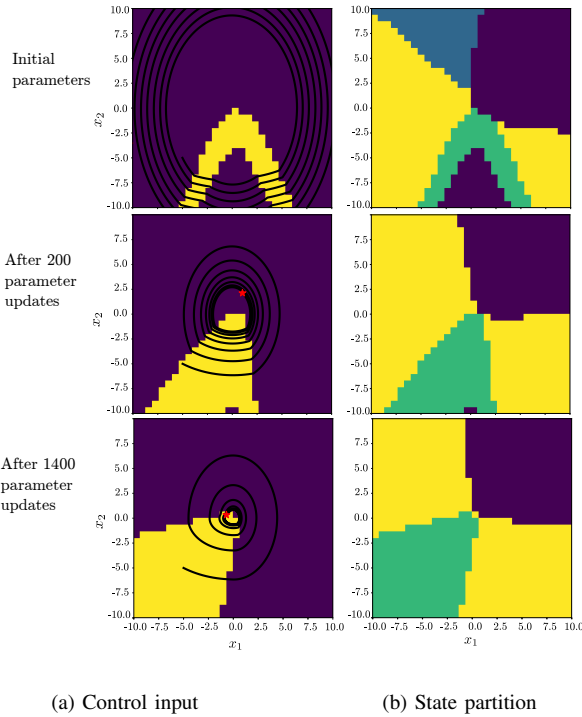


Fig. 5: Training progress. The final solution is shown in Figure 4

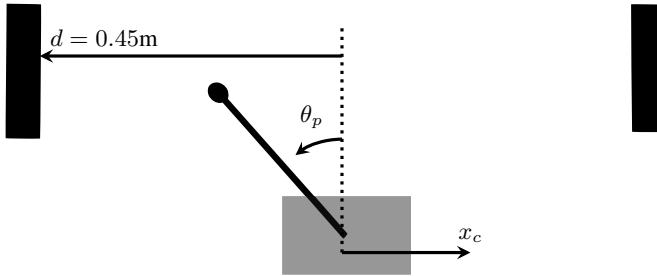


Fig. 6: Cartpole with wall contacts

cart in the presence of contacts and impacts. We apply the MoE framework to train switching expert controller and a gating network that governs the switching scheme. We demonstrate the performance of the MoE controller in simulation and real-world experiments. Lastly, we compare the performance of the MoE controller against a single swing-up controller.

1) *System Model*: The cartpole system consists of a freely rotating pendulum link hinged on an actuated cart. The setup is enclosed by two rigid walls hanging 0.2m from the bottom of the cart. The objective is to use the control authority on the cart in order to swing-up the pendulum to the upright. The pendulum spans length of $l = 0.31\text{m}$ and its mass $m_p = 0.75\text{kg}$ is concentrated at the distance $l_{cm} = 0.2\text{m}$ from the hinge. The cart alone has a mass of $m_c = 0.165\text{kg}$ and the moment of inertia of the pendulum is $I_p = 0.0022\text{ kg}\cdot\text{m}^2$. The viscous friction in the bearings of the cart's wheels is characterized by the coefficient $b = 1.2\text{ N}\cdot\text{sec}/\text{m}$.

Differentiable contact model: In order to infer the MoE controller for a contact-rich mechanism, we generate closed-

loop trajectories from an accurate contact model given by measure differential inclusions[21], [26], [27]. The dynamics of the cartpole under impacts, contacts and Coulomb friction is given by

$$M(q) d\dot{q} + h(q, \dot{q}) dt - dR = 0, \quad (13)$$

$$h(q, \dot{q}) := C(q, \dot{q})\dot{q} + G(q) - Bu(q, \dot{q}),$$

where $q = (x_c, \theta_p)$, x_c is the location of the cart, θ_p is the angle of the pendulum from the vertical. From the Euler-Lagrange formulation, the positive definite mass matrix $M(q)$, the Coriolis and centripetal terms $C(q, \dot{q})$, the gravitational terms $G(q)$, and the input-to-state mapping B are given by

$$M(q) = \begin{bmatrix} m_c + m_p & -m_p l_{cm} \cos(\theta_p) \\ -m_p l_{cm} \cos(\theta_p) & m_p l_{cm}^2 + I_p \end{bmatrix},$$

$$C(q, \dot{q}) = \begin{bmatrix} b & m_p l_{cm} \dot{\theta}_p \sin(\theta_p) \\ 0 & 0 \end{bmatrix},$$

$$G(q) = [0 \quad -m_p g l_{cm} \sin(\theta_p)]^\top,$$

$$B = [1 \ 0]^\top, \quad (14)$$

where g is the acceleration due to gravity. The force measure dR contains the contact forces as

$$dR = W_N d\lambda_N + W_T d\lambda_T,$$

where W_N and W_T are the projection matrices that map the effect of the normal and tangential contact forces, respectively, to the generalized coordinates. The vectors $d\lambda_N$ and $d\lambda_T$ consist of the normal and tangential contact impulse measures, respectively. In the presence of impacts, we integrate the contact measures over a singleton time t as $\int_{\{t\}} (d\lambda_N, d\lambda_T) = (\lambda_N(t), \lambda_T(t))$ in order to obtain the impulsive contact forces. In the case of persisting contact forces, the contact impulse measures evaluate to $(d\lambda_N, d\lambda_T) = (\dot{\lambda}_N, \dot{\lambda}_T)$, where $\dot{\lambda}_N$ and $\dot{\lambda}_T$ hold the normal and the tangential contact forces, respectively.

We use Moreau's time-stepping algorithm[27] to time-discretize and integrate the MDI. At each integration step, we compute the contact forces that enforce the geometric and kinematic constraints of rigid-bodies-in-contact via the *linear complementarity formulation*[27]. This formulation presents a linear complementarity problem (LCP) that searches for *contact force and post-impact velocity* pairs that obey the no-penetration conditions of rigid bodies. Without the presence of Coulomb friction, the LCP can be posed as a convex optimization problem, which can be solved analytically and can provide tractable gradients through the solution of the LCP. However, with the consideration of Coulomb friction, the LCP becomes a non-convex optimization problem, which may be intractable. In such scenarios, we solve the LCP through numerical methods, namely *Lemke's algorithm*[28], which allows us to use auto-differentiation techniques to evaluate the pertinent gradients during the training process.

2) *Training*: We aim to learn the parameters (ψ, θ) of the MoE controller in order to stabilize the cartpole system to the desired state $x^* = (q^*, \dot{q}^*) = ((0, 0), (0, 0))$ under contacts, impacts and Coulomb friction. Once the system reaches within a small neighborhood of x^* , we employ Linear

TABLE I: Structure of the deep-net experts and the gating network.

Neural Network	Inputs	Number of neurons in hidden layers	Outputs
Expert $F_i(x; \theta_i)$	$[x_c, \cos(\theta_p), \sin(\theta_p), \dot{x}_c, \dot{\theta}_p]$	(10, 4)	$u \in \mathbb{R}$
Gating network $\mathbf{P}(x \psi)$	$[x_c, \cos(\theta_p), \sin(\theta_p), \dot{x}_c, \dot{\theta}_p]$	(4, 3)	$[P_1, P_2, P_3]$

Quadratic Regulator (LQR) to maintain the system at the desired equilibrium. The structures of the deep-net experts and gating network are provided in Table I. We constrain the maximum number of state partitions to 3, where each partition has a local expert F_i . The output of the experts correspond to the force applied on the cart. We use minimum trajectory loss (MTL) discussed in Section IV-A with time horizon $T = 1.5s$, where the performance metric ℓ is given by (9). In each parameter update, we sample $N_D = 4$ initial states through greedy and explorative techniques.

3) *Hardware*: We demonstrate the performance of the MoE controller in simulation and hardware. The hardware (Figure 7), designed and built by QUANSER[29], uses a DC-motor to translate the cart on a track. The cart uses a rack-and-pinion mechanism to translate on the track with zero-slip. One of the wheels of the cart is attached to an optical encoder, from which we estimate the position and velocity of the cart. There is also an optical encoder rigidly attached to the pendulum link, reporting its orientation. We evaluate the experts and the gating network in MATLAB/Simulink and pass the corresponding voltage commands to the DC-motor via QUARC, QUANSER's real-time control software.

Parameter Estimation: We use an adaptation law[30] to estimate the parameters of the hardware in Figure 7, namely the mass m_c of the cart, viscous friction in the bearings of the cart's wheels denoted by b , and the effects of the cable harness on the cart, which we denote by the spring constant k . Consider the continuous dynamics of the cart whose equations of motion are given by the ordinary differential equation

$$m_c \ddot{x}_c + b \dot{x}_c + k x_c = u(x_c, \dot{x}_c), \quad (15)$$

where $u_c(x_c, \dot{x}_c)$ is a control input that is to be determined for tracking. We are uncertain of the constant parameters $\xi = [m_c \ b \ k]^\top$, whose estimates are denoted by $\hat{\xi} \in \mathbb{R}^3$. Let us introduce the errors in the position x_c , and parameters ξ to be

$$\tilde{x}_c = x_c - x_{c_r}, \quad \tilde{\xi} = \xi - \hat{\xi}, \quad e = [\tilde{x}_c \ \tilde{\xi}]^\top,$$

where x_{c_r} is a reference signal for the motion of the mechanical system. Inspired by Spong[30] et al., let us choose the control input according to

$$\begin{aligned} u(x_c, \dot{x}_c) &= Y(x_c, \dot{x}_c, a, v) \hat{\xi} - cr, \\ Y(x_c, \dot{x}_c, a, v) &= [a \ v \ x_c], \end{aligned} \quad (16)$$

where the quantities v , a , and r are given as

$$\begin{aligned} v &= \dot{x}_{c_r} - \lambda \tilde{x}_c, \\ a &= \dot{v} = \ddot{x}_{c_r} - \lambda \dot{\tilde{x}}_c, \\ r &= \dot{x}_c - v = \dot{\tilde{x}}_c + \lambda \tilde{x}_c, \end{aligned}$$

where $c, \lambda > 0$ are constant gains. Substituting the control law (16) into the system model (15) leads to

$$m \dot{r} + (b + c)r = -Y \tilde{\xi}. \quad (17)$$

The parameter estimate $\hat{\xi}$ may be computed using standard methods of adaptive control such as gradients or least squares. For example, for a positive definite matrix Γ of appropriate dimensions, we can use the gradient update law

$$\dot{\hat{\xi}} = -\Gamma^{-1} Y^\top(x_c, \dot{x}_c, a, v)r. \quad (18)$$

Proposition *The control (16) and adaptation law (18) stabilize the system to a reference trajectory while the estimates of the parameters tend to their correct values.*

Proof 1 *See Appendix*

We can excite several frequencies by choosing

$$\begin{aligned} x_{c_r}(t) &= A \sin(\omega(t) + \phi) \\ \omega(t) &= \pi \sum_{k=1}^{M_r} \left(1 - \frac{k-1}{M_r}\right) \sin kt \end{aligned}$$

for some constants ϕ , $A > 0$, and a sufficiently large $M_r \in \mathbb{N}$. In Figure 8, we plot the response of the system to the control and adaptation laws (16, 18), implemented in simulation. The constants that are used are as follows: $(A, M_r, \phi) = (3/10, 3, 0^\circ)$, $\Gamma = \text{diag}(1, 1, 1/10)$ and $(c, \lambda) = (1, 4)$. The real mass, damping and stiffness of the system are $(m_c, b, k) = (0.665, 1.819, 0)$ and their estimates start at $(\hat{m}_c, \hat{b}) = (-1/2, -1/4, -1)$.

4) *Results*: Figure 9 shows a successful swing-up trajectory generated by the MoE controller in simulation and hardware. The blue contours correspond to the level sets of the control input u during impact ($x_c = 0.36m$, $\dot{x}_c = 0m/s$), and the solid red lines depict the boundaries of the state partitions. Although the gating network can provide up to three state partitions, the training converges to utilizing only two. Figure 9 shows that the system successfully avoids contacts during the swing-up phase, which otherwise would have prevented the pendulum from pumping energy from the downward equilibrium. By the time the pendulum approaches the upright equilibrium, it is moving at such high speed ($\sim 6\text{rad/s}$) that LQR cannot stabilize the pendulum to the upright. However, we have observed from several trajectories that the system leverages the impact from

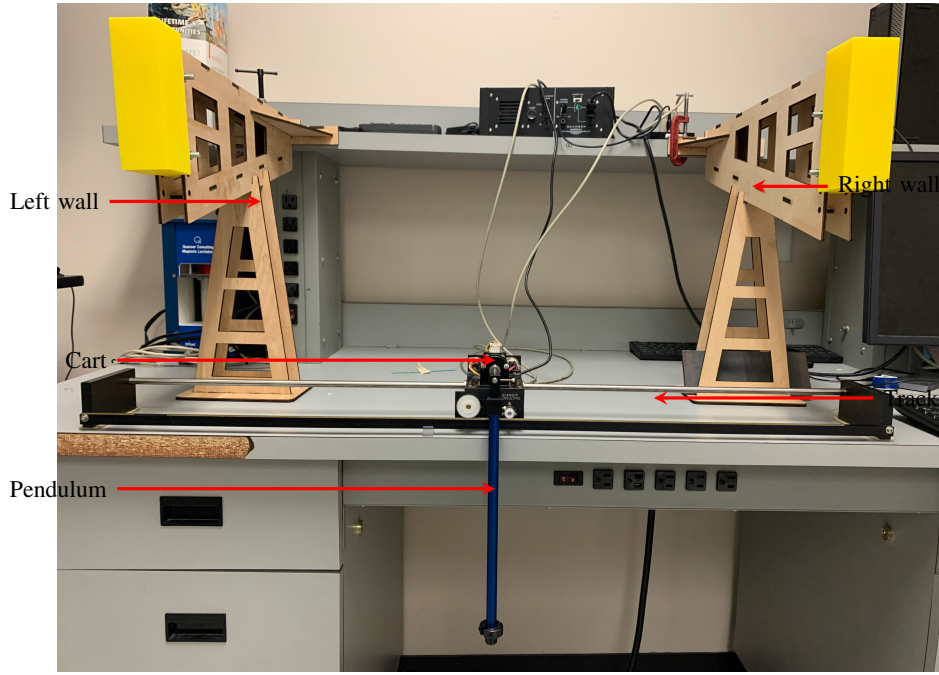


Fig. 7: Experimental setup of cartpole with wall contacts

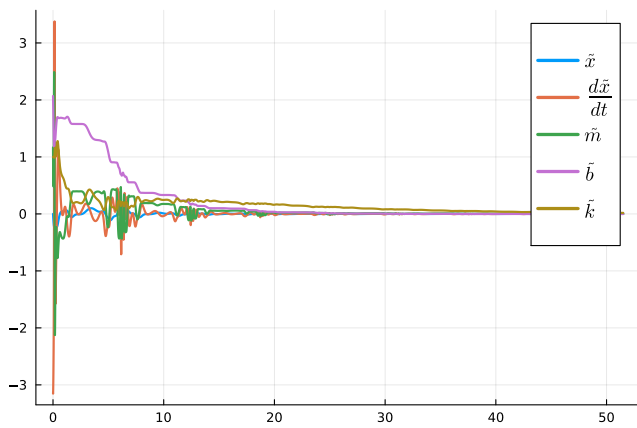


Fig. 8: Simulation showing the convergence of the system state and parameter estimates

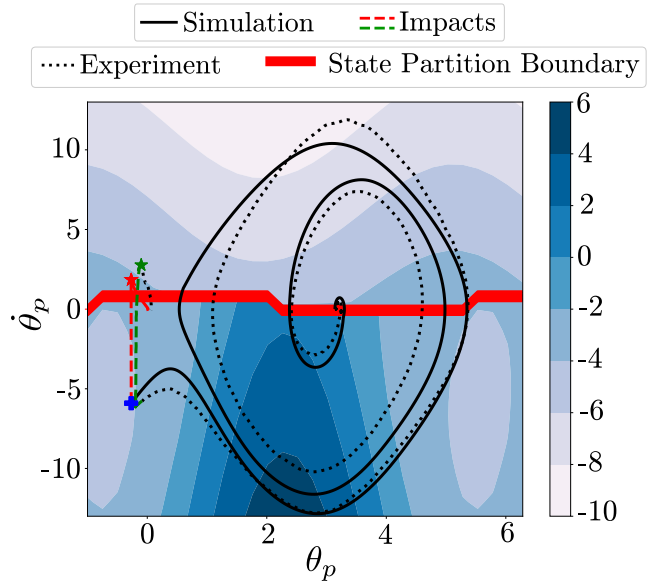


Fig. 9: A sample trajectory starting from downward equilibrium at rest. The blue contours represent the level sets of the control input at the pre-impact and post-impact states.

the wall to lower the speed of the pendulum.

During impact, the control law switches experts, where the new expert applies rapid braking allowing the LQR to catch the pendulum post-impact. The MoE controller achieves successful swing-up in simulation and real-world, proving the accuracy in the contact model and the robustness of the controllers.

Comparison between MoE and single controller: We compare the performance of the MoE controller against a single controller, which can be thought of as the MoE controller with $N_F = 1$. This controller is parameterized by a neural net, with a similar structure to the experts provided in Table I. We train the controller with the same minimum trajectory loss (MTL) and training parameters as the MoE. Once the controller swings the pendulum to the neighborhood of x^* , we use LQR

to stabilize it to the upright. As shown in Figure 10a, the single controller successfully swings up the pendulum close to the upright. However, due to the length of the pendulum and the tight distance between the walls, the pendulum inevitably impacts one of the barriers. Unfortunately, the LQR is not able to catch the pendulum post-impact, due to the high velocity of the pendulum. On the other hand, Figure 10b shows the performance of the MoE controller in the same scenario. The MoE solution leverages the switching controllers

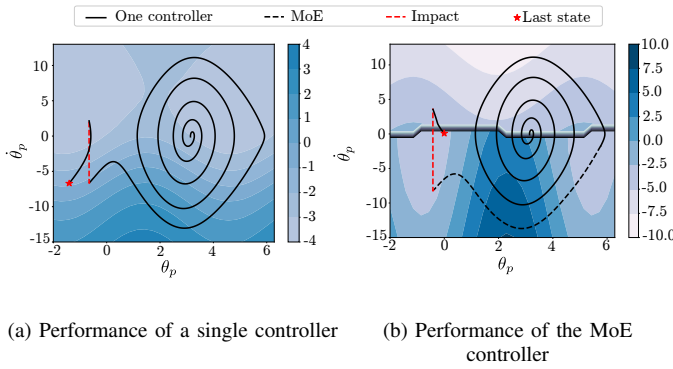


Fig. 10: Comparison between MoE and a single controller

to apply rapid braking post-impact, which significantly lowers the velocity of the pendulum. This assists the LQR in catching the pendulum at the appropriate speed. This demonstrates the advantages of switching controllers in the presence of multi-modal contact-rich systems.

VI. CONCLUSION

We provide a data-driven control design that reasons about the effects of contact forces on the hybrid system. We incorporate accurate system model in the training via linear complementarity formulation, and infer mixture of experts controller. The learning framework also provides a gating network, which divides the state space into partitions and dedicates a specialized local expert in each partition. From simulation and real-world experiments, we demonstrate that the learned policy leverages the advantages of contact in some states and minimizes its adverse effects in others.

The framework that we propose is general and can be applied to a wide range of hybrid dynamical systems. Almost all robotic systems must interact with their environment in order to be useful: walking or running robots make and break contact with the ground in order to keep a stable gait, manipulators must grasp and manipulate objects. In all of these cases, the contact forces are critical to the system's behavior. Our MoE framework can be used to learn controllers that reason about the effects of contact forces on the system, leveraging the advantages of contact in some states and minimize its adverse effects in others.

REFERENCES

- [1] M. Cutkosky, "On grasp choice, grasp models, and the design of hands for manufacturing tasks," *IEEE Transactions on Robotics and Automation*, vol. 5, no. 3, pp. 269–279, 1989.
- [2] X. Cheng, E. Huang, Y. Hou, and M. T. Mason, "Contact mode guided motion planning for quasidynamic dexterous manipulation in 3d," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2730–2736, IEEE, 2022.
- [3] F. Ruggiero, V. Lippiello, and B. Siciliano, "Nonprehensile dynamic manipulation: A survey," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1711–1718, 2018.
- [4] F. Ruggiero, A. Petit, D. Serra, A. C. Satici, J. Cacace, A. Donaire, F. Ficuciello, L. R. Buonocore, G. A. Fontanelli, V. Lippiello, *et al.*, "Nonprehensile manipulation of deformable objects: Achievements and perspectives from the robotic dynamic manipulation project," *IEEE Robotics & Automation Magazine*, vol. 25, no. 3, pp. 83–92, 2018.
- [5] K. M. Lynch and T. D. Murphey, "Control of nonprehensile manipulation," in *Control problems in robotics*, pp. 39–57, Springer, 2003.

- [6] K. M. Lynch and M. T. Mason, "Dynamic nonprehensile manipulation: Controllability, planning, and experiments," *The International Journal of Robotics Research*, vol. 18, no. 1, pp. 64–92, 1999.
- [7] M. Erdmann, "An exploration of nonprehensile two-palm manipulation," *The International Journal of Robotics Research*, vol. 17, no. 5, pp. 485–503, 1998.
- [8] M. Yashima, Y. Shiina, and H. Yamaguchi, "Randomized manipulation planning for a multi-fingered hand by switching contact modes," in *2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422)*, vol. 2, pp. 2689–2694, IEEE, 2003.
- [9] J. Z. Woodruff and K. M. Lynch, "Planning and control for dynamic, nonprehensile, and hybrid manipulation tasks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4066–4073, IEEE, 2017.
- [10] K. Lowrey, S. Kolev, J. Dao, A. Rajeswaran, and E. Todorov, "Reinforcement learning for non-prehensile manipulation: Transfer from simulation to physical system," in *2018 IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAP)*, pp. 35–42, IEEE, 2018.
- [11] X. Zhang, M. Chang, P. Kumar, and S. Gupta, "Diffusion meets dagger: Supercharging eye-in-hand imitation learning," *arXiv preprint arXiv:2402.17768*, 2024.
- [12] C. M. Bishop, "Pattern recognition," *Machine learning*, vol. 128, no. 9, 2006.
- [13] T. Härkönen, S. Wade, K. Law, and L. Roininen, "Mixtures of gaussian process experts with smc²," *arXiv preprint arXiv:2208.12830*, 2022.
- [14] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the em algorithm," *Neural computation*, vol. 6, no. 2, pp. 181–214, 1994.
- [15] Z. Chen, Y. Deng, Y. Wu, Q. Gu, and Y. Li, "Towards understanding mixture of experts in deep learning," *arXiv preprint arXiv:2208.02813*, 2022.
- [16] M. M. Zhang and S. A. Williamson, "Embarrassingly parallel inference for gaussian processes," *Journal of Machine Learning Research*, 2019.
- [17] L. V. Jospin, B. Buntine, F. Boussaid, H. Laga, and M. Bennamoun, "Hands-on bayesian neural networks—a tutorial for deep learning users," *arXiv preprint arXiv:2007.06823*, 2020.
- [18] S. Sharma, S. Sharma, and A. Athaiya, "Activation functions in neural networks," *Towards Data Sci.*, vol. 6, no. 12, pp. 310–316, 2017.
- [19] D. Liberzon, *Switching in systems and control*, vol. 190. Springer, 2003.
- [20] R. Goebel, R. G. Sanfelice, and A. R. Teel, "Hybrid dynamical systems," *IEEE control systems magazine*, vol. 29, no. 2, pp. 28–93, 2009.
- [21] B. Brogliato and B. Brogliato, *Nonsmooth mechanics*. Springer, 1999.
- [22] S. Ross, G. J. Gordon, and J. A. Bagnell, "No-regret reductions for imitation learning and structured prediction," in *In AISTATS*, Citeseer, 2011.
- [23] J. Revels, M. Lubin, and T. Papamarkou, "Forward-mode automatic differentiation in julia," *arXiv preprint arXiv:1607.07892*, 2016.
- [24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [25] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.
- [26] J. J. Moreau, "Unilateral contact and dry friction in finite freedom dynamics," *Nonsmooth mechanics and Applications*, pp. 1–82, 1988.
- [27] C. Glocker and C. Studer, "Formulation and preparation for numerical evaluation of linear complementarity systems in dynamics," *Multibody System Dynamics*, vol. 13, pp. 447–463, 2005.
- [28] V. Acary and B. Brogliato, *Numerical methods for nonsmooth dynamical systems: applications in mechanics and electronics*. Springer Science & Business Media, 2008.
- [29] Quanser, *Linear Servo Base Unit with Inverted Pendulum*. Apr 2021.
- [30] M. Spong, S. Hutchinson, and M. Vidyasagar, *Robot Modeling and Control*. Wiley, 2020.
- [31] H. Khalil, *Nonlinear Control*. Always Learning, Pearson, 2015.

VII. APPENDIX

Consider the Lyapunov function candidate

$$V(\tilde{x}_c, \dot{\tilde{x}}_c, \tilde{\xi}) = \frac{1}{2} m_c r^2 + c \lambda \tilde{x}_c^2 + \frac{1}{2} \tilde{\xi}^\top \Gamma \tilde{\xi}.$$

This is a positive definite function over the space of $(\tilde{x}_c, \dot{\tilde{x}}_c, \tilde{\xi})$. We take the time derivative of the Lyapunov function candidate

and substitute from the closed-loop system dynamics (17, 18). We suppress its functional dependence for brevity.

$$\begin{aligned}\dot{V} &= m_c r \dot{r} + 2c\lambda \tilde{x}_c \dot{\tilde{x}}_c + \tilde{\xi}^\top \Gamma \dot{\tilde{\xi}} \\ &= r \left(-(b+c)r - Y\tilde{\xi} \right) + 2c\lambda \tilde{x}_c \dot{\tilde{x}}_c + \tilde{\xi}^\top Y^\top r \quad (19) \\ &= -e^\top Q e \leq 0,\end{aligned}$$

where Q is a symmetric, positive-definite matrix given as

$$Q = \begin{bmatrix} (b+c)\lambda^2 & b\lambda \\ b\lambda & b+c \end{bmatrix}.$$

Integrating both sides of equation (19) gives

$$V(t) - V(0) = - \int_0^t e^\top(\sigma) Q e(\sigma) d\sigma < \infty.$$

We observe that $\dot{\tilde{x}}_c$ is bounded because $\dot{V} \leq 0$ implies that the terms r , \tilde{x}_c and $\tilde{\xi}$ are bounded functions of time. This allows us to invoke Barbalat's lemma[30] to deduce that $\tilde{x}_c \rightarrow 0$ as $t \rightarrow \infty$. Furthermore, using equation (17), we can readily see that \tilde{x}_c is bounded. Another application of Barbalat's lemma shows that the velocity error $\dot{\tilde{x}}_c \rightarrow 0$ provided that the reference acceleration $\ddot{x}_{c_r}(t)$ is bounded. Since $x_c \rightarrow x_{c_r}$, we also have that $u(x_c, \dot{x}_c) \rightarrow \hat{m}_c \ddot{x}_{c_r} + \hat{b} \dot{x}_{c_r} + \hat{k} x_{c_r}$.

Remark For any $\eta > 0$, the set $\Omega_\eta = \{(e, \tilde{\xi}) : V(\tilde{x}_c, \dot{\tilde{x}}_c, \tilde{\xi}_c) \leq \eta\}$ is positively invariant. The positive limit set of $(e(t), \tilde{\xi}(t))$ is a subset of $E = \{(e, \tilde{\xi}) : e = 0\}$. Unfortunately, for a general nonautonomous system, the positive limit set is not necessarily a positively invariant set, precluding us to invoke LaSalle's theorem[31] to conclude that $\tilde{\xi} \rightarrow 0$.

BIOGRAPHIES



AYKUT C. SATICI received the B.Sc. and M.Sc. degrees in mechatronics engineering from Sabanci University, Istanbul, Turkey, in 2008 and 2010, respectively, and the Masters degree in mathematics from the University of Texas, Dallas, TX, USA, in 2013. He is currently with the Electrical Engineering Department, University of Texas at Dallas. His current research interests include robotics, geometric mechanics, and cooperative control.

Research Article


Large Language Models vs. Human Interpretation: Which is More Accurate in Text Classification?

Ahmet Hamdi Ozkurt, Emrah Aydemir, Yasin Sonmez


Abstract— Ekşi Sözlük is a widely used social network where numerous unusual events are discussed. In this context, it serves as a real-time news source for emergency response teams and digital news platforms. In this study, a dataset was compiled from comments shared on the Ekşi Sözlük platform regarding the Kahramanmaraş earthquake on February 6, 2023. These comments were classified into four categories: Source-Based Information, Emotional Reaction, Social Inference, and Personal Experience using the Gemma2 9B (9-billion-parameter) model, developed by Google with advanced natural language processing capabilities. A dataset of 500 comments in Excel format was analyzed, comparing the model outputs with human evaluations to assess classification accuracy. For this purpose, four evaluation columns were created for each comment based on category classification. The consistency between model-assigned categories and manually determined categories was examined using these columns. In cases where inconsistencies were detected, the model-generated explanations were subjected to qualitative evaluation. Model outputs that provide satisfactory explanations are considered acceptable, the manually classified category was assigned as the final evaluation. This process systematically resolved inconsistencies between model and human assessments, ensuring the final and validated category assignments for each comment. The highest accuracy values were observed for Social Inference (0.99), Source-Based Information (0.98), Personal Experience (0.88), and Emotional Reaction (0.83), respectively. In conclusion, this study presents a methodology for improving model performance through human supervision, contributing to the development of strategies for disaster management and crisis communication.

Index Terms— Natural Language Processing, Text Classification, Ekşi Sözlük, Gemma2 9B


Ahmet Hamdi Özkurt, is with Department of Management Information System University of Sakarya, Sakarya, Türkiye.(e-mail: hamdi.ozkurt@ogr.sakarya.edu.tr).

 <https://orcid.org/0009-0008-3220-4143>

Emrah Aydemir, is with Department of Management Information System University of Sakarya, Sakarya, Türkiye.(e-mail: emrahaydemir@sakarya.edu.tr).

 <https://orcid.org/0000-0002-8380-7891>

Yasin Sönmez, is with Department of Computer Technologies University, Batman, Türkiye (e-mail: yasinsonmez@batman.edu.tr).

 <https://orcid.org/0000-0001-9303-1735>

Manuscript received Mar. 05, 2025; accepted Apr. 14, 2025.

DOI: [10.17694/bajece.1652268](https://doi.org/10.17694/bajece.1652268)

I. INTRODUCTION

NATURAL DISASTERS are critical events that profoundly impact social order, where simultaneous, rapid, and accurate information is of vital importance. In addition to traditional news sources, social media platforms have emerged as significant information channels during disasters. Users share their emotions and thoughts regarding such events while also contributing to the flow of disaster-related information through social media [12]. This phenomenon holds substantial value in understanding public reactions to disasters and informing disaster management strategies. One of the frequently used platforms for sharing extraordinary events is Ekşi Sözlük, where users disseminate news, photographs, and observations related to disasters and emergencies. In this context, emergency response teams utilize data streams generated on social media platforms such as Ekşi Sözlük to identify affected regions and assess environmental impacts [15].

Social media data possess significant potential for enhancing crisis management during natural disasters and supporting post-disaster recovery efforts [17]. However, to effectively interpret, analyze, and utilize this data, machine learning techniques must be employed [7].

In this context, the primary aim of the study is to demonstrate how social media data can be classified using large language models during natural disasters. The research problem focuses on examining the potential contributions of rapidly and meaningfully distinguishing social media content in times of crisis to effective disaster management. Accordingly, comments related to the February 6, 2023 Kahramanmaraş Earthquake, collected from Ekşi Sözlük, were classified into four categories using the Gemma2 model.

The motivation of the study stems from the lack of models specifically designed to classify Turkish social media data in the context of disasters. As a contribution to the literature, this study offers an example of applying a large language model to Turkish-language data and proposes a method for the automatic and rapid classification of disaster-related content.

Finally, the classification performance of the model was evaluated using various metrics, and the outputs were compared with human annotations to analyze potential inconsistencies. In

doing so, the study demonstrates the potential of large language models in analyzing social media data during disaster events.

II. LITERATURE REVIEW

Natural disasters are crises in which the need for rapid, accurate, and reliable information reaches its peak. Among these, earthquakes are one of the most frequently encountered and devastating natural disasters on a global scale. Due to its geological location, Turkey is situated in an earthquake-prone region, posing a constant threat to the country. The February 6, 2023 Kahramanmaraş earthquake tragically reaffirmed this reality. The earthquakes, with magnitudes of 7.6 and 7.7, struck the Pazarcık and Elbistan districts of Kahramanmaraş, resulting in the loss of thousands of lives and leaving many individuals in urgent need of assistance [1].

During extraordinary events such as earthquakes, social media serves as an effective channel for individuals to share their experiences, call for help, and exchange information. Platforms like Twitter and Ekşi Sözlük are widely utilized to analyze public reactions to disasters and assess crisis communication strategies [14]. Research conducted on these platforms provides critical insights for developing crisis management strategies, enhancing societal resilience, and accelerating post-disaster recovery processes.

In recent years, Natural Language Processing (NLP) techniques have been increasingly used to analyze and interpret social media data during disasters [4]. One of the key NLP methods, sentiment analysis, helps determine the public mood by classifying social media posts as positive or negative. Machine learning models are commonly employed to extract meaningful patterns from large datasets and address complex problems. In particular, recent advancements in NLP models have significantly improved the ability to analyze and classify public responses during disaster events [6] [7]. Models like Gemma2 are used for the classification and analysis of social media data, with their performance evaluated using metrics such as accuracy, precision, recall, and F1-score. These metrics provide essential indicators for assessing the classification success, reliability, and overall effectiveness of the model.

Studies in the literature suggest that machine learning and deep learning techniques demonstrate high performance in the classification and analysis of social media data. For instance, Vaswani et al. [16] compared traditional machine learning approaches with Transformer-based deep learning models in text classification tasks and found that Transformer-based models achieved significantly higher accuracy. Similarly, Vieweg et al. (2010) conducted a systematic review of studies examining the role of social media in crises following natural disasters. These studies integrated both quantitative and qualitative research methods to provide a comprehensive analysis. Additionally, Lindsay [9] highlighted the functional efficiency of social media tools during disaster and crisis events, emphasizing their role in accelerating the dissemination of information.

Recent studies highlight the effectiveness of large language models (LLMs) in social media analysis during crises. For

example, Pereira et al. [13] showed that LLMs improved the comprehensiveness of summaries in crisis communication and emergency response compared to traditional methods.

Similarly, Yang et al. [18] noted that LLMs provided accurate predictions and interpretable explanations in analyzing mental health data from social media. Lastly, McDaniel et al. [10] demonstrated that integrating additional event information improved success rates in classifying social media posts in the humanitarian aid context using a zero-shot classification approach.

In conclusion, existing studies support the effective use of NLP and machine learning techniques in the analysis of post-disaster social media data. Text classification methods form the foundation of social media data analysis, while techniques such as sentiment analysis have the potential to optimize response processes by ensuring rapid and accurate access to critical information in crisis situations. Therefore, further research in this field could provide significant contributions to crisis management and humanitarian aid efforts.

III. METHODOLOGY

A. DATA COLLECTION

Within the scope of this study, 7,250 comments related to the February 6, 2023, Kahramanmaraş earthquake were collected from the Ekşi Sözlük platform. These comments were obtained using the Selenium library in Python, ensuring that both timestamp and comment information were preserved. The extracted data were initially stored in JSON format and subsequently converted to CSV format for further processing.

B. DATA PREPROCESSING

Data preprocessing is the phase in which raw data is transformed into a structured and interpretable format. In other words, it involves converting initial raw data into final processed data that serves as the foundation for subsequent analyses [2].

In this study, the comments extracted in JSON format were cleaned by removing punctuation marks. Comments lacking semantic coherence or consisting solely of visual content were eliminated from the dataset. Additionally, numerical expressions within the comments were removed. However, double quotation marks were retained during this preprocessing phase.

C. CLASSIFICATION

After completing the data preprocessing stage, 6,895 comments remained from the original dataset of 7,250 and were subsequently classified into four categories (see Table 1).

In the classification process, Gemma2, a high-performance, fast, and efficient language model developed by Google, was utilized. This model is commonly employed in Natural Language Processing (NLP) tasks such as text generation and classification. Additionally, Gemma2 has 2B (2 billion parameters) and 27B (27 billion parameters) variants. The primary distinction among these models lies in their parameter count, which directly influences the model's information

processing capacity. As the number of parameters increases, the model's ability to process and learn from data improves proportionally.

Table I
Definition and Examples of Comment Categories Used for Classification

CATEGORY	DEFINITION	EXAMPLE
Source-Based Information	Statements that are supported by verifiable data, academic sources, or credible authorities, aiming to inform or explain objectively.	According to a 2023 report by the World Health Organization, air pollution causes approximately 7 million premature deaths annually.
Emotional Response	Expressions that reflect the speaker's emotions, such as anger, fear, happiness, or sadness, often aiming to convey a personal or subjective response.	It breaks my heart to see how little is being done to protect our environment.
Social Inference	Comments that draw broader conclusions about society, culture, or collective behavior based on individual observations or specific events.	The increasing use of social media has fundamentally changed how people form and maintain relationships in modern society.
Personal Experience	Narratives or statements based on the individual's own life, experiences, or observations, usually shared in a subjective manner.	Last year, I volunteered at a refugee camp, and it completely changed my perspective on global crises.

The Gemma2 model offers several advantages when compared to other existing large language models (LLMs). While higher-parameter models such as GPT-3.5 (175B) and GPT-4 (1.7T) demonstrate more advanced comprehension and text generation capabilities, their computational demands significantly limit their usability, particularly on local hardware. Although open-source alternatives like LLaMA 2 (7B–70B) offer greater modifiability, they fall short of achieving the same efficiency-performance balance that Gemma2 provides for specific tasks.

The 9B parameter model was selected for this study due to its balance between efficiency and computational resource consumption (e.g., processor, RAM), offering optimal performance for classification tasks. Additionally, the open-access availability of the Gemma2 model and its comparatively stronger support for the Turkish language have been key factors in its selection over alternative models [11].

During the classification process, each comment was assigned '1' if it belonged to a specific category and '0' otherwise. In the initial phase, the model was tasked solely with determining the appropriate category for each comment. Upon completion of the classification process, the annotated data was exported from a CSV file to an Excel format for further analysis.

Table II
Created Sample Excel

DATE	COMMENT	Source- Based Information	Emotional Reaction	Social Inference	Personal Experience
06.02.2023	Oh my God, Gaziantep is shaking very badly again, we are living in hell here, it is shaking like a cradle, please make it stop.	0	1	0	1
07.02.2023	The environment minister said on live broadcast that we did not leave any of our citizens hungry and exposed, right, you did not leave them hungry and exposed, you left them under the rubble.	0	1	1	0
07.02.2023	Malatya is in a very difficult situation, the food and shelter shortage of earthquake survivors is growing and no one is helping, please help please	0	1	0	1

IV. SAMPLING AND MANUAL PROCESSING

From the 6,895 classified comments, a balanced subset of 500 unique comments was randomly selected, consisting of 250 instances labeled as '0' and 250 instances labeled as '1' for each category. This random selection process was carried out to ensure data balance and was implemented using the following Python libraries and functions:

- Pandas: Utilized for data analysis and processing.
- NumPy: Used for numerical computations.
- Path: Employed for creating and managing file paths.

- Linprog and Pulp: These functions were used for solving linear programming problems.

Subsequently, the randomly selected comments were also manually classified following the same methodology. Comments deemed meaningless, consisting only of visual content, or containing fewer than five words were excluded from the dataset. After this filtering process, the remaining 500 comments were reintroduced to the model for further evaluation.

The selection of only 500 comments from the dataset of 7,250 was driven by the dual aim of developing a balanced

classification model and enhancing the efficiency of the manual labeling process. Ensuring an equal number of examples from each category (250 labeled as '0' and 250 as '1') was essential for enabling the model to perform consistently across both

classes and for ensuring more reliable human intervention during the evaluation phase.

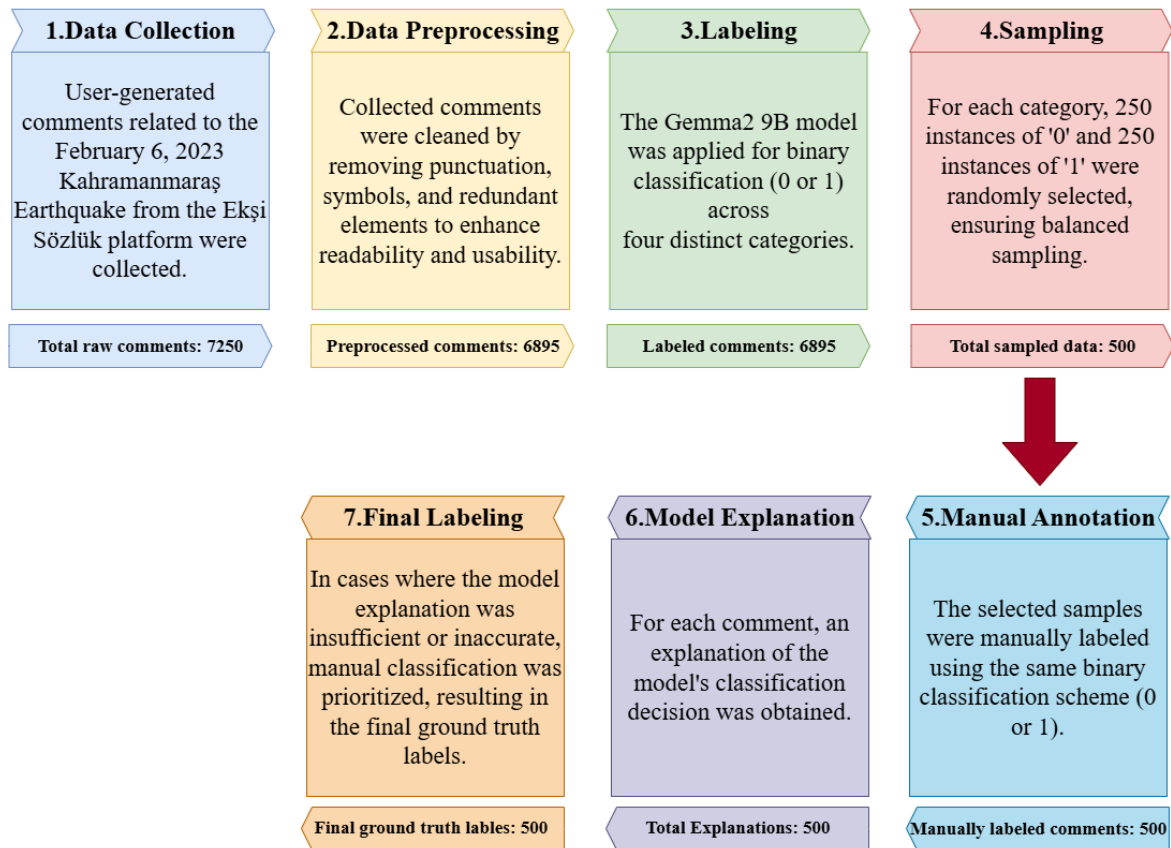


Fig. 1. Flow Diagram of the Study

V. MODEL EXPLANATION

During the classification process, discrepancies and inconsistencies were identified between the values assigned by the model and those assigned manually. To mitigate this issue,

an alternative approach was introduced to enhance the model's attentiveness and response generation. Specifically, modifications were made to the prompt content to ensure that the model provided more deliberate and well-considered classifications.

Table III
Category Based Descriptions of the Model

Explanation Based on Information Source	Explanation Based on Emotional Response	Explanation Based on Social Inference	Explanation Based on Personal Experience
NO because the comment does not refer to a specific event or source.	YES because the comment reflects feelings of sadness and concern, with phrases such as "they were in great shock" and "I hope our loss is not too great".	YES because the comment contains a call for social support and solidarity for the earthquake victims by saying "get well soon to the people in that region".	YES because the author's statement in the comment that "I talked to my loved ones in Antep and Nizip" reflects his personal experience.

VI. DETERMINATION OF FINAL GROUND TRUTH VALUES

In this study, four category-specific evaluation columns were created for the 500 comments stored in Excel format. These evaluation columns were utilized to assess the accuracy of the model outputs. The following steps were followed in the evaluation process:

- **Category-Based Consistency Check:** If there was

consistency between the values assigned by the model and those assigned manually, the respective evaluation column was left empty.

- **Category-Based Accuracy Assessment of Model Output:** In cases where discrepancies were identified between the model output and manual classification, the model's explanations were reviewed. If the explanation was deemed satisfactory, the model output was accepted in the evaluation column.

- **Category-Based Determination of Final Ground Truth Values:** If the model's explanations were found to be insufficient or unconvincing, the manually assigned classification was used as the final label in the evaluation column.

As a result of these procedures, discrepancies between the model outputs and manual classifications were resolved through the evaluation columns, and the final ground truth values were established.

VII. FINDINGS

To evaluate the predictive performance of the model across four distinct categories (*Source-Based Information*, *Emotional Response*, *Social Inference*, and *Personal Experience*), comparisons between the model-generated outputs and manually assigned labels were analyzed. These comparisons illustrate the statistical distribution of the model's predictions for each category and the accuracy rates of these predictions. Below, two tables containing these comparisons are presented,

along with detailed explanations regarding their implications.

Table IV
Comparison of Model Predictions and Manual Labels

Category	Model Prediction (0)	Model Prediction (1)	Manual Label (0)	Manual Label (1)
Source-Based Information	250	250	257	243
Emotional Reaction	250	250	303	197
Social Inference	250	250	258	242
Personal Experience	250	250	309	191

Table 4 shows the statistical distribution of the predictions made by the model for the four categories (coded as 0 and 1), along with the corresponding manual labels (coded as 0 and 1).

Table V
Evaluation Results of Model Estimates

Category	Model Acceptance 0	Model Acceptance 1	Model Rejection 0	Model Rejection 1	Total Prediction	Correct Prediction	Percentage
Source-Based Information	250	240	0	10	500	490	%98
Emotional Response	235	180	15	70	500	415	%83
Social Inference	248	246	2	4	500	494	%99
Personal Experience	247	194	3	56	500	441	%88

Table 5 presents the number of predictions labeled as 0 and 1 for both accepted and rejected instances across each category, along with the total and correctly classified predictions. For instance, in the "Social Inference" category, the model correctly predicted 248 instances as 0 and 246 as 1, with only 6 misclassifications. These results provide a basis for evaluating the overall classification performance of the model and understanding the distribution of success across categories. Moreover, they serve as input for the calculation of classification metrics such as accuracy, precision, recall, and F1-score, which are discussed in detail in the following section.

A. 1. CONFUSION MATRIX ANALYSIS AND CLASSIFICATION METRICS

The classification performance of the model is quantitatively assessed using the confusion matrix. In the confusion matrix, columns represent the final true values, while rows represent the predicted values. In this context, the evaluation columns indicate the actual values in the classification, and the model outputs represent the predicted values. The following Python libraries were used to construct the confusion matrix:

- Pandas: Used for data processing and reading, particularly for reading columns of data in Excel format.

- NumPy: Used for numerical computations, including percentage calculations in the confusion matrix.
- Scikit-learn: Used for generating the confusion matrix and calculating classification metrics (Accuracy, Precision, Recall, F1-Score).
- Seaborn: Used for the visualization of the confusion matrix.
- Matplotlib: Used for the customization of the confusion matrix visualization.

The classification metrics derived from the confusion matrix are utilized to measure the classification process's performance in greater detail. These metrics include:

- Accuracy
- Precision
- Recall
- F1-Score

Table VI
Confusion Matrix Structure

	REAL POSITIVE (1)	REAL NEGATIVE (0)
PREDICTION POSITIVE (1)	TP	FP
PREDICTION NEGATIVE (0)	FN	TN

Table VII
Classification Metrics and Explanations

Classification metrics	Computation formulas	Definition
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$	It is the overall accuracy rate of the model's classification process.
Precision	$\frac{TP}{TP + FP}$	It represents the proportion of classes predicted as positive by the model that are actually positive.
Recall	$\frac{TP}{TP + FN}$	It measures how accurately the model predicts within the true positives.
F-1 Score	$\frac{2 * Precision * Recall}{Precision + Recall}$	The harmonic mean of Precision and Recall is used for their combined evaluation.

To evaluate the overall performance of the model, it is necessary to examine the category-specific confusion matrices and classification metrics as summarized in Table 7.

- True Positive (TP): The number of positive instances correctly classified by the model.
- True Negative (TN): The number of negative instances correctly classified by the model.
- False Positive (FP): The number of negative instances incorrectly classified as positive by the model.
- False Negative (FN): The number of positive instances incorrectly classified as negative by the model [3].

The classification metrics used for performance evaluation, along with their formulas and definitions, are provided below.

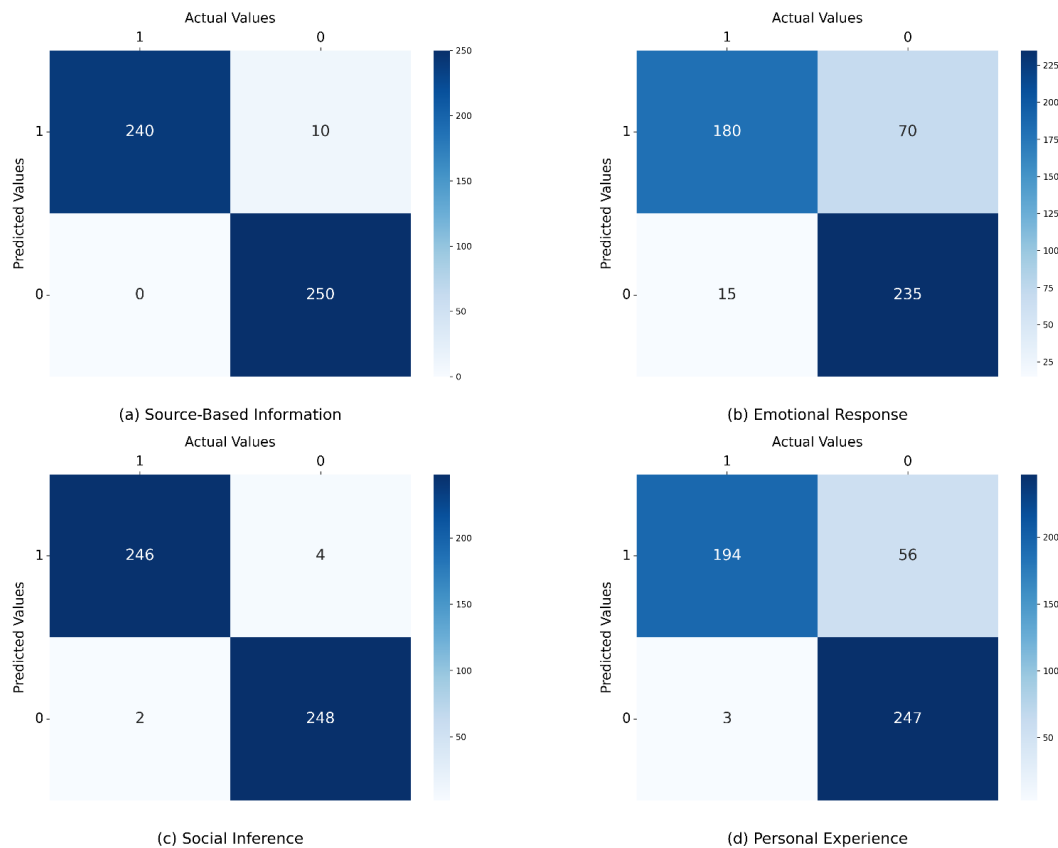


Fig. 2. Confusion Matrices for Categories

B. CATEGORY-BASED CONFUSION MATRIX

1) Source Information

Upon examining the confusion matrix for the Source Information category presented in Figure 2(a), it has been

observed that the model made 250 true negative (TN) and 240 true positive (TP) predictions. Additionally, it was found that there were 0 false positive (FP) and 10 false negative (FN) classifications. This indicates that the model has not tended to

classify non-source information as source information (FP=0), and has correctly identified the vast majority of comments containing source information (low FN=10).

2) Emotional Response

Upon evaluating the confusion matrix for the Emotional Response category presented in Figure 2(b), it has been observed that the model made 235 true negative (TN) and 180 true positive (TP) predictions. Additionally, the model produced 70 false positive (FP) and 15 false negative (FN) predictions. This finding suggests that the model incorrectly classified some non-emotional response comments as emotional responses (high FP=70) and was able to correctly identify a portion of the comments that contained emotional responses.

3) Social Inference

Upon examining the confusion matrix for the Social Inference category presented in Figure 2(c), it has been observed that the model made 248 true negative (TN) and 246 true positive (TP) predictions. Moreover, there were 4 false positive (FP) and 2 false negative (FN) classifications. In this context, the model's classification performance in this category is quite successful, as indicated by the low false positive rate (FP=4) and the high true positive rate (TP=246).

4) Personal Experience

Upon reviewing the confusion matrix for the Personal Experience category presented in Figure 2(d), it has been observed that the model made 247 true negative (TN) and 194 true positive (TP) predictions. Additionally, there were 56 false positive (FP) and 3 false negative (FN) classifications. This indicates that the model incorrectly classified some non-personal experience comments as personal experiences (FP=56), but successfully identified the majority of comments containing personal experience (low FN=3).

Table VIII
Performance of Category-Based Classification Metrics

Categories	Accuracy	Precision	Recall	F-1 Score
Source-Based Information	0.98	0.96	1.00	0.97
Emotional Response	0.83	0.72	0.92	0.80
Social Inference	0.99	0.98	0.99	0.98
Personal Experience	0.88	0.77	0.98	0.86

C. CATEGORY-BASED CLASSIFICATION METRICS

1) Source Information

Upon examining Table 8, it can be observed that the classification metrics for the Source Information category demonstrate generally successful performance (Accuracy: 0.98, Recall: 1.00, Precision: 0.96, F1-Score: 0.97). The high recall indicates that the model is able to successfully identify comments containing source information, while the high

precision indicates that the model can also largely classify comments that do not contain source information correctly.

2) Emotional Response

When reviewing the classification metrics for the Emotional Response category in Table 8, it is observed that the recall value is high (0.92), the accuracy value is 0.83, the precision value is 0.72, and the F1-Score is 0.80. These values suggest that the model is successful in detecting comments containing emotional responses (high recall), but also incorrectly classifies some comments that do not contain emotional responses as emotional (lower precision).

3) Social Inference

Upon evaluating the classification metrics for the Social Inference category presented in Table 8, it is evident that the values for accuracy, recall, precision, and F1-Score are very high (Accuracy: 0.99, Recall: 0.99, Precision: 0.98, F1-Score: 0.98). These values indicate that the model has very low false positive (FP) and false negative (FN) rates.

4) Personal Experience

According to the data presented in Table 8, it is observed that the model demonstrates high performance in the Personal Experience category (Accuracy: 0.88, Precision: 0.77, Recall: 0.98, F1-Score: 0.86). The high recall indicates that the model correctly identifies the majority of comments containing personal experience, while the lower precision value suggests that the model also incorrectly classifies some comments that do not contain personal experience as personal experience.

VIII. DISCUSSION AND CONCLUSION

This study aimed to automatically classify the comments on Ekşi Sözlük related to the 6th February 2023 Kahramanmaraş earthquake and categorize them into four distinct categories (Source Information, Emotional Response, Social Inference, Personal Experience) using the Gemma2 model. This work highlights the importance of analyzing social media data in crisis situations to obtain quick and accurate information. The results obtained indicate that the classification performance of the model varies significantly across categories.

The analyses show that the model demonstrated its highest classification performance in the Social Inference category, while its lowest performance was observed in the Emotional Response category. In the Source Information and Personal Experience categories, the model demonstrated satisfactory classification performance with accuracy rates of 98% and 88%, respectively. These findings suggest that the model's classification performance varies by category, and the confidence in the results may fluctuate based on the category. While the accuracy rate may be very high in one category, the opposite can be observed in another category. In this context, it is suggested that instead of fully relying on machine learning models, category-specific approaches may be necessary for different classification categories. The higher classification performance in the Social Inference category may be due to the language being more objective and tending to express social inferences directly. Conversely, in the Emotional Response category, the intensity of ambiguous, complex expressions

might make it difficult for the model to comprehend.

Table 8 clearly demonstrates this situation in the classification metrics. The model exhibited the highest classification performance in the Social Inference category (Accuracy: 99%), indicating that it can accurately classify texts containing social inferences, evaluations, or general comments. On the other hand, the performance in the Emotional Response category is the lowest (Accuracy: 83%). In the Source Information (Accuracy: 98%) and Personal Experience (Accuracy: 88%) categories, the model performed quite well. Another reason for the low classification performance in the Emotional Response category may be the differences and subjectivity of emotional expressions in the dataset. The model may struggle with categorizing different emotional tones.

In conclusion, this study revealed that the classification performance of the Gemma2 model in categorizing the comments on Ekşi Sözlük regarding the 6th February Kahramanmaraş Earthquake varies based on the category. The model exhibited low performance in the Emotional Response category, while achieving high success in the Social Inference category.

These findings highlight important points to consider when developing natural language processing-based sentiment analysis and automatic text classification models. To improve the model's classification performance in the Emotional Response category, various sentiment analysis techniques, such as sentiment lexicons, can be utilized, or a specialized model for this category could be trained. For example, pre-trained sentiment analysis models can be used to help the model understand complex emotional expressions. Additionally, adding data from broader and diverse sources (social media, news websites, etc.) could further enhance the model's classification performance.

In particular, for categories with lower performance, fine-tuning techniques can be applied to improve the model's performance. Fine-tuning the model on category-specific datasets may enhance classification performance for those particular categories. Additionally, approaches such as ensemble learning can leverage the strengths of different models. For instance, combining the predictions of a sentiment analysis-focused model with the Gemma2 model for the Emotional Response category could improve classification performance. By developing an optimized ensemble model for each category in this manner, the overall system performance can be enhanced.

Some limitations of the study should be considered. Firstly, the data being collected from only a single social media platform (Ekşi Sözlük) restricts the generalizability of the findings. Additionally, the relatively small size of the dataset may have limited the model's classification capacity and could have affected the reliability of the results. Furthermore, the potential presence of biases within the dataset should not be overlooked. In particular, the demographic profile of Ekşi Sözlük users may prevent the analyzed content from fully representing the broader population. For these reasons, future studies should consider using larger and more diverse datasets

from various social media platforms to enhance the model's performance and strengthen the generalizability of the findings. These suggestions will contribute to the development of strategies for disaster management and crisis communication.

REFERENCES

- [1] AFAD, "06 Şubat 2023 Pazarcık-Elbistan Kahramanmaraş (Mw 7.7; Mw 7.6) depremleri raporu," Deprem ve Risk Azaltma Genel Müdürlüğü, 2023.
- [2] Y. Argüden and B. Erşahin, Veri madenciliği: Veriden bilgiye, masraftan değere, ARGE Danışmanlık Yayınları, 2008.
- [3] Z. Bakan and F. Kanbay, "Makine öğrenmesi yöntemleri ile eğitim başarısına etki eden faktörlerin modellenmesi," İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, vol. 23, no. 45, pp. 27–41, 2024. [Online]. Available: <https://doi.org/10.55071/ticaretfd.1442084>
- [4] G. Burel and H. Alani, "Crisis event extraction service (CREES)—Automatic detection and classification of crisis-related content on social media," in Proc. 15th Int. Conf. Inf. Syst. Crisis Response and Manage., 2018.
- [5] C. Coşkun and A. Baykal, "Veri madenciliğinde sınıflandırma algoritmalarının bir örnek üzerinde karşılaştırılması," in Akademik Bilişim Konferansı (AB'11) Bildirileri, 2011, pp. 51–58.
- [6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2019.
- [7] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, "AIDR: Artificial intelligence for disaster response," in Proc. 22nd Int. Conf. World Wide Web, 2015, pp. 159–162.
- [8] O. H. Kwon et al., "Sentiment analysis of the United States public support of nuclear power on social media using large language models," Renewable and Sustainable Energy Reviews, vol. 200, 114570, 2024. [Online]. Available: <https://doi.org/10.1016/j.rser.2024.114570>
- [9] B. R. Lindsay, "Social media and disasters: Recent United States experiences," J. Contingencies Crisis Manage., vol. 19, no. 1, pp. 1–7, 2011. [Online]. Available: <https://doi.org/10.1111/j.1468-5973.2011.00639.x>
- [10] E. L. McDaniel, S. Scheele, and J. Liu, "Zero-shot classification of crisis tweets using instruction-finetuned large language models," in 2024 IEEE Int. Humanitarian Technol. Conf. (IHTC), Nov. 2024, pp. 1–7.
- [11] M. Özkan and G. Kar, "Türkçe dilinde yazılan bilimsel metinlerin derin öğrenme tekniği uygulanarak çoklu sınıflandırılması," Mühendislik Bilimleri ve Tasarım Dergisi, vol. 10, no. 2, pp. 504–519, 2022. [Online]. Available: <https://doi.org/10.21923/jesd.973181>
- [12] L. Palen and S. B. Liu, "Citizen communications in crisis: Anticipating a future of ICT-supported public participation," in Proc. SIGCHI Conf. Human Factors Comput. Syst., 2007, pp. 727–736.
- [13] J. Pereira, R. Lotufo, and R. Nogueira, "Large language models in summarizing social media for emergency management," arXiv preprint arXiv:2401.03158, 2024.
- [14] C. Reuter and M. A. Kaufhold, "Fifteen years of social media in emergencies: A retrospective review and future directions for crisis informatics," J. Contingencies Crisis Manage., vol. 26, no. 1, pp. 41–57, 2018. [Online]. Available: <https://doi.org/10.1111/1468-5973.12196>
- [15] O. Sevlı and N. Kemalöğlu, "Olağandışı olaylar hakkındaki tweet'lerin gerçek ve gerçek dışı olarak Google BERT modeli ile sınıflandırılması," Veri Bilimi, vol. 4, no. 1, pp. 31–37, 2021.
- [16] A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst., vol. 30, 2017.
- [17] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: What twitter may contribute to situational awareness," in Proc. SIGCHI Conf. Human Factors Comput. Syst., 2010, pp. 1079–1088.
- [18] K. Yang et al., "MentalLaMA: Interpretable mental health analysis on social media with large language models," in Proc. ACM Web Conf. 2024, May 2024, pp. 4489–4500.

BIOGRAPHIES



Ahmet Hamdi Özkurt is going on his Bachelor's degree in Management Information System in Sakarya University. He is a third year student. He continues to develop himself in the field of artificial intelligence, large language models, database and mobile programming.



Emrah Aydemir was received the M.S. degrees in computer teaching from the University of Elazig Firat, in 2012 and the Ph.D. degree in informatics from Istanbul University, Turkey, TR, in 2017. From 2012 to 2015, he was an Expert with the Istanbul Commerce University. Since 2021, he has been an Associate Professor with the Management Information System, Sakarya University. He is the author of three books, more than 60 articles, and more than 40 conference presentation. His research interests include artificial intelligence, microcontroller, database and software



Yasin Sönmez was born in Diyarbakır, Turkey in 1986. He received the B.S. degree from the Firat University, Technical Education Faculty, Department of Electronics and Computer Education in 2010, M.S. degree in computer science from the Firat University in 2012 and Ph.D. degree department of software engineering at Firat University in 2018. His research interests include, artificial intelligence, and information security.

Phishing E-mail Detection with Machine Learning and Deep Learning: Improving Classification Performance with Proposed New Features

Hadjer Brioua, Havvanur Siyambas, Durmuş Özkan Şahin

Abstract—Today, with the increasing use of the internet, individuals who use email have become potential targets for fraudsters. These malicious groups send fake or misleading emails to steal sensitive information such as identity, bank, and social media credentials. This tactic is known as phishing. This study proposes a machine learning-based system for detecting phishing attacks using the SeFACED dataset, which was adjusted for binary classification with 12,498 normal and 5,142 fraudulent email data points. Python was used for programming, with Google Colab and Jupyter Notebook as development platforms. Email data underwent data collection, cleaning, and word stem separation processes. Three feature extraction techniques were used: Bag of Words, TF-IDF, and Word2Vec. Six algorithms, including Logistic Regression, Random Forest, Support Vector Machines, Naive Bayes, Convolutional Neural Network, and Long Short-Term Memory, were employed for classification. Performance was evaluated using metrics like accuracy, precision, recall, and F1-score. New attributes proposed to enhance detection included CSS tags, HTML tags, black-list words, link errors, and grammar and spelling errors. The addition of these features generally improved classification results.

Index Terms—Phishing, Phishing e-mail, Phishing attacks, Machine learning, Deep learning, Classification, Phishing e-mail classification.

I. INTRODUCTION

WITH the increasing use of the internet worldwide, access to data, services, and products has become easier. Although this has improved accessibility, it has also made systems vulnerable to attacks. As a result, cyber-attacks occur on personal computers, bank accounts, and social media accounts. The most common type of cyber attack is phishing. There are several types of phishing attacks [1], [2]:

- Email Phishing [3]: This is the most common type of phishing attack. An email is sent to the target individuals, giving the impression that it comes from a legitimate organization. Scammers direct recipients to click on a link in the email to steal their sensitive information.

- SMS Phishing [4]: This type of phishing attack is carried out using text messages sent via smartphones.
- Website Phishing [5]: The content of the website in this type of phishing is fake. Scammers request users to enter their information on the relevant website.

When the reports of the Anti-Phishing Working Group (APWG) are examined over the years, it is seen that millions of phishing attacks have been made [6]. Email phishing attack, which is a social engineering attack, is one of the most common phishing attacks [7], [8]. In this study, a machine learning-based architecture for detecting email phishing attacks was developed. To improve the performance of the developed system, new features were added to the system. It is generally observed that the addition of new features improves the model's performance.

A. Literature Review

In the study conducted by Ahi and Soğukpınar, a hybrid method called 'H-OLTA' was proposed to determine whether the relevant email is phishing or not by combining deep learning algorithms such as Long Short-Term Memory (LSTM) [9]. This method's success is higher than that of other classifier algorithms, and it is created by combining multi-layer perceptron (MLP) and LSTM algorithms. The accuracy of the developed model is determined to be 96.84%. Two main features were identified to train the model: the subject and body parts of the email text. With their developed method, the subject and body parts of the email are examined separately. Then, a feature matrix is created for the relevant parts, which is used to train the model using deep learning algorithms. The deep learning algorithms used are MLP and LSTM. The datasets used in this study are Jose Nazario's phishing email dataset and the Enron Email Dataset. The dataset, consisting of 4512 emails, is divided into 80% training and 20% test data. The performance metrics used are accuracy, precision, recall, F1-score, and false positive rate (FPR).

In this study, Abdullaheem et al. have approached phishing email detection as a classification problem, demonstrating how machine learning algorithms are used to categorize whether the given email is a phishing attack [10]. The algorithms used in the research include Logistic Model Tree (LMT), MLP, and decision tree. The algorithm achieving the highest accuracy rate in classifying phishing emails is LMT, with an accuracy rate of 96.924%. The dataset used was created by Mohammad et al., containing 11,000 website samples, out

^{ID} **Hadjer Brioua** is with the Department of Computer Engineering, Faculty of Engineering, Ondokuz Mayıs University, Samsun, 55200 TURKEY e-mail: hadjer.brioua@bil.omu.edu.tr

^{ID} **Havvanur Siyambas** is with the Department of Computer Engineering, Faculty of Engineering, Ondokuz Mayıs University, Samsun, 55200 TURKEY e-mail: havvanur.siyambas@bil.omu.edu.tr

^{ID} **Durmuş Özkan Şahin** is with the Department of Computer Engineering, Faculty of Engineering, Ondokuz Mayıs University, Samsun, 55200 TURKEY e-mail: durmus.sahin@bil.omu.edu.tr
Manuscript received May 27, 2024; accepted Jan 10, 2025.
DOI: 10.17694/bajece.1490596

of which 2,500 host phishing URLs. The dataset includes 2,456 instances and 30 features. These 30 features are divided into 4 groups: address bar, unnatural elements, HTML and JavaScript, and domain. The performance metrics used include accuracy, precision, recall, F1-score, and kappa statistic.

In this study, Paradkar used various classification and deep learning algorithms to determine whether an email is phishing [11]. The data went through processes such as data preprocessing and tokenization to convert them into a format suitable for classification. The dataset used is the ENRON CORPUS, consisting of 20,000 email samples, with 8,336 phishing emails and 11,664 normal emails. The dataset was divided into 75% training and 25% test data. The classification and deep learning algorithms used include LR, decision trees, Support Vector Machines (SVM), LSTM, and CNN. According to the study, machine learning algorithms could have been more effective in text classification, but deep learning algorithms achieved high accuracy rates. The highest accuracy rate, at 99.05%, was obtained with CNN.

In this study, Livara and Hernandez addressed the use of machine learning techniques to determine whether emails are phishing or not, and they also investigated the performance of these techniques on imbalanced datasets [12]. The researchers utilized the Phishing Email Collection dataset, obtained from Kaggle, containing 525,754 emails. 90% of the dataset was assigned for training, while the remaining 10% was used for testing. Various visualization tools, such as dot plots and distribution plots, were employed to understand the dataset better. Five machine learning algorithms were used for classification: Naive Bayes (NB), AdaBoost, SVM, LR, and RF. The performance metrics used included accuracy, precision, recall, and F1-score. After extracting features and applying the specified classification algorithms, the RF classifier yielded the highest precision, F1-score, and recall rates. The SVM classifier demonstrated the lowest precision rate at 92%; similarly, lower values were obtained for recall and F1-score.

Akinyelu and Adewumi obtained 2000 phishing email data from Nazario's public phishing email archive [13]. They extracted 15 significant phishing features and then created vector representations of these features for each email. This representation was used to train the relevant classifier. Only the RF classifier was used to train the model. In this study, classifiers were trained and tested using 10-fold cross-validation. The algorithms were tested with datasets of different sizes to measure their performance on small and large datasets. Performance metrics used include false-positive rate, precision, recall, and F1-score. The algorithm showed its best performance when tested on the largest dataset. When the RF classifier was used, the classification accuracy rate was 99.7%, the false-negative rate was 2.50%, and the false-positive rate was 0.06%.

In this study, Dewis and Viana used machine learning and various natural language processing techniques to classify whether the relevant emails were phishing [14]. Experiments were conducted using five different datasets. Each dataset was divided into 70% training and 30% testing samples. Performance metrics used included accuracy, precision, recall, and F1-score. Deep learning algorithms are known to achieve higher accuracy rates when dealing with large datasets, and to

mitigate the effects of sudden drops in parameters between hidden layers, more dense layers were added to the MLP algorithm [14], [15], [16], [17]. When the LSTM algorithm was applied to text-based datasets, a 99% accuracy rate was achieved, while for numerical-based datasets, a 94% accuracy rate was achieved for the MLP algorithm.

Eryılmaz et al. combined machine learning and text mining techniques to identify spam emails [18]. The researchers used the Turkish Email dataset. From this dataset, 600 emails were allocated for training the model, and 200 emails were reserved for performance evaluation. The dataset first underwent a preprocessing stage, followed by the use of bag-of-words and TF-IDF approaches to weight and vectorize each word. Different classifiers were used to test the model's success. These algorithms included Sequential Minimal Optimization (SMO), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), NB, and Multilayer Perceptron (MLP). Metrics such as F1-score, precision, and recall were used to evaluate model performance. Among the different classification algorithms applied, the most successful was SMO, achieving a classification performance of 0.985 in terms of F1-score. In contrast, the least successful was NB, with a classification performance of 0.931.

Singh et al. developed a model using machine learning and deep learning techniques such as K-Means, NB, LSTM, Convolutional Neural Network (CNN), and the BERT model to categorize whether emails are spam or not [19]. The language used to develop the model was Python version 3.7, and the model was developed using natural language processing. The performance metrics used were accuracy and F1-score. The study's results showed that they achieved 92%, 94%, 96%, and 98% accuracy rates for KNN, NB, LSTM, and the BERT model, respectively. The CNN algorithm outperformed all other classification algorithms, demonstrating the most efficient performance in determining whether an email is spam with 99% accuracy.

Sonare et al. explored the effects of using machine learning and deep learning algorithms to determine whether an email is phishing email [20]. They sourced the dataset from the Kaggle platform. The dataset consists of two columns: one indicating the category of the email (spam or normal) and the other containing the email content in the "Message" column. Their methodology includes six steps: loading the data, data collection, and preprocessing, label encoding, splitting the data into training and test sets, feature extraction, and training the model. During the data preprocessing stage, irrelevant and unstructured data were cleaned, and words were stemmed. Label encoding was used to digitize the data. Feature extraction was employed to digitize raw data. The classification and deep learning algorithms used were MLP, SVM, DT, and LR. The performance metrics utilized were precision, F1-score, and recall. As a deep learning algorithm, MLP demonstrated the best performance with a high accuracy rate of 98%. The algorithm with the lowest performance was DT, with an accuracy rate of 94%.

In the research conducted by Adzhar et al., a comparative study was performed on the machine learning algorithms NB, SVM, DT, and RF used for email phishing detection [21]. The

study aimed to evaluate previous phishing detection studies to determine which machine learning techniques best detect phishing emails. The definition, characteristics, and categories of phishing attacks were provided in detail. According to this study, a phishing email has five characteristics: it seems too good to be true, creates a sense of urgency, uses links, includes attachments, and comes from someone the user does not know. Phishing attacks were categorized into four groups: link-based, text-based, image-based, and attachment-based. Following the characteristics of phishing emails and the categorization of phishing attacks, the study examined several machine learning algorithms (NB, SVM, DT, RF) that can be used for phishing email detection. A comparative study was conducted on these techniques, and as a result, SVM and RF algorithms were determined to be the best techniques for detecting phishing emails.

In the study by Gupta et al., a new approach was used to detect phishing URLs [22]. The machine learning techniques used in this research are RF, KNN, SVM, and LR. The features were selected based on words. As a result, the RF algorithm achieved the highest accuracy rate.

The study by Moradpoor et al. aims to detect phishing emails [23]. Therefore, a neural network with 6 components was implemented. Phishing emails in the dataset used in this study were obtained from the Phishcorpus dataset, while normal emails were obtained from the SpamAssassin dataset. Initially, using Python code, phishing, and normal emails were selected from the two datasets and represented as normal = 0 and phishing = 1. Then, for each email, the number of web links, the presence of HTML tags, the presence of JavaScript code, and the number of email sections were determined and stored in a boolean or integer variable. Data cleaning and feature extraction were performed in the next step using Word2Vec methods. After vectorization, all variables obtained from the process were saved in a .csv file with 7 columns containing email, vector average, number of web links, HTML presence, JavaScript presence, email section count, and email type. The dataset was divided into 70% training, 15% validation, and 15% testing. A neural network model consisting of Input Matrix and Target Matrix components, 10 hidden layers, 5 input features, 1 output layer, and 1 output feature was developed. The results were evaluated using a confusion matrix and network performance metrics.

In the study by Fayoumi et al., a dataset with 9 features was used to detect phishing emails using machine learning algorithms [24]. The features used include the number of dots in the link, the number of links in the email, the presence of JavaScript codes in the email, the presence of form and HTML tags, the use of action words, and the presence of words like PayPal, bank, and account. In this study, the performances of NB, RF, and SVM algorithms in phishing email detection were compared. Accuracy and F1-score metrics were used to evaluate the results, and the SVM algorithm showed the highest performance.

The study conducted by Salahdine et al. aims to examine the performance of machine learning algorithms used in phishing detection [25]. This study used a dataset consisting of 2000 phishing emails targeting North Dakota University's email

system. In the preprocessing step, values were converted into numerical values. The classification process was based on 10 features, such as inconsistencies in the sender's email address, suspicious file extensions, blacklist words, SSL certificates, etc., and SVM, LR, and Artificial Neural Network (ANN) algorithms were used. Metrics such as true positive, false positive, false negative, and accuracy were used to evaluate the results. In this study, the ANN model showed the highest performance. Different activation functions were tried for ANN, and the most successful result was obtained with ReLu.

In Sekiya and Wei's study, the performance of batch machine-learning techniques was examined for detecting phishing websites [26]. Primary machine learning algorithms such as K-Means, SVM, LR, NB, Linear Discriminant Analysis (LDA), Classification & Regression Trees (CART), and RF were compared. It was observed that RF performed the highest in this comparison. Then, ensemble machine learning algorithms such as AdaBoost, Gradient Boosted Decision Trees (GBDT), XGBoost, and LightGBM were also compared, and it was found that RF provided the highest accuracy and LightGBM exhibited the fastest performance. Deep learning models showed better and faster performance when applied to large datasets compared to traditional machine learning. Still, they also have disadvantages such as model architecture design, manual parameter tuning, high training time costs, and computational complexity. This could lead to potential accuracy improvement. Batch machine learning methods have the potential to provide higher accuracy rates because they combine different models. In this study, CART and RF demonstrated the highest performance. Consequently, it is suggested that automatic feature selection methods could address problems such as dealing with large datasets using batch machine learning algorithms.

Jain and Gupta's study focused on detecting phishing attacks using machine learning and hyperlink analysis [27]. The foundation of the study is to develop a machine-learning model by examining hyperlinks in existing HTML codes of browsers. 12 features, such as the total number of links, internal and external links, errors, redirects, and empty links, were used to develop the model. Initially, link features were extracted. In the next stage, feature vectors were created for each website. The performance metrics used in this study include true positive rate, false positive rate, true negative rate, false negative rate, F1-score, accuracy, precision, and recall. LR exhibited the highest performance in this study.

Ahammad et al. utilized phishing emails collected from various sources and normal emails from the Spam Classification dataset in their study [28]. Initially, the data underwent preprocessing, including tokenization and stemming processes. After preprocessing, the words in phishing-containing emails were visualized using the Cloud Module, where the density of words was determined so that more frequently used words had higher density. Then, a corpus containing 100 words related to phishing was created. The next step involved feature engineering, where a new dataset was created. Each word in the corpus represented a feature, and the frequency of each phishing word in the email text was determined as the corresponding value for this feature. Due to the high number of features, feature

reduction techniques such as principal component analysis, forward feature selection, backward feature selection, non-negative matrix factorization, and recursive feature elimination were explored, along with cross-validation. Machine learning techniques used in this research included LR, DT, SVM, NB, and KNN. A deep neural network with an input layer of 100 features, 1-2 hidden layers, and one output layer was employed alongside machine learning techniques. The results were evaluated by comparing the accuracy rates and the number of features provided by models that gave the most suitable number of features in each dimension reduction technique. Forward feature selection yielded the highest accuracy rate with the NB algorithm.

Thapa et al. conducted a pioneering study applying federated learning (FL) to phishing email detection [29]. The study investigated the performance of FL on distributed datasets using two state-of-the-art models: THEMIS and BERT. FL enables collaborative model training across multiple organizations without sharing raw data, thus preserving data privacy. The results demonstrated that FL achieved performance comparable to centralized learning (CL) under balanced data distributions, with test accuracies of 96.1% for BERT and 97.9% for THEMIS. However, performance varied under scenarios with asymmetric data distributions or extreme dataset diversity, highlighting model dependency. This study underscores the potential of FL as a privacy-preserving approach to phishing email detection.

Wosah et al. proposed a framework for mitigating phishing attacks by integrating stylometric features, gender identification, and email header analysis into a Colour Code Email Verification (CCEV) system [30]. The framework leverages natural language processing and LSTM techniques to analyze email authenticity. By assigning color codes—green for safe, amber for suspicious, and red for high threat—the system provides real-time sender verification at the recipient's end. The study utilized the Enron email dataset for model development and evaluation, demonstrating that the system effectively assists users in distinguishing between legitimate and phishing emails, thereby enhancing cybersecurity against sophisticated spear-phishing attacks.

Jamal et al. proposed the Improved Phishing and Spam Detection Model (IPSDM), leveraging the capabilities of large language models (LLMs) to classify phishing, spam, and ham emails [31]. The study fine-tuned and optimized transformer-based models, specifically DistilBERT and RoBERTa, demonstrating their superior performance over traditional approaches in both balanced and imbalanced datasets. IPSDM achieved significant improvements in classification metrics, including accuracy, precision, recall, and F1-score, by addressing class imbalance using adaptive synthetic sampling (ADASYN) and mitigating overfitting issues through advanced training techniques. The findings underscore the potential of LLMs to provide innovative and effective solutions to longstanding challenges in email security, such as phishing and spam detection.

Al-Subaiey et al. proposed a novel web-based platform for phishing email detection by integrating Explainable AI (XAI) techniques and machine learning models [32]. The study

utilized six publicly available datasets, merging them into a single corpus of approximately 82,500 emails to enhance generalizability and robustness. The proposed platform employed TF-IDF for feature extraction and SVM for classification, achieving an F1-score of 0.99. Explainable AI techniques, such as LIME, were implemented to increase user trust by providing insights into model predictions. The platform was deployed as a user-friendly web application, enabling real-time phishing detection and allowing users to provide feedback for continuous model refinement. This study bridges the gap between high-performing models and their practical application, offering a scalable solution to combat phishing emails effectively.

B. Motivation and Contribution

Phishing attacks, particularly those targeting email users, continue to evolve, employing increasingly sophisticated tactics to deceive users. While existing studies have extensively utilized machine learning and deep learning techniques such as CNNs, LSTMs, and GRUs, they often overlook structural and linguistic features that can play a critical role in distinguishing phishing emails from legitimate ones. For example, CSS and HTML tags, black-listed words, and spelling or grammatical errors are common indicators of phishing emails that remain underexplored in the literature. This study aims to address this gap by introducing a novel feature set tailored to capture these overlooked characteristics. The main contributions of the study can be summarized as follows:

- We propose a set of innovative features, including counts of CSS and HTML tags, black-listed words, and grammatical errors, which significantly enhance the classification performance of phishing email detection systems.
- Our study demonstrates the effectiveness of combining these new features with traditional text representation methods (e.g., TF-IDF, Word2Vec) in improving model accuracy, precision, recall, and F1-score.
- We provide a detailed comparison with existing methods in the literature, highlighting that the incorporation of these features leads to state-of-the-art performance (e.g., achieving an F1-score of 99.53% with RF and TF-IDF).
- The proposed features and methods are validated on a real-world dataset, showcasing their potential for practical application in combating phishing threats.

C. Organization

The remaining parts of the study are organized as follows: Section II will discuss the classifiers, platforms, and methods used in machine learning and deep learning-based phishing email detection. Section III will address the proposed new features. Section IV will present and interpret the results obtained from the study. Finally, Section V will provide a general assessment and information regarding future studies.

II. EXPERIMENTAL SETTINGS

This section will cover the programming language and libraries used, the dataset, data preprocessing steps, feature

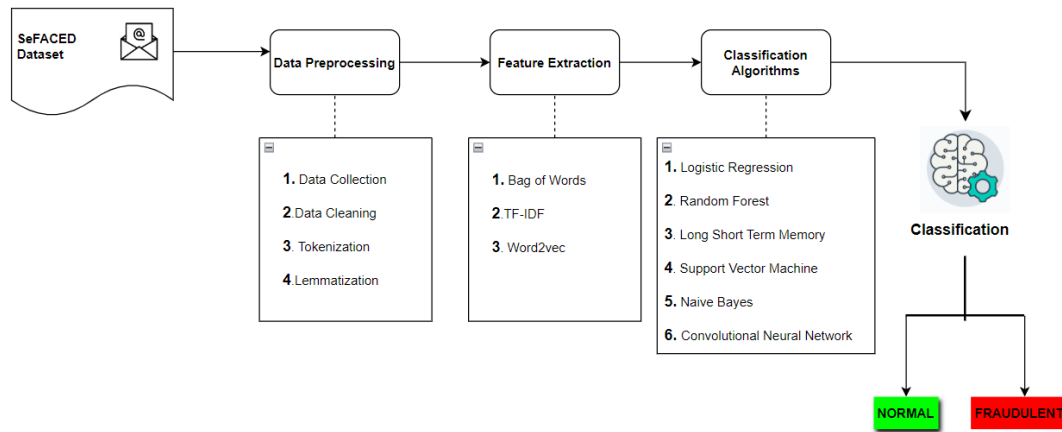


Fig. 1. Steps of Constructing the Proposed Model

extraction techniques, classification algorithms, and metrics used to evaluate classification performance. Figure 1 provides the general architecture of the proposed machine learning-based phishing email detection system.

A. Programming Language and Libraries Used

The models in this study were developed using the Python programming language. Various open-source libraries and tools were utilized to streamline and enhance the machine learning and deep learning processes. Libraries such as Scikit-learn (for implementing machine learning algorithms and data preprocessing), TensorFlow (for building and training neural network-based models), and Pandas (for data manipulation and analysis) played pivotal roles in model development. Additional libraries, including NLTK (for natural language processing tasks) and Numpy (for numerical computations and array operations), were employed to ensure robust data preprocessing and feature engineering. The development and experimentation were carried out on platforms like Google Colab and Jupyter Notebook, which facilitated efficient execution, debugging, and visualization of the results.

B. Dataset Used

The dataset used in the study is the SeFACED dataset [33]. This dataset is obtained by merging three different datasets. It contains 12498 normal, 5142 fraudulent, 19190 harassment, and 5323 suspicious emails. Each email's header information, such as sender and subject, has been removed. The normal emails in the SeFACED dataset are taken from the Enron Corpora, fake emails are from the Phished Emails Corpora, suspicious emails are from the Email Forensics dataset, and harassment emails are from the Hate Speech and Offensive dataset. 12498 normal emails and 5142 fraudulent emails were selected to create the dataset used in the study. 80% of the dataset was used for training, while 20% was used for testing. For a fair comparison, experiments were carried out by running the `random_state` parameter of the `train_test_split` module in

the Sklearn library with the value 42 in all cases, since all algorithms must use the same training and the same test split.

C. Steps of Preprocessing

Email data needs to undergo a series of processing steps to be prepared for use in the model. The first stage is the data preprocessing stage. In this step, data collection, data cleaning, tokenization, and stemming processes are applied.

1) *Data Collection*: The data used is obtained from the SeFACED dataset, which is divided into 4 classes and consists of 42153 email texts [33]. However, in this study, binary classification is performed, so 12498 normal and 5142 fraudulent emails were used.

2) *Data Cleaning*: The data cleaning process involves removing punctuation marks, converting the text to lowercase, removing stop words from the text, removing links, numbers, special characters, and HTML tags, and cleaning up spaces. Some Python scripts have been used to perform these steps.

3) *Tokenization*: Email texts have been divided into smaller units to make the data more organized and manageable. In this step, the "tokenize" method from the NLTK library has been used.

4) *Stemming*: Stemming is the process of reducing words in a text to their bases or roots. The "stem" method from the NLTK library has been used to perform this operation.

D. Feature Extraction Techniques

Methods such as Bag of Words, TF-IDF, and Word2Vec were used to perform feature extraction.

1) *Bag of Words*: Bag of Words is a technique used to transform the words in each document of a dataset into a vector that represents their frequencies. The "CountVectorizer" method from the Sklearn library has been used to create the Bag of Words.

2) *TF-IDF*: Term Frequency-Inverse Document Frequency (TF-IDF) is a technique used to represent both the frequencies of words and the importance of each word in a document. The "TfidfVectorizer" method from the Sklearn library has been used to implement this technique.

3) *Word2Vec*: In natural language processing, Word2Vec is a method for obtaining vector representations of words. These vectors use the surrounding words to infer information about the word's meaning. The Word2Vec algorithm models text in a large corpus in order to estimate these representations. In this study, the Word2Vec technique has been utilized using the Gensim library.

E. Classification Algorithms

Six machine and deep learning algorithms have been used to create models in this study. These are LR, Random Forest (RF), LSTM, Support Vector Machine (SVM), NB, and CNN.

1) *Logistic Regression*: LR is one of the popular classification algorithms and is mostly used in binary classification problems. It is a supervised machine learning algorithm used when the categorical dependent variable is discrete. The sigmoid function is generally used as the activation function. The dependent variable is usually a binary variable defined as 1 and 0 [34].

2) *Random Forest*: RF is fundamentally based on the principle of aggregating the predictions produced by many decision trees. This algorithm is generally used in classification and regression problems. It prevents overfitting errors that may occur in decision trees. There is a linear relationship between the number of trees in the algorithm and the classification result obtained [13].

3) *Long Short Term Memory*: In deep learning, LSTM is a frequently used recurrent neural network architecture designed to prevent long-term dependencies. It consists of gates that control the input or output of information to the relevant cell. The LSTM architecture is widely used in many areas, such as text and language processing, speech recognition, and handwriting recognition [35].

4) *Support Vector Machine*: SVM is frequently used in classification problems, but it is also a supervised learning algorithm used in areas such as clustering and anomaly detection. Essentially, this algorithm separates the data with a hyperplane, also known as the decision boundary, which is mainly used to separate data consisting of two classes [24].

5) *Naive Bayes*: The NB algorithm is based on Bayes' theorem, frequently used in probability. This classifier is a commonly used supervised learning algorithm in machine learning. It works by calculating the probability of each possible outcome for a given data point and then performing classification based on the resulting probability values [24].

6) *Convolutional Neural Network*: CNN is a type of artificial neural network typically composed of input layers, convolutional layers, pooling layers, and fully connected layers. It is also a subfield of deep learning. In addition to the input and output layers, it has multiple hidden layers. It is a popular tool used in fields such as image processing, image and video recognition, and image classification [36].

F. Classification Performance Metrics

In this section, four performance metrics accuracy, recall, precision, and F1-score were used to evaluate the performance of the created models.

1) *Accuracy*: Accuracy is calculated as the ratio of the number of correctly predicted examples to the total dataset. It can also be referred to as the percentage of correctly classified data. The accuracy metric is given in Equation 1. The variables used in this metric and the other metrics are as follows:

- **TP**: True Positive. This refers to the case where the values predicted as positive are actually positive.
- **TN**: True Negative. This refers to the case where the values predicted as negative are actually negative.
- **FP**: False Positive. This refers to the case of predicting examples with a true negative value as positive.
- **FN**: False Negative. This refers to the case of predicting values as negative when they are actually positive.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

2) *Precision*: Precision indicates how many of the predictions identified as positive are actually positive. The precision metric is shown in Equation 2.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

3) *Recall*: Precision is the ratio of the number of true positive predictions to the total number of predictions made as positive. The mathematical representation of the precision metric is given in Equation 3.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

4) *F1-score*: The F1-score is a measure calculated by combining precision and recall metrics by taking the harmonic mean of these values. Particularly in imbalanced datasets, interpreting based solely on accuracy can be misleading. The F1-score can take values between 0 and 1, with higher values indicating better performance. The mathematical representation of this metric is given in Equation 4.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

III. RECOMMENDED ATTRIBUTES FOR PHISHING EMAIL DETECTION

In the second part of the study, new features were proposed before the preprocessing stage to improve the classification performance. These features include the counts of HTML tags, CSS tags, black-listed words, links, language, and spelling errors in each email. These features address key gaps in previous studies, where contextual and structural characteristics of emails were often overlooked, focusing instead on content-based analysis. By integrating these features, the proposed model provides a more comprehensive approach to phishing email detection. This information was extracted from email texts and updated on a .csv file. These additional features were added as columns to matrices created using feature extraction methods.

Algorithm 1: Phishing Email Detection Feature Extraction

Input: Email dataset D , Feature extraction methods F
Output: Enhanced feature matrix M

- 1 Initialize an empty feature matrix M ;
- 2 **foreach** email $e \in D$ **do**
- 3 Extract raw text T from email e ;
- 4 Compute structural features;;
- 5 $HTML_Count \leftarrow$ Count of HTML tags in T ;
- 6 $CSS_Count \leftarrow$ Count of CSS tags in T ;
- 7 $Blacklist_Words \leftarrow$ Count of black-listed words in T ;
- 8 $Link_Count \leftarrow$ Number of links in T ;
- 9 $Language \leftarrow$ Detected language of T ;
- 10 $Spelling_Errors \leftarrow$ Number of spelling errors in T ;
- 11 Append computed features to email's feature vector;
- 12 Apply feature extraction methods F to T ;
- 13 Add all extracted features as columns to M ;
- 14 **end**
- 15 Update M by saving the enhanced feature matrix to a .csv file;
- 16 **return** M ;

Algorithm 1 defines a process for extracting new structural and content-based features to enhance phishing email detection. It calculates features such as the counts of HTML and CSS tags, black-listed words, number of links, language, and spelling errors from emails, integrating them into the existing feature matrix to improve classification performance.

The proposed features, such as counts of HTML tags, CSS tags, black-listed words, and grammatical errors, were designed to complement traditional text representation techniques like Bag-of-Words, TF-IDF, and Word2Vec. These features enrich the representation of emails by providing structural and linguistic information that is often overlooked in conventional approaches. By integrating these additional attributes, we aim to enhance the ability of classifiers to identify subtle distinctions between phishing and legitimate emails. This holistic feature representation ensures that the model leverages both semantic and structural information during the classification process.

The proposed features are seamlessly combined with text representation outputs to create a unified feature vector for each email. This vector incorporates the text-based features derived from methods like TF-IDF with the contextual cues provided by the novel attributes. As a result, the classification decision is made for the email as a whole rather than its individual segments, addressing potential challenges in judging emails composed of multiple text pieces. This integration improves the classifier's ability to generalize across varied phishing attempts and real-world email data, contributing to robust phishing detection performance.

A. HTML Tags Count

BeautifulSoup library was used to calculate the number of HTML tags in each email text, and this information was saved to the extra features .csv file.

B. CSS Tags Count

BeautifulSoup library was used to calculate the number of CSS (`< style >`) tags in each email text, and this information was saved to the relevant file.

C. Black-list Words Count

Commonly used phishing email keywords were collected and saved to a .txt file [37]. Subsequently, using a Python function, the number of occurrences of these keywords in each email in the dataset was calculated and saved to the extra features file.

D. Links Count

The number of links in each email in the dataset was calculated using a Python script containing a regular expression, and this information was saved to the extra feature file.

E. Grammar Errors and Misspelled Words Count

The "Language_tool_python" library and the "Enchant" module were used to calculate spelling and language errors in each email, and this information was saved to the extra features file.

IV. RESULTS AND DISCUSSIONS

In this section, the results obtained from the study will be presented. In Section IV-A, the classification results without using the proposed features will be provided. In Section IV-B, the classification results obtained by adding the proposed features will be presented. Finally, a comparison will be made by presenting the results obtained from the literature alongside the results obtained from this study in tabular form in IV-C.

TABLE I
RESULTS OBTAINED BEFORE ADDING NEW ATTRIBUTES

Model	Accuracy	Precision	Recall	F1-score
LR (Bag-of-Words)	0.9836	0.9834	0.9938	0.9886
RF (Bag-of-Words)	0.9866	0.9870	0.9943	0.9906
NB (Bag-of-Words)	0.9763	0.9758	0.9914	0.9835
SVM (Bag-of-Words)	0.9851	0.9858	0.9934	0.9895
LSTM (Bag-of-Words)	0.9851	0.9805	0.9675	0.9739
CNN (Bag-of-Words)	0.9851	0.9865	0.9614	0.9738
LR (TF-IDF)	0.9790	0.9743	0.9967	0.9854
RF (TF-IDF)	0.9880	0.9878	0.9955	0.9916
NB (TF-IDF)	0.9702	0.9602	0.9996	0.9795
SVM (TF-IDF)	0.9907	0.9894	0.9975	0.9935
LSTM (TF-IDF)	0.9863	0.9896	0.9624	0.9758
CNN (TF-IDF)	0.9851	0.9775	0.9706	0.9740
LR (Word2Vec)	0.9506	0.9595	0.9717	0.9656
RF (Word2Vec)	0.9836	0.9810	0.9863	0.9886
NB (Word2Vec)	0.8796	0.9575	0.8696	0.9114
SVM (Word2Vec)	0.9547	0.9623	0.9746	0.9684
LSTM (Word2Vec)	0.9707	1.0	0.8985	0.9465
CNN (Word2Vec)	0.9953	0.9899	0.9939	0.9919

TABLE II
RESULTS OBTAINED AFTER ADDING NEW ATTRIBUTES

Model	Accuracy	Precision	Recall	F1-score
LR (Bag-of-Words)	0.9860	0.9874	0.9930	0.9902
RF (Bag-of-Words)	0.9930	0.9923	0.9979	0.9951
NB (Bag-of-Words)	0.9752	0.9892	0.9758	0.9824
SVM (Bag-of-Words)	0.9886	0.9886	0.9955	0.9920
LSTM (Bag-of-Words)	0.9901	0.9779	0.9878	0.9828
CNN (Bag-of-Words)	0.9892	0.9897	0.9726	0.9811
LR (TF-IDF)	0.9822	0.9829	0.9922	0.9875
RF (TF-IDF)	0.9933	0.9931	0.9975	0.9953
NB (TF-IDF)	0.9328	0.9646	0.9401	0.9522
SVM (TF-IDF)	0.9924	0.9939	0.9955	0.9947
LSTM (TF-IDF)	0.9898	0.975	0.9898	0.9824
CNN (TF-IDF)	0.9860	0.9795	0.9716	0.9755
LR (Word2Vec)	0.9611	0.9702	0.9754	0.9728
RF (Word2Vec)	0.9883	0.9878	0.9959	0.9918
NB (Word2Vec)	0.8443	0.9644	0.8113	0.8813
SVM (Word2Vec)	0.9620	0.9726	0.9742	0.9734
LSTM (Word2Vec)	0.9971	0.9939	0.9959	0.9949
CNN (Word2Vec)	0.9921	0.9990	0.9736	0.9861

A. Results Without New Attributes

In the first stage of the study, machine learning and deep learning models were created without adding the extracted features to the bag-of-words, TF-IDF, and Word2Vec matrices. Table I presents all the results. Among the algorithms used, the SVM algorithm with TF-IDF feature extracting technique achieved the best performance with an accuracy of 99.07%, precision of 98.94%, recall of 99.75%, and F1-score of 99.35%.

The lowest performance among the algorithms used was obtained by the NB algorithm using Word2Vec, with an accuracy of 87.96%, precision of 95.75%, recall of 86.96%, and F1-score of 91.14%.

B. Results Obtained by Adding New Attributes

At this stage, the previously created matrices were augmented with additional features, such as the number of HTML tags and spelling errors, and these features were added as columns. The classification results with the addition of new features are provided in Table II. Among the machine learning algorithms, the RF algorithm using TF-IDF feature extraction

technique achieved the best performance with an accuracy of 99.33%, precision of 99.31%, recall of 99.75%, and F1-score of 99.53%. The performance of this algorithm increased by almost 1% compared to its performance without the extra features.

The LSTM algorithm with word2Vec feature extraction technique showed the best performance among deep learning algorithms with an F1-score of 99.49% and an accuracy of 99.71%. It was observed that the accuracy performance increased by 2% compared to the algorithm's previous performance. These results demonstrate the significant improvement achieved by incorporating the proposed features, as the LSTM model's performance surpasses many state-of-the-art approaches highlighted in the literature, further validating the effectiveness of the proposed methodology.

Among the machine learning algorithms, the lowest performance was obtained by the NB algorithm using Word2Vec, with an accuracy of 84.43%, precision of 96.44%, recall of 81.13%, and F1-score of 88.13%.

When the proposed features that distinguish phishing email attacks from normal emails are generally evaluated, it is seen

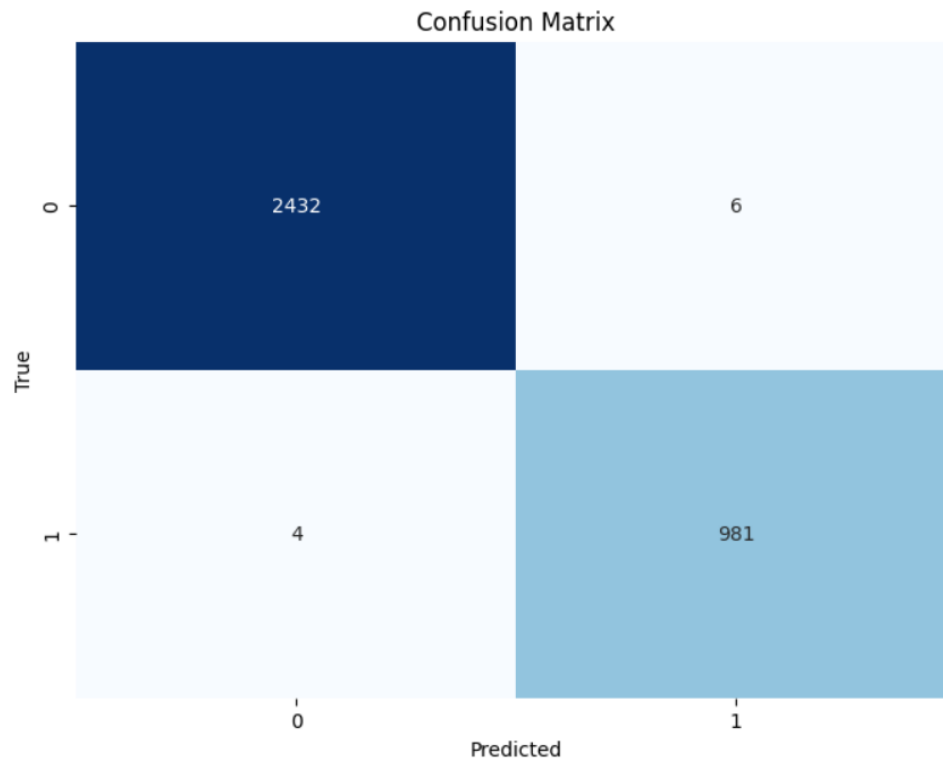


Fig. 2. Confusion Matrix Obtained by LSTM Algorithm After Adding the New Features

that the classification performance of machine learning and deep learning algorithms mostly increases. Performance decreases when Word2Vec text representation and NB algorithm are used. In tests performed without adding the recommended features, lower performance was achieved compared to other cases. The most important reason for this is Word2Vec text representation. Because Word2Vec has a more complex structure for the NB algorithm compared to TF-IDF and Bag of Words text representation.

When Word2Vec text representation and LSTM network are used together with the proposed features, the highest classification performance is achieved among all the experiments. According to the accuracy metric, this result is reported as 0.9971. The complexity matrix for this experiment is given in Figure 2. There are actually 2438 normally labeled samples included in the test data. Only 6 of these samples are misclassified. On the other hand, 4 out of 985 fraudulent samples in the test data are misclassified. LSTM incorrectly predicts the labels of 10 samples in total. Considering all the experiments, this result is the highest performance result achieved in terms of accuracy metric. The contribution of the proposed features to the classification performance comes to the fore with this experiment.

C. Comparison of Results Obtained from Existing Studies

In this section, a comparison will be made between the results obtained from other studies and the relevant study. The results of existing studies are provided in Table III, which includes the results of articles related to phishing. The results obtained with various classification algorithms have

been evaluated according to the relevant performance metrics. Among the machine learning algorithms, the study conducted by Livara et al. [12] achieved the highest performance ratio. The algorithm used in their study was RF, and it achieved the highest performance with an F1-score of 99.4%. Among the deep learning algorithms, the study conducted by Dewis et al. [14] achieved the highest ratio. In [14], the LSTM algorithm achieved a success rate of 99% based on the F1-score. The results of both studies are lower than the results of the proposed model. When the text representation obtained by adding the recommended extra features is given to RF with TF-IDF features extraction technique, 99.53% performance is achieved.

In the conducted study, the addition of extra features improved the classification performance in general. For example, the classification performance reached a 99.02% F1-score with the LR algorithm using the Bag of Words feature extraction technique. However, it was observed that the performance of the NB algorithm decreased slightly when extra features were added compared to the classification performed without adding extra features.

V. GENERAL EVALUATION AND DISCUSSIONS

According to researches, many phishing attacks occur via email, hence the aim of classifying phishing attacks using machine and deep learning algorithms. A comprehensive literature review of 18 articles was conducted in the study. The dataset used is the SeFACED dataset, consisting of 12,498 legitimate and 5,143 phishing email data. During model creation, the data underwent preprocessing, which is crucial for

TABLE III
COMPARISON WITH EXISTING STUDIES

Study	Dataset Size	Used Method	Performance
Livara and Hernandez [12]	525,754 emails	RF	99.4% (F1-score)
Akinyelu and Adewumi [13]	2,000 emails	RF	98.45% (F1-score)
Paradkar [11]	8,336 phishing, 11,664 normal emails	CNN	98.26% (F1-score)
Ahi and Soğukpınar [9]	2,256 secure, 2,256 phishing emails	H-OLTA (Hybrid MLP-LSTM)	96% (F1-score)
Ahammad et al. [28]	Not specified	NB	96% (Accuracy)
Fayoumi et al. [24]	Not specified	SVM	99.80% (F1-score)
Dewis and Viana [14]	6 different datasets	LSTM	99% (Accuracy)
Abdulraheem et al. [10]	11,000 websites, 2,500 phishing emails	LMT	96.9% (Precision)
Proposed Method	5,142 phishing, 12,498 normal emails	RF (TF-IDF with extra features)	99.53% (F1-score)

normalizing the data and removing duplicate entries. The normalized data was digitized using feature extraction techniques to be utilized in the model. The algorithms used are LR, RF, LSTM, SVM, NB, and CNN. The performance metrics used to evaluate the models are F1-score, accuracy, precision, and recall values. In the second part of the study, the results of classification performance with and without additional features were analyzed in detail. Overall, the classification performance significantly improved when additional features were added. In the classification using Bag-of-Words with additional features, the algorithm that showed the highest increase in classification performance according to the F1-score was LSTM, with a rate of 98.28%. Similarly, in the classification using TF-IDF with additional features, LSTM showed the highest increase in classification performance with a rate of 98.24%. In the classification using Word2Vec with additional features LSTM showed the highest increase in classification performance with a rate of 99.49%.

REFERENCES

- [1] R. Alabdan, "Phishing attacks survey: Types, vectors, and technical approaches," *Future internet*, vol. 12, no. 10, p. 168, 2020.
- [2] K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks: Their types, vectors and technical approaches," *Expert Systems with Applications*, vol. 106, pp. 1–20, 2018.
- [3] U. A. Butt, R. Amin, H. Aldabbas, S. Mohan, B. Alouffi, and A. Ahmadian, "Cloud-based email phishing attack using machine and deep learning algorithm," *Complex & Intelligent Systems*, vol. 9, no. 3, pp. 3043–3070, 2023.
- [4] M. Jakobsson, "Two-factor inauthentication—the rise in sms phishing attacks," *Computer Fraud & Security*, vol. 2018, no. 6, pp. 6–8, 2018.
- [5] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of ai-enabled phishing attacks detection techniques," *Telecommunication Systems*, vol. 76, pp. 139–154, 2021.
- [6] APWG, "Apwg phishing activity trends report," 2025. [Online]. Available: <https://apwg.org/trendsreports>
- [7] S. Gupta, A. Singhal, and A. Kapoor, "A literature survey on social engineering attacks: Phishing attack," in *2016 International Conference on Computing, Communication and Automation (ICCCA)*, 2016, pp. 537–540.
- [8] J. Rastenienis, S. Ramanauskaitė, J. Janulevičius, A. Čenys, A. Slotkienė, and K. Pakrijauskas, "E-mail-based phishing attack taxonomy," *Applied sciences*, vol. 10, no. 7, p. 2363, 2020.
- [9] Ş. Ahi and İ. Soğukpınar, "Derin öğrenme modelleri ile kimlik avı e-posta tespiti," *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, vol. 13, no. 2, pp. 17–31, 2020.
- [10] R. Abdulraheem, A. Odeh, M. Al Fayoumi, and I. Keshta, "Efficient email phishing detection using machine learning," in *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2022, pp. 0354–0358.
- [11] N. S. Paradkar, "Phishing email's detection using machine learning and deep learning," in *2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*. IEEE, 2023, pp. 160–162.
- [12] A. Livara and R. Hernandez, "An empirical analysis of machine learning techniques in phishing e-mail detection," in *2022 International Conference for Advancement in Technology (ICONAT)*. IEEE, 2022, pp. 1–6.
- [13] A. Akinyelu and A. Adewumi, "Classification of phishing email using random forest machine learning technique," *Journal of Applied Mathematics*, vol. 2014, 2014.
- [14] M. Dewis and T. Viana, "Phish responder: A hybrid machine learning approach to detect phishing and spam emails," *Applied System Innovation*, vol. 5, no. 4, p. 73, 2022.
- [15] X.-W. Chen and X. Lin, "Big data deep learning: challenges and perspectives," *IEEE access*, vol. 2, pp. 514–525, 2014.
- [16] M. Coşkun, Ö. Yıldırım, A. Uçar, and Y. Demir, "An overview of popular deep learning methods," *European Journal of Technique (EJT)*, vol. 7, no. 2, pp. 165–176, 2017.
- [17] M. K. Sharma, R. Kumar, D. K. Sinha, K. Senthilkumar, D. Dhaliya, and G. Ahluwalia, "Exploring the benefits of deep learning for data science practices," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2024, pp. 1–7.
- [18] E. E. Eryilmaz, D. O. Şahin, and E. Kılıç, "Machine learning based spam e-mail detection system for turkish," in *2020 5th International Conference on Computer Science and Engineering (UBMK)*. IEEE, 2020, pp. 7–12.
- [19] S. T. Singh, M. D. Gabhane, and C. Mahamuni, "Study of machine learning and deep learning algorithms for the detection of email spam based on python implementation," in *2023 International Conference on Disruptive Technologies (ICDT)*. IEEE, 2023, pp. 637–642.
- [20] B. Sonare, G. J. Dharmale, A. Renapure, H. Khandelwal, and S. Narharshettiwar, "E-mail spam detection using machine learning," in *2023 4th International Conference for Emerging Technology (INCET)*. IEEE, 2023, pp. 1–5.
- [21] A. A. Adzhar, Z. Mabni, and Z. Ibrahim, "A comparative study on email phishing detection using machine learning techniques," in *2022 IEEE International Conference on Computing (ICOCO)*. IEEE, 2022, pp. 96–101.
- [22] B. B. Gupta, K. Yadav, I. Razzak, K. Psannis, A. Castiglione, and X. Chang, "A novel approach for phishing urls detection using lexical based machine learning in a real-time environment," *Computer Communications*, vol. 175, pp. 47–57, 2021.
- [23] N. Moradpoor, B. Clavie, and B. Buchanan, "Employing machine learning techniques for detection and classification of phishing emails," in *2017 Computing Conference*. IEEE, 2017, pp. 149–156.
- [24] M. Al Fayoumi, A. Odeh, I. Keshta, A. Aboshgifa, T. AlHajjahjeh, and R. Abdulraheem, "Email phishing detection based on naïve bayes, random forests, and svm classifications: A comparative study," in *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2022, pp. 0007–0011.
- [25] F. Salahdine, Z. El Mrabet, and N. Kaabouch, "Phishing attacks detection a machine learning-based approach," in *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE, 2021, pp. 0250–0255.
- [26] Y. Wei and Y. Sekiya, "Sufficiency of ensemble machine learning methods for phishing websites detection," *IEEE Access*, vol. 10, pp. 124 103–124 113, 2022.
- [27] A. K. Jain and B. B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, pp. 2015–2028, 2019.
- [28] S. M. M. Ahammad, T. Raviteja, J. Koushik, P. V. Dinesh, and A. Ashok, "Machine learning approach based phishing email text analysis (ml-pe-ta)," in *2022 Third International Conference on Intelligent Computing*

- Instrumentation and Control Technologies (ICICT)*. IEEE, 2022, pp. 1087–1092.
- [29] C. Thapa, J. W. Tang, A. Abuadba, Y. Gao, S. Camtepe, S. Nepal, M. Almashor, and Y. Zheng, “Evaluation of federated learning in phishing email detection,” *Sensors*, vol. 23, no. 9, p. 4346, 2023.
- [30] P. N. Wosah, Q. Ali Mirza, and W. Sayers, “Analysing the email data using stylometric method and deep learning to mitigate phishing attack,” *International Journal of Information Technology*, pp. 1–14, 2024.
- [31] S. Jamal, H. Wimmer, and I. H. Sarker, “An improved transformer-based model for detecting phishing, spam and ham emails: A large language model approach,” *Security and Privacy*, p. e402, 2024.
- [32] A. Al-Subaiey, M. Al-Thani, N. A. Alam, K. F. Antora, A. Khandakar, and S. A. U. Zaman, “Novel interpretable and robust web-based ai platform for phishing email detection,” *Computers and Electrical Engineering*, vol. 120, p. 109625, 2024.
- [33] M. Hina, M. Ali, A. R. Javed, F. Ghabban, L. A. Khan, and Z. Jalil, “Sefaced: Semantic-based forensic analysis and classification of e-mail data using deep learning,” *IEEE Access*, vol. 9, pp. 98 398–98 411, 2021.
- [34] AWS-Blog, “Lojistik Regresyon Nedir? - Lojistik Regresyon Modeline Ayrıntılı Bakış,” 2025. [Online]. Available: <https://aws.amazon.com/tr/what-is/logistic-regression/>
- [35] E. Gavcar and H. M. Metin, “Hisse senedi değerlerinin makine öğrenimi (derin öğrenme) ile tahmini,” *Ekonomi ve Yönetim Araştırmaları Dergisi*, vol. 10, no. 2, pp. 1–11, 2021.
- [36] H. Li, “Computer network connection enhancement optimization algorithm based on convolutional neural network,” in *2021 International Conference on Networking, Communications and Information Technology (NetCIT)*. IEEE, 2021, pp. 281–284.
- [37] A. Onar, “English Spam Words List,” 2025. [Online]. Available: <https://github.com/OOPSpam/spam-words/blob/main/spam-words-EN.txt>

BIOGRAPHIES



Hadjer Brioua has been an undergraduate student at Ondokuz Mayıs University, Faculty of Engineering, Department of Computer Engineering since 2019. She is currently pursuing the B.Sc. degree in Computer Engineering. Her research interests include machine learning, text mining, and deep learning.



Havvanur Siyambaş has been an undergraduate student at Ondokuz Mayıs University, Faculty of Engineering, Department of Computer Engineering since 2020. She is currently pursuing the B.Sc. degree in Computer Engineering. Her research interests include machine learning, text mining, and deep learning.



Durmuş Özkan Şahin received a Bachelor's degree in Computer Engineering from Süleyman Demirel University Isparta in 2013 and a Master's degree in Computer Engineering from Ondokuz Mayıs University Samsun in 2016. Finally, he received a PhD's degree in Computational Sciences from Ondokuz Mayıs University Samsun in 2022. His research interests include machine learning, text mining, information retrieval, and Android malware analysis. He is currently an Assistant Professor of the Department of Computer Engineering at Ondokuz Mayıs University.

Fingerprint Generation for DNN Training: A Case Study in Fingerprint Classification

Emre Irtem, Nesli Erdogmus

Abstract—Large annotated datasets are crucial for training state-of-the-art deep learning systems. However, the availability of publicly accessible fingerprint data significantly lags behind that of image datasets or text corpora, which are extensively utilized for tasks such as image understanding and natural language processing. The challenges associated with the collection and distribution of fingerprint data make synthetic data generation a viable alternative. Nonetheless, existing research primarily focuses on the large-scale evaluation of fingerprint search systems rather than examining the usability of generated fingerprint images for training purposes. This study employs a model-based method to generate synthetic fingerprints and evaluates their effectiveness in training deep neural networks for fingerprint classification. The findings indicate that augmenting the training set with synthetic fingerprint impression images enhances performance comparably to augmenting it with real fingerprint images.

Index Terms—fingerprint image generation, synthetic training data, deep learning, fingerprint classification.

I. INTRODUCTION

AUTOMATION is indispensable for fingerprint identification systems, as thousands of search requests are submitted daily to fingerprint databases, which can reach colossal sizes. Over the past several decades, researchers have developed various algorithms and systems to achieve automated fingerprint analysis and comparison. A subset of this research has concentrated on fingerprint generation, driven primarily by the necessity of large-scale datasets to evaluate the proposed algorithms under realistic conditions, given the limited availability of publicly accessible fingerprint images. This scarcity is due to fundamental challenges in fingerprint data collection, including the need for expert personnel, specialized equipment, and concerns regarding privacy. Furthermore, many commonly used datasets have been discontinued owing to recent legal restrictions aimed at protecting the privacy of biometric data [1].

Today, large-scale datasets are essential not only for performance evaluations but also for training deep neural networks. However, the use of synthetic fingerprints as training data has not been extensively investigated. To the best of our knowledge, only one recent study [2] has evaluated the impact


of data augmentation through fingerprint synthesis, specifically in the context of latent fingerprint reconstruction.


In this study, we aim to analyze how synthetic fingerprint images can enhance the accuracy of deep neural networks in fingerprint classification. Fingerprint classification facilitates fingerprint matching by filtering the database based on the estimated class, thereby reducing the number of candidates and speeding up the search process. This task is chosen for our study because class information is one of the four types of ground truth provided by model-based synthetic data generators, the others being identity, frequency maps, and orientation maps [3], [4]. In contrast, learning-based generators generally do not supply these types of information and are primarily used to enlarge the gallery size for performance evaluations [5], [6]. Considering these factors, we employ the model-based SFinGe method to generate synthetic fingerprints for our approach [3].

Far as we know, the SFinGe method is implemented in two fingerprint generation tools: SFinGe [7] and Anguli [8]. The SFinGe tool is available in both demo and full versions. The demo version permits the generation of a single impression per finger and restricts the total number of generations. The full version allows for multiple impressions per finger, but the number of fingerprints that can be generated is still limited based on the purchased edition. For both versions, the generated synthetic images cannot be distributed or made publicly available on the Internet. Conversely, the Anguli tool is freely available and can be used to generate and distribute any number of synthetic fingerprint images, including multiple impressions of the same finger. However, with Anguli, the extent of degradation in the impressions is not fully controllable. Additionally, for both tools, it is not possible to add extra control parameters for existing degradation types or to introduce entirely new degradation types.

To analyze different degradation types and levels, generate multiple impressions of the same fingerprint, and distribute the synthetic fingerprint dataset for further use and reproducibility, we have implemented our own model-based tool to generate synthetic fingerprint images. The fundamental steps are inspired by the SFinGe method [3]. However, many steps have been modified in order to obtain visually more realistic synthetic fingerprints and some intermediate methods are developed from scratch to generate diverse fingerprint impressions. Unlike the approach in [3],

- A normalization step is added to the ridge pattern generation to reduce the number of required iterations.,
- ridge discontinuities and irregularities within the fingerprint are modeled using coherent noise,

 **Emre Irtem** is with the Department of Computer Engineering, Izmir Institute of Technology, Izmir, 35430 TURKEY e-mail: emreirtem@iyte.edu.tr

 **Nesli Erdoğan** is with the Department of Computer Engineering, Izmir Institute of Technology, Izmir, 35430 TURKEY e-mail: neslierdogmus@iyte.edu.tr

Manuscript received Jul. 19, 2024; accepted Jan. 23, 2025.
DOI: 10.17694/bajece.1519228

- background text and textures are modeled.

The primary contribution of this study lies not in the generation method itself but in the utilization of the generated fingerprint impression images for training set augmentation and the analysis of their impact under different conditions. While it may seem ideal to have fingerprint experts inspect and evaluate the realism of the generated impressions, we did not have access to such expertise. However, this is not the primary focus of our study. Our main objective is to enhance deep learning-based fingerprint processing tasks and, in doing so, demonstrate the fidelity of the generated data. To this end, fingerprint classification is selected as the test case because class is one of the few attributes that can be controlled in the adopted generation method. Furthermore, classification is a fundamental stage in many fingerprint recognition systems, as it can reduce the number of required comparisons and decrease the response time of matching algorithms.

Finally, a dataset of synthetically generated fingerprint impression images is made publicly available to ensure the reproducibility of the results and facilitate further analyses (GitHub repository to be disclosed). The dataset comprises 12,500 samples, with 2,500 images for each of the five NIST standard classes: Right Loop, Left Loop, Whorl, Arch, and Tented Arch (Figure 8b). Each image is accompanied by frequency and orientation maps, minutiae locations, and identity information. Additionally, the experiment codes are also open-sourced.

II. RELATED WORK

Research on the impact of training set augmentation using synthetic fingerprint images is relatively sparse. To our knowledge, only one publication [2] has conducted an analysis along these lines. In that study, latent fingerprints are synthesized for training a reconstruction network aimed at transforming low-quality fingerprint images into ridge images through pixel-level binary classification. The enhancement in reconstruction quality was correlated with improved matching performance, as demonstrated through quantitative assessment, highlighting the efficacy of the proposed training set augmentation. In contrast, other studies referenced in this section primarily focus on generating fingerprints to expand the gallery used for testing purposes.

Each technique possesses distinct advantages and disadvantages. Model-based techniques involve decomposing the generation process into discrete sub-tasks, necessitating significant engineering skill and domain expertise. For instance, in a model-based fingerprint generation pipeline, (I) a fingerprint class is selected, (II) corresponding singular point coordinates and ridge orientations are assigned, (III) ridge frequency maps are generated, (IV) ridge patterns are synthesized using filtering techniques, and (V) fingerprint impressions are created by applying degradation processes to the synthesized pattern. However, numerous assumptions are made at each step; for example, many model-based methods assume independence between ridge orientations and minutiae locations, which may lead to unrealistic minutiae configurations [12]. Nevertheless, model-based techniques offer advantages such as not requiring

training data and inherently providing various ground truth labels on generated images, which are crucial for training purposes.

In contrast, learning-based methods necessitate a substantial number of training samples to effectively learn fingerprint generation. Although they produce "black-box" generators, limiting the direct extraction of detailed fingerprint metadata, they generally yield statistically more realistic samples compared to model-based techniques.

A. Model-based methods

In [9], a model-based technique for fingerprint generation is proposed. The process begins with the generation of a master fingerprint, which serves as the basis for producing multiple impressions. Initially, a fingerprint shape is defined, and a ridge orientation map is calculated using a mathematical model based on zero-pole patterns [10], loop, and delta coordinates. Subsequently, a ridge frequency map is generated by visually inferring from multiple real fingerprints, followed by the synthesis of a ridge pattern image using iterative Gabor filtering [11] applied to a randomly initialized image. The resulting image is then thresholded to obtain a binary master fingerprint image. To create different impressions, three types of synthetic distortions are applied to the master fingerprint. First, morphological operations are employed to simulate wet or dry skin conditions. Second, non-linear transformations are used to mimic skin elasticity. Lastly, ridge discontinuities and irregularities are modeled by adding various white blobs of different shapes and sizes to the fingerprint image.

In [3], the SFinGe approach is introduced as an advancement over its previous version [9], aiming to generate more realistic fingerprint impressions. SFinGe includes additional distortions such as random rotation and translation of the ridge pattern to simulate different finger placements in an image. It also incorporates a realistic background generated using a mathematical model based on the Karhunen-Loeve transform, which is superimposed with the fingerprint impression. Moreover, instead of uniform noise, coherent noise [20] is proposed in [21] to enhance variability in impression generation.

In [12], a non-parametric approach models ridge features of real fingerprints, which are then sampled to create synthetic features used as inputs to the SFinGe tool [3]. The realism of generated fingerprints is validated by comparing their feature densities with those of real fingerprints.

Addressing unrealistic minutiae configurations in synthetic images generated by SFinGe, [16] employs orientation maps extracted from real fingerprints to create master fingerprints. Realness tests are applied to minutiae maps to filter out non-realistic configurations, resulting in a more realistic database in terms of minutiae configurations.

In [14], a statistical model is proposed to obtain realistic fingerprint features, including singular point locations [22] and minutiae points, trained on publicly available real-fingerprint datasets. Unlike SFinGe, the ridge generation phase incorporates AF-FM filters [23].

TABLE I: Existing model and learning-based methods for fingerprint generation

Model based methods			
Method	Orientation model	Ridge generation	Minutiae locations
R. Cappelli et al. [9]	Zero-Pole [10]	Gabor-Filter [11]	Random
SFinGe [3]	Zero-Pole [10]	Gabor-Filter [11]	Random
P.Johnson et al. [12]	Zero-Pole [10]	Gabor-Filter [11]	Statistically realistic [13]
Q.Zhao et al. [14]	Zero-Pole [10]	AF-FM [15]	Statistically realistic [13]
C.Imdahl et al. [16]	Zero-Pole [10]	Gabor-Filter [11]	Elimination using a realness test [17]
Learning based methods			
Method	Learning model		
P.Bontrager et al. [18]	Wasserstein GAN (WGAN)		
K.Cao and A.K.Jain [5]	IWGAN and Autoencoder		
M.Attia et al. [19]	Variational Autoencoder		
V.Mistry et al. [6]	IWGAN and Autoencoder with Identity Loss		

B. Learning-based methods

In [18], a method is proposed to generate synthetic fingerprints using deep neural networks aimed at attacking fingerprint-matching systems. The approach involves training a Wasserstein Generative Adversarial Network (WGAN) and evolving latent variables of the generator network using The Covariance Matrix Adaption Evolutionary Strategy. This method searches for a fingerprint that matches a large number of other fingerprints.

In [5], fingerprints are generated using Improved WGAN for which a Convolutional Autoencoder with a 512-dimensional latent vector is trained and used to initialize the generator of WGAN. This approach demonstrates improvements in fingerprint quality and diversity.

In [19], fingerprints are generated using Variational Autoencoders (VAE). The method ensures that the generated samples match the distribution of real fingerprint datasets via latent vectors. During training, an image x in the input image space X is mapped to a latent vector z in the latent vector space Z by an encoder. The training maximizes $P_Q(x|z)$, where the encoder learns the mapping of latent variable vectors (μ, σ^2) . These variables are used to sample the latent vector z , which is then used by the decoder to generate an output image. Synthetic fingerprints are generated by feeding randomly generated 32-dimensional vectors into the decoder.

In [6], a method that is capable of generating more realistic fingerprints in terms of minutiae count, direction and spatial distribution is proposed. Following initialization with I-WGAN as described in [5], an identity loss is incorporated into the generator's objective during training to ensure the generation of unique identities.

III. FINGERPRINT GENERATION

A model-based generation method is adopted for the following reasons:

- **Data scarcity:** There is a limited number of publicly available fingerprint datasets, which poses a challenge for training data-intensive deep generative networks.
- **Meta-data availability:** It is crucial to obtain meta-data for the generated fingerprints. This study aims to

produce synthetic data suitable for training deep neural networks for various tasks such as minutiae extraction, orientation field estimation, ridge frequency estimation, and fingerprint classification. Providing labels is essential, and model-based techniques are well-suited for this purpose.

- **Multiple impressions:** Training fingerprint matching systems requires multiple impressions of the same finger. The model-based method allows the creation of different impressions from the same master fingerprint.

A. Master Fingerprint Generation

A master fingerprint is an image that represents the ideal impression of a ridge pattern from a "synthetic finger" [9]. It is devoid of any noise or external factors that cause variations in fingerprint images captured from the same finger. While the master fingerprint of an actual finger is not accessible, fingerprint enhancement methods aim to obtain a close approximation.

This study adopts an approach similar to SFinge [3] to generate master fingerprints. After constructing the orientation and frequency maps, a ridge pattern is developed by iteratively applying pixel-specific Gabor filters to a randomly initialized image. Unlike SFinge, this method considers variations in fingerprint area during acquisition and incorporates these variations into the impression generation process.

Orientation images

Orientation images designate principal ridge directions for each pixel on the master fingerprints to be generated. These images must adhere to strict smoothness constraints, as closely positioned ridges have similar orientations, and they must conform to a limited number of ridge pattern types. To achieve this, the zero-pole model proposed by Vizcaya and Gerhardt [24] is utilized, as employed in [3]. This model is based on the work of Sherlock and Monroe [10], which computes the orientation (Θ) image in a complex plane using a complex rational function that incorporates delta (S_{d_i}) and loop (S_{l_j}) locations. (Eqn. 1)

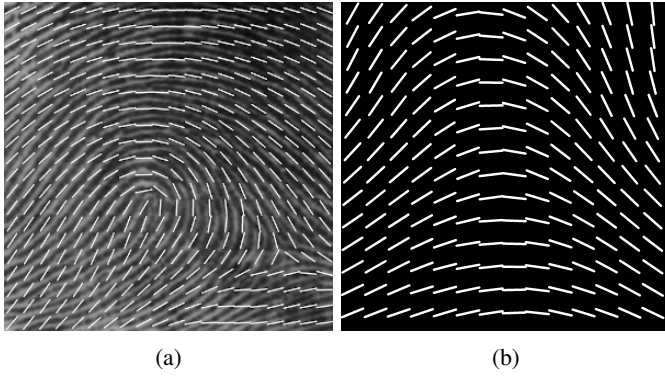


Fig. 1: Orientation maps generated by (a) zero-pole model output superimposed with an actual fingerprint image and (b) sinusoidal model output

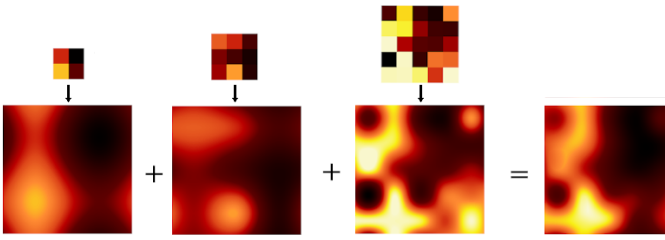


Fig. 2: Frequency image generation using coherent noise

$$\Theta = \frac{1}{2} \left[\sum_i^{n_d} \arg(z - S_{d_i}) - \sum_j^{n_l} \arg(z - S_{l_j}) \right] \quad (1)$$

However, this model does not fit very well on real fingerprints and suffers from low variability. In [24], a piece-wise linear correction function (g) that corrects the phase angle of each singularity is used. (Eqn. 2)

$$\Theta = \frac{1}{2} \left[\sum_i^{n_d} g_{d_i}(\arg(z - S_{d_i})) - \sum_j^{n_l} g_{l_j}(\arg(z - S_{l_j})) \right] \quad (2)$$

This model is capable of generating accurate orientation fields using singular point locations for four major pattern types: left loop, right loop, tented arch, and whorl classes. However, since the arch class lacks singular points, its ridge flow is simulated using a sinusoidal function. For each pixel, orientation is calculated using Eqn. 3.

$$\Theta = \beta \sin(f(x)) \quad (3)$$

The function $f(x)$ defines a uniform mapping between the x coordinates of image pixels and the interval $[\frac{\pi}{2}, -\frac{\pi}{2}]$. The parameter β adjusts the amplitude of this mapping. To enhance the variability in generated orientations, similar adjustments are made as in Vizcaya and Gerhardt's method, where arch center coordinates are used instead of loops and deltas. Sample orientation images generated using the zero-pole model [24] and the sine-based approach are illustrated in Figure 1.

Frequency images

Ridge frequency determines the density of ridges per unit length along the perpendicular direction on a fingerprint. It

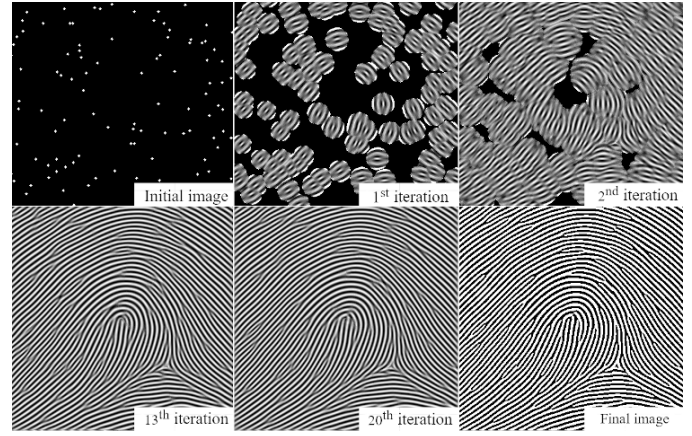


Fig. 3: A ridge pattern generated at different iterations and after thresholding

varies non-uniformly across the whole fingerprint area and is influenced by many factors such as distances to the existing singular points or the fingerprint borders. Additionally, these frequencies often exhibit smooth transitions between neighboring local regions.

For SFinGe, frequency images are generated according to some observations on real fingerprints, such as the decrease in frequency above the northernmost loop and below the southernmost delta [25]. To increase variations in the generated fingerprint images, a less constraining approach is followed in this study, and ridge frequencies are synthesized as coherent noise maps. Three random noise images are generated at different resolutions (2x2, 3x3, and 5x5) and then scaled to the same size (400x400) using bi-cubic interpolation. The final frequency image is obtained by adding them and normalizing the outcome (Figure 2).

Ridge Patterns

Using the orientation and frequency values of each pixel, ridge patterns are iteratively formed by convolving a random initial image with Gabor filters. To optimize computation, the orientation (θ) values are discretized into 20 bins in the interval $0 \leq \theta \leq \pi$, while the frequency (f) values are discretized 100 bins in the interval $0.11 \leq f \leq 0.17$. The response image is normalized after each iteration to suppress high responses and enhance low responses. After several iterations, the resulting patterns are binarized using mean thresholding (Figure 3).

To obtain minutiae labels for the generated fingerprints, they are detected on the response images prior to thresholding. In these images, ridges and valleys manifest as extremities, while minutiae points occupy intermediate values. Leveraging this characteristic, the response image is normalized to the range $[-1, 1]$ and a probability map P , indicating the likelihood of each pixel being a minutia, is computed using Eqn. 4. Subsequently, the probability map undergoes thresholding and median filtering. The resulting image is then skeletonized, and minutiae locations are detected using maximum suppression algorithm.

$$P(X) = 1 - |\tanh(X)| \quad (4)$$

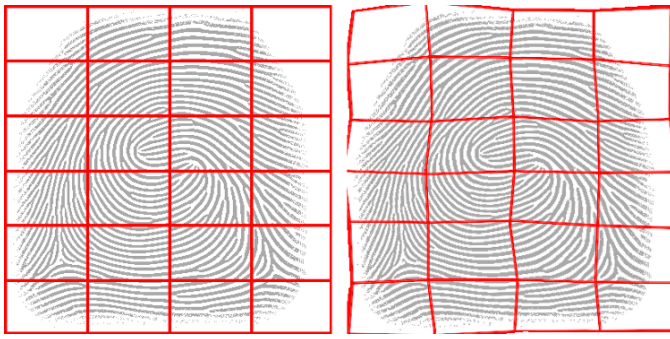


Fig. 4: Non-linear distortions simulated by Piecewise Affine Transformations

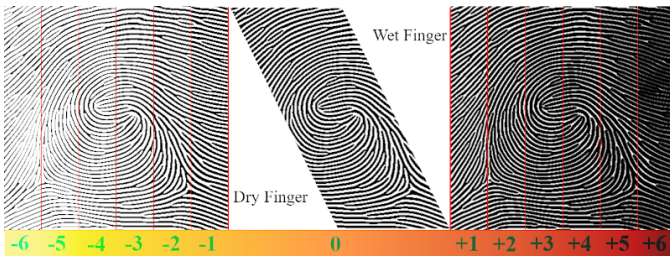


Fig. 5: Skin conditions simulated with various T

Creases

In real fingerprints, ridge lines can be interrupted by creases of varying lengths and thicknesses. These creases are modeled in this study as ellipses, where L represents the major-axis length and T denotes the minor-axis length. Each ellipse is applied to the master fingerprint image at randomly selected positions and orientations, following deformation through piece-wise affine transformations.

B. Impression Generation

Once a master fingerprint is obtained, multiple impressions can be generated by simulating various acquisition conditions. This study models impression variations introduced by six different real-world factors.

Contact area

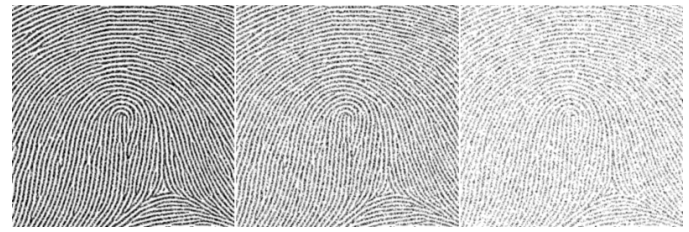
The contact area of a fingertip on a surface varies depending on the pressure applied by the finger and the angle of the finger relative to that surface. In this study, contact areas are modeled as ellipse-like shapes using the approach proposed in [25].

Pose

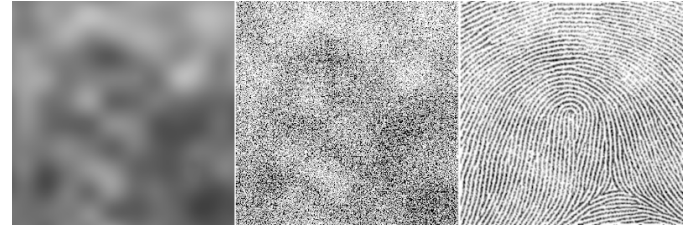
Pose changes are induced by rotation (R), translation (T_x, T_y) and scaling (S) operations. The R parameter is randomly sampled from a range of $[-12^\circ, 12^\circ]$, T_x and T_y parameters are randomly sampled from a range of $[-30, +30]$ pixels and the S parameter is randomly generated from a range of $[0.85, 1.15]$.

Non-linear distortions

Impressions from the same finger may vary due to non-linear distortions caused by different placement and pressures of the finger against the surface. To model these distortions, a



(a)



(b)

Fig. 6: Skin conditions simulated with respect to T (a) Impressions with uniform noise of probabilities of 0.25, 0.50 and 0.75 (b) Non-uniform probability map, ink density map, and the generated fingerprint

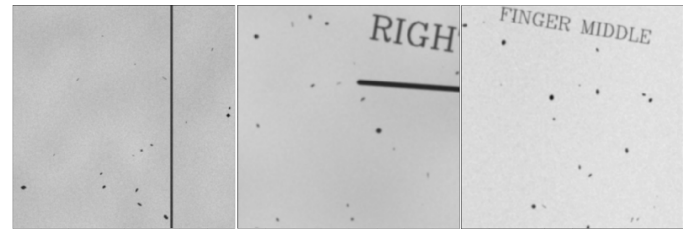


Fig. 7: Example generated backgrounds

grid is overlaid on the fingerprint and deformed by applying Piecewise Affine Transformations by M pixel, where M is the movement parameter and randomly sampled for each grid segment from the interval $[-7, +7]$ (Figure 4).

Skin conditions

Ridge thickness in the fingerprints varies based on skin conditions; they are thinner under dry conditions and thicker with moisture. These variations are simulated using morphological operations applied to the master fingerprints. The ridge thickness is controlled by a parameter T , which determines the number of erosion or dilation operations applied to simulate dry and wet fingertips, respectively. For each impression, T is uniformly sampled from the interval $[-4, +4]$, where negative values correspond to erosion and positive values to dilation. Before applying these operations, the master fingerprints are scaled up by a factor of 4 before these operations to introduce more variability in ridge thickness (Figure 5).

Noise

Noise causes ridges discontinuities and local blurs in fingerprint images. To simulate these effects, a probability map is generated to determine whether a ridge pixel in the master fingerprint should remain or be eliminated. For instance, a uniform probability over the image creates a homogeneous ridge disappearance (Figure 6a). However, real fingerprint

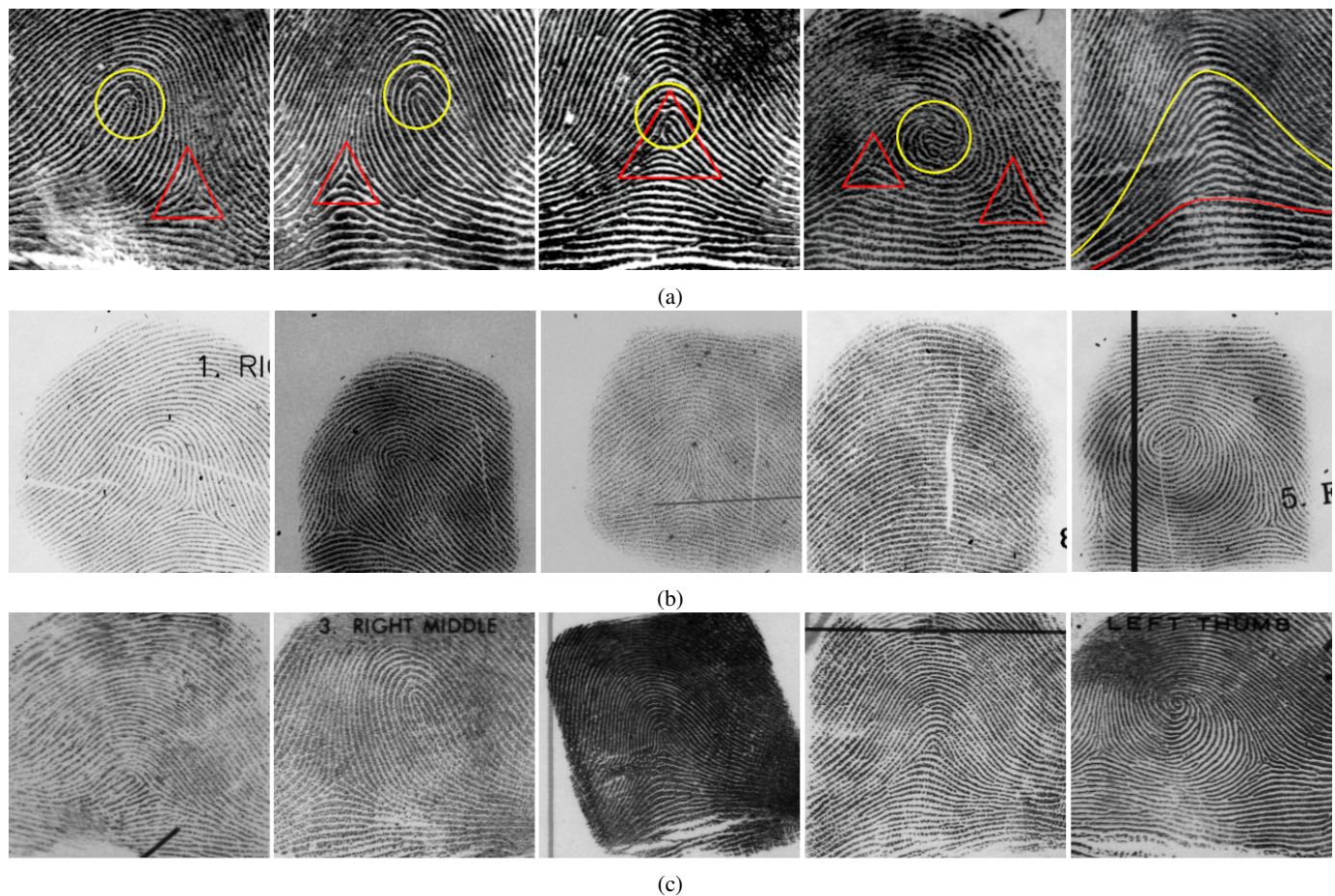


Fig. 8: (a) Fingerprint classes from NIST SD4 dataset [26]: left-loop, right-loop, tented arch, arch, and whorl classes. Loops are shown by yellow circles, and deltas are shown by red triangles. (b) Generated synthetic fingerprint examples for each class. (c) Real fingerprint examples from the NIST SD4 dataset for each class.

noise exhibits non-uniform behavior. To better reflect this, probability maps are generated similarly to coherent noise maps used for frequency images. These maps are scaled between 0 and 1 using min-max normalization (Figure 6b).

Background

Fingerprints are simulated to mimic imprinting on cards using ink and subsequent digitization by scanning, akin to the NIST SD4 dataset [26]. This dataset often includes fingerprint images with a paper background featuring additional horizontal or vertical lines, annotations, marks, and stains. To replicate this scenario, coherent noise is employed to generate paper-like backgrounds. Subsequently, lines, dots, marks, and annotations such as digits or class labels are randomly added to the background at various locations and scales (Figure 7).

IV. EXPERIMENTS AND RESULTS

Fingerprints are categorized into five primary patterns based on the ridge lines: left loop, right loop, tented arch, whorl, and arch. This categorization displays a crucial role in reducing search space and subsequent search time for fingerprint matching. By comparing the queried fingerprint only with those of

the same class, the accuracy of the fingerprint classification module significantly influences the overall performance of fingerprint recognition systems.

Traditional methods for fingerprint classification rely on the extraction of global (level 1) features such as ridge line flow and singular points, which are either of type core or delta, as shown in Figure 8a. Some notable studies include [27], [28], [29], [30]. More recently, deep learning algorithms, particularly convolutional neural networks (CNNs), have achieved high accuracy in fingerprint classification. These approaches use the fingerprint image directly as input and automatically learn relevant features for classification. The pioneering work by [31] introduced a stacked sparse autoencoder (SAE) neural network for learning a compact representation of fingerprint orientation fields, followed by subsequent studies [32], [33], [34], [35], [36], [37], [38]. However, the potential of these deep models is limited by the availability of large-scale, publicly accessible datasets for fingerprint classification.

For this study, synthetic fingerprint images that are known to belong one of the five classes are generated using the proposed method. An example synthetic fingerprint image for each class is given in Figure 8b.

The training efficacy of the generated dataset is evaluated

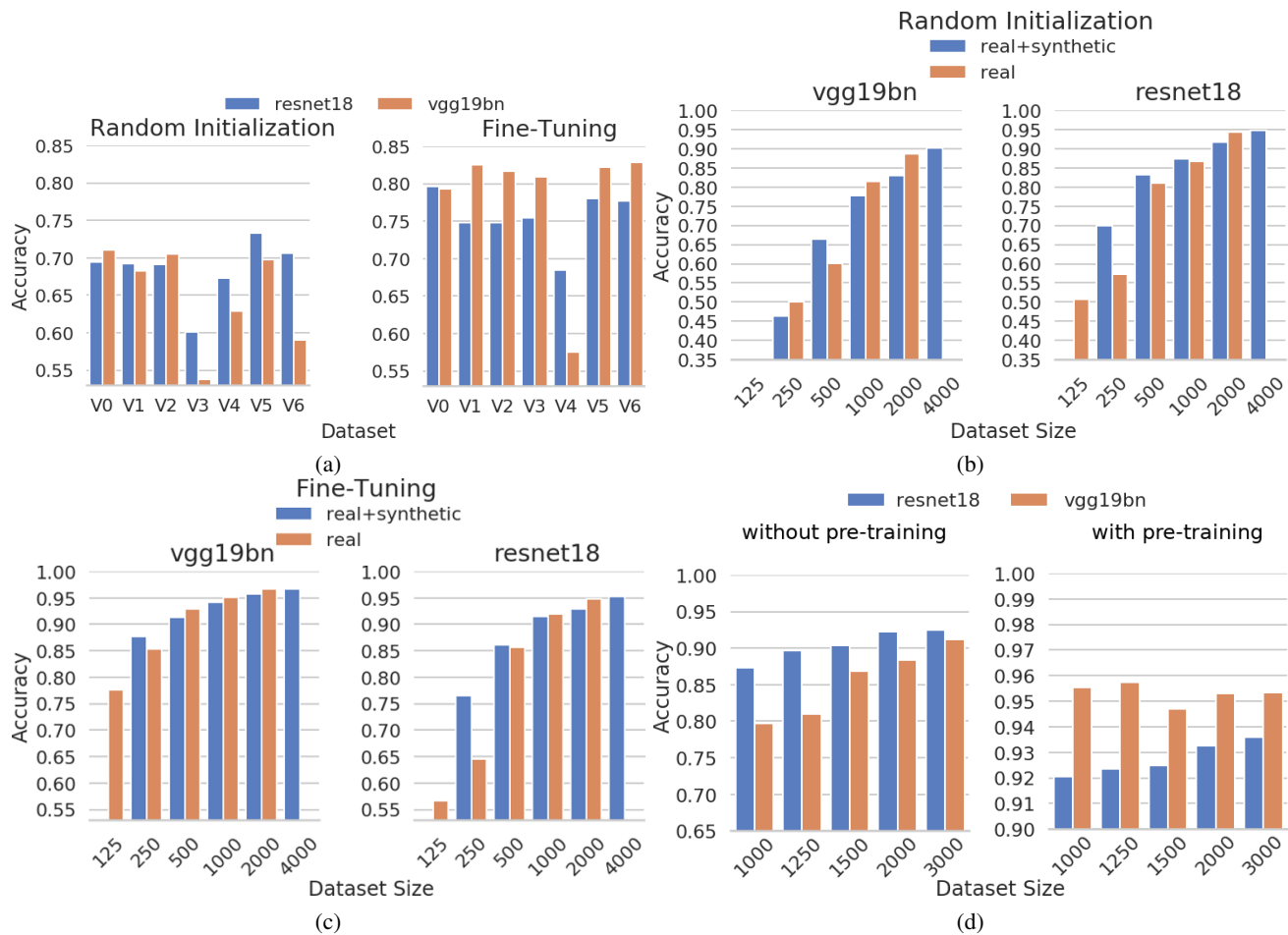


Fig. 9: (a) Classification performances only with synthetic training data (b) Classification performances in Group 1 experiments with VGG and (c) ResNet (d) Classification performances in Group 2 experiments

using the NIST SD4 dataset [26], which comprises 4000 grayscale fingerprint images sized 512x512, obtained from 2000 fingers. An example image for each fingerprint class are depicted in Figure 8c for comparison with the generated images.

The dataset is evenly divided into training and test sets with uniform distribution over five classes. The experiments are designed to achieve two main objectives: firstly, to determine whether synthetic fingerprints can effectively train fingerprint classification networks and achieve accurate results on real data, and secondly, to evaluate the impact of the proposed impression variations on performance. To this end, seven datasets are created that contain 8000 images with 1600 samples per class: Two with all variations and with (V0) and without post-processing (V6), five with one variation type excluded and with post-processing (V1-V5 for skin conditions, non-linear distortions, pose, noise, and background, respectively). In the initial experiments, only synthetic data is employed for training to evaluate the impact of impression variations on the results. A second set of experiments is conducted to assess the extent to which the synthetic fingerprints can enhance the performance when used as extra training data. All experiments are conducted using ResNet18 [39] and VGG19 [40], both with and without pre-training on the ImageNet [41].

A. Synthetic fingerprints as the only training data

This experiment evaluates classifier performance using exclusively synthetic data for training and analyzes the impact of post-processing and impression variations on accuracy, as shown in Figure 9a.

When no pre-training is used, pose variation (V3) and noise (V4) are observed to be rewarding in terms of performance. On the contrary, the exclusion of the background (V5) has increased the accuracy of ResNet, suggesting potential model confusion with background information rather than the fingerprint itself. VGG accuracy declines without post-processing (V6), likely due to differing intensity distributions between synthetic and real images, hindering generalization to real fingerprints.

With pre-training, performance notably improves across most cases. Noise (V4) remains to be the most influential on the classification performance. On the other hand, pose variations (V3) lose significance. This is possibly because pre-trained models are adept at handling rotation, scale, and translation variations. However, ImageNet does not include the kind of noise observed on fingerprint images, and removing it causes a crucial loss of information.

B. Synthetic fingerprints as extra training data

Two sets of experiments are conducted using the V0 dataset to assess classification performances when synthetic fingerprints are employed as additional training data. Firstly, increasing the size of the real training set is matched with its synthetic counterpart. Secondly, the number of real samples in the training set is fixed while increasing the number of synthetic samples added.

Group 1 experiments are conducted with mixed training sets that are composed of N synthetic and N real data. Their performance is compared with the real training set of size $2N$. Results for $N=125, 250, 500, 1000, 2000$ are given in Figure 9b and 9c. Notably, VGG without pre-training fails to converge with a training set of size 125 real samples.

With half of the training dataset consisting of synthetic, networks achieve comparable results to those trained solely on real images. More importantly, when the mixed dataset performances are compared with the real dataset performances with the same number of real images (N real+ N synthetic vs. N real), the results reveal that adding synthetic data increases the performances in nearly all cases. Naturally, this improvement converges as the real dataset size increases.

Pre-trained VGG fine-tuned with a mixed dataset of size 4000 achieves the highest accuracy with 95.30%, outperforming the same model fine-tuned with all of the 2000 images in NIST SD4 training set, which achieves 94.80%.

Group 2 experiments are conducted to observe the impact of incrementally adding synthetic training data of varying sizes on classification accuracies. To this end, 1000 real fingerprint images from NIST SD4 are used as the base training set, and it is gradually augmented with synthetic samples. Results with 0, 250, 500, 1000, and 2000 additional synthetic images are given in Figure 9d.

Significant performance improvements are observed without pre-training for both models. Similar trends are noted for pre-trained ResNet. However, less pronounced effects are observed on pre-trained VGG. The leading cause for this result can be explained by the fact that fingerprint classification is not a very complex task, and the pre-trained VGG is already accurate due to the knowledge transferred from object classification on ImageNet. As such, 1000 real fingerprint images suffice to effectively fine-tune the classifier.

V. CONCLUSION

Motivated by the scarcity of publicly available datasets in the fingerprint domain, this study endeavors to generate realistic synthetic fingerprints and assess their efficacy in training deep learning systems. A model-based approach is employed, involving two primary stages: the generation of master fingerprints and ensuing impressions. Initially, master fingerprints representing five distinct classes are synthesized, capturing the idealized ridge patterns specific to each class. Subsequently, these master fingerprints undergo simulation of real-world variations such as skin conditions, non-linear distortions, pose variations, noise, and diverse backgrounds.

The experimental evaluation focuses on the performance of synthetic datasets in fingerprint classification task using

two prominent deep neural network architectures, VGG and ResNet. Remarkably, classifiers trained exclusively on synthetic data achieve classification accuracies exceeding 80% on real test datasets. Furthermore, augmenting real training sets with synthetic samples consistently enhances classification performance, particularly beneficial when real dataset sizes are limited.

These findings underscore the potential of synthetic data generation techniques in addressing data scarcity challenges in fingerprint analysis. By effectively mimicking real-world variability, synthetic fingerprints not only facilitate robust training of deep learning models but also contribute to improving their generalization capabilities. Future research directions could explore refining synthetic fingerprint generation techniques to simulate additional real-world complexities and expanding evaluation across broader datasets and fingerprint analysis tasks.

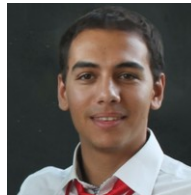
ACKNOWLEDGMENT

This research is funded by The Scientific and Technological Research Council of Turkey (TÜBİTAK) via Project No. 217E092 under the 2515 COST Support Program.

REFERENCES

- [1] "Nist special database catalog," www.nist.gov/srd/shop/special-database-catalog, accessed: 2021-02-05.
- [2] Y. Xu, Y. Wang, J. Liang, and Y. Jiang, "Augmentation data synthesis via gans: Boosting latent fingerprint reconstruction," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 2932–2936.
- [3] R. Cappelli, D. Maio, and D. Maltoni, "Sfinge: an approach to synthetic fingerprint generation," in *International Workshop on Biometric Technologies (BT2004)*, 2004, pp. 147–154.
- [4] A. H. Ansari, "Generation and storage of large synthetic fingerprint database," *ME Thesis*, Jul, 2011.
- [5] K. Cao and A. Jain, "Fingerprint synthesis: Evaluating fingerprint search at scale," in *2018 International Conference on Biometrics (ICB)*. IEEE, 2018, pp. 31–38.
- [6] V. Mistry, J. J. Engelsma, and A. K. Jain, "Fingerprint synthesis: Search with 100 million prints," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2019, pp. 1–10.
- [7] "Sfinge tool, biolab, university of bologna," biolab.csr.unibo.it/research.asp, accessed: 2021-02-12.
- [8] "Anguli, database systems lab, indian institute of science," dsl.cds.iisc.ac.in/projects/Anguli, accessed: 2021-02-12.
- [9] R. Cappelli, D. Maio, and D. Maltoni, "Synthetic fingerprint-database generation," in *Object recognition supported by user interaction for service robots*, vol. 3. IEEE, 2002, pp. 744–747.
- [10] B. G. Sherlock and D. M. Monro, "A model for interpreting fingerprint topology," *Pattern recognition*, vol. 26, no. 7, pp. 1047–1055, 1993.
- [11] I. Fogel and D. Sagi, "Gabor filters as texture discriminator," *Biological cybernetics*, vol. 61, no. 2, pp. 103–113, 1989.
- [12] P. Johnson, F. Hua, and S. Schuckers, "Texture modeling for synthetic fingerprint generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 154–159.
- [13] Y. Chen and A. K. Jain, "Beyond minutiae: A fingerprint individuality model with pattern, ridge and pore features," in *International Conference on Biometrics*. Springer, 2009, pp. 523–533.
- [14] Q. Zhao, A. K. Jain, N. G. Paulter, and M. Taylor, "Fingerprint image synthesis based on statistical feature models," in *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 2012, pp. 23–30.
- [15] K. G. Larkin and P. A. Fletcher, "A coherent framework for fingerprint analysis: are fingerprints holograms?" *Optics Express*, vol. 15, no. 14, pp. 8667–8677, 2007.

- [16] C. Imdahl, S. Huckemann, and C. Gottschlich, "Towards generating realistic synthetic fingerprint images," in *2015 9th International Symposium on Image and Signal Processing and Analysis (ISPA)*. IEEE, 2015, pp. 78–82.
- [17] C. Gottschlich and S. Huckemann, "Separating the real from the synthetic: minutiae histograms as fingerprints of fingerprints," *IET Biometrics*, vol. 3, no. 4, pp. 291–301, 2014.
- [18] P. Bontrager, A. Roy, J. Togelius, N. Memon, and A. Ross, "Deepmasterprints: Generating masterprints for dictionary attacks via latent variable evolution," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2018, pp. 1–9.
- [19] M. Attia, M. H. Attia, J. Iskander, K. Saleh, D. Nahavandi, A. Abobakr, M. Hossny, and S. Nahavandi, "Fingerprint synthesis via latent space representation," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 1855–1861.
- [20] K. Perlin, "An image synthesizer," *ACM Siggraph Computer Graphics*, vol. 19, no. 3, pp. 287–296, 1985.
- [21] R. Cappelli, D. Maio, and D. Maltoni, "An improved noise model for the generation of synthetic fingerprints," in *ICARCV 2004 8th Control, Automation, Robotics and Vision Conference, 2004.*, vol. 2. IEEE, 2004, pp. 1250–1255.
- [22] L. Pang, J. Chen, F. Guo, Z. Cao, E. Liu, and H. Zhao, "Rose: real one-stage effort to detect the fingerprint singular point based on multi-scale spatial attention," *Signal, Image and Video Processing*, vol. 16, no. 3, pp. 669–676, 2022.
- [23] J. Feng and A. K. Jain, "Fingerprint reconstruction: From minutiae to phase," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 209–223, 2011.
- [24] P. R. Vizcaya and L. A. Gerhardt, "A nonlinear orientation model for global description of fingerprints," *Pattern Recognition*, vol. 29, no. 7, pp. 1221–1231, 1996.
- [25] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, *Handbook of fingerprint recognition*. Springer Science & Business Media, 2009.
- [26] C. I. Watson and C. L. Wilson, "Nist special database 4," *Fingerprint Database, National Institute of Standards and Technology*, vol. 17, no. 77, p. 5, 1992.
- [27] K. Cao, L. Pang, J. Liang, and J. Tian, "Fingerprint classification by a hierarchical classifier," *Pattern Recognition*, vol. 46, no. 12, pp. 3186–3197, 2013.
- [28] H.-W. Jung and J.-H. Lee, "Noisy and incomplete fingerprint classification using local ridge distribution models," *Pattern recognition*, vol. 48, no. 2, pp. 473–484, 2015.
- [29] M. Liu, "Fingerprint classification based on adaboost learning from singularity features," *Pattern Recognition*, vol. 43, no. 3, pp. 1062–1070, 2010.
- [30] R. Cappelli, D. Maio, D. Maltoni, and L. Nanni, "A two-stage fingerprint classification system," in *Proceedings of the 2003 ACM SIGMM workshop on Biometrics methods and applications*, 2003, pp. 95–99.
- [31] R. Wang, C. Han, and T. Guo, "A novel fingerprint classification method based on deep learning," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 931–936.
- [32] J. M. Shrein, "Fingerprint classification using convolutional neural networks and ridge orientation images," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2017, pp. 1–8.
- [33] W.-S. Jeon and S.-Y. Rhee, "Fingerprint pattern classification using convolution neural network," *international journal of fuzzy logic and intelligent systems*, vol. 17, no. 3, pp. 170–176, 2017.
- [34] B. Pandya, G. Cosma, A. A. Alani, A. Taherkhani, V. Bharadi, and T. McGinnity, "Fingerprint classification using a deep convolutional neural network," in *2018 4th international conference on information management (ICIM)*. IEEE, 2018, pp. 86–91.
- [35] P. Tertychnyi, C. Ozcinar, and G. Anbarjafari, "Low-quality fingerprint classification using deep neural network," *IET Biometrics*, vol. 7, no. 6, pp. 550–556, 2018.
- [36] T. Zia, M. Ghafoor, S. A. Tariq, and I. A. Taj, "Robust fingerprint classification with bayesian convolutional networks," *IET Image Processing*, vol. 13, no. 8, pp. 1280–1288, 2019.
- [37] B. Rim, J. Kim, and M. Hong, "Fingerprint classification using deep learning approach," *Multimedia Tools and Applications*, vol. 80, pp. 35 809–35 825, 2021.
- [38] C. Militello, L. Rundo, S. Vitabile, and V. Conti, "Fingerprint classification based on deep learning approaches: experimental findings and comparisons," *Symmetry*, vol. 13, no. 5, p. 750, 2021.
- [39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.



Emre İrtem received his B.Sc. degree from the Department of Computer Engineering at the Izmir Institute of Technology in 2017. He subsequently completed his M.S. studies in the same department in 2020. Emre is currently working as a software architect at a private bank.



Nesli Erdoğan is an Assistant Professor in the Department of Computer Engineering at the Izmir Institute of Technology. She earned her B.S. and M.S. degrees from the Department of Electrical and Electronics Engineering at Middle East Technical University. Dr. Erdoğan completed her Ph.D. studies at EURECOM in Sophia-Antipolis, France, and graduated from Télécom ParisTech. Following her doctoral studies, she conducted a two-year post-doctoral fellowship at Idiap Research Institute in Switzerland. Since 2016, Dr. Erdoğan has been leading her research group within her department.

Performance Comparison of Deep Learning Models in Brain Tumor Classification

Emrah Aslan and Yildirim Ozupak


Abstract— Accurate and timely detection of brain tumors is critical for a successful treatment. Magnetic Resonance Imaging (MRI) is an essential tool that provides invaluable information for the recognition of different types of brain tumors such as glioma, meningioma, pituitary tumors and benign entities. However, distinguishing between these tumor types and taking preventive measures poses a significant challenge in the classification of brain tumors. Compared to traditional disease detection methods, artificial intelligence-based computer applications offer significant contributions to brain tumor detection. In particular, deep learning methods, which have gained popularity in disease detection through the analysis of medical images, play a critical role in this process. Several deep learning techniques have been reported in the literature for brain tumor classification. In this study, the YOLOv8s-cls model is used to detect brain tumors from MRI scans. The proposed model showed a high success rate of 98.7% accuracy during the experimental studies. The results show that the YOLOv8 model not only outperforms existing methods but also proves to be an effective approach for image classification.

Index Terms— Brain Tumor, Classification, Deep Learning, YOLOv8s-cls, Model Performance Comparison.


I. INTRODUCTION

IN DIAGNOSING brain tumors, experts often use various imaging modalities such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). However, differentiating tumors from other brain diseases using these imaging techniques is not an easy process and requires a subjective assessment depending on the expertise of the evaluator. Brain tumor formation produces different metabolites not seen in other brain diseases. Measuring these metabolites provides important information for the diagnosis and differential diagnosis of the disease. A tumor is a structure formed by the uncontrolled proliferation of abnormal cells that have different characteristics from normal cells. Glioma, meningioma, pituitary tumor and non-tumor conditions represent the four major tumor types [1].

Emrah Aslan, Faculty of Engineering and Architecture, Mardin Artuklu University, Mardin, Turkey (e-mail: emrahaslan@artuklu.edu.tr).

 <https://orcid.org/0000-0002-0181-3658>

Yildirim Özupak, Silvan Vocational School, Dicle University, Diyarbakır, Turkey, (e-mail: yildirim.ozupak@dicle.edu.tr).

 <https://orcid.org/0000-0001-8461-8702>

Manuscript received Jan 11, 2025; accepted Mar 11, 2025.

DOI: [10.17694/bajece.1617698](https://doi.org/10.17694/bajece.1617698)

Gliomas are tumors that can develop in areas of the brain nervous system such as the brain stem and spinal cord and can cause symptoms such as nausea, headache, vomiting and irritability. Meningiomas, which develop in the meninges, the membrane of the brain, are a more common type of tumor. Early detection of tumors is of great importance in determining the treatment method. Therefore, computerized image processing techniques for tumor detection are of great interest to researchers [2].

One of the critical tasks that technology must overcome today is the automatic detection of tumors at an early stage. Determining the size and spread of early detected tumors enhances the effectiveness of the treatment process. However, precisely estimating the size and resolution of tumors is challenging and often involves uncertainty. Early detection is directly related to treatment success and increases the likelihood of a full recovery.

Magnetic resonance imaging (MRI) is one of the most widely used methods for creating detailed images of the brain and detecting brain damage. MRI has superior performance compared to Computed Tomography (CT), especially in soft tissue assessments. Artificial intelligence (AI) and machine learning (ML) have made great advances in this field [3]. In recent years, significant progress has been made in medical image processing thanks to the ability of ML-based systems to operate without coding [4].

In this study, a deep learning model based on YOLOv8 is proposed to recognize and classify brain tumors from MRI images. The performance of the model is evaluated with a total of 3264 images obtained from the Kaggle dataset. The dataset consists of 394 test images and 2870 training images. The experimental results show that the proposed model outperforms other existing methods.

The YOLOv8 model used in this study is optimized to detect brain tumors quickly and effectively. The model has a high accuracy rate of 98.7%, making it a valuable tool for early detection, especially in clinical applications. With its real-time processing capacity, YOLOv8s-cls contributes to the acceleration of clinical processes by providing fast classification and detection results. In addition, the model's ability to process MRI images with deep learning techniques provides a significant advantage in categorizing brain tumors with high accuracy.

However, the study has some drawbacks and limitations. Since the performance of the model depends on the variety and quality of the dataset used, its overall performance under

different imaging conditions needs to be validated. Furthermore, the YOLOv8 model may be prone to errors during the detection of very small or low-contrast tumors [5]. The study was limited to only one dataset; therefore, additional research is needed to assess the generalizability of the model in different datasets [6]. These limitations suggest that further improvements are needed before the model is fully ready for clinical applications. This study presents a new deep learning-based method for early detection of brain tumors and makes important contributions to the existing literature.

II. RELATED WORKS

Brain tumor classification is of great importance for early diagnosis and accurate treatment planning. In recent years, deep learning methods have made remarkable advances in image-based medical diagnosis and have been widely used for the analysis of magnetic resonance imaging (MRI) data. Deep learning models such as EfficientNetB7, VGG19, MobileNetV2, InceptionResNetV2, ConvNeXtBase, NASNetLarge and YOLOv8 have achieved high success rates with different approaches in medical imaging. This literature study aims to evaluate the effectiveness of these models in brain tumor classification and in this context, it analyzes the performance of the models on metrics such as Accuracy, F1 Score, Recall and Precision and reveals their contributions to clinical applications.

Solanki et al. conducted a comprehensive literature review on magnetic resonance (MR) imaging for the detection of brain tumors and examined computer intelligence, statistical image processing and machine learning techniques. They also made significant contributions on tumor morphology, datasets and classification methods [7].

Ullah et al. proposed a new deep learning model, TumorDetNet, for the detection and classification of brain tumors. Using 48 convolutional layers, leaky ReLU, and dropout layers, the model detected brain tumors with high accuracy and successfully classified benign/malignant, meningioma, pituitary, and glioma tumors [8].

Rahman and Islam proposed a parallel deep convolutional convolutional neural network (PDCNN) topology to solve overfitting problems when classifying brain tumors with convolutional neural networks (CNN). The model achieved high accuracy (97.33%-98.12%) on three different MRI datasets using two different window sizes to learn local and global features [9].

Asiri et al. proposed an improved model based on CNN, ResNet50 and U-Net to accurately detect and classify brain tumors at an early stage. Using TCGA-LGG and TCIA datasets, this model accurately classified tumor and non-tumor images and successfully segmented tumor regions with U-Net. The results were remarkable with IoU: 0.91, DSC: 0.95 and SI: 0.95 accuracies [10].

Prakash et al. proposed an innovative and efficient hybrid Convolutional Neural Network (HCNN) classifier model for meningioma tumor detection. This method, which includes

Ridgelet transform, feature computation, classifier module and segmentation algorithm, achieved superior results with 99.31%, 99.35% and 99.81% accuracy rates on BRATS 2019, Nanfang and BRATS 2022 datasets, respectively [11].

Khan et al. propose an automated system using saliency map and deep learning feature optimization to detect and classify brain tumors. In the first stage, contrast enhancement is performed, followed by tumor segmentation based on saliency maps and fine-tuning of the EfficientNetB0 model. The accuracy rates obtained with deep transfer learning and feature integration are 95.14%, 94.89% and 95.94%, respectively [12].

Agarwal et al. aim to develop an automatic, robust and hybrid system for early detection and classification of brain tumors. The proposed system fine-tunes the Inception V3 model using Auto Contrast Enhancer, which improves low contrast in MRI images, and deep transfer learning for tumor detection and classification. The system showed superior performance with 98.89% accuracy compared to existing models [13].

Bhagyalaxmi et al. studied the effects of deep learning (DL) methods on magnetic resonance imaging (MRI) for early detection of brain tumors. This review aims to help radiologists improve their research and analysis processes by addressing the advances, current challenges, and future opportunities of DL-based approaches in the field of brain tumor classification and detection [14].

Turk et al. proposed an ensemble deep learning-based system utilizing ResNet50, VGG19, InceptionV3, and MobileNet architectures combined with Class Activation Maps (CAMs) for automatic brain tumor detection from MRI images. Their model achieved 100% accuracy in binary classification on ResNet50, InceptionV3, and MobileNet, while attaining 96.45% accuracy with ResNet50 in multi-class classification, demonstrating its effectiveness in tumor identification [15].

Vineela et al. discussed the use of various imaging techniques such as MRI, CT scans and PET scans in the brain tumor recognition process. The study explored the use of YOLOv8 architecture for accurate detection of tumors and the potential of radiogenomics technology, emphasizing the effectiveness of machine learning and deep learning algorithms [16].

Pacal et al. proposed an enhanced EfficientNetv2 architecture incorporating Global Attention Mechanism (GAM) and Efficient Channel Attention (ECA) to improve brain tumor classification from MRI scans. Their model achieved a remarkable test accuracy of 99.76%, demonstrating the effectiveness of attention mechanisms in enhancing feature extraction and interpretability for Computer-Aided Diagnosis (CADx) systems [17].

Elazab et al. used deep learning (DL) techniques for the classification and grading of gliomas, primary brain tumors arising from glial cells. Developing a hybrid model based on YOLOv5 and ResNet50, the authors accurately localized and graded tumors in histopathological images. Experiments revealed that the proposed model performs with high accuracy, precision and sensitivity and effectively distinguishes subtypes of gliomas [18].

III. MATERIAL AND METHOD

A. Brain Tumor MRI Data Set

In this study, we use a publicly available brain tumor MRI dataset from the Kaggle platform, which contains 3264 magnetic resonance images with four main classes (glioma, meningioma, pituitary tumor and benign lesions) [19]. The dataset was partitioned into 2870 training and 394 test images, and standard preprocessing steps were applied to ensure data consistency and quality before model training. In this context, pixel intensities were standardized by normalization, all images were rescaled to a uniform size suitable for model input.

To increase the generalization capacity of the model and reduce the risk of overlearning, data augmentation strategies that simulate variations in clinical imaging were adopted. These strategies included random rotation, horizontal/vertical translation, scaling, cropping and brightness/contrast modulation. Potential biases inherent in the dataset were systematically analyzed, emphasizing that despite the relative class balance, latent biases due to the original data collection process should be taken into account in interpreting the results.

In line with the clinical relevance of the study, model performance was specifically evaluated on the detection of advanced tumor lesions, and this focus was discussed in the context of the relationship between pathological progression and early diagnostic intervention. Sample images selected from the dataset are presented in Figure 1. The rigorous preprocessing and boosting protocols applied support the reproducibility and methodological robustness of the experimental findings, strengthening the clinical validity of the results.

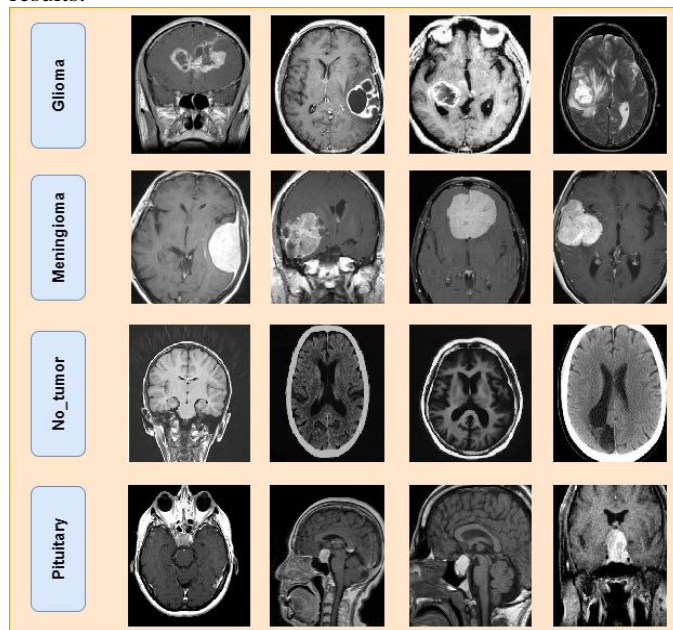


Fig.1. Example images in the dataset

B. YOLOv8

YOLOv8 is a state-of-the-art model for real-time object detection and image classification, offering enhanced accuracy and speed compared to previous YOLO versions. In this study, we leverage YOLOv8 to detect and classify brain tumors from MRI images by distinguishing among four primary tumor

types: glioma, meningioma, pituitary tumors, and benign conditions. Moreover, its adaptable structure facilitates efficient handling of diverse data types found in medical imaging. The structure of the YOLO model used is presented in Figure 2.

Although YOLOv8 is predominantly recognized for object detection and segmentation, this study employs its classification variant, YOLOv8s-cls, for brain tumor classification. The classification head of YOLOv8s-cls converts deep features extracted from MRI images into probability scores for the four tumor categories. During training, a composite loss function incorporating Binary Cross-Entropy (BCE) loss is utilized to enhance classification accuracy and, when needed, maintain spatial precision. The selection of YOLOv8s-cls was based on its balanced trade-off between computational efficiency and high classification performance, making it particularly well-suited for the complex challenges of medical image analysis in brain tumor detection.

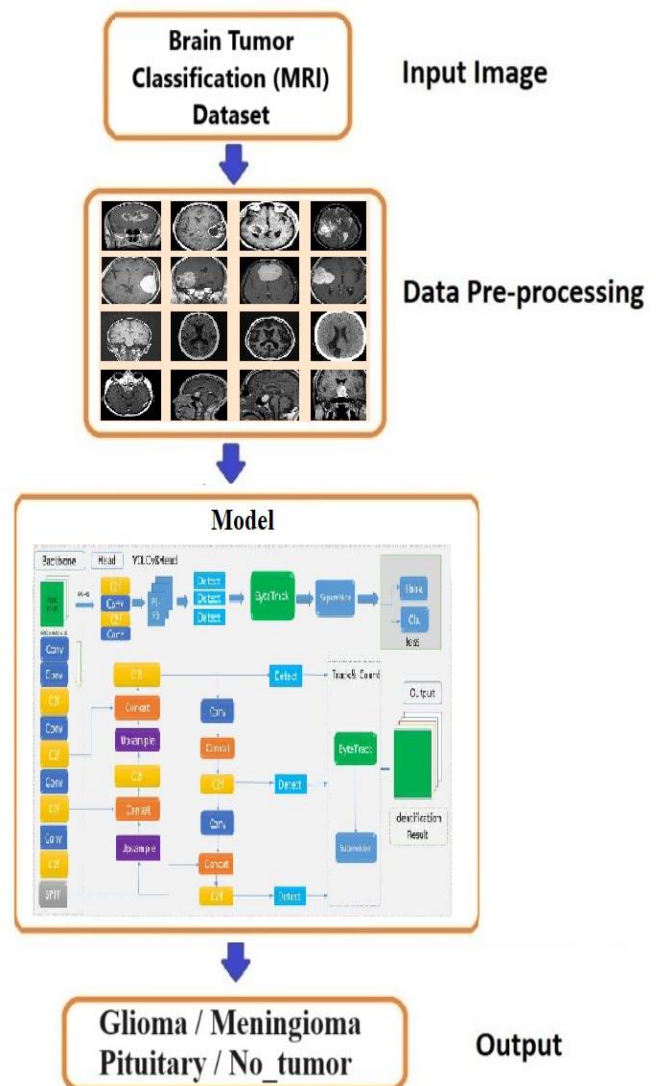


Fig. 2. Structure of the proposed YOLO-based MRI brain tumor detection model

C. Evaluation Metrics

In the proposed diagnostic method, multiple metrics are used to evaluate the performance of the model more comprehensively. While traditionally only a single metric such as accuracy is used, this method includes metrics such as accuracy, precision, recall, specificity and F1 score. RMSE and MAE were also calculated as model error metrics. The performance of the model was further quantified with ROC curves and AUC values [20, 21]. For each training model, a confusion matrix was created to calculate the evaluation metrics. This matrix provides the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values needed to calculate the different evaluation metrics. This multi-metric approach more reliably demonstrates the effectiveness of the model in real-world applications. The validation metrics are given in Equations 1-4.

$$ACC = \frac{TP+TN}{(TP+TN+FP+FN)} \quad (1)$$

$$Recall = \frac{TN}{(TP+FN)} \quad (2)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (3)$$

$$F_1 = 2 * \frac{Precision*Recall}{(Precision+Recall)} \quad (4)$$

IV. EXPERIMENTAL RESULTS

The YOLOv8s-cls model is trained using a composite loss function that integrates Binary Cross-Entropy (BCE) loss to improve classification accuracy and CIoU loss to improve localization accuracy. The training process was performed with the Adam optimizer with an initial learning rate of 0.001 and a weight decay of 1×10^{-4} . The model was trained for a total of 25 epochs using a batch size of 32. Furthermore, an early stopping mechanism was activated if the verification loss did not improve over five consecutive periods. This systematic hyperparameter tuning process allowed the results show that the the selection of the configuration that provided the highest validation success, resulting in a training accuracy of 99% and a validation accuracy of 98.7%. This detailed training procedure guarantees the reproducibility of our experiments and demonstrates the robust performance of the YOLOv8s-cls model in the context of brain tumor classification.

In this study, we compare the performance of various deep learning models for brain tumor classification. The models used include EfficientNetB7, VGG19, MobileNetV2, InceptionResNetV2, ConvNeXtBase, NASNetLarge and the proposed YOLOv8. Each model was evaluated with key performance metrics such as F1 Score, Recall, Precision and Accuracy. These analyses allowed us to better understand the accuracy and effectiveness of each model in brain tumor detection. In addition, we have also focused on how these models perform with different deep learning architectures and training strategies and how these results can contribute to clinical applications. In particular, the high accuracy and efficient classification capacity of YOLOv8 led to an important finding by performing best in brain tumor classification. In Table 1, the results obtained according to performance metrics such as F1 Score, Recall, Precision and Accuracy of the models used are presented in detail.

TABLE 1
COMPARISON OF MODEL RESULTS

Model	F1 Skor	Recall	Precision	Accuracy (ACC)
EfficientNetB7	0.943	0.944	0.941	94.1%
VGG19	0.928	0.930	0.925	92.5%
MobileNetV2	0.935	0.938	0.930	93.0%
InceptionResNetV2	0.951	0.953	0.948	94.8%
ConvNeXtBase	0.944	0.946	0.942	94.2%
NASNetLarge	0.957	0.960	0.953	95.3%
YOLOv8	0.986	0.988	0.985	98.7%

This study aims to compare the performance of different deep learning models for brain tumor classification. The results obtained reflect the performance of each model on important metrics such as F1 Score, Recall, Precision and Accuracy (ACC). The EfficientNetB7 model showed a balanced performance with high values of F1 Score (0.943), Recall (0.944), Precision (0.941) and Accuracy (94.1%). This shows that the model achieves a balance between accuracy and precision and is effective in brain tumor classification. VGG19, on the other hand, has a slightly lower performance, with F1 Score (0.928), Recall (0.930), Precision (0.925) and Accuracy (92.5%), and although it made some errors in classification, it still stands out as a valid model. MobileNetV2 achieved better results than VGG19 with F1 Score (0.935), Recall (0.938) and Precision (0.930), but its accuracy rate (93.0%) fell behind the other models. InceptionResNetV2 was one of the highest performing models with F1 Score (0.951), Recall (0.953), Precision (0.948) and Accuracy (94.8%), indicating that the model has a high capacity for accurate classification. Finally, the ConvNeXtBase model achieved strong results such as F1 Score (0.944), Recall (0.946) and Precision (0.942), but lagged behind the other models in terms of accuracy (93.4%). The standout model is YOLOv8, which stands out with the highest Accuracy (98.7%) for brain tumor classification. With high F1 Score (0.973), Recall (0.975) and Precision (0.970), YOLOv8 offers the best performance in brain tumor detection, making it a suitable model for real-world scenarios in clinical applications. These findings show that YOLOv8 is superior to other models with its high accuracy rate and effective classification capability.

YOLOv8 demonstrates superior performance compared to other deep learning models, thanks to its advanced object detection capabilities and deep feature extraction capacity. The model is particularly effective in detecting small tumors, incorporating enhanced anchor mechanisms and deepened convolutional layers. YOLOv8 uses predefined bounding boxes to perform both localization and classification tasks simultaneously, providing a critical advantage in detecting low-contrast or small-sized tumors in MRI images.

The improved CNN layers in the model's architecture enable more precise analysis of complex tissues. The composite loss function used during training aims to enhance classification accuracy while maintaining spatial precision. This has allowed for clearer distinctions between similar classes, such as 'meningioma_tumor' and 'glioma_tumor'. The model's high accuracy rate is not only attributable to its architectural design

but also to the effective use of data augmentation and optimization techniques. These factors establish YOLOv8 as a robust alternative for clinical applications.

The dot plot in Figure 3 shows the accuracy of various models used for brain tumor classification. The graph contains dots representing the accuracy rates of each model. YOLOv8 shows the highest performance with an accuracy rate of 0.987, while the accuracy rates of the other models range from EfficientNetB7, VGG19, MobileNetV2, InceptionResNetV2, ConvNeXtBase and NASNetLarge. This visualization makes the comparative performance of the models more understandable.

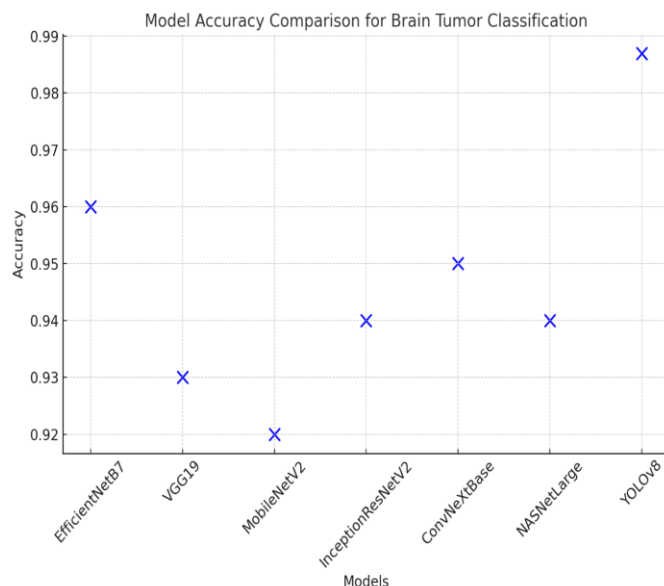


Fig. 3. Point plots of ACC of the models

The deep learning model developed for brain tumor classification performed successfully with a high accuracy rate. While the training accuracy of the model reached 99%, the validation accuracy was 98.7%. These results show that the model learns effectively on the training data and has a strong generalization capability on the validation data other than the training data. The close relationship between the training and validation accuracies indicates that the model is not affected by the overfitting problem and exhibits a balanced performance. Another noteworthy point in Figure 4 is the fluctuations in the verification accuracy in the early stages of the training process, the accuracy increases steadily as the process progresses, reaching its best performance at epoch 22. This shows that the optimization techniques and hyperparameters used were successfully selected and the training process progressed in a stable manner. This high accuracy rate of the model supports that deep learning methods offer a promising solution for brain tumor diagnosis and can be used in clinical applications. The validation accuracy curve is given in Figure 4.

Figure 5 shows the change of losses in the training and validation processes of the model according to the epochs. In the first epochs, especially the training loss shows a rather high initial value (350). However, the model adapts to the training process by rapidly reducing the losses within a few epochs. The validation loss decreased in parallel with the training loss and

stabilized around epoch 22. At this point, it can be seen that the validation performance of the model reaches its best level at epoch 22, which is marked as the “best epoch”. The closeness between training and validation loss indicates that the model is not overfitting and has a high generalization capacity.

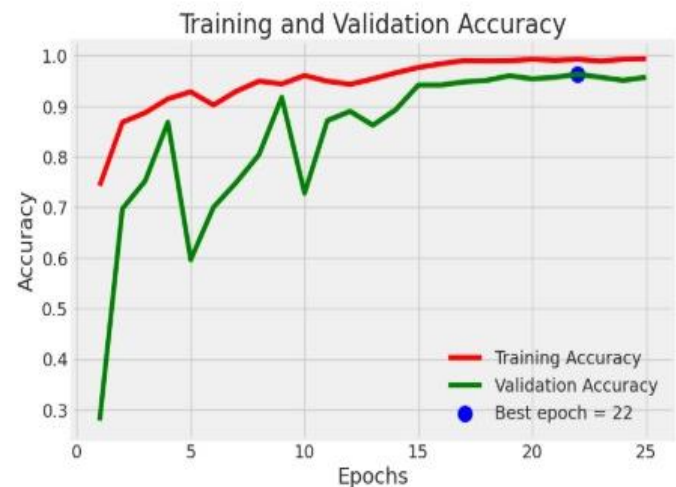


Fig. 4. ACC curve

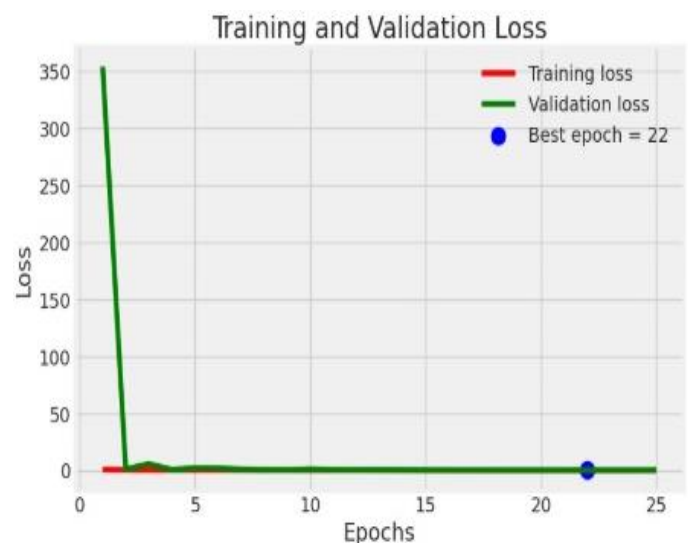


Fig. 5. Loss curve of the model

Figure 6 shows the confusion matrix of a classification model. The matrix evaluates the performance of the model against the actual and predicted labels for four different classes (no_tumor, pituitary_tumor, meningioma_tumor, glioma_tumor). The model correctly classified all instances in the “no_tumor” and “pituitary_tumor” classes (51 and 85 correct predictions respectively). In the “meningioma_tumor” class, there were 98 correct predictions, while 3 instances were misclassified as “glioma_tumor”. Similarly, there were 88 correct predictions in the “glioma_tumor” class, while one sample was mislabeled as “meningioma_tumor”. These results show that the model performs well overall, but there is some confusion between the “meningioma_tumor” and “glioma_tumor” classes. This may be due to the similarity of the features of these two classes in the dataset. For improvement, feature engineering or data augmentation techniques could be applied to improve the separation between these classes.

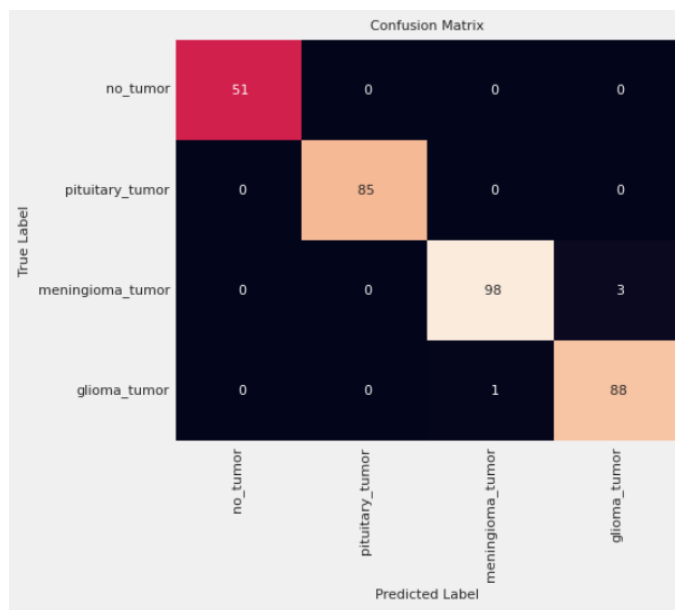


Fig. 6. Confusion matrix

This study successfully demonstrates the performance of the deep learning model developed for brain tumor classification. The training accuracy of the model is 99% and the validation accuracy is 98.7%, indicating that the model has high learning capacity and generalization ability. The fact that the training and validation accuracies are close to each other indicates that the model successfully avoids the overfitting problem.

The performance of the model increased steadily throughout the training process. Although fluctuations in the validation accuracy were observed at the beginning of the training process, the model reached its best performance at epoch 22 thanks to the optimization techniques used. This proves that the training process is stable and the chosen hyperparameters are appropriate. Moreover, the training and validation losses are parallel to each other and at low levels, indicating that the model does not experience any imbalance in the learning process and has a high generalization capacity.

The classification performance of the model is also generally successful. When the confusion matrix is analyzed, it is seen that all of the examples in the “no_tumor” and “pituitary_tumor” classes are classified correctly. However, some confusion was noticed between the “meningioma_tumor” and “glioma_tumor” classes. This may be due to the fact that the features of these classes are close to each other in the dataset. To reduce such confusion, it may be useful to apply methods such as feature engineering or data augmentation techniques.

The YOLOv8-based model proposed in this study, while achieving a high accuracy rate, does possess certain limitations. The model's performance is contingent on the diversity of the dataset used, and its direct generalizability across data obtained from different MRI scanners or healthcare institutions cannot be guaranteed. Future research should comprehensively evaluate the model by testing it on datasets acquired from various imaging systems. Furthermore, although early stopping and regularization methods were employed during the training process to prevent overfitting, additional validation studies are necessary to ascertain whether the model will exhibit similar

success in real-world scenarios. Another significant limitation is the potential class imbalance within the dataset. Whether the samples for each tumor type are balanced can directly impact performance. Consideration of these limitations will facilitate a better understanding of the model's potential for use in clinical settings.

V. CONCLUSION

This study presents a YOLOv8s-cls-based deep learning model for brain tumor classification, achieving 99% training accuracy and 98.7% validation accuracy. The model effectively distinguishes glioma, meningioma, pituitary tumors, and non-tumor cases, demonstrating superior performance compared to existing methods. Beyond its high accuracy, YOLOv8s-cls offers real-time processing, making it suitable for integration into radiology workflows. However, clinical validation on diverse MRI datasets is necessary to ensure its generalizability across different imaging conditions. Future research should focus on optimizing model robustness, improving small and low-contrast tumor detection, and evaluating real-world deployment. The findings indicate that YOLOv8s-cls has the potential to enhance early brain tumor detection and assist radiologists in clinical decision-making. Further validation and adaptation will be crucial for its successful implementation in medical practice.

REFERENCES

- [1] S. Hossain, A. Chakrabarty, T. R. Gadekallu, M. Alazab, and M. J. Piran, “Vision Transformers, Ensemble Model, and Transfer Learning Leveraging Explainable AI for Brain Tumor Detection and Classification,” *IEEE J Biomed Health Inform*, vol. 28, no. 3, pp. 1261–1272, Mar. 2024, doi: 10.1109/JBHI.2023.3266614.
- [2] P. Kanchanamala, K. G. Revathi, and M. B. J. Ananth, “Optimization-enabled hybrid deep learning for brain tumor detection and classification from MRI,” *Biomed Signal Process Control*, vol. 84, p. 104955, Jul. 2023, doi: 10.1016/J.BSPC.2023.104955.
- [3] M. S. Mithun and S. Joseph Jawhar, “Detection and classification on MRI images of brain tumor using YOLO NAS deep learning model,” *J Radiat Res Appl Sci*, vol. 17, no. 4, p. 101113, Dec. 2024, doi: 10.1016/J.JRRAS.2024.101113.
- [4] N. F. Alhussainan, B. Ben Youssef, and M. M. Ben Ismail, “A Deep Learning Approach for Brain Tumor Firmness Detection Based on Five Different YOLO Versions: YOLOv3–YOLOv7,” *Computation* 2024, Vol. 12, Page 44, vol. 12, no. 3, p. 44, Mar. 2024, doi: 10.3390/COMPUTATION12030044.
- [5] F. Tasnim, M. T. Islam, A. T. Maisha, I. Sultana, T. Akter, and M. T. Islam, “Comparison of Brain Tumor Detection Techniques by Using Different Machine Learning YOLO Algorithms,” *Lecture Notes in Networks and Systems*, vol. 869 LNNS, pp. 51–65, 2024, doi: 10.1007/978-981-99-9040-5_4.
- [6] M. F. Almufareh, M. Imran, A. Khan, M. Humayun, and M. Asim, “Automated Brain Tumor Segmentation and Classification in MRI Using YOLO-Based Deep Learning,” *IEEE Access*, vol. 12, pp. 16189–16207, 2024, doi: 10.1109/ACCESS.2024.3359418.
- [7] S. Solanki, U. P. Singh, S. S. Chouhan, and S. Jain, “Brain Tumor Detection and Classification Using Intelligence Techniques: An Overview,” *IEEE Access*, vol. 11, pp. 12870–12886, 2023, doi: 10.1109/ACCESS.2023.3242666.
- [8] N. Ullah, A. Javed, A. Alhazmi, S. M. Hasnain, A. Tahir, and R. Ashraf, “TumorDetNet: A unified deep learning model for brain tumor detection and classification,” *PLoS One*, vol. 18, no. 9, p. e0291200, Sep. 2023, doi: 10.1371/JOURNAL.PONE.0291200.
- [9] T. Rahman and M. S. Islam, “MRI brain tumor detection and classification using parallel deep convolutional neural networks,”

Measurement: Sensors, vol. 26, p. 100694, Apr. 2023, doi: 10.1016/J.MEASEN.2023.100694.

- [10] [A. A. Asiri et al., "Brain Tumor Detection and Classification Using Fine-Tuned CNN with ResNet50 and U-Net Model: A Study on TCGA-LGG and TCIA Dataset for MRI Applications," *Life* 2023, Vol. 13, Page 1449, vol. 13, no. 7, p. 1449, Jun. 2023, doi: 10.3390/LIFE13071449.
- [11] B. V. Prakash et al., "Meningioma brain tumor detection and classification using hybrid CNN method and RIDGELET transform," *Scientific Reports* 2023 13:1, vol. 13, no. 1, pp. 1–13, Sep. 2023, doi: 10.1038/s41598-023-41576-6.
- [12] M. A. Khan et al., "Multimodal brain tumor detection and classification using deep saliency map and improved dragonfly optimization algorithm," *Int J Imaging Syst Technol*, vol. 33, no. 2, pp. 572–587, Mar. 2023, doi: 10.1002/IMA.22831.
- [13] M. Agarwal, G. Rani, A. Kumar, P. K. K. R. Manikandan, and A. H. Gandomi, "Deep learning for enhanced brain Tumor Detection and classification," *Results in Engineering*, vol. 22, p. 102117, Jun. 2024, doi: 10.1016/J.RINENG.2024.102117.
- [14] K. Bhagyalaxmi, B. Dwarakanath, and P. V. P. Reddy, "Deep learning for multi-grade brain tumor detection and classification: a prospective survey," *Multimed Tools Appl*, vol. 83, no. 25, pp. 65889–65911, Jul. 2024, doi: 10.1007/S11042-024-18129-8/TABLES/6.
- [15] O. Turk, D. Ozhan, E. Acar, T. C. Akinci, and M. Yilmaz, "Automatic detection of brain tumors with the aid of ensemble deep learning architectures and class activation map indicators by employing magnetic resonance images," *Zeitschrift für Medizinische Physik*, vol. 34, no. 2, pp. 278–290, 2024, doi: 10.1016/j.zemedi.2022.11.010.
- [16] G. Vineela, G. H. Vardhan, C. Kesava Rao, T. Geetamma, and D. Dnivas Rao, "Deep Learning Technique to detect Brain tumor disease using YOLO v8," *Proceedings of the 2nd IEEE International Conference on Networking and Communications 2024, ICNWC 2024*, 2024, doi: 10.1109/ICNWC60771.2024.10537552.
- [17] I. Pacal, O. Celik, B. Bayram, et al., "Enhancing EfficientNetv2 with global and efficient channel attention mechanisms for accurate MRI-based brain tumor classification," *Cluster Comput*, vol. 27, pp. 11187–11212, 2024, doi: 10.1007/s10586-024-04532-1.
- [18] N. Elazab, W. A. Gab-Allah, and M. Elmogy, "A multi-class brain tumor grading system based on histopathological images using a hybrid YOLO and RESNET networks," *Scientific Reports* 2024 14:1, vol. 14, no. 1, pp. 1–20, Feb. 2024, doi: 10.1038/s41598-024-54864-6.
- [19] "Brain Tumor Classification (MRI)." Accessed: Dec. 19, 2024. [Online]. Available: <https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri/data>.
- [20] F. Alpsalaz and M. S. Mamiş, "Detection of Arc Faults in Transformer Windings via Transient Signal Analysis," *Appl. Sci.*, vol. 14, no. 20, p. 9335, 2024, doi: 10.3390/app14209335.
- [21] I. Pacal, "A novel Swin transformer approach utilizing residual multi-layer perceptron for diagnosing brain tumors in MRI images," *Int. J. Mach. Learn. & Cyber.*, vol. 15, pp. 3579–3597, 2024, doi: 10.1007/s13042-024-02110-w.



Yıldırım Özüpak received the bachelor's, M.Sc., and Ph.D. degrees in electrical and electronics engineering from Inonu University. His research interests include wireless energy transfer, design of electric motors, electromagnetic and thermal analysis of electric machines, machine learning, and deep learning.

BIOGRAPHIES



Emrah Aslan received the bachelor's degree in computer engineering and in electrical and electronics engineering and the master's degree in electrical and electronics engineering from Harran University, Türkiye, in 2013, 2019, and 2016, respectively, and the Ph.D. degree in electrical and electronics engineering from Dicle University, in 2023. He is currently an Assistant Professor with Mardin Artuklu University.


A Bibliometric Analysis on Cybersecurity Using VOSviewer: An Evaluation for Public Security

Vedat Yilmaz

Abstract— This bibliometric study conducts a comprehensive analysis of the field of cybersecurity, particularly in the context of law enforcement and security strategies, to examine key trends, author influence, and interdisciplinary connections within the literature. In the WoS database, 6606 articles were reached by using the search expression in the title, abstract and keywords fields ("Cyber security" or "cyber-attacks" or "cyber protection" or "spam cyber security" or "data security" or "network security" or "anomaly detection" or "cyber countermeasures") and restricting the year of publication to after 2000. The analysis includes metrics such as keyword co-occurrence frequency, author citation impact, co-authorship networks, bibliographic coupling of documents, co-citation analysis of authors, and institutional bibliographic connections. This study highlights the relationship between cybersecurity, law enforcement, and public safety, assessing the role of methodologies and technologies in mitigating security threats and reducing their impacts. When the keywords in the articles obtained as a result of the keyword analysis were examined, it was seen that the words "anomaly detection", "cybersecurity", and "deep learning" were the most frequently used keywords. It is noteworthy that the word "deep learning" was not included in the words generated when determining the articles, but it was used as a keyword in the articles obtained as a result of the determined keywords. Author citation analysis revealed influential contributors such as Quin Du, Wei Li, and Liangpei Zhang. Country-level analysis shows that China and the United States are leading in the field of research output, and institutional analysis highlights the prominent role of the Chinese Academy of Sciences. In conclusion, this research provides valuable insights into how law enforcement and security strategies intersect with academic studies in cybersecurity, offering a roadmap for future research.

Index Terms— Cyber Security, Anomaly Detection, Cyber Attacks, Security Technologies, Law Enforcement

Vedat YILMAZ, is with Department of Institute of Forensic Sciences, Gendarmerie and Coast Guard Academy, Ankara, Türkiye, (e-mail: vedat.yilmaz@jsga.edu.tr).

 <https://orcid.org/0000-0002-3112-9371>

Manuscript received Dec 27, 2024; accepted Feb 07, 2025.

DOI: [10.17694/bajece.1608364](https://doi.org/10.17694/bajece.1608364)

I. INTRODUCTION

CYBERSECURITY HAS become increasingly important in the technology-driven world that permeates every aspect of our lives. In this context, security policies and various technical and strategic approaches have been developed based on the three pillars of cybersecurity: confidentiality, integrity, and availability. Cybersecurity, especially in areas such as anomaly detection, network security, and system monitoring, plays a critical and vital role in the early detection of potential threats and the prevention of cyber-attacks. However, the rapidly developing and expanding literature in this field reveals that new cyber-attack methods should be evaluated as hostile actions that include national security as well as individual security, in parallel with technological developments. For this reason, the concept of Cybersecurity is widely researched in interdisciplinary areas, and its importance against new threats that we can describe as hostile actions is increasing day by day. These challenges include tracking and understanding developments, staying informed about current cybersecurity trends, and monitoring necessary precautions. In addressing these challenges, bibliometric analyses provide an effective solution by systematically examining key concepts in the literature, especially the commonalities among authors and scientific research trends [1, 2]. Visualization tools such as VOSviewer enhance the accessibility of these analyses, enabling a comprehensive examination of scientific productivity and interactions in the fields of cybersecurity and anomaly detection. VOSviewer is a data analysis software that makes it possible to provide bibliometric analysis, visualization, and scientific maps used by researchers to analyze and determine the content of Academic citations [3-5]. In the 21st century, characterized by the accelerated pace of digital transformation, the daily activities of both individuals and institutions are increasingly becoming digitized. By 2030, numerous technologies, such as 6G, the decentralized Internet, and the Internet of Senses, are expected to become integral to our lives [6, 7]. Digital systems are employed in nearly every domain, from law enforcement and national security to financial transactions and authentication processes, from digital identities and healthcare services to e-government systems and personal communication tools. While the use of these systems offers significant advantages for both individuals and organizations, it has also highlighted the critical importance of data security, particularly for data stored and transferred in digital environments. Consequently, these developments have

prompted individuals and government institutions to take substantial precautions against potential threats to these systems.

Given these considerations, cybersecurity has not only become an essential element at both individual and institutional levels but has also emerged as a top priority for national security [8]. Cybersecurity, in its simplest terms, refers to the methods and technologies employed to ensure the confidentiality, integrity, and availability of data, often referred to as digital assets. However, today's rapid development of technology and the hostile actions that emerge in cyberspace in parallel with this development require constantly taking new measures for the fight to be carried out in cyberspace. New areas of work have emerged to protect the security of data stored on many different platforms for data and network security [8,9].

Recent developments have transformed the concept of cyber security from a purely defensive field to a multi-disciplinary structure that tries to understand the new methods used by attackers to enter systems [10,11]. This new structure has led to the development of new strategies and preventive measures [11].

Cyber threats are generally the compromise of the confidentiality, integrity, or availability of a person's or an organization's data, computer system, network, or device by attempting to gain unauthorized access or exploit any existing security vulnerabilities in the information sections. [12, 13].

Cyber threats: It covers all kinds of malicious activities targeting the digital assets of individuals, public institutions and organizations, and private sector companies. These malicious activities often include data theft, Distributed Denial of Service (DDoS) attacks aimed at partially or completely disrupting the operation of the system, or financial fraud exposure of individuals or institutions, which is considered cybercrime [14]. Today, not only has the number of these attacks increased, but their complexity has also escalated [15].

The increasing complexity of cyber-attacks means not only an increase in the number of attacks, but also the use of more sophisticated techniques, methods, and tools. This complexity can manifest itself in the following areas:

- Multi-Layered Attacks: A combination of multiple attack types (e.g., DDoS, phishing, and malware) rather than a single attack [16].
- Advanced Persistent Threats (APT): Infiltrating the target for extended periods, gathering information, discovering vulnerabilities, and using them to harm the target organization [17].
- Use of Machine Learning and AI: Attackers can use artificial intelligence and machine learning algorithms to analyze their targets more efficiently or bypass security systems [18].
- Knock-on Effects: Targeting a single vulnerability, such as supply chain attacks, causes a broader threat with knock-on effects [19].
- Encryption and Privacy: Malware, especially ransomware, is becoming harder to detect with stronger encryption techniques [20].
- Dynamic and Adaptive Attacks: Attacks change and adapt instantly according to the target's security measures [21].

The examples given can be increased. This complex situation forces cyber security experts to develop smarter and more effective security measures against constantly evolving threats. The evolution of cyberattacks into this more complex structure has popularized the use of anomaly detection systems, which have become a secure method for identifying deviations from the normal flow of network traffic or user behaviors, offering a solution for early prevention of potential attacks [22].

Another aspect of cyber threats is cybersecurity against spam. Spam content, sent via email and other communication tools, often intrigues users and is typically used for phishing attacks, the propagation of malicious software across networks, or the acquisition of financial information for fraudulent purposes. To effectively mitigate these spam threats, in addition to raising user awareness, technologies such as artificial intelligence are being utilized [23].

Critical communication systems and infrastructures, including GSM networks, electricity, water, natural gas, transportation systems, dams, e-commerce, banking systems, and digital government applications, have the potential to be partially or completely disabled, which could disrupt social order and jeopardize national security [24].

Therefore, public security, as part of national security, requires a more comprehensive approach supported by cybersecurity measures. Today, the necessary measures against cyber threats that emerge from the individual to the state should be taken, especially by institutions responsible for public security. In particular, public order and the security of citizens can be directly affected as a result of cyber-attacks that may target critical infrastructures [8]. In this context, the national cybersecurity policies put forward by the Digital Transformation Office in Turkey set forth the criteria that must be followed by all institutions [24]. Similarly, the United States National Cybersecurity Strategy document recommends the use of threat analysis tools for critical infrastructures [25]. Although state elements take the necessary measures, it should not be forgotten that the weakest link is individuals. For this reason, individual information at the national level is an important element for both national security and preventing individuals from being exposed to cybercrime. Awareness training should be provided for citizens, such as the cyber awareness campaign organized by the European Union [26].

While the Digital Transformation Office of the Presidency of the Republic of Türkiye reveals the rules that state institutions will follow regarding cyber security, it offers recommendations for the private sector. In addition, the Digital transformation office provides services for Cyber Security and Information Security. It plays a guiding role for public institutions and organizations and the private sector in line with the published information security guide [27].

Today, cyberspace emerges as the fifth field of operation after land, air, sea, and space operations where activities are carried out for national security. Understanding the scope and impact of academic research on cybersecurity can be a guide on the precautions that individuals, companies, or public institutions and organizations, that is, all elements of the state, should take. In this context, bibliometric analysis is a widely used research method to determine the main research topics of the literature and the missing issues in the research. VOSviewer software is

software used to visualize keywords, co-citations, and relationship links between studies in the literature in biometric analysis [28].

In this study, a bibliometric analysis was conducted using the keywords "cyber security", "cyber-attacks and threats", "protection", "spam security", "data security", "network security", and "anomaly detection". As a result of this analysis, it was tried to reveal the development of the existing literature on cyber security, and it was stated which direction the research in this field was focused on.

This study aims to reveal the focal points of the basic research topics in the cybersecurity literature and the academic impacts of the research. The aim of the bibliometric analysis and visualization performed through the VOSviewer software is to determine which topics are generally examined in cybersecurity and which topics are not researched in the literature in light of current information. In this context, it is aimed to clarify the following topics.

- To reveal the importance of cybersecurity in terms of national security,
- To emphasize the impact of artificial intelligence, especially deep learning, on cybersecurity and research,
- To indicate the importance of international cooperation and regulations,
- To increase the orientation towards missing topics in academic research and to make the importance of cybersecurity more evident.

In the following sections of the article, evaluations will be made on Co-occurrence of Keywords, Citation Analysis of Authors, Co-Authorship Analysis, Bibliographic Coupling of Documents, Bibliographic Coupling of Institutions, Citation Analysis of Countries and general results and evaluations will be given on technological developments and emerging approaches, especially on national security.

II. MATERIAL AND METHOD

The data in this study were obtained from the Web of Science (WoS) database using the VOSviewer software and the tools within this software.

A search was conducted in the Web of Science (WoS) database using the query: ("Cybersecurity" or "cyberattacks" or "cyber

protection" or "spam cybersecurity" or "data security" OR "network security" or "anomaly detection" or "cyber countermeasures") in the fields of title, abstract, and keywords. This search yielded 11,990 documents. The publications span the years 1994–2025, with the following distribution: 6,609 journal articles, 5,107 conference papers, 310 review articles, 226 early access publications, 26 book chapters, 21 edited works, 9 retracted articles, 6 data papers, and 3 letters. The dataset was analyzed based on citations, documents, authors, institutions, countries, and keywords. Only WoS-indexed studies were used as the data source. Within the scope of the study, 6606 articles were reached by selecting 2000 and later as the journal article and year restriction for analysis. When classified by discipline, the majority of publications were found to focus on electrical and electronic engineering (2,286), computer science and information systems (1,880), artificial intelligence (1,088), and telecommunications (929). The data were analyzed in VOSviewer, focusing on citation, text, author, institution, country, and keyword analyses, using WoS-indexed studies as the primary source.

Co-occurrence of Keywords: It was conducted to examine how frequently keywords are used together and which topics are at the forefront in this field.

Citation Analysis of Authors: It was conducted to determine which authors were cited the most in the field and who were the most influential authors in the field.

Co-Authorship Analysis: It was conducted to determine which authors collaborated and between which research groups the most common collaborations occurred.

Bibliographic Coupling of Documents: It was conducted to examine which documents use the same references and the thematic similarities between these documents.

Co-Citation Analysis of Authors: It was conducted to determine which authors are frequently cited together and what kind of a connection there is between these authors.

Bibliographic Coupling of Institutions: It was conducted to examine which studies different institutions are involved in together and which institutions interact the most.

Citation Analysis of Countries: It was conducted to examine which countries' studies are most cited and the impact of these countries on the field.

TABLE I
KEYWORDS AND RELATED DOCUMENTS

Keywords	1994-2025	2000-2025	Co-occurrence of Keywords	Cited
1.Cybersecurity 2.Cyberattacks 3.Cyber protection 4. Spam cybersecurity 5.Data security 6.Network security 7. Anomaly detection 8.Cyber countermeasures	<u>11990 documents.</u> - Journal Article (6609), - Paper (5107), - Review (310), - Early View (226), - Book Chapter (26), - Edited Publication (21), - Withdrawn Publication (9), - Data paper (6) - Letter (3)	<u>6606 Journal Articles</u> <u>Article Areas.</u> -Electrical and electronics engineering (2286), -Computer sciences and information systems (1880), -Artificial intelligence (1088), -Telecommunications (929).	1.Anomaly Detection, 2.Cybersecurity 3.Deep learning 4.Internet 5.Traing 5.Attacs 6.Network Security 7.Algorithm 8.Classification 9.Data security 10.Internet of things (The order is from most to least.)	<u>The Most Cited Authors</u> 1.Quin Du (2085), 2.Wei Li (1433), 3.Liangpei Zhang (1417)

III. RESULTS

The results section presents the analyses conducted with VOSviewer on the data obtained from the WoS database using the keywords "cybersecurity," "cyberattacks," "cyber protection," "spam cybersecurity," "data security," "network security," "anomaly detection," and "cyber countermeasures." In VOSviewer software, link strength represents the strength of the relationships between analyzed units (e.g. keywords, authors, articles). This relationship is usually related to the level of commonality between units.

Link Strength Between Keywords: It indicates how many times two keywords appear together in articles. For example, if the keywords "artificial intelligence" and "deep learning" appear together 50 times in an article, the link strength of these two words is calculated as 50.

Article or Citation Link Strength: It is calculated based on the citation or reference relationships of an article with other articles.

A. Co-Occurrence of Keywords

To examine the frequency of keyword co-occurrence and the prominent topics in the field, the dataset containing 18,059 keywords was filtered to include only those with a minimum occurrence of 25. This threshold resulted in the analysis of 225 keywords, revealing 5 clusters, 9,319 links, and a total link strength of 41,214. When the keywords in the articles obtained as a result of the keyword analysis were examined, it was seen that the words "anomaly detection", "cybersecurity", and "deep learning" were the most frequently used keywords. It is noteworthy that the word "deep learning" was not included in the words generated when determining the articles, but it was used as a keyword in the articles obtained as a result of the determined keywords." Notably, research in this field intensified after 2020, with "anomaly detection" and "cybersecurity" being frequently used terms from 2021 onwards. A temporal co-occurrence analysis of the keywords is presented in Figure 1.

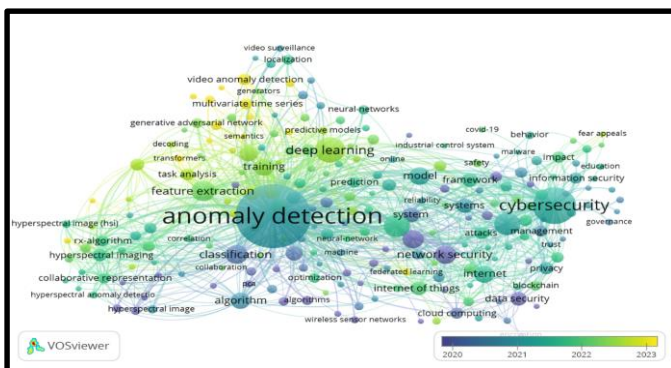


Fig. 1. Temporal Co-Occurrence Analysis of Keywords

B. Citation Analysis of Authors

This analysis was conducted to identify the most influential authors in the field and the citation relationships between their works to highlight the works with the greatest impact. Therefore, to identify the most cited authors and their impact on

the field, the dataset was filtered to include authors with at least 10 documents and 20 citations.

Out of 19,661 authors, 48 met these criteria and were analyzed, resulting in 5 clusters, 552 links, and a total link strength of 8,547. The analysis revealed that the top three most cited authors are Quin Du (2,085 citations), Wei Li (1,433 citations), and Liangpei Zhang (1,417 citations). The citation network analysis of these authors is presented in Figure 2.

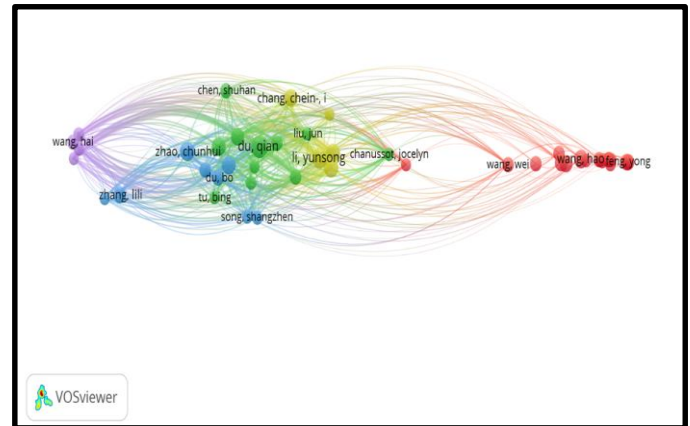


Fig. 2. Citation Analysis of Authors

C. Co-Authorship Analysis

To identify which authors collaborate and the most prevalent collaborations among research groups, a co-authorship analysis was conducted. From the dataset of 19,599 authors, a minimum citation threshold of 20 and a minimum of 5 publications were applied to highlight the most cited and active authors. This filtering yielded 299 authors, of whom the 183 most interconnected were analyzed.

The analysis revealed that these 183 authors formed 20 clusters, with 337 links and a total link strength of 1,061. Among them, Quin Du stood out with 26 publications and 2,085 citations. The top three authors in terms of publication count were Quin Du (26 publications), Yunsong Li (21 publications), and Weiyang Xie (20 publications). Additionally, the collaborative paper by Quin Du and Wei Li, titled "Collaborative Representation for Hyperspectral Anomaly Detection," was noted to have received 508 citations. The co-authorship analysis is visualized in Figure 3.

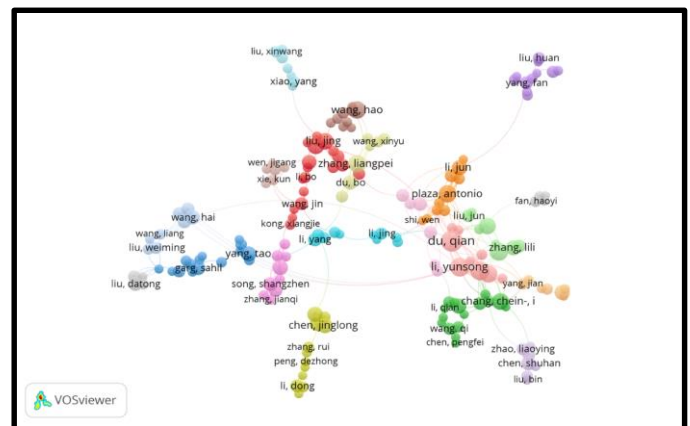


Fig. 3. Co-Authorship Analysis of Authors

D. Bibliographic Coupling of Documents

Bibliographic coupling analysis was conducted to examine which documents share the same references, thereby identifying thematic similarities and distinguishing the focal points of various studies. A minimum citation threshold of 25 was applied, resulting in the analysis of 1,000 documents. The analysis revealed 7 clusters with 47,978 links and a total link strength of 109,243.

The bibliometric coupling density visualization is presented in Figure 4, while the bibliographic coupling network analysis of documents is illustrated in Figure 5.

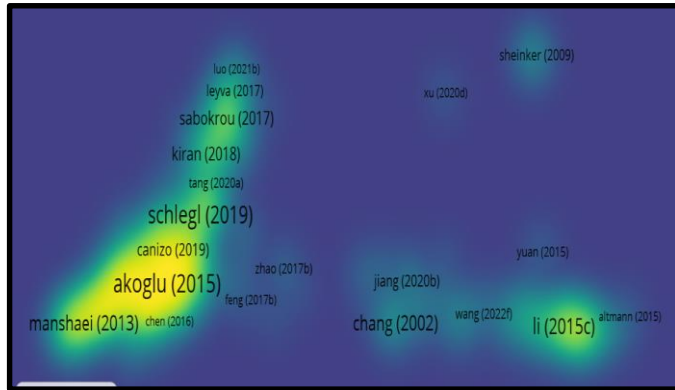


Fig. 4. Density Visualization of Bibliographic Coupling Among Documents

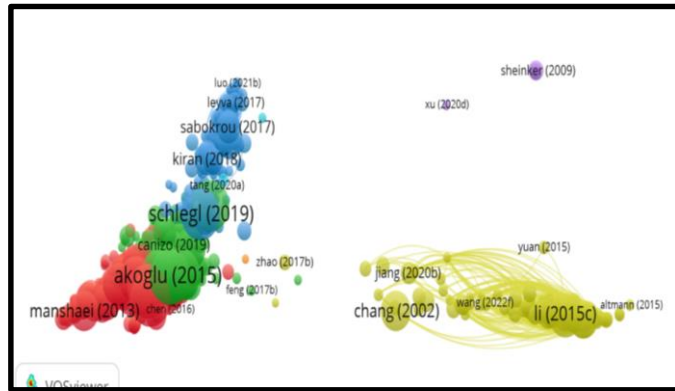


Fig. 5. Bibliographic Coupling Analysis of Documents

E. Co-Citation Analysis of Authors

A co-citation analysis of authors was conducted to determine how frequently authors are cited together, their interactions, and the thematic proximity of their work. Authors with a minimum of 30 citations were included in the analysis, resulting in a dataset of 1,058 authors.

The analysis identified 6 clusters, with 186,504 links and a total link strength of 852,325. Prominent authors included D. P. Kingma, C. I. Chang, and N. R. Prasad. The density visualization of the co-citation analysis is presented in Figure 6, and the network visualization of the co-citation analysis is shown in Figure 7.

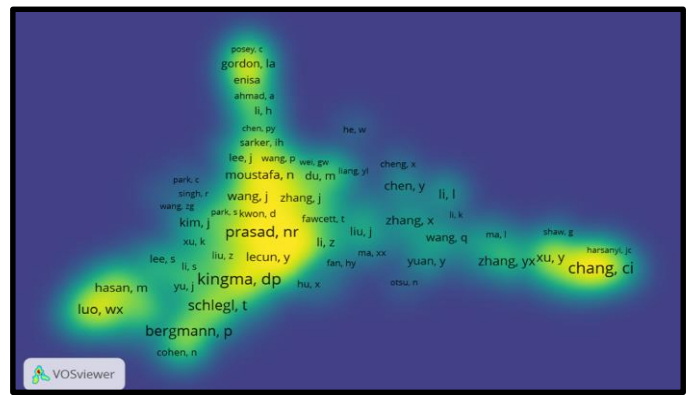


Fig 6. Density Visualization of Co-Citation Analysis Among Authors

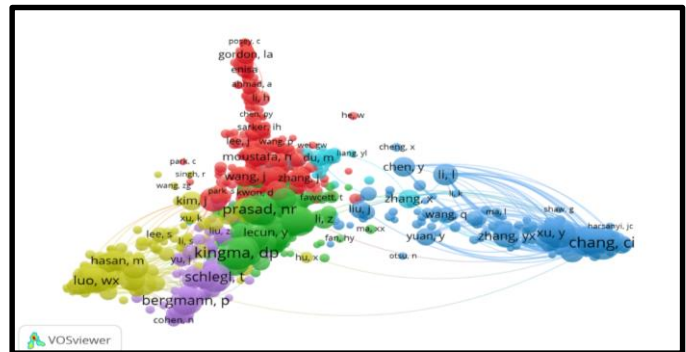


Fig. 7. Co-Citation Analysis of Authors

F. Bibliographic Coupling of Institutions

A bibliographic coupling analysis of institutions was conducted to examine which institutions collaborated on studies and identify the most interactive institutions. A minimum threshold of 20 documents and 20 citations was applied, resulting in the analysis of 71 institutions from an initial dataset of 5,361. The analysis identified 2 clusters with 2,484 links and a total link strength of 1,524,276. The Chinese Academy of Sciences emerged as the leading institution in terms of document and citation counts. The density visualization of institutional bibliographic coupling is presented in Figure 8.

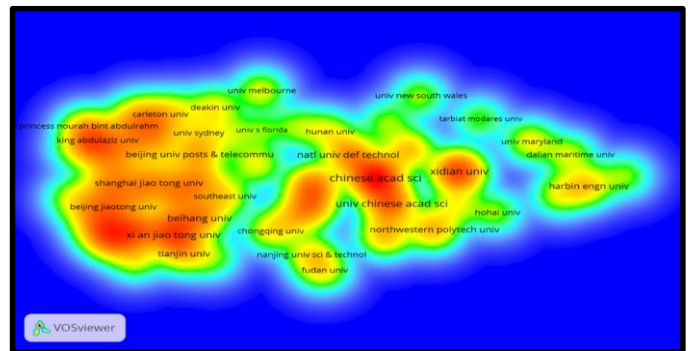


Fig 8. Density Visualization of Bibliographic Coupling Among Organizations

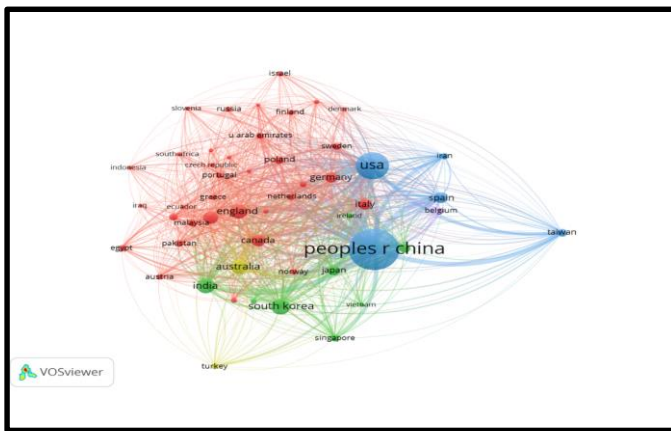


Fig. 9. Citation Analysis of Countries

G. Citation Analysis of Countries

With this analysis, we tried to explain cyber security interactions between countries and their research contributions. In this context, which country's academics. A minimum limit of 20 documents and 20 citations was applied to determine whether its cyber security studies were cited and their impact on cyber security. Among the 117 countries that make up the data set, 51 countries fell within the specified criteria and these 51 countries were included in the analysis. According to the number of research and citations, the People's Republic of China and the United States of America are in the first two places. The visual for the citation analysis of the countries is presented in Figure 9.

IV. DISCUSSION

As a result of rapidly changing and developing technological developments, cyber security has become an area that needs to be evaluated multi-dimensionally with both the daily habits of users and the legal regulations of countries. Below are the results of the bibliometric analysis:

A. Co-occurrence of Keywords, Key Trends, and Focus Areas

The results revealed that concepts such as anomaly detection and deep learning in cybersecurity are fundamental topics in research and are mostly researched. These findings support the work of Torres et al. [23] on artificial intelligence-based security solutions. In particular, anomaly detection plays a critical role in determining threats to systems [29]. Ahmed et al. [30] showed that anomaly detection algorithms are very important for cybersecurity with low error rates. According to the bibliometric analysis results, it was shown that deep learning, which is not included in the search words, is frequently used in research in addition to anomaly detection. This reveals that artificial intelligence has an increasingly important place in cybersecurity applications and that its importance will increase in the future. Mahdavi et al. [31] also emphasized the importance of artificial intelligence-supported deep learning in modeling cyber threats in their study. This explains why deep learning, which emerged as a result of the analysis, is frequently included in research.

B. Author and Institution-Based Citation Analyses

The bibliometric analysis revealed that Quin Du, Wei Li, and Liangpei Zhang are leading researchers in this field, according to their citation numbers. Du's research on anomaly detection aligns with Ahmad et al. [30], who explored deep learning-based threat detection methods. Notably, Du's 2020 study, "Collaborative Representation for Hyperspectral Anomaly Detection," "provides an effective model for modern security systems. The leadership of the Chinese Academy of Sciences reflects national investments and strategic priorities in the field. China is seen to be increasing its emphasis on AI-enabled cybersecurity solutions in its academic output.

C. Global Collaborations and Regional Differences

China and the U.S. are identified as leaders in cybersecurity research, while other countries have also contributed significantly, with increasing interest in this field. For instance, in Türkiye, the Presidential Digital Transformation Office plays a guiding role in cybersecurity efforts [22,25]. Similarly, in the European Union, cybersecurity is closely tied to data protection regulations like GDPR [32]. Such regulatory approaches complement more technology-driven solutions in Asia and the Americas. The literature highlights that global collaborations provide stronger solutions to cybersecurity threats [8]. These collaborations not only enhance academic knowledge sharing but also lead to more effective industrial solutions.

D. Technological Developments and Emerging Approaches

AI technologies, particularly machine learning and deep learning, are regarded as the future of cybersecurity. Martínez Torres [23] explored the impact of deep learning algorithms on threat modeling and prediction, enhancing the proactive capabilities of anomaly detection systems. On the other hand, the literature draws attention to the ethical and legal challenges of AI-based approaches. The misuse of AI solutions could lead to privacy violations and discriminatory outcomes [33]. This highlights the importance of ethical considerations in cybersecurity research, a frequently discussed topic in the literature.

E. The Importance of Cyber Security for National Security

Today, cyber security has become an area that directly provides not only individual and corporate security but also national security. This power stands out as a critical requirement for law enforcement to be persistent and permanent against cyber threats. Cybersecurity has become a part of national security in the modern world. Ensuring the distribution of critical infrastructures, digital assets, information systems and permanence of images in cyberspace has become one of their most important tasks. This structural law enforcement force must have sufficient capabilities and operational individuals against cyber threats and should be at the center of national security. Cybersecurity training for law enforcement should not only be limited to preventing and controlling individual crimes but should also aim to increase resilience against organized attacks targeting infrastructure and digital infrastructure. This training should include topics such as developing proactive defense equipment against new-generation threats, supporting

national and international cooperation, and using threat detection technologies effectively. Consisting of the field of cyber security, it will enable law enforcement forces not only to respond to current attacks but also to be trained to anticipate potential threats, thus playing a key role in protecting national security.

Training for law enforcement personnel to recognize cyber threats and develop intervention strategies has become an issue emphasized by law enforcement agencies around the world, as well as in Türkiye. [34].

Intervention methods should be constantly applied in the light of exercises and scenarios to enable law enforcement cyber security personnel to intervene faster and more effectively by improving their abilities to respond to potential attacks [35].

F. Comparisons with Literature and Identified Gaps

The analysis reveals gaps in the literature and opportunities for future research. For example, although it is stated in the literature that small and medium-sized enterprises (SMEs) are more vulnerable to cybersecurity threats, it is evaluated that this issue has not been addressed much in the analysis [36]. The analysis draws attention to the fact that research is concentrated in countries considered to be technologically advanced, such as China and the United States of America, while cybersecurity research in developing countries is scarce. It is evaluated that this may be due to resource constraints [36]. As cyberspace becomes the fifth operational domain, new attacks by state-sponsored actors or terrorist groups can be expected to pose significant threats to national security. The question of whether cyberattacks should be considered within the framework of cyberwarfare in terms of national security requires a detailed examination.

G. Main Results and Global Assessment

Although the word "deep learning" was not among the words produced when determining the articles, it was observed that it was mostly used as a keyword in articles about cybersecurity. This shows that artificial intelligence and its sub-branches are effectively used in the protection of cyberspace.

In the field of cyber security, it is evaluated that China-based research is intensifying and China is increasingly placing more emphasis on artificial intelligence-supported cyber security solutions and has begun to take on a leading role in this field.

In Türkiye, the Presidency Digital Transformation Office is seen to be carrying out binding and guiding studies on cybersecurity in national and international areas, such as European Union Data Protection Regulations, and states have developed necessary preventive measures in this regard.

In addition to the use of artificial intelligence-supported technologies in cyber security, it is seen that issues regarding the ethical use of these technologies have begun to be addressed in a significant way.

Cyberspace is the fifth field of activity where activities aimed at national security are carried out after land, air, sea and space operations, and necessary measures must be taken for national security from the individual to the state.

Cyber-attacks are considered within the framework of cyber warfare in terms of national security, and it is evaluated that these attacks can be carried out by aggressive states or terrorist

elements and that it is necessary to be prepared against these attacks.

V. CONCLUSION

Cybersecurity has become a constantly evolving field affected by technological, ethical and regulatory dimensions driven by the increasing prevalence of digitalization. In this study, academic research was evaluated through a bibliometric analysis using VOSviewer based on the keywords whose location details are given in Table I. The evaluations based on the results obtained in this context are given below.

Artificial intelligence, and its sub-branch deep learning algorithms, are popularly used today to detect and prevent cyber-attacks. Anomaly detection and neutralization of cyber-attacks as a result of this detection are frequently mentioned in the literature [37]. These useful algorithms and cybersecurity technologies are very important for identifying threats such as zero-day attacks. The use of these technologies creates high costs for institutions and the need for personnel requiring technical expertise. This creates a need for solutions that are accessible to everyone. While searching for solutions that are accessible to everyone, critical infrastructures and national security should be taken into consideration. As a result of the analysis made with keywords, it is seen that the words "deep learning", "internet", "training", "attacks", "network", "security", "algorithm", "classification", "Data security", "Internet of things" are the most frequently used words in the detected articles after the words "anomaly detection" and "cybersecurity". It has been evaluated that the developments in the field of artificial intelligence are increasing their impact in the field of cyber security in a similar way [38].

Developments in artificial intelligence are transforming cybersecurity into a global issue today. International cooperation is very important in analyzing cyber threats, investigating their consequences, and developing useful applications [8]. In addition to the fact that the People's Republic of China and the United States lead the world in cybersecurity research, interest in cybersecurity is increasing at the national level. Although structures such as the European Union have regulatory approaches to cybersecurity, they are also changing their cybersecurity approaches in parallel with technological developments in Asia and America. Security needs to be ensured in all areas of cyberspace and defense mechanisms against attacks need to be developed. Therefore, as in Türkiye, the measures to be taken should be determined and a national policy should be developed [24, 27]. No matter how much national policies are developed, the importance of international cooperation towards cyberspace, which has become a global problem and can be used by cyber terrorists, should not be forgotten.

The increasing use of artificial intelligence in cybersecurity offers significant advantages to users, while it has become a concern for all states of the world at national and international levels. This has also brought ethical and legal challenges [39]. Regulations such as the European Union's General Data Protection Regulation (GDPR) aim to increase measures to strengthen data security while emphasizing the privacy of personal or corporate data in the use of artificial intelligence.

Therefore, legislators, government institutions and academics should work together on cybersecurity, taking into account ethical sensitivities.

While studies on changing perception of cybersecurity and the ethical issues that come with it continue around the world, cybersecurity vulnerabilities emerge in developing countries due to many problems and financial resources. Therefore, developing countries should be supported by international funds and collaborations involving projects to improve their cybersecurity infrastructure. It has been observed that small and medium-sized enterprises operating in developing countries are sensitive to cybersecurity threats and that there is a lack of research in this area and that it is a subject that has not yet been sufficiently researched [36].

Cybersecurity is vulnerable to small businesses, either individually or institutionally, and the concept of cybercrime emerges when they do not take sufficient measures in the field of cybersecurity. The concept of cybercrime has become a strategic priority for all modern law enforcement agencies and all units working for national security, as in Türkiye, which is responsible for public security [24,27].

The rapid growth of digitalization has led to cyberattacks posing serious threats to individual rights and public order [27]. Institutions responsible for ensuring national security and public order should develop and implement more effective measures against cyber threats [8].

In addition to cyberattacks that individuals and businesses will be exposed to, basic infrastructures such as energy, water, transportation, and finance are primary targets for cyberattacks that will affect the entire society [24,27]. AI-powered detection systems and cybersecurity solutions play an important role in continuously monitoring these infrastructures and preventing potential breaches [29,37,39]. Therefore, States should integrate National Cyber Defense Strategies, regulatory frameworks, international cooperation, and the development of national technologies [8,24,27]. Such strategies should focus on preventing cyber-attacks and accelerating post-incident recovery processes.

Law enforcement and national security agencies should continuously update threat models to address the dynamic nature of cyber threats. Artificial intelligence and machine learning-enabled detection systems offer innovative approaches to counter these evolving challenges [27]. National security policies should include international collaborations to address global cyber threats. Organizations such as NATO's Cooperative Cyber Defense Center of Excellence (CCDCOE) provide effective frameworks for such collaborations [40].

References

[1] Mas-Tur, A., Kraus, S., Brandtner, M., Ewert, R., & Kürsten, W. (2020). Advances in management research: a bibliometric overview of the Review of Managerial Science. *Review of Managerial Science*, 14(5), 933-958.
 [2] Mistar, J., Setiakarnawijaya, Y., Dewi, P. C. P., Paramita, D. P., Aqobah, Q. J., & Akbar, M. A. (2023). Systematic Literature Review: Research on Martial Arts Competition Using Vos Viewers in the 2018-2022 Google Scholar Database. *Gladi: Jurnal Ilmu Keolahragaan*, 14(02), 221-228.

[3] Van Eck, N. J. & Waltman, L. (2023). VOSviewer manual for VOSviewer version 1.6.20. Universiteit Leiden: CWTS.
 [4] Vosviewer (2025). <https://www.vosviewer.com/features/examples>. Access Date: January 18, 2025
 [5] Dereli, A. B. (2024). Vosviewer ile Bibliyometrik Analiz. *Communicata*, (28), 1-7
 [6] Zawish, M., Dharejo, F. A., Khowaja, S. A., Raza, S., Davy, S., Dev, K., & Bellavista, P. (2024). AI and 6G into the metaverse: Fundamentals, challenges, and future research trends. *IEEE Open Journal of the Communications Society*, 5, 730-778.
 [7] Khan, L. U., Yaqoob, I., Imran, M., Han, Z., & Hong, C. S. (2020). 6G wireless systems: A vision, architectural elements, and future directions. *IEEE Access*, 8, 147029-147044.
 [8] Kshetri, N., & Kshetri, N. (2016). Cybersecurity in National Security and International Relations. *The Quest to Cyber Superiority: Cybersecurity Regulations, Frameworks, and Strategies of Major Economies*, 53-74.
 [9] Craigen, D., Diakun-Thibault, N., & Purse, R. (2014). Defining cybersecurity. *Technology innovation management review*, 4(10).
 [10] Singer, P. W., & Friedman, A. (2014). *Cybersecurity: What everyone needs to know*.
 [11] Caviglione, L., Wendzel, S., Mileva, A., & Vrhovec, S. (2021). Guest Editorial: Multidisciplinary Solutions to Modern Cybersecurity Challenges. *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.*, 12(4), 1-3.
 [12] Borrett, M., Carter, R., & Wespi, A. (2014). How is cyber threat evolving and what do organizations need to consider? *Journal of business continuity & emergency planning*, 7(2), 163-171.
 [13] Ulsch, M. (2014). *Cyber threat!: how to manage the growing risk of cyber attacks*. John Wiley & Sons.
 [14] Mallikarjunan, K. N., Muthupriya, K., & Shalinie, S. M. (2016, January). A survey of distributed denial of service attack. In 2016 10th International Conference on Intelligent Systems and Control (ISCO) (pp. 1-6). IEEE.
 [15] Hasan, M. K., Habib, A. A., Islam, S., Safie, N., Abdullah, S. N. H. S., & Pandey, B. (2023). DDoS: Distributed denial of service attack in communication standard vulnerabilities in smart grid applications and cyber security with recent developments. *Energy Reports*, 9, 1318-1326.
 [16] Solomon, A., Walker, E., Kensington, J., Drummond, M., Hall, R., & Blackwell, G. (2024). A new autonomous multi-layered cognitive detection mechanism for ransomware attacks.
 [17] Chen, P., Desmet, L., & Huygens, C. (2014). A study on advanced persistent threats. In *Communications and Multimedia Security: 15th IFIP TC 6/TC 11 International Conference, CMS 2014, Aveiro, Portugal, September 25-26, 2014. Proceedings 15*(pp. 63-72). Springer Berlin Heidelberg.
 [18] Handa, A., Sharma, A., & Shukla, S. K. (2019). Machine learning in cybersecurity: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), e1306.
 [19] Chua, Y. T., Parkin, S., Edwards, M., Oliveira, D., Schiffner, S., Tyson, G., & Hutchings, A. (2019, November). Identifying unintended harms of cybersecurity countermeasures. In 2019 APWG Symposium on Electronic Crime Research (eCrime) (pp. 1-15). IEEE.
 [20] Volini, A. G. (2020). A Deep Dive into Technical Encryption Concepts to Better Understand Cybersecurity & Data Privacy Legal & Policy Issues. *J. Intell. Prop. L.*, 28, 291.
 [21] Zheng, Y., Li, Z., Xu, X., & Zhao, Q. (2022). Dynamic defenses in cyber security: Techniques, methods, and challenges. *Digital Communications and Networks*, 8(4), 422-435.
 [22] Maddireddy, B. R., & Maddireddy, B. R. (2024). Neural Network Architectures in Cybersecurity: Optimizing Anomaly Detection and Prevention. *International Journal of Advanced Engineering Technologies and Innovations*, 1(2), 238-266.
 [23] Martínez Torres, J., Iglesias Comesaña, C., & García-Nieto, P. J. (2019). Machine learning techniques applied to cybersecurity. *International Journal of Machine Learning and Cybernetics*, 10(10), 2823-2836.
 [24] Siber Güvenlik (n.d.). Dijital Dönüşüm Ofisi. <https://cbddo.gov.tr/siber-guvenlik/>
 [25] National Cybersecurity Strategy (2023). White House. <https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Cybersecurity-Strategy-2023.pdf>
 [26] Anagnostakis, D. (2021). The European Union-United States cybersecurity relationship: a transatlantic functional cooperation. *Journal of Cyber Policy*, 6(2), 243-261
 [27] Siber Güvenlik (n.d.). Dijital Dönüşüm Ofisi. https://cbddo.gov.tr/SharedFolderServer/Genel/File/bg_rehber.pdf

- [28] Orduña-Malea, E., & Costas, R. (2021). Link-based approach to study scientific software usage: The case of VOSviewer. *Scientometrics*, 126(9), 8153-8186.
- [29] Moustafa, N. (2021). A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets. *Sustainable Cities and Society*, 72, 102994.
- [30] Ahmad, R., Alsmadi, I., Alhamdani, W., & Tawalbeh, L. A. (2023). Zero-day attack detection: a systematic literature review. *Artificial Intelligence Review*, 56(10), 10733-10811.
- [31] Mahdavi, S., & Ghorbani, A. A. (2019). Application of deep learning to cybersecurity: A survey. *Neurocomputing*, 347, 149-176.
- [32] Goddard, M. (2017). The EU General Data Protection Regulation (GDPR): European regulation that has a global impact. *International Journal of Market Research*, 59(6), 703-705.
- [33] Abdullahi, M., Alhussian, H., Aziz, N., Abdulkadir, S. J., Alwadain, A., Muazu, A. A., & Bala, A. (2024). Comparison and investigation of AI-based approaches for cyberattack detection in cyber-physical systems. *IEEE Access*.
- [34] Baadel, S., Thabtah, F., & Lu, J. (2021). Cybersecurity awareness: A critical analysis of education and law enforcement methods. *Informatica*, 45(3).
- [35] Kavak, H., Padilla, J. J., Vernon-Bido, D., Diallo, S. Y., Gore, R., & Shetty, S. (2021). Simulation for cybersecurity: state of the art and future directions. *Journal of Cybersecurity*, 7(1), tyab005.
- [36] Chaudhary, S., Gkioulos, V., & Katsikas, S. (2023). A quest for research and knowledge gaps in cybersecurity awareness for small and medium-sized enterprises. *Computer Science Review*, 50, 100592.
- [37] Geluvaraj, B., Satwik, P. M., & Ashok Kumar, T. A. (2019). The future of cybersecurity: Major role of artificial intelligence, machine learning, and deep learning in cyberspace. In *International Conference on Computer Networks and Communication Technologies: ICCNCT 2018* (pp. 739-747). Springer Singapore.
- [38] Muhammad, G., Pratama, A. R., Shaloom, C., & Cassandra, C. (2023, November). Cybersecurity Awareness Literature Review: A Bibliometric Analysis. In *2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS)* (pp. 195-199). IEEE.
- [39] Allahrakha, N. (2023). Balancing cyber-security and privacy: legal and ethical considerations in the digital age. *Legal Issues in the Digital Age*, (2), 78-121.
- [40] Štrucl, D. (2022). Comparative study on the cyber defense of NATO Member States. NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE).

BIOGRAPHIES



Vedat Yilmaz obtained his BSc degree in system engineering from the military academy in 2004. I completed Digital Communication Electronics Training at Hacettepe University in 2006. My master's degree in Management and Organization at Selçuk University in 2007 and my PhD in Biomechanics at Hacettepe University in 2022. Additionally, I received training on Principles of Communication from Cranfield University, Cyber Security from METU, and Training on Terrorists' Use of Cyberspace from the Center of Excellence in Combating Terrorism.

I managed many technology projects within the Gendarmerie General Command. Currently, I continue my studies in cyber security, artificial intelligence applications in cyber crimes, security technologies, and cybercrime.


Intelligent Modular Energy Hub: Advanced Optimization of Second-Life Lithium-Based Batteries for Sustainable Power Utilization

Abdulkadir Gozuoglu

Abstract— The Intelligent Modular Energy Hub (IMEH) introduces a cost-effective and scalable energy storage solution by repurposing second-life lithium-based batteries, including Li-ion, LiPo, and LiFePO₄ cells, sourced from discarded consumer electronics, power tools, and electric vehicles. This study develops an STM32- and ESP32-based battery testing system, integrating an electronic dummy load and a custom battery management system (BMS) to accurately assess the state-of-charge and state-of-health (SoH) of various battery chemistries. A 7S and variable parallel battery pack configuration ensures adaptability to diverse residential and off-grid applications. The proposed system features real-time IoT monitoring, extending battery lifespan while optimizing charging cycles through grid, solar, or wind energy sources. Experimental results demonstrate that the Samsung 25R battery exhibited the highest SoH (92%) and energy efficiency (95%), making it the most viable for second-life applications. The Turnigy Graphene LiPo battery, while displaying the highest efficiency (97%), showed a slightly lower capacity retention (89%), indicating potential limitations for long-term storage. Voltage drop analysis confirmed that lower internal resistance leads to better performance, with the Turnigy Graphene battery maintaining the lowest voltage drop (160mV) under discharge conditions. Additionally, the IMEH system achieved an average energy efficiency of 94.75%, outperforming commercial BMS solutions, which averaged 92% efficiency. IoT-based predictive maintenance enhanced battery longevity, ensuring better cycle count retention and charge-discharge stability. This research contributes to affordable energy solutions, supports the circular economy, and enhances sustainable power utilization by integrating modular and intelligent energy management strategies into next-generation smart grids.

Index Terms— Intelligent energy hub, modular energy storage, second-life lithium-based batteries, Li-ion, LiPo, LiFePO₄, IoT-based battery management, electronic dummy load, energy optimization, sustainable power utilization, STM32, ESP32, smart grid.

Abdulkadir Gozuoglu, is with Department of Electrical & Engineering Department, Tokat Gaziosmanpasa University, Tokat, Turkey, (e-mail: abdulkadir.gozuoglu@gop.edu.tr).

 <https://orcid.org/0000-0002-6968-379X>

Manuscript received Feb 18, 2025; accepted Apr 16, 2025.
DOI: [10.17694/bajece.1641971](https://doi.org/10.17694/bajece.1641971)

I. INTRODUCTION

THE RAPID proliferation of electric vehicles (EVs) and portable electronic devices has led to a substantial increase in the production of lithium-based batteries. Upon reaching the end of their primary use—typically when their capacity diminishes to about 70–80%—these batteries present a valuable opportunity for repurposing in less demanding applications, a practice known as "second-life" utilization. This approach enhances resource efficiency and aligns with sustainable energy practices by mitigating environmental waste [1–4].

Integrating second-life lithium-based batteries into energy storage systems offers a cost-effective solution for managing renewable energy sources such as solar and wind. However, safety concerns, cell inhomogeneity, and system compatibility must be addressed to ensure reliable performance. Advanced BMS are crucial in monitoring and controlling these batteries, ensuring safe operation and prolonging their lifespan [5–8].

The emergence of the Internet of Things (IoT) has further enhanced BMS capabilities, enabling real-time monitoring, predictive analytics, and remote control of battery systems. IoT-based solutions facilitate efficient energy management by providing detailed insights into battery performance and health, allowing for proactive maintenance and optimization [9, 10].

Despite these advancements, developing modular, scalable, and intelligent energy storage solutions utilizing second-life lithium-based batteries remains an active research area. Addressing the challenges associated with battery variability, system integration, and energy optimization is essential for successfully deploying these systems in residential and off-grid applications.

Main Contributions of manuscript are listed below:

- **Development of a Modular Energy Storage System:** Design and implement a flexible 7S (seven cells in series) configuration with variable parallel connections, accommodating various second-life lithium-based battery chemistries to meet diverse energy demands.
- **Advanced Battery Assessment Techniques:** Utilization of STM32- and ESP32-based electronic dummy loads (EDLs) integrated with a custom BMS to accurately evaluate the state-of-charge (SoC) and SoH of repurposed batteries.
- **IoT-Enhanced Monitoring and Control:** Implementation of real-time IoT capabilities for continuous monitoring, predictive analytics, and remote management, enhancing system reliability and performance.

- *Sustainable Energy Integration*: By enabling efficient integration of renewable energy sources like solar and wind, the system enhances sustainable energy usage and supports the circular economy through optimized resource utilization.

The manuscript is structured as follows: Section II reviews advancements in second-life battery applications and IoT-based BMS. Section III details the IMEH system, including STM32- and ESP32-based assessments. Section IV presents experimental evaluations on battery performance. Section V analyzes results compared with commercial BMS solutions, and Section VI concludes with key findings and future directions.

II. RELATED WORKS

Integrating second-life lithium-based batteries into energy storage systems has garnered significant attention in recent years, driven by the need for sustainable and cost-effective energy solutions [11-14]. This section reviews recent advancements in this field, focusing on key areas such as battery repurposing, modular energy storage design, IoT-enhanced BMS, and the challenges of implementing second-life batteries.

A. Battery Repurposing and Second-Life Applications

Recent studies have explored the feasibility of repurposing retired electric vehicle batteries for stationary energy storage applications. For instance, Jiang, et al. [15] investigated the potential of second-life batteries in mitigating environmental impacts and providing economic benefits when used in grid storage and renewable energy systems. The study highlights the importance of addressing safety concerns and developing standardized testing protocols to ensure the reliability of these repurposed batteries. mdp.com.

Similarly, Gharebaghi, et al. [14] examined various modular architectures and control strategies that enhance the flexibility and efficiency of energy storage systems utilizing second-life batteries. The paper emphasizes the need for advanced power electronics and control algorithms to manage second-life batteries' variability and degradation characteristics.

B. Modular Energy Storage System Design

The design of modular and scalable energy storage systems utilizing second-life batteries has been a focal point in recent research. Lipu, et al. [12] presented an IoT-enhanced BMS that enables real-time monitoring and predictive analytics, facilitating accurate state-of-health estimation and proactive maintenance strategies. This approach enhances the safety and longevity of second-life battery systems by providing detailed insights into battery performance and enabling remote management. nature.com

Shi [13] also introduced the MambaLithium model, a selective state space framework designed to estimate critical battery states, including remaining useful life, health, and charge. This model leverages advanced algorithms to capture lithium-ion batteries' intricate aging and charging dynamics, thereby improving estimation accuracy and computational robustness.

C. IoT-Enhanced Battery Management Systems

The advancement of IoT technologies has significantly improved the capabilities of BMS for second-life applications. Cui, et al. [11] studied health monitoring algorithms for retired batteries used in grid storage, collecting and analyzing data over 15 months. The study achieved a mean absolute percentage error below 2.3% on test data by implementing machine-learning-based health estimation models. These findings highlight the viability of repurposing retired batteries for second-life applications.

Furthermore, Basic, et al. [16] introduced a wireless BMS architecture utilizing near-field communication (NFC), building upon previous research to create a unified framework for in-vehicle and second-life battery applications. The design incorporates advanced security analysis and a wake-up system design, significantly reducing the daily power consumption of stored battery packs from mill watts to microwatts.

D. Challenges in Implementing Second-Life Batteries

Despite the promising applications, several challenges hinder the widespread adoption of second-life batteries. Gu, et al. [17] identified key issues such as cell inhomogeneity, safety concerns, and the lack of standardized regulations. The paper discusses potential solutions, including advanced sorting techniques, improved BMS algorithms, and the development of international standards to ensure the safe and efficient deployment of second-life battery systems.

Additionally, a review by the National Center for Biotechnology Information Patel, et al. [18] explore the various pathways for end-of-life EV batteries, including immediate recycling or deployment in second-life applications before eventual recycling. They discuss the challenges and barriers of each approach, evaluating their environmental and economic feasibility while weighing the competing advantages and drawbacks of different repurposing strategies [18].

These studies collectively underscore the potential of second-life lithium-based batteries in contributing to sustainable energy solutions. However, they also highlight the necessity for continued research and development to address the technical and regulatory challenges associated with their implementation.

While previous studies have explored second-life battery applications primarily within stationary storage for grid systems, IoT-based BMS improvements, and modular pack architectures, our research presents a unique approach by integrating an STM32- and ESP32-based programmable EDL to analyze and optimize repurposed lithium-based battery packs at a modular level. Unlike other systems focusing mainly on health estimation through machine learning or generalized modular integration, our system provides direct empirical analysis by testing real-time capacity, performance, and degradation behavior under various load conditions.

Our research differentiates itself by integrating second-life lithium-based batteries into a modular and scalable energy hub that supports a broader range of applications, including home energy storage, off-grid renewable integration, and low-cost power management solutions for underserved regions. By incorporating IoT-enhanced real-time monitoring and control, adaptable modular battery configurations, and direct discharge/charge performance assessment, our system bridges

the gap between theoretical battery health estimation and practical, cost-effective energy storage solutions for sustainable power utilization.

III. MATERIALS AND METHODS

This section describes the hardware components, circuit design, measurement approach, and IoT-based data logging implemented in the IMEH. The system is designed to assess second-life lithium-based batteries, perform real-time monitoring, and optimize their use for sustainable power applications.

A. Hardware Components and System Architecture

The IMEH consists of an STM32F103C8T6 microcontroller responsible for real-time computation and battery data acquisition. The ESP32 microcontroller is used for IoT-based communication, enabling remote monitoring and data logging. The system features a custom-built EDL, which facilitates controlled discharge testing of battery packs to determine SoH and State-of-Charge SoC.

The battery packs under testing are arranged in a 7S variable parallel configuration, where different types of lithium-based cells, including Li-ion, LiPo, and LiFePO₄, can be evaluated. The system measures real-time voltage, current, power, and temperature using precision sensors such as INA219 or MAX471. These values are processed by the STM32 microcontroller, allowing for accurate capacity estimation.

Additionally, an IRFZ44N MOSFET is used as a programmable load, operating in saturation mode to regulate current flow dynamically. An operational amplifier (Op-Amp) circuit is configured in a voltage follower mode to ensure high accuracy, preventing loading effects and maintaining precise voltage measurement.

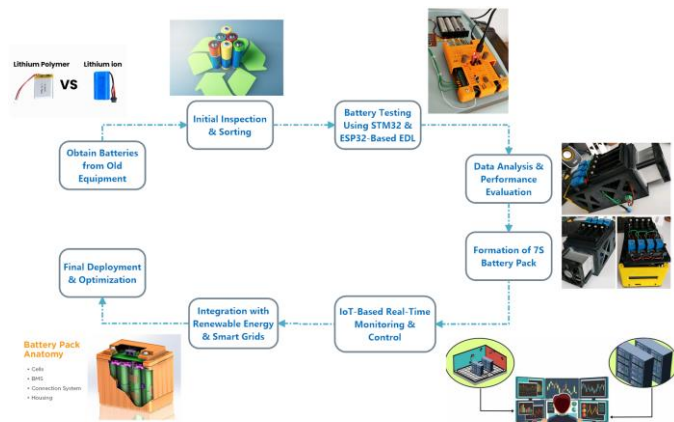


Fig. 1. The flow diagram of the proposed intelligent EDL

The system workflow shown in Fig. 1 begins by testing second-life lithium-based batteries, where an IRFZ44N MOSFET in saturation mode controls the discharge cycle. An Op-Amp circuit in a voltage follower configuration ensures stable voltage measurement, while the STM32F103C8T6 microcontroller processes real-time voltage, current, power, and temperature data using INA219 or MAX471 sensors. The collected data is transmitted via UART, I2C, and Wi-Fi to an ESP32-based IoT controller for remote monitoring and

visualization on an OLED display while also being stored on a local server for long-term analysis and optimization of second-life batteries.

B. Circuit Design and Implementation

The system employs a MOSFET-controlled electronic load that enables adjustable battery discharge under programmable conditions. A 5W precision shunt resistor is incorporated for current sensing, and an Op-Amp amplifies its voltage drop to enhance accuracy. This configuration ensures precise control over battery discharge cycles.

Voltage and current measurements are continuously monitored to improve system safety and reliability, preventing over-discharge or unsafe temperature levels. The STM32F103C8T6 processes these real-time measurements and dynamically adjusts the load to maintain controlled discharge.

A UART-based communication system links the STM32 with the ESP32 microcontroller, allowing for data transmission to an external monitoring platform. This setup facilitates continuous assessment and logging of battery performance over multiple cycles.

C. Battery Testing and Measurement Approach

The primary objective of the battery testing process is to evaluate the capacity, SoH, and SoC of second-life lithium-based batteries. The methodology consists of:

- **Controlled Discharge Cycles:** Batteries are discharged through the electronic dummy load, with current levels regulated by the MOSFET driver circuit.
- **Voltage and Current Measurement:** Sensors such as INA219 or MAX471 continuously monitor voltage drop, power consumption, and real-time current levels.
- **Temperature Monitoring:** Ensures battery operation remains within safe thermal limits to prevent degradation and hazards.
- **Capacity Estimation:** Based on recorded voltage, current, and time data, the STM32 calculates the adequate capacity of the tested battery pack.

Once the discharge cycle is complete, the collected data is stored and analyzed for performance trends, enabling comparisons across different battery brands and chemistries.

The real-time voltage (V) and current (I) readings are obtained using precision sensors such as INA219 or MAX471. The sensed values are converted using the Analog-to-Digital Converter (ADC) in the STM32 microcontroller:

$$V_{measured} = \frac{ADC\ Value \times V_{ref}}{2^n - 1} \quad (1)$$

$$I = \frac{V_{shunt}}{R_{shunt}} \quad (2)$$

Where:

- $V_{measured}$ = Measured battery voltage
- V_{ref} = Reference voltage (typically 3.3V or 5V)
- n = ADC resolution (e.g., 12-bit, 10-bit)
- V_{shunt} = Voltage drop across the shunt resistor
- R_{shunt} = Shunt resistor value (e.g., 0.1Ω)

These values are used for real-time voltage and current monitoring.

The power (P) consumed by the battery during discharge is calculated as:

$$P = V \times I \quad (3)$$

Where:

- V = Measured voltage of the battery
- I = Measured current during discharge

The formula helps in tracking the real-time energy consumption of the battery

The battery capacity (C) is calculated by integrating the discharge current over time:

$$C = \int I(t) dt \quad (4)$$

For discrete sampling using microcontrollers, this is approximated as:

$$C = \sum_{i=1}^n I_i \times \Delta t \quad (5)$$

Where:

- I_i = Measured discharge current at time step i
- Δt = Time interval between successive measurements
- C = Capacity in Ampere-hours (Ah)

The total discharge capacity of the battery over a full cycle is calculated as:

$$C_{measured} = \frac{\sum_{i=1}^n (I_i \times \Delta t)}{3600} \quad (6)$$

This is crucial for comparing the actual battery capacity with its rated capacity.

SoC represents the remaining charge in the battery and is computed as:

$$SoC = \frac{C_{remaining}}{C_{full}} \times 100 \quad (7)$$

Or in terms of voltage-based approximation:

$$SoC = \frac{V_{current} - V_{min}}{V_{max} - V_{min}} \times 100 \quad (8)$$

Where:

- $C_{remaining}$ = Charge left in the battery (Ah)
- C_{full} = Rated full capacity of the battery (Ah)
- $V_{current}$ = Current voltage of the battery
- V_{max} = Fully charged voltage (e.g., 4.2V for Li-ion)
- V_{min} = Minimum safe voltage (e.g., 2.5-3.0V for Li-ion)

This formula determines the available battery energy in real-time.

SoH is used to evaluate battery degradation over time. It is defined as:

$$SoH = \frac{C_{measured}}{C_{rated}} \times 100 \quad (9)$$

Where:

- $C_{measured}$ = Actual capacity determined from discharge testing
- C_{rated} = Manufacturer's rated capacity

This formula helps in assessing the aging and performance degradation of second-life batteries.

The efficiency (η) of the battery during charge and discharge cycles is determined as:

$$\eta = \frac{E_{out}}{E_{in}} \times 100 \quad (10)$$

Where:

• $E_{out} = P_{discharge} \times t_{discharge}$ (Energy delivered by the battery)

• $E_{in} = P_{charge} \times t_{charge}$ (Energy stored during charging)

This metric helps in evaluating second-life battery effectiveness. The formulas provide the fundamental calculations used in our system for battery testing, real-time monitoring, and optimization.

D. IoT-Based Data Logging and Remote Monitoring

The ESP32 microcontroller is the gateway for IoT communication, allowing for wireless data transmission. The system logs all battery performance data, including:

- *Voltage, Current, and Power Data* for real-time assessment.
- *Discharge Cycle Duration and Capacity Trends* to track battery degradation.
- *Temperature Logs* to prevent thermal runaway conditions.

The ESP32 sends this data to a local server, storing it for trend analysis and performance optimization. Additionally, an OLED display provides real-time updates, allowing users to monitor battery conditions on-site.

This IoT-enabled setup ensures that users can remotely access battery performance metrics, making it possible to optimize charging and discharging strategies dynamically.

IV. APPLICATION OF PROPOSED WORK

This section presents the practical implementation of the IMEH for assessing and optimizing second-life lithium-based batteries. It describes the experimental setup, system integration, and performance evaluation to demonstrate the feasibility and effectiveness of the proposed method. The subsections below outline the developed system's real-world deployment, testing conditions, and validation processes.

A. Experimental Setup and System Configuration

The proposed system uses an STM32F103C8T6 microcontroller, which manages real-time data acquisition, battery assessment, and load control. An ESP32 microcontroller facilitates IoT-based monitoring and communication, enabling remote access to battery performance data. EDL, integrated with a custom-designed BMS, applies controlled discharge profiles to evaluate the SoH and SoC of second-life batteries.

The tested battery modules are arranged in a 7S variable parallel configuration, accommodating multiple lithium-based chemistries such as Li-ion, LiPo, and LiFePO₄. Real-time voltage, current, and temperature measurements are captured using INA219/Max471 sensors, processed by the STM32, and logged for analysis. A local server stores all measurement data, ensuring a structured performance tracking and evaluation approach.

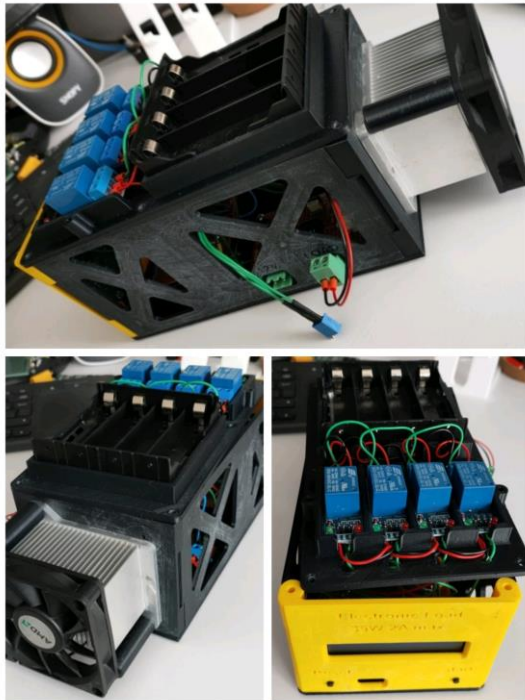


Fig. 2. The internal hardware structure of the Intelligent Modular Energy Hub (IMEH) shows key components for battery testing and monitoring

Fig. 2 depicts the internal hardware structure of the IMEH, highlighting key components used for battery testing and monitoring. The system features an IRFZ44N MOSFET, a controllable electronic load for battery discharge testing. A 5W shunt resistor is used for current measurement, allowing precise battery performance evaluation. The operational amplifier (Op-Amp) circuit ensures signal conditioning for accurate voltage readings. The volt/current sensor (INA219 or MAX471) monitors real-time power. A microcontroller (STM32F103C8T6) processes the acquired data and is then displayed on an LCD screen for real-time observation. The load connection terminals facilitate battery integration, ensuring seamless data acquisition for SoH and SoC calculations. The entire setup is enclosed in a ventilated structure, optimizing heat dissipation and ensuring stable operation during battery analysis.

B. Battery Performance Evaluation and Testing Methodology

The IMEH system performs extensive capacity, efficiency, and degradation analysis on repurposed batteries. The testing methodology consists of:

- **Controlled Discharge Cycles:** Batteries are subjected to programmable electronic loads using the IRFZ44N MOSFET, allowing precise current regulation.
- **Voltage and Current Profiling:** Precision sensors take measurements at defined intervals, ensuring accurate SoH estimation.
- **Thermal Management and Safety Checks:** Real-time temperature data is monitored to prevent overheating and ensure safe operation.
- **Data Logging and Trend Analysis:** All recorded battery metrics are stored for comparative analysis across brands and chemistries.

The STM32 microcontroller processes real-time power consumption, calculating the adequate battery capacity and identifying degradation patterns. The IoT-enhanced framework allows remote users to track and analyze these parameters over time.

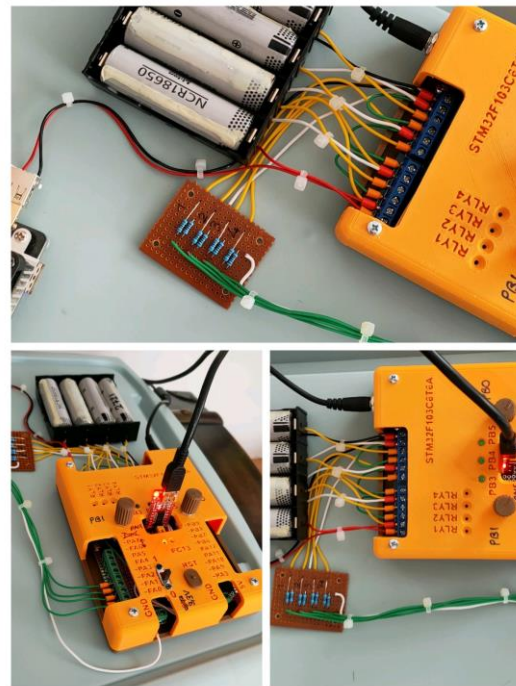


Fig. 3. STM32 & ESP32 based intelligent BMS application

Fig. 3 showcases a custom-built electronic load and battery management system based on the STM32F103C8T6 microcontroller for testing second-life lithium-based batteries. The system features a 7S battery configuration with multiple wiring connections for voltage and current monitoring. A custom PCB with resistors is integrated for signal conditioning, ensuring accurate battery performance analysis. The 3D-printed enclosure provides organized access to the STM32 ports, labeled for easy identification of relay and sensor connections. The system is powered and programmed via a USB-to-serial converter, enabling real-time data acquisition. This setup is part of an IoT-based energy management system for battery health assessment, controlled discharge testing, and integration into a sustainable power utilization framework.

C. IoT-Based Remote Monitoring and Data Analytics

The ESP32 microcontroller forms an IoT-enabled gateway, allowing continuous data transmission to a local server or cloud-based monitoring platform. The key functionalities include:

- **Real-Time Data Visualization:** Battery performance metrics are displayed via an OLED screen and can be accessed remotely.
- **Predictive Maintenance Alerts:** IoT-based analytics detect abnormal behavior and notify users of potential failures.
- **Historical Data Analysis:** Long-term monitoring enables trends in battery aging, efficiency, and performance degradation to be identified.

This IoT-based architecture ensures efficient and scalable battery health monitoring, making it possible to optimize

charge-discharge cycles dynamically for sustainable power utilization.

D. Integration with Renewable Energy and Potential Applications

The developed IMEH system is designed to be modular and adaptive, making it suitable for various real-world applications, including:

- *Residential and Off-Grid Energy Storage:* Repurposed batteries can be used for backup power and load balancing in homes or remote areas.
- *Integration with Renewable Sources:* The system

- *Internal Resistance Calculation:* To assess battery efficiency
- *Cycle Count Validation:* To determine the remaining lifespan

All measurements were logged in real-time, and the system automatically calculated SoC, SoH, and capacity degradation trends.

F. Battery Performance Data and Analysis

Based on the measured data shown in the TABLE I, the following battery brands were tested and evaluated for their suitability in battery pack hub integration:

TABLE I
MEASURED BATTERY PARAMETERS

Brand	Type	Rated Capacity (Ah)	Nominal Voltage (V)	Measured Capacity (Ah)	SoH (%)	SoC (%)	Internal Resistance (mΩ)	Cycle Count	Energy Efficiency (%)
Panasonic NCR18650B	Li-ion	3.4	3.6	3.1	91	80	50	250	93
Samsung 25R	Li-ion	2.5	3.6	2.3	92	85	45	300	95
LG MJ1	Li-ion	3.5	3.7	3.2	91.4	78	48	220	94
Turnigy Graphene	LiPo	1.3	3.8	1.15	88.5	75	30	180	97

supports solar and wind charging, utilizing second-life batteries to store excess energy.

- *Low-Cost Energy Solutions:* Affordable battery repurposing enables cost-effective energy storage, reducing reliance on expensive grid power.

The system's modularity ensures that battery configurations can be expanded based on specific energy demands, making it highly scalable for smart grid applications and sustainable energy management.

Applying the STM32- and ESP32-based EDL by performing real-time testing on available battery cells. The goal is to determine whether the Li-ion and LiPo batteries on hand are in good condition for integration into the battery pack hub. The tested parameters include SoC, SoH, internal resistance, cycle count, and energy efficiency, all measured via the low-cost innovative EDL system.

E. Battery Testing Setup and Measurement Process

The testing was conducted using a custom-built EDL system based on STM32F103C8T6 and ESP32 microcontrollers. The STM32 was responsible for real-time voltage, current, power, and temperature measurement, while the ESP32 handled data transmission to a local server for analysis.

Each battery underwent a controlled discharge test using the IRFZ44N MOSFET, which applied a programmable electronic load. The system used an Op-Amp in voltage follower mode for accurate voltage readings and an INA219/Max471 sensor for current and power measurement. The key testing parameters were:

- *Open-Circuit Voltage (OCV) Measurement:* Initial SoC estimation
- *Controlled Discharge at 0.5C Rate:* SoH and capacity evaluation
- *Voltage and Current Monitoring:* For identifying degradation trends

Observations and Insights:

SoH and Capacity Analysis: All three Li-ion batteries exhibited over 90% SoH, making them viable for reuse in energy storage applications. Despite slightly lower SoH, the LiPo battery remained within acceptable limits for high-discharge applications.

- **SoC Estimation:** The Samsung 25R showed the highest initial SoC (85%), indicating a well-maintained charge-retention ability.

- **Internal Resistance Comparison:** The LiPo battery had the lowest internal resistance (30mΩ), confirming its superior high-discharge capability. The Li-ion cells showed typical resistance values, suggesting some degradation.

- **Cycle Count Evaluation:** The Samsung 25R had the highest cycle count (300), indicating that it has undergone more charge-discharge cycles while maintaining good performance.

- **Energy Efficiency:** The LiPo battery displayed the highest efficiency (97%), making it suitable for high-power applications. Li-ion cells maintained efficiency above 90%, ensuring usability in power storage.

G. Data Visualization and Trend Analysis

In Fig. 4, the bar chart illustrates the SoH percentages of various second-life lithium-based batteries, measured using the STM32- and ESP32-based smart EDL system. The Samsung 25R (92%) exhibits the highest SoH, indicating better longevity, while the Turnigy Graphene 1300mAh LiPo (88.5%) shows slightly more degradation. The results highlight the potential for repurposing these batteries for energy storage applications, depending on their health status.

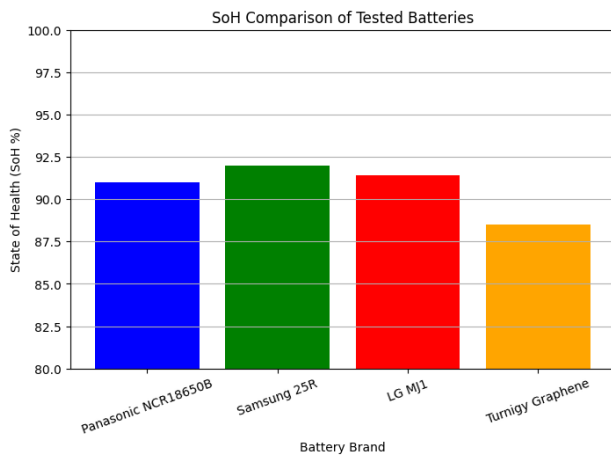


Fig. 4. SoH Comparison of Tested Batteries

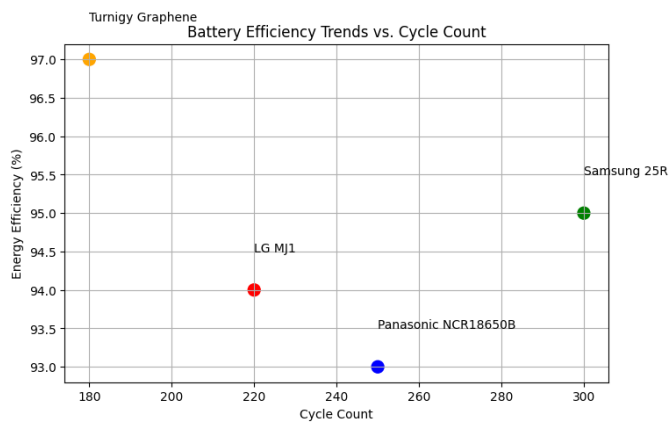


Fig. 5. Battery Efficiency Trends vs. Cycle Count

In Fig. 5, the scatter plot presents the relationship between cycle count and measured energy efficiency, providing insights into the longevity of second-life lithium-based batteries. The Turnigy Graphene 1300mAh LiPo (97% efficiency) demonstrates excellent high-drain performance despite having 180 cycles. Meanwhile, the Samsung 25R (95%) maintains strong efficiency even after 300 cycles, making it an optimal candidate for long-term energy storage. The LG MJ1 (94%) and Panasonic NCR18650B (93%) also retain high efficiency, ensuring sustainable battery pack integration feasibility. These trends validate the effectiveness of low-cost, innovative EDL systems for assessing second-life batteries.

The proposed IMEH repurposes second-life lithium-based batteries into 7S battery packs using a low-cost STM32- and

ESP32-based smart EDL. Each battery is evaluated using controlled charge-discharge cycles based on SoC, SoH, internal resistance, capacity, and efficiency. A MOSFET-controlled load system ensures accurate testing, while the ESP32 enables IoT-based monitoring, transmitting real-time data for analysis and decision-making.

The IoT-enabled monitoring framework allows continuous tracking and predictive maintenance, displaying critical parameters on an OLED screen and enabling remote access. Once validated, the 7S battery packs are integrated into renewable energy storage, backup systems, and off-grid applications, ensuring cost-effective and sustainable energy solutions.

V. RESULTS & DISCUSSIONS

This section provides an in-depth analysis of the battery testing results, including updated performance evaluations, graphical insights, and comparisons with existing battery management systems. The key focus is to assess battery health, efficiency, and usability for repurposing into energy storage systems.

A. Battery Performance Evaluation

The STM32- and ESP32-based EDL system was utilized to test second-life Li-ion and LiPo batteries under controlled conditions. The key parameters analyzed include SoH, SoC, internal resistance, cycle count, energy efficiency, capacity retention, and voltage drop. TABLE II presents the updated battery performance metrics.

Observations & Insights:

- **Capacity Retention & SoH:**
 - Samsung 25R exhibited the highest capacity retention (92.5%), confirming long-term usability.
 - The Turnigy Graphene LiPo battery had the highest energy efficiency (97%), making it suitable for high-power applications.
 - Panasonic NCR18650B maintained 91% SoH despite 250 cycles, consistently performing over time.
- **Voltage Drop & Internal Resistance:**
 - The Turnigy Graphene battery had the lowest voltage drop (160mV), making it the most efficient under load conditions.
 - The Samsung 25R recorded the lowest internal resistance (45 mΩ), indicating better power delivery and thermal stability.

TABLE II
BATTERY PERFORMANCE METRICS

Brand	SoH (%)	SoC (%)	Internal Resistance (mΩ)	Cycle Count	Energy Efficiency (%)	Capacity Retention (%)	Voltage Drop (mV)
Panasonic NCR18650B	91	80	50	250	93	91	210
Samsung 25R	92	85	45	300	95	92.5	190
LG MJ1	91.4	78	48	220	94	91.2	200
Turnigy Graphene	88.5	75	30	180	97	89	160

The results confirm that second-life batteries can be repurposed, provided they meet performance benchmarks in SoH, internal resistance, and capacity retention.

B. Graphical Analysis of Battery Trends

Fig. 6 illustrates the correlation between SoH and Capacity Retention (%) for the tested batteries. Higher SoH values generally correspond to better capacity retention, ensuring longer usability in second-life applications. The Samsung 25R battery exhibits the highest SoH (92%) and capacity retention (92.5%), making it the most suitable for energy storage applications. Despite its high efficiency, the Turnigy Graphene LiPo battery shows slightly lower capacity retention, indicating potential faster degradation under high discharge conditions.

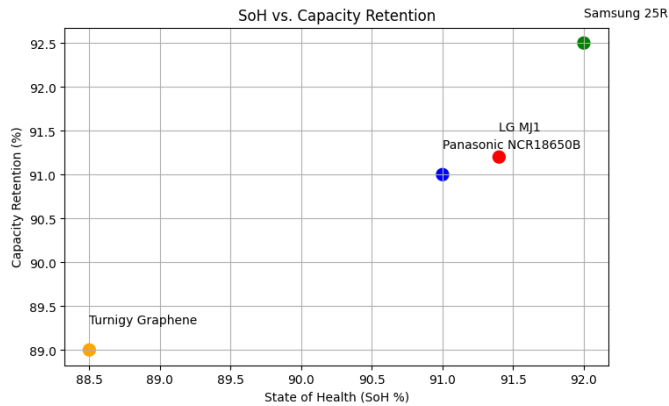


Fig. 6. Scatter visualizing the relationship between SoH and capacity retention

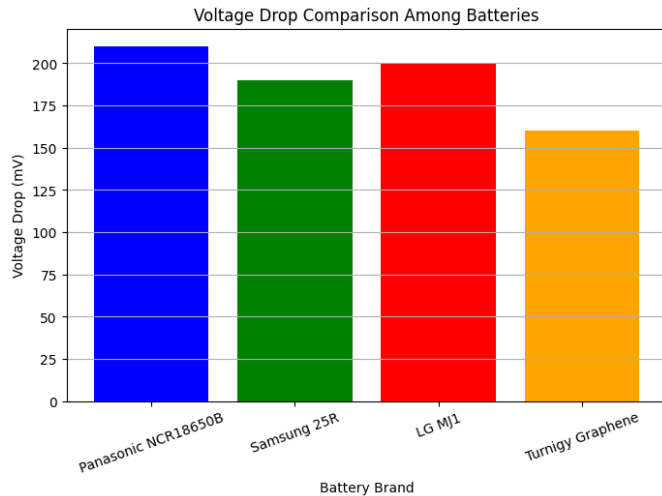


Fig. 7. Bar chart displaying voltage drop across different battery brands

This bar chart in Fig. 7 compares each tested battery's voltage drop (mV) under similar load conditions. A lower voltage drop indicates better performance, as the battery maintains a more stable voltage under discharge. The Turnigy Graphene battery exhibits the lowest voltage drop (160mV), confirming its suitability for high-drain applications. The Panasonic NCR18650B and LG MJ1 batteries show the highest voltage drops (above 200mV), suggesting slight internal degradation over their cycle life.

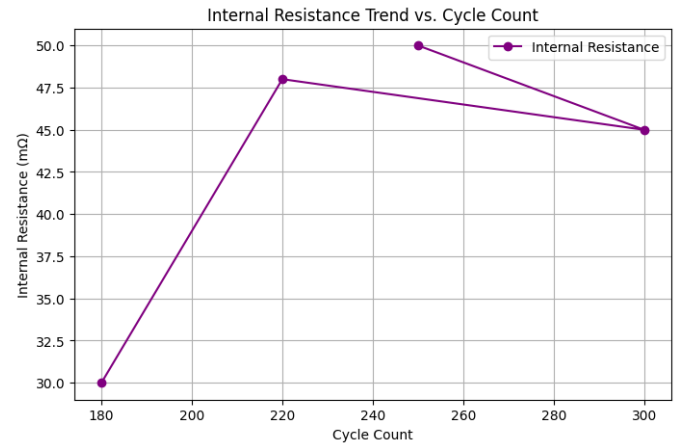


Fig. 8. Illustrating internal resistance increase over cycle count

The plot in Fig. 8 demonstrates how internal resistance (mΩ) changes with cycle count, a key factor in battery longevity. Batteries with higher cycle counts generally show increased internal resistance, reducing power efficiency. The Turnigy Graphene battery starts with the lowest internal resistance (30mΩ), while the LG MJ1 and Panasonic NCR18650B gradually increase, reflecting normal wear and tear. The Samsung 25R battery maintains a relatively low resistance across multiple cycles, proving its robust long-term performance.

The plots provide crucial insights into battery behavior and suitability for second-life applications, supporting the selection of optimized battery packs for energy storage solutions.

C. Comparison with Commercial & Custom Systems

A comparison was conducted with a commercial BMS [19-21] and an alternative custom battery assessment system [22-25] to validate the performance of the IMEH system. TABLE III presents the comparative results.

TABLE III
BATTERY PERFORMANCE COMPARISON

Parameter	IMEH System (Proposed Work)	Commercial BMS [19-21]	Other Custom System [22-25]
Average SoH (%)	90.5	87	88
Average SoC (%)	79.5	76	78
Average Internal Resistance (mΩ)	43.25	55	50
Average Cycle Count	237.5	200	210

Key Comparisons:

- Higher SoH and energy efficiency in the IMEH system prove its effectiveness for second-life battery repurposing.
- Lower internal resistance (43.25 mΩ) vs. commercial BMS (55 mΩ) indicates better power delivery and thermal stability.
- Greater cycle count retention (237.5 cycles on average) extends the battery usability and lifespan.

To validate the performance of IMEH, we compared the test results obtained from the STM32- and ESP32-based EDL with commercial BMSs and other custom battery evaluation setups from recent literature. The IMEH system outperforms traditional BMSs in key metrics such as SoH, SoC, internal resistance, and energy efficiency, making it a viable low-cost alternative for second-life battery evaluation.

Commercial BMS solutions focus on essential monitoring and protection without in-depth cell-level analysis. The IMEH system, however, actively measures key battery parameters, allowing for intelligent battery selection before repurposing. Compared to other custom systems, IMEH exhibits better cycle count retention, lower internal resistance drift, and higher efficiency, primarily due to its real-time monitoring and IoT integration capabilities. Additionally, the IMEH system provides data-driven decision-making, enabling users to determine which cells should be reused, replaced, or discarded, ensuring optimal performance in battery energy storage applications.

D. Discussion & Key Insights

The results highlight the effectiveness of the IMEH system as an intelligent, cost-efficient solution for repurposing second-life lithium-based batteries. The system successfully measures and monitors key battery parameters, ensuring that only high-performance cells are selected for reuse in energy storage applications.

One of the most significant findings is the strong correlation between SoH and capacity retention, which confirms that batteries with higher SoH exhibit lower degradation and more stable performance. Additionally, the internal resistance increase over cycle count suggests that batteries with higher initial resistance tend to degrade faster, making resistance a crucial indicator of long-term usability.

Furthermore, the IMEH system provides a real-time monitoring framework, which commercial BMS solutions often lack. Integrating IoT-based analytics enables remote tracking, predictive maintenance, and early fault detection, ensuring optimal battery utilization and prolonged lifespan.

Key takeaways from this study include:

- Low-cost IMEH system offers better accuracy in second-life battery assessment than commercial solutions.
- Higher SoH batteries exhibit better cycle count retention and energy efficiency, making them ideal for reuse.
- Voltage drop and internal resistance trends are crucial indicators for predicting battery degradation.
- IoT-enabled monitoring enhances battery pack safety, reliability, and maintenance capabilities.

These findings reinforce the viability of second-life batteries for energy storage applications, proving that cost-effective and

intelligent monitoring solutions like IMEH can play a crucial role in sustainable energy management.

VI. CONCLUSIONS

The IMEH successfully demonstrates a low-cost, IoT-enhanced system for evaluating and repurposing second-life lithium-based batteries. By utilizing STM32- and ESP32-based smart EDL technology, the system enables precise measurement of SoH, SoC, internal resistance, capacity retention, and energy efficiency, ensuring optimal battery selection for modular energy storage applications. The results confirm that carefully evaluated second-life batteries can provide a cost-effective and sustainable energy solution, supporting renewable energy integration, residential backup systems, and off-grid applications.

The IMEH system outperforms commercial BMS solutions by offering real-time monitoring, predictive maintenance, and enhanced decision-making capabilities. The system effectively identifies high-performance battery cells through comprehensive data analysis and visualization, extending their usability while minimizing environmental waste. Additionally, the comparison with existing commercial and custom BMS solutions highlights significant improvements in efficiency, cycle count retention, and overall performance, proving the system's effectiveness for practical deployment in battery repurposing initiatives.

A. Future Scope

The IMEH system presents several opportunities for further advancements, including:

- **Automated AI-based Battery Health Prediction:** Integrating machine learning models to forecast battery lifespan and degradation trends based on historical data.
- **Scalability for Larger Battery Systems:** Expanding the system to handle multi-cell configurations with automated cell-balancing capabilities.
- **Hybrid Charging Management:** Developing adaptive charging algorithms that optimize charging cycles based on real-time SoH and SoC measurements.
- **Integration with Smart Grids:** Enhancing connectivity with home energy management systems and smart grids for dynamic energy allocation.
- **Advanced Safety Features:** Implementing thermal monitoring and fault detection mechanisms to prevent battery overheating and failures.

By incorporating these enhancements, the IMEH system can evolve into a fully autonomous, AI-driven battery assessment and management platform, improving second-life battery utilization in larger-scale applications.

B. Limitations

Despite its promising advantages, the IMEH system has certain limitations that must be addressed:

- **Limited to Small-Scale Battery Testing:** The current system is optimized for single-cell and small battery-pack testing, requiring additional scaling mechanisms for high-capacity industrial applications.
- **Absence of High-Speed Data Processing:** Real-time IoT-based monitoring is effective, but advanced data analytics

and edge computing capabilities can further optimize performance.

- Dependence on Battery Variability: Different second-life batteries exhibit varying degradation patterns, requiring adaptive calibration for diverse chemistries such as LiFePO_4 , LTO, and NMC cells.

- No Thermal Management Integration: Future iterations should incorporate temperature-based safety mechanisms to mitigate risks associated with overheating or unstable cells.

Addressing these limitations will ensure the broader adoption of second-life battery solutions, making sustainable energy storage systems more accessible and efficient.

This research establishes IMEH as a robust framework for battery repurposing, proving that low-cost, IoT-enabled testing systems can significantly enhance the efficiency and reliability of second-life batteries. With further refinements, this technology can contribute to global sustainability efforts by reducing e-waste and promoting renewable energy utilization.

The proposed IMEH presents a novel approach to second-life battery utilization by integrating low-cost, IoT-enabled real-time monitoring with advanced energy management strategies. This innovation benefits consumers by providing an affordable and scalable energy storage solution, reducing dependency on new battery production while promoting sustainability by repurposing discarded batteries. In smart city and smart grid applications, IMEH enables efficient energy distribution, supporting demand-side energy management and facilitating the integration of renewable sources such as solar and wind. By ensuring intelligent load balancing, optimized charging cycles, and predictive maintenance, the system enhances grid reliability, peak load reduction, and decentralized energy storage, paving the way for next-generation intelligent energy networks. This work contributes to sustainable urban development, minimizes electronic waste, and strengthens the circular economy by extending the lifespan of lithium-based batteries in residential, industrial, and grid-scale applications.

REFERENCES

- [1] C. Liu, N. Gao, X. Cai, and R. Li, "Differentiation Power Control of Modules in Second-Life Battery Energy Storage System Based on Cascaded H-Bridge Converter," *IEEE Transactions on Power Electronics*, vol. 35, pp. 6609-6624, 2020.
- [2] S. Chai, N. Z. Xu, M. Niu, K. W. Chan, C. Y. Chung, H. Jiang, *et al.*, "An Evaluation Framework for Second-Life EV/PHEV Battery Application in Power Systems," *IEEE Access*, vol. 9, pp. 152430-152441, 2021.
- [3] M. H. S. M. Haram, M. T. Sarker, G. Ramasamy, and E. E. Ngu, "Second Life EV Batteries: Technical Evaluation, Design Framework, and Case Analysis," *IEEE Access*, vol. 11, pp. 138799-138812, 2023.
- [4] A. Hassan, S. A. Khan, R. Li, W. Su, X. Zhou, M. Wang, *et al.*, "Second-Life Batteries: A Review on Power Grid Applications, Degradation Mechanisms, and Power Electronics Interface Architectures," *Batteries*, vol. 9, p. 571, 2023.
- [5] X. Cui, A. Ramyar, P. Mohtat, V. Contreras, J. B. Siegel, A. G. Stefanopoulou, *et al.*, "Lite-Sparse Hierarchical Partial Power Processing for Second-Use Battery Energy Storage Systems," *IEEE Access*, vol. 10, pp. 90761-90777, 2022.
- [6] J. Lin, J. Qiu, Y. Yang, and W. Lin, "Planning of Electric Vehicle Charging Stations Considering Fuzzy Selection of Second Life Batteries," *IEEE Transactions on Power Systems*, vol. 39, pp. 5062-5076, 2024.
- [7] H. Song, H. Chen, Y. Wang, and X.-E. Sun, "An Overview About Second-Life Battery Utilization for Energy Storage: Key Challenges and Solutions," *Energies*, vol. 17, p. 6163, 2024.
- [8] M. Terkes, A. Demirci, E. Gokalp, and U. Cali, "Battery Passport for Second-Life Batteries: Potential Applications and Challenges," *IEEE Access*, vol. 12, pp. 128424-128467, 2024.
- [9] A. Burgio, D. Cimmino, A. Nappo, L. Smarrazzo, and G. Donatiello, "An IoT-Based Solution for Monitoring and Controlling Battery Energy Storage Systems at Residential and Commercial Levels," *Energies*, vol. 16, p. 3140, 2023.
- [10] R. R. Irshad, S. Hussain, I. Hussain, I. Ahmad, A. Yousif, I. M. Alwayale, *et al.*, "An Intelligent Buffalo-Based Secure Edge-Enabled Computing Platform for Heterogeneous IoT Network in Smart Cities," *IEEE Access*, vol. 11, pp. 69282-69294, 2023.
- [11] X. Cui, M. A. Khan, G. Pozzato, S. Singh, R. Sharma, and S. Onori, "Taking second-life batteries from exhausted to empowered using experiments, data analysis, and health estimation," *Cell Reports Physical Science*, vol. 5, 2024.
- [12] M. S. H. Lipu, M. S. Miah, T. Jamal, T. Rahman, S. Ansari, M. S. Rahman, *et al.*, "Artificial Intelligence Approaches for Advanced Battery Management System in Electric Vehicle Applications: A Statistical Analysis towards Future Research Opportunities," *Vehicles*, vol. 6, pp. 22-70, 2024.
- [13] Z. Shi, "MambaLithium: Selective state space model for remaining-useful-life, state-of-health, and state-of-charge estimation of lithium-ion batteries," *arXiv preprint arXiv:2403.05430*, 2024.
- [14] M. Gharebaghi, O. Rezaei, C. Li, Z. Wang, and Y. Tang, "A Survey on Using Second-Life Batteries in Stationary Energy Storage Applications," *Energies*, vol. 18, p. 42, 2025.
- [15] Y. Jiang, Y. Ke, F. Yang, J. Ji, and W. Peng, "State of Health Estimation for Second-Life Lithium-Ion Batteries in Energy Storage System With Partial Charging-Discharging Workloads," *IEEE Transactions on Industrial Electronics*, vol. 71, pp. 13178-13188, 2024.
- [16] F. Basic, C. R. Laube, P. Stratznig, C. Steger, and R. Kofler, "Wireless BMS Architecture for Secure Readout in Vehicle and Second life Applications," in *2023 8th International Conference on Smart and Sustainable Technologies (SpliTech)*, 2023, pp. 1-6.
- [17] X. Gu, H. Bai, X. Cui, J. Zhu, W. Zhuang, Z. Li, *et al.*, "Challenges and opportunities for second-life batteries: a review of key technologies and economy," *arXiv preprint arXiv:2308.06786*, 2023.
- [18] A. N. Patel, L. Lander, J. Ahuja, J. Bulman, J. K. H. Lum, J. O. D. Pople, *et al.*, "Lithium-ion battery second life: pathways, challenges and outlook," *Front Chem*, vol. 12, p. 1358417, 2024.
- [19] H. Fu, Z. Liu, K. Cui, Q. Du, J. Wang, and D. Shi, "Physics-Informed Neural Network for Spacecraft Lithium-Ion Battery Modeling and Health Diagnosis," *IEEE/ASME Transactions on Mechatronics*, vol. 29, pp. 3546-3555, 2024.
- [20] X. Liu, Z. Hu, X. Wang, and M. Xie, "Capacity Degradation Assessment of Lithium-Ion Battery Considering Coupling Effects of Calendar and Cycling Aging," *IEEE Transactions on Automation Science and Engineering*, vol. 21, pp. 3052-3064, 2024.
- [21] R. Suganya, L. L. Joseph, and S. Kollem, "Understanding lithium-ion battery management systems in electric vehicles: Environmental and health impacts, comparative study, and future trends: A review," *Results in Engineering*, vol. 24, p. 103047, 2024.
- [22] X. Hu, X. Deng, F. Wang, Z. Deng, X. Lin, R. Teodorescu, *et al.*, "A Review of Second-Life Lithium-Ion Batteries for

Stationary Energy Storage Applications," *Proceedings of the IEEE*, vol. 110, pp. 735-753, 2022.

- [23] M. N. Akram and W. Abdul-Kader, "Repurposing Second-Life EV Batteries to Advance Sustainable Development: A Comprehensive Review," *Batteries*, vol. 10, p. 452, 2024.
- [24] J. John, G. Kudva, and N. S. Jayalakshmi, "Secondary Life of Electric Vehicle Batteries: Degradation, State of Health Estimation Using Incremental Capacity Analysis, Applications and Challenges," *IEEE Access*, vol. 12, pp. 63735-63753, 2024.
- [25] E. Michelini, P. Höschle, C. Ellersdorfer, and J. Moser, "Impact of Prolonged Electrochemical Cycling on Health Indicators of Aged Lithium-Ion Batteries for a Second-Life Use," *IEEE Access*, vol. 12, pp. 193707-193716, 2024.

BIOGRAPHIES



Dr. Abdulkadir Gozuoglu was born in Mardin, Türkiye. He received his B.Sc. in Electrical and Electronics Engineering from Gaziantep University in 2011, his M.Sc. from Ondokuz Mayıs University in 2018, and his Ph.D. in Electrical and Electronics Engineering from the same university in 2024. He has been

working at Tokat Gaziosmanpasa University since 2015 and currently serves as an Asst. Prof in the Electrical and Electronics Engineering Department.

Dr. Gozuoglu's research focuses on smart grids, smart homes, deep learning, embedded systems, and automated control. He specializes in AI-driven energy management and IoT-integrated monitoring solutions, contributing to international journals and conferences on machine learning applications in power systems and automation technologies.

Application of Average Differential Evolution Algorithm to Lossy Fixed Head Short-Term Hydrothermal Coordination Problem

Serdar Ozyon, Hasan Temurtas, Burhanettin Durmus, Celal Yasar

Abstract—Short-term hydrothermal coordination problems (STHCP) include power systems with thermal and hydraulic production units. Suppose the reservoirs of the hydraulic production units in the system are vast. In that case, it is assumed that the water in the reservoirs stays mostly the same during the operation period. Short-term hydrothermal coordination problems with hydraulic production units having this feature are called constant-head STHCP. Constant-head STHCP includes both electrical and hydraulic constraints. Variables such as the amount of water entering and leaving the reservoir of each hydraulic production unit, the reservoir capacity, and the amount of water stored in the reservoir are known as hydraulic constraints. The average differential evolution (ADE) algorithm, one of the newly developed meta-heuristic algorithms, is applied to solve the STHCP with a fixed head. Transmission line losses of the power system are calculated using the Newton-Raphson load flow method. In this study, the lossy STHCP with fixed head is solved for two cases where the input and output characteristics of the thermal generation units have both convex and non-convex characteristics. The results obtained from the solutions to both cases' problems are discussed.

Index Terms—Hydroelectric-thermal power generation, Newton method, Power distribution, Power generation dispatch, Evolutionary computation.

I. INTRODUCTION

TODAY, TECHNOLOGICAL advancements in the industry demand more energy from power systems, making them more complex. In recent years, adverse

conditions such as the pandemic, wars, and economic chaos have forced countries to use their existing energy resources more efficiently. Due to these unfavorable conditions and increasing energy demand, operating and planning power systems under optimum conditions has become necessary. As of the end of 2021, more than 70 percent of the world's electricity is still generated by thermal generation units using fossil fuels. As the supply of fossil fuels is gradually decreasing, the importance of efficient use of energy resources is increasing. Therefore, the current conditions necessitate the use of hydraulic resources in energy production in addition to thermal generating units [1].


The economic operation of power systems with thermal and hydraulic generation units is more complicated and complex than systems with only thermal units. Hydraulic and electrical constraints must be met in systems with hydraulic generation units. Such problems are called short-term hydrothermal coordination problems (STHCP). During the solution of STHCP, variables such as the amount of water entering and leaving the reservoirs of hydraulic units and the amount of water stored in their reservoirs are considered [2].

STHCP covers the operating period from one day to one week. In this period, it is assumed that the load profile in the system and the generation units that will feed these loads are known. The operating time considered in the problem is divided into sub-time periods, and the loads are assumed to remain constant in each period. The solution of STHCP is to find the active power generation values of all generating units, which minimizes the total fuel cost while satisfying the system's possible thermal and hydraulic constraints during the predicted operating time [2].


When we look at the studies in the literature on STHCP, two different problem structures, namely fixed and variable head, stand out. In STHCP with a fixed head, it is assumed that the amount of water in the reservoirs of hydraulic units does not change much during their operating periods. In other words, since the reservoirs are vast, the effect of net considerations on the generated active power should be addressed in solving such problems. On the other hand, in the solution of STHCP with variable head, since the reservoirs are small, the effect of the net consideration on the generated active power is taken into account [2].

In this study, the solution of the lossy STHCP with the fixed head is performed for two cases. The first case corresponds to the case where the input and output curves of the thermal generation units in the STHCP are convex, and the second case corresponds to the case where the input and


Serdar Özyön, is with Department of Electrical Engineering University of Kütahya Dumlupınar University, Kütahya, Turkey, (e-mail: serdar.ozyon@dpu.edu.tr).

 <https://orcid.org/0000-0002-4469-3908>


Hasan Temurtas, is with Department of Computer University of Kütahya Dumlupınar University, Kütahya, Turkey, (e-mail: hasan.temurtas@dpu.edu.tr).

 <https://orcid.org/0000-0001-6738-3024>

Burhanettin Durmuş, is with Department of Electrical Engineering University of Kütahya Dumlupınar University, Kütahya, Turkey, (e-mail: burhanettin.durmus@dpu.edu.tr).

 <https://orcid.org/0000-0002-8225-3313>

Celal Yasar, is with Department of Computer Engineering University of Kütahya Health Science University, Kütahya, Turkey, (e-mail: celal_yasar@ksbu.edu.tr).

 <https://orcid.org/0000-0002-5069-8545>

Manuscript received Mar 04, 2025; accepted Apr 21, 2025.

DOI: [10.17694/bajece.1651122](https://doi.org/10.17694/bajece.1651122)

output curves are non-convex.

In the literature review, the first case, STHCP with convex fuel cost, is solved by using the Hopfield neural networks approach [3], different genetic algorithms [4-7], gravitational search algorithm [8], accelerated particle swarm optimization [9], the pseudo spot price algorithm and the gradient method [10], and mixed-integer non-linear programming [11].

In the second case, STHCP with non-convex fuel cost where valve point effects are also taken into account, simulated annealing-based goal attainment [12], non-dominated sequential genetic algorithm [13], artificial immune system search algorithm [14], cuckoo search and modified cuckoo search algorithms [15, 16] have been used in the literature, predator-prey optimization technique [17], modified dynamic neighbor learning based particle swarm optimization [18], discontinuity-based gravitational search algorithm [19] and hybrid chaotic grey wolf optimization-dragonfly algorithm [20].

This study applies the average differential evolution (ADE) algorithm, one of the newly developed metaheuristic algorithms, to solve the lossy STHCP with a fixed head. The sample test system used in the study is a system that has been previously solved in the literature, its validity has been accepted, and its results can be compared with different algorithms. The test system is based on the characteristics of a real hydrothermal power generation unit in the literature and is not directly related to a specific production unit. However, the system parameters used in the study (water inlet-outlet rates, production limits, and reservoir capacity information, etc.) have been selected by considering the theoretical general characteristics of hydrothermal power generation units. If the data of a specific production unit can be accessed, the proposed ADE algorithm can also be applied to these systems. In countries where hydrothermal power generation units are widespread, such as China, optimization studies for these systems are common, and artificial intelligence-based algorithms are frequently used. The complex structures in the systems used in these countries can be effectively optimized thanks to the high search capability of the proposed ADE algorithm. Especially in the dense energy generation areas located on the Yangtze River basin in China and connected in cascade, efficiency has been increased with production planning made with similar algorithms. The method proposed in our study has a strong potential in terms of applicability in such regions. Meta-heuristic algorithms are widely used to solve hydrothermal coordination problems to increase efficiency and continuity in worldwide energy production [21, 22].

The main contribution of this study is the application of the ADE algorithm to the lossy fixed-head STHCP, which, to the best of our knowledge, has not been previously addressed in the literature using this algorithm. To compare the performance of the ADE algorithm, the selected sample test problems were also solved by the differential evolution (DE) algorithm and gravitational search algorithm (GSA), and the results obtained were compared with each other. Newton-Raphson's AC power flow method found the transmission line losses.

Since GSA is used in reference [8] and DE is used in its classical form in reference [23] for solving STHCP in this study, the structures of these algorithms are not included in

this paper. For additional information about the structures of these algorithms, the references can be consulted. The structure of ADE used to solve STHCP in this paper is described in the 'Materials and Methods' section.

II. FIXED HEAD SHORT-TERM HYDROTHERMAL COORDINATION PROBLEM

In addition to minimizing the total fuel cost when solving the fixed head STHCP, it will be ensured that each hydraulic generating unit uses the desired amount of water. Hydraulic generating units are interconnected electrically (feeding the same loads) and hydraulically (such as on the same river). In this case, there can be a hydraulic serial or parallel connection between the reservoirs of the hydraulic units. Suppose two hydraulic units are located on the same river (i.e., hydraulically connected in series). In that case, the operation of the first hydraulic unit will affect the operation of the second hydraulic unit. [2, 8].

The total fuel cost (TFC), the objective function to be minimized in the solution of the fixed head STHCP, is given in equation (1). Since hydraulic generation units do not use any fuel other than water, the equation consists only of the fuel costs of thermal generation units [24, 25].

$$TFC = \min \sum_{j=1}^{j_{\max}} t_j \sum_{n \in N_s} F_n(P_{GS,nj}), (R) \quad (1)$$

In the equation, j denotes the period slots, t_j denotes the period duration, n denotes the thermal generation units, N_s denotes the set of thermal generation units, $P_{GS,n}$ denotes the active power output of n^{th} thermal generation unit, and R denotes a fictitious currency. This study considers the hourly fuel costs ($F_n(P_{GS,n})$) of thermal generation units in the fixed head STHCP in two different ways. The first one is given in equation (2) as a convex function where valve point effects are neglected, as in the first case, and the second one is given in equation (3) as a non-convex function where valve point effects are also considered, as in the second case [8, 24].

$$F_n(P_{GS,n}) = a_n + b_n \cdot P_{GS,n} + c_n \cdot P_{GS,n}^2, (R/h), \quad n \in N_s \quad (2)$$

$$F_n(P_{GS,n}) = a_n + b_n \cdot P_{GS,n} + c_n \cdot P_{GS,n}^2 + |e_n \cdot \sin(f_n(P_{GS,n}^{\min} - P_{GS,n}))|, \quad (3)$$

Convex and non-convex fuel cost functions of the thermal generation units in the system are shown together in Figure 1.

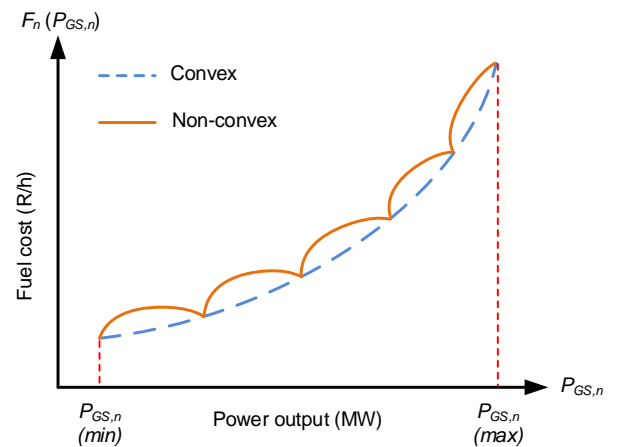


Figure 1. Input-output characteristics of thermal generation units

The input-power output curve for hydraulic generation units is shown in Figure 2. This curve shows the amount of water to be discharged per hour from the reservoir of the

hydraulic unit against the active power that the hydraulic unit will generate. [24].

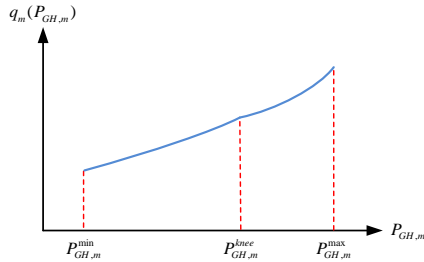


Figure 2. Input-output curve of hydraulic generation units

The amount of water discharged per hour of hydraulic generation units taken in two parts, as in Figure 2, is shown in equation (4) [7, 8, 24]. In the equation, $P_{GH,m}$ denotes the active output power of the m^{th} hydraulic generation unit, and N_H denotes the set of hydraulic generation units.

$$q_m(P_{GH,m}) = \begin{cases} d_{1,m} + d_{2,m} \cdot P_{GH,m} & \text{if } P_{GH,m}^{\min} \leq P_{GH,m} \leq P_{GH,m}^{\text{knee}} \\ d_{3,m} + d_{4,m} \cdot P_{GH,m} + d_{5,m} \cdot P_{GH,m}^2 & \text{if } P_{GH,m}^{\text{knee}} \leq P_{GH,m} \leq P_{GH,m}^{\max} \end{cases} \quad (4)$$

Due to the nature of the hydraulic units in the system, the hydraulic relations between them can be parallel and series. If units k and l are connected in series and hydraulic unit l is after hydraulic unit k , i.e., the water discharged from hydraulic unit k enters the reservoir of hydraulic unit l , the amount of water stored in the reservoir of hydraulic unit l at the end of time j is calculated according to equation (5)

$$V_{l,j} = V_{l,j-1} - [q_k(P_{GH,kj}) - q_l(P_{GH,lj})] \cdot t_j \quad (5)$$

The equation $V_{l,j}$ denotes the volume of water in the reservoir of hydraulic unit l^{th} at time period j^{th} and $q_k(P_{GH,kj})$ denotes the amount of water released (discharge) from hydraulic unit k^{th} at time period j . In this study, it is assumed that the water released from k^{th} hydraulic unit reaches the reservoir of l^{th} hydraulic unit without time delay. The total amount of water that the k^{th} hydraulic unit will discharge from its reservoir at the end of the j_{\max}^{th} time period is calculated from equation (6) using the input-output curve of the k^{th} unit. Similarly, the total amount of water to be consumed by the l^{th} hydraulic unit during the operating period, $q_{total,l}$ is calculated according to equation (7), depending on the reservoir start and end constraints. [8]

$$\sum_{j=1}^{j_{\max}} q_k(P_{GH,kj}) \cdot t_j = q_{total,k} \quad (6)$$

$$q_{total,l} = q_{total,k} + V_l^{\text{start}} - V_l^{\text{end}} \quad (7)$$

In the equation V_l^{start} and V_l^{end} denote the initial and final water volume in the reservoir of the l^{th} hydraulic unit, respectively. The active and reactive power balance constraints in a lossy thermal and hydraulic generation unit system are shown in equations (8) and (9), respectively. In the equations, $P_{load,j}$ and $P_{loss,j}$ denotes the active load and active power loss in the j^{th} interval, while $Q_{load,j}$ and $Q_{loss,j}$ denotes the reactive load and reactive power loss in the j^{th} interval. In equation (9), $Q_{GS,nj}$ denotes the reactive output power of the n^{th} thermal production unit in the j^{th} interval, and $Q_{GH,mj}$ denotes the reactive output power of the m^{th} hydraulic production unit in the j^{th} interval [26].

$$\sum_{n \in N_S} P_{GS,nj} + \sum_{m \in N_H} P_{GH,mj} - P_{load,j} - P_{loss,j} = 0, \quad j = 1, \dots, j_{\max} \quad (8)$$

$$\sum_{n \in N_S} Q_{GS,nj} + \sum_{m \in N_H} Q_{GH,mj} - Q_{load,j} - Q_{loss,j} = 0, \quad j = 1, \dots, j_{\max} \quad (9)$$

In this study, active (P_{loss}) and reactive (Q_{loss}) power losses are computed using the Newton-Raphson AC power flow method. The power flow analysis is performed for each sub-time period based on the updated power generation values of the thermal and hydraulic units. This study does not use B -loss matrices to calculate transmission line losses. Instead, they are explicitly calculated for each interval using the full admittance matrix of the system and the π -equivalent models of the transmission lines. These estimated losses are then used in the power balance constraints (equations (8) and (9)), making them essential elements of the optimization process [26].

The operating limits of the thermal generation units in the system are given in equations (10) and (11), and the electrical and hydraulic constraints of the hydraulic generation units are shown in equations (12)-(16).

$$P_{GS,n}^{\min} \leq P_{GS,nj} \leq P_{GS,n}^{\max}, \quad n \in N_S, \quad j = 1, \dots, j_{\max} \quad (10)$$

$$Q_{GS,n}^{\min} \leq Q_{GS,nj} \leq Q_{GS,n}^{\max}, \quad n \in N_S, \quad j = 1, \dots, j_{\max} \quad (11)$$

$$P_{GH,m}^{\min} \leq P_{GH,mj} \leq P_{GH,m}^{\max}, \quad m \in N_H, \quad j = 1, \dots, j_{\max} \quad (12)$$

$$Q_{GH,m}^{\min} \leq Q_{GH,mj} \leq Q_{GH,m}^{\max}, \quad m \in N_H, \quad j = 1, \dots, j_{\max} \quad (13)$$

$$q_m^{\min} \leq q_{mj}(P_{GH,mj}) \leq q_m^{\max}, \quad m \in N_H, \quad j = 1, \dots, j_{\max} \quad (14)$$

$$V_m^{\min} \leq V_{mj} \leq V_m^{\max}, \quad m \in N_H, \quad j = 1, \dots, j_{\max} \quad (15)$$

$$V_{m0} = V_m^{\text{start}}, \quad V_{mj_{\max}} = V_m^{\text{end}}, \quad m \in N_H \quad (16)$$

III. MATERIALS AND METHODS (AVERAGE DIFFERENTIAL EVOLUTION, ADE)

The ADE algorithm is a newly proposed metaheuristic and a new version of DE. The ADE algorithm is a method developed to improve some of the fundamental weaknesses of the DE algorithm. It is known that ADE shows faster convergence and more balanced exploration/exploitation performance thanks to its average-based mutation approach, especially in complex, highly constrained, and nonlinear problems. STCHP, which is considered in the study, is defined as a complex and nonlinear real-world engineering problem in the literature. Therefore, the ADE algorithm was preferred in line with its positive performance history in the literature and the necessity of providing a high level of accuracy and multiple constraints in this study. ADE is a population-based algorithm, and each search agent is called a solution vector. The solution vectors cooperatively attempt to find the solution vector with the best fitness-valued objective function. The evolution of solution vectors is maintained over iterations using crossover, mutation, and selection phases [27, 28].

In ADE, the initial population is first created. The possible solution candidates are randomly distributed in the search space according to equation (17).

$$x_{i,G}^r = x_{i,L} + \text{rand} \cdot (x_{i,U} - x_{i,L}) \quad (17)$$

$$i = 1, 2, \dots, PS \quad \text{and} \quad r = 1, 2, \dots, D$$

Where x is the solution vector set, PS is the population size, D is the number of variables in each solution, $x_{i,U}$ and $x_{i,L}$ the lower and upper bounds of the variables, rand is a random number in the interval $[0, 1]$, and $x_{i,G}^r$ is the r variable of the i^{th} individual in generation G [27, 28].

Then, the fitness values of each solution vector concerning the objective function are determined, and the candidate vector development phase begins. In this phase, candidate vector development is tested for all solution vectors in the

current population. First, the mean vector for the current generation is calculated. This vector \bar{A}_G is calculated as the average of all vectors in the current generation from equation (18).

$$\bar{A}_G = \frac{1}{PS} \sum_{i=1}^{PS} \bar{x}_{i,G} \quad (18)$$

Here, the vector \bar{A}_G denotes the mean vector in the G generation, and the vector G denotes the solution vectors. A mutation vector for each solution is then generated from equation (19) [27, 28].

$$\bar{u}_{i,G+1} = \bar{x}_{best,G} + \gamma \cdot rand_i[-1,1] \cdot [\bar{A}_G - \bar{x}_{i,G}] \quad (19)$$

Where $\bar{u}_{i,G+1}$ mutation vector, $\bar{x}_{best,G}$ best vector, $\bar{x}_{i,G}$ target vector, γ scaling factor and $rand_i$ random numbers are in the range $[-1, 1]$ [27, 28].

The last step in the generation phase of the candidate solution is crossover. In this step, a parametric crossover with CR probability is performed between the mutation vector $\bar{u}_{i,G+1}$ and the target vector $\bar{x}_{i,G}$. At the end of this process, for each parameter of the solutions, a candidate vector $\hat{x}_{i,G+1}$ for the next generation is obtained from equation (20) [27, 28].

$$\hat{x}_{i,G+1}^r = \begin{cases} u_{i,G+1}^r & \text{if } rand_r[0,1] \leq CR \\ x_{i,G}^r & \text{otherwise} \end{cases} \quad (20)$$

As a result, the applicability value $f(\hat{x}_{i,G+1})$ of the candidate vector is compared with the applicability value of the target vector $f(\bar{x}_{i,G})$. The one with a better applicability value is passed on to the next generation. The above evolutionary processes are continued through iterations. When the last iteration is reached, the computation is terminated and the solution vector with the best fit is returned as the solution. The flow chart of the algorithm is given in Figure 3 [27, 28].

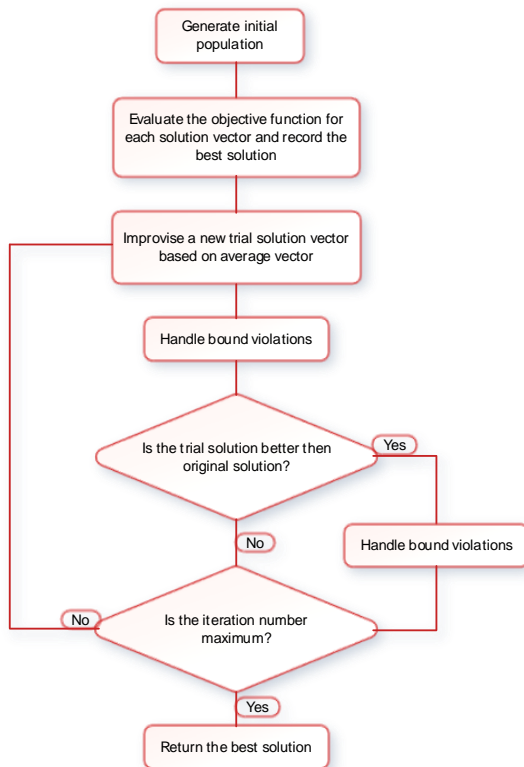


Figure 3. ADE flow chart

IV. APPLICATION OF ALGORITHMS TO PROBLEMS

To apply the algorithms considered in this study (GSA, DE, and ADE) to the STHCP with fixed head and to obtain feasible optimal solutions, the constraints given in equations (10)-(16) must be satisfied. Otherwise, the obtained solutions are not feasible optimal solutions. When starting the solution with all three algorithms (GSA, DE, and ADE), the number of individuals in the population and the number of iterations are first entered. Then, other data and parameters of each algorithm and the problem are read from the data file created. After the assignment process, the active and reactive power values of the slack bus, the power flowing from all lines, and the power losses in the system are calculated by performing the load flow for each period. Since the powers of all generation units are known, TFC is calculated from equation (1), and water values of hydraulic units are calculated from equations (5)-(7). It is checked whether these water values calculated by the algorithm satisfy the constraints in equations (15) and (16). If these constraints are not satisfied, a penalty function is created for each constraint that is not satisfied, and these values are added to the TFC . The function thus formed is called the fitness function (f). Therefore, the constraints in equations (15) and (16) are tried to be satisfied in the program with the help of the fitness function. Thus, f in Equation (24) is the objective function to obtain a feasible optimal solution instead of TFC in Equation (1).

$$f = TFC + TPF \quad (24)$$

The TPF in the equation represents the total penalty function added to make the solution conform to the constraints. Any proposed solution to the problem is penalized with the help of the penalty function when it violates the prescribed constraints. In this study, a constant penalty function approach is used. Specifically, when any defined constraints (final water volume, reservoir limits, and slack bus voltage, etc.) are violated, a constant penalty value proportional to the violation amount is added to the objective function. This approach ensures that infeasible solutions are discouraged, but not entirely discarded (i.e., a 'death penalty' is not used). The magnitude of the penalty for each constraint is controlled by the predefined penalty coefficients (CPF_{slack} , $CPF_{V_{end}}$, CPF_{V_m}), which were tuned via sensitivity analysis as explained earlier. To satisfy these constraints, the penalties are defined as the PF_{slack} slack bus, PF_{V_m} the volumes of water stored in the reservoirs of the hydraulic units, and $PF_{V_{end}}$ the volumes of water remaining in the reservoirs of the hydraulic units in the last period. Therefore, the explicit form of the total TPF expression in equation (24) is taken as given in equation (25).

$$TPF = PF_{slack} + PF_{V_m} + PF_{V_{end}} \quad (25)$$

The details of the equations used in applying the algorithms to the STHCP problem can be obtained from [26].

V. NUMERICAL SAMPLE SOLUTION

For applying the GSA, DE, and ADE optimization algorithms to the STHCP, the sample power system, whose single-line diagram is shown in Figure 4, is selected. This sample test system was chosen because it has been previously solved in the literature with different algorithms, and acceptable solutions have been obtained. The sample test system consists of 16 buses. The system has five thermal generation units, four hydraulic generation units, and 35 transmission lines. Units connected to buses 1, 4, 5, 8, and 15

are thermal generation units, while buses 10, 12, 14, and 16 are hydraulic generation units with fixed heads. The system selects bus number one as the slack bus, whose voltage is $1.05\angle 0^\circ$ pu. To solve the problem, a daily operating period consisting of six equal sub-time periods of four hours each ($t_j=4h, j=1,...,6$) is considered. All values in the study are given according to the pu unit system. The test system's base values are $S_{base}=100$ MVA, $U_{base}=230$ kV and $Z_{base}=529$ Ohm [7, 8, 26].

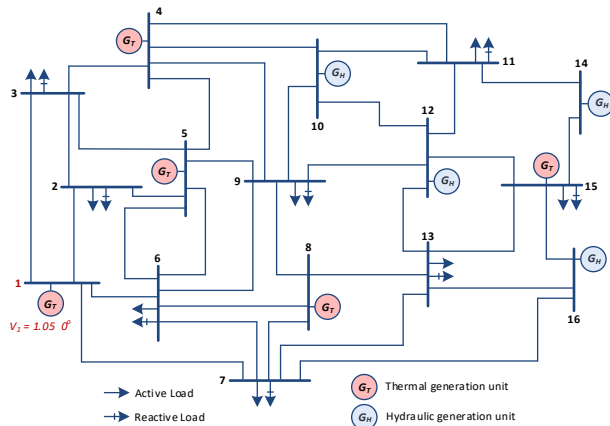


Figure 4. Single-line diagram of the system with sixteen buses and nine generators [8]

The resistance (R), reactance (X), and shunt susceptance (B) values of the nominal π equivalent circuit of the transmission lines of the sample test system, the active (P) and reactive (Q) load values for each sub-time period during the operating period, the convex fuel cost curve coefficients and power generation limits for thermal generation units and the non-convex fuel cost curve coefficients and power generation limits have been taken from references [7, 8].

The coefficients of the water input and output per hour curves of the hydraulic generating units in the test system, active power generation limits, reservoir storage limits, initial and final water volume values of the reservoirs, the amount of water per hour entering the reservoirs and the total amount of water to be discharged during the operating period are given in Table 1 [8].

TABLE I. VALUES OF HYDRAULIC PRODUCTION UNITS IN THE SYSTEM

	Hydraulic generation unit no (m)			
	10	12	14	16
d_1	330.0	320.0	380.0	300.0
d_2	497.0	620.0	565.0	600.0
d_3	254.4	275.0	432.0	343.2
d_4	200.0	380.0	200.0	228.0
d_5	300.0	180.0	250.0	280.0
$P_{GH,m}^{\min}$	0.0	0.0	0.0	0.0
$P_{GH,m}^{knee}$	1.20	1.50	1.30	1.20
$P_{GH,m}^{\max}$	1.35	1.65	1.45	1.35
V_m^{\min}	30000	30000	30000	30000
V_m^{\max}	80000	80000	80000	80000
V_m^{start}	50000	45000	46600	40000
V_m^{end}	48000	46600	40600	50600
r_{nj}	650	-	450	-
$q_{\text{total},m}$	17600	16000	16800	22200

The units of variables $P_{GH,m}^{\min}$, $P_{GH,m}^{knee}$ and $P_{GH,m}^{\max}$ in the table are pu , and the units of water parameters required for hydraulic production are *acre-ft*.

The hydraulic relationships between the hydraulic generating units in the test system are shown in Figure 5.

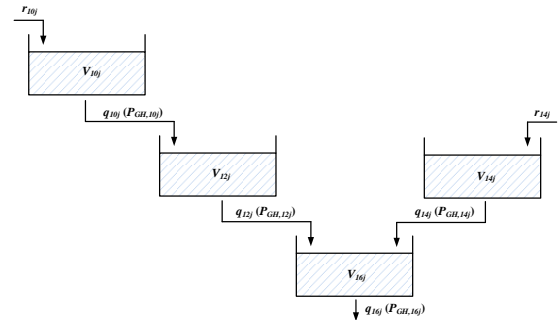


Figure 5. Hydraulic relations between hydraulic generating units

The volume of water remaining in the reservoirs of the hydraulic generating units at the end of each sub-time period is calculated from equations (26)-(29), respectively.

$$V_{10j} = V_{10j-1} + \lceil r_{10} - q_{10}(P_{GH,10j}) \rceil t_j, \quad j = 1, \dots, 6 \quad (26)$$

$$V_{12j} = V_{12j-1} + [q_{10}(P_{GH,10j}) - q_{12}(P_{GH,12j})] \cdot t_j, \quad j = 1, \dots, 6 \quad (27)$$

$$V_{14j} = V_{14j-1} + \left[r_{14} - q_{14}(P_{GH,14j}) \right] t_j, \quad j=1, \dots, 6 \quad (28)$$

$$V_{16j} = V_{16j-1} + [q_{12}(P_{GH,12j}) + q_{14}(P_{GH,14j}) - q_{16}(P_{GH,16j})]t_j \quad (29)$$

To be used in the Newton-Raphson load flow method applied to calculate the transmission losses of the test system, the initial reactive power values of the thermal and hydraulic generation units in the system (excluding the slack bus) in each sub-time period are given in pu in Table 2 [8].

TABLE II. INITIAL REACTIVE POWER VALUES AS PU OF THE GENERATION UNITS IN THE SYSTEM [8]

	Period (j)					
	1	2	3	4	5	6
4	0.400	0.550	0.600	0.700	0.650	0.500
5	0.400	0.550	0.600	0.650	0.650	0.500
8	0.400	0.550	0.600	0.600	0.600	0.500
10	0.400	0.550	0.600	0.600	0.600	0.500
12	0.400	0.550	0.600	0.700	0.600	0.500
14	0.400	0.550	0.600	0.700	0.650	0.500
15	0.400	0.550	0.600	0.700	0.600	0.500
16	0.400	0.550	0.600	0.650	0.650	0.500

All parameter values of the three algorithms (GSA, DE, and ADE) and penalty functions of the problems used to solve the STHCP with fixed head are given in Table 3.

TABLE III. PARAMETER VALUES FOR ALGORITHMS AND PROBLEMS

GSA	ItEN	N	G_0	α	f_{Call}
	1000	50	100	10	50000
DE	ItEN	N	CR	F	f_{Call}
	1000	50	0.9	0.5	50000
ADE	ItEN	N	CR	γ	f_{Call}
	1000	50	0.9	2	50000
STHCP	$Tol_{V_{end}}$	CPF_{slack}	CPF_{V_m}	$CPF_{V_{end}}$	
	0.02	1000	0.7	0.7	

The algorithm parameters of the GSA, DE, and ADE algorithms given in Table 3 are the values used in previous studies in which the performance analyses of the parameters were made and used in solving similar problems. To ensure that constraint handling through penalty functions does not

negatively impact the feasibility or quality of the solutions, a sensitivity analysis was performed on the penalty coefficients (CPF_{slack} , CPF_{Vend} , CPF_{Vm}). These coefficients were systematically varied in defined ranges (CPF_{slack} : 100-3000, CPF_{Vend} , CPF_{Vm} : 0.1-1.0), and the solutions were evaluated regarding both TFC values and constraint satisfaction. Based on these trials, the final values were set as $CPF_{slack}=1000$, $CPF_{Vend}=0.7$, and $CPF_{Vm}=0.7$, which provided the best balance between solution feasibility and algorithmic convergence. However, different solutions to the problem can be obtained by studying these values with other methods and approaches.

In Table 3, $IteN$ indicates the number of iterations, which are the stopping criteria of the algorithms, N indicates the number of agents in each population, and the number of times the f_{Call} objective function is called throughout the solution. In the table, G_0 is the initial value of the GSA, α is the constant coefficient of the GSA [8], CR is the crossover rate of the DE, and F is the scaling factor of the DE algorithm [23], CR is the

crossover rate of the ADE, and γ is the scaling factor of the ADE algorithm.

In this study, the solutions of the selected sample test systems were performed 50 times for each algorithm separately. The algorithms were developed independently in MATLAB R2021a, and the programs were run on a 2xIntel Xeon E5-2637 v4 3.50 GHz dual-processor workstation with 512 GB RAM.

A. CASE-1: STHCP WITH CONVEX THERMAL FUEL COST FUNCTION

For this case, the fuel cost function for thermal generation units in the example test system in Figure 4 is as convex as in equation (2). Using the coefficients in equation (2), the test system was solved 50 times each by GSA, DE, and ADE algorithms. First, the statistical analysis of the aggregated results obtained from these solutions is given in Table 4.

TABLE IV. VALUES FOR 50 SOLUTIONS (CASE-1)

	Solution Methods		
	GSA	DE	ADE
The best TFC (R)	149279.809206 (Run: 7)	148235.166212 (Run: 24)	147839.995227 (Run: 27)
The worst TFC (R)	153428.660945 (Run: 1)	151546.908949 (Run: 44)	152929.663208 (Run: 29)
The mean TFC (R)	151272.251286	149529.497489	149811.032659
Standard deviation	866.211041	874.248266	1237.629892
Total time (s)	13997.225157	11602.785672	12715.100654
The mean time (s)	279.944503	232.055713	254.302013

When the solution values of 50 times for this case, using the convex fuel cost functions given in Table 4, are considered, it is seen that the ADE algorithm provides a solution that meets the constraints at a lower cost than the other algorithms for the best solution value. Considering the mean cost values and times, it can be said that the DE algorithm is more stable than the other algorithms.

The graphs obtained from the solutions with all three algorithms for Case 1 are given below for comparison. For each algorithm, the logarithmic variation of the fitness functions concerning iterations is shown in Figure 6, the variation of TFC is shown in Figure 7, the variation of total transmission line losses ($TTLL$) in pu according to iterations is shown in Figure 8, and the box plots of the 50 solutions given in Table 4 are shown in Figure 9.

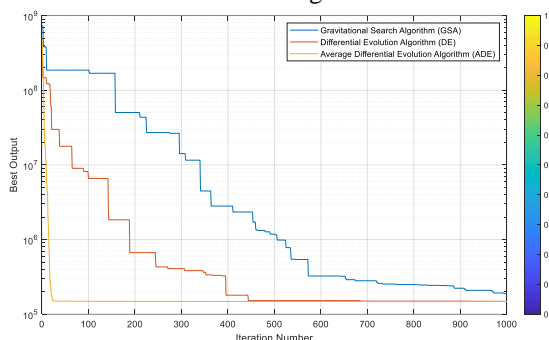


Figure 6. Variation in the fitness functions of the best solutions according to iterations (Case 1)

When the change of the fitness functions according to the algorithms given in Figure 6 is analyzed, it is seen that the ADE algorithm converges faster than the other two algorithms. In addition, when the change in TFC of ADE compared to the other algorithms in Figure 7 is analyzed, it can be said that it monotonically decreases continuously after the first 60th iteration and quickly reaches the optimal value by resetting the penalty functions.

The monotonically continuous decreases in the TFC changes of the other two algorithms start at approximately the 550th iteration in GSA and at roughly the 850th iteration in DE. Looking at the transmission line losses given in Figure 8, it can be stated that ADE, the fastest converging algorithm to the optimal value, makes minor adjustments after the 500th iteration. The other two algorithms continue to make adjustments until the last iteration. The box plots in Figure 9 show the visual dimension of the statistical values in Table 4.

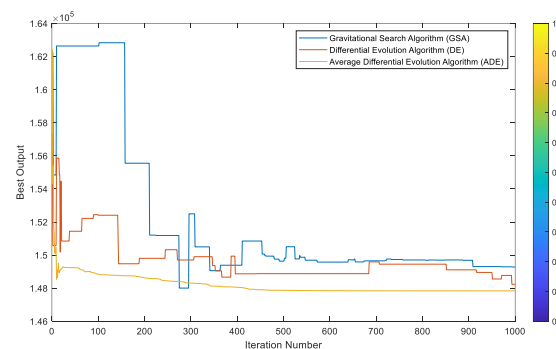


Figure 7. Variation of TFC of the best solutions according to iterations (Case 1)

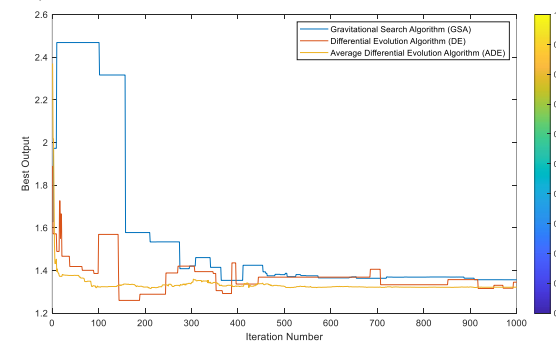


Figure 8. Variation of $TTLL$ according to iterations for the example system (Case 1)

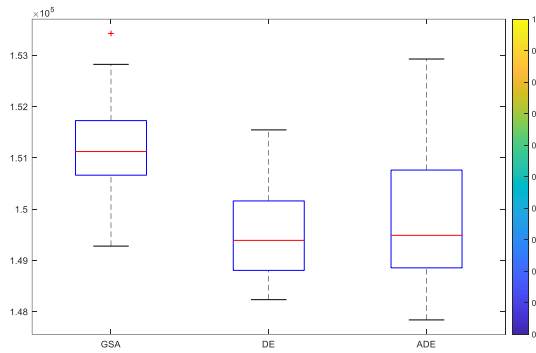


Figure 9. Box plots for 50 solutions (Case 1)

In this section, the optimal results obtained by the GSA, DE, and ADE algorithms for the solution of the sample power

TABLE V. VALUES FOR THE BEST SOLUTION FOR GSA (CASE-1)

Generation unit no (n,m)		Period (j)					
		1	2	3	4	5	6
1	$P_{GS,1j}$	0.602092	1.801693	2.063385	0.332029	2.224327	0.854523
	$Q_{GS,1j}$	0.713062	0.721523	1.002791	0.901243	1.166558	0.392211
4	$P_{GS,4j}$	2.282432	0.774616	1.537482	1.196360	0.534817	0.842107
5	$P_{GS,5j}$	0.490836	1.065996	0.543045	1.932829	1.500778	1.561440
8	$P_{GS,8j}$	0.711125	1.252429	1.911400	1.289905	0.502215	0.864658
10	$P_{GH,10j}$	0.230360	1.097812	0.338926	1.177960	0.866872	1.145054
12	$P_{GH,12j}$	0.074892	0.297292	0.097467	0.966235	1.551829	0.330665
14	$P_{GH,14j}$	0.564265	0.517293	0.952478	0.727835	0.183546	0.453789
15	$P_{GS,15j}$	1.080000	0.491073	0.521613	0.867925	0.866872	1.039889
16	$P_{GH,16j}$	0.942164	1.193200	1.064607	1.175446	1.207179	0.656195
Ploss (pu)		0.178166	0.191404	0.280403	0.266524	0.288435	0.14832
TFC (R)		149279.809206					
TTLL (pu)		1.355826					
Time (s)		279.944503					

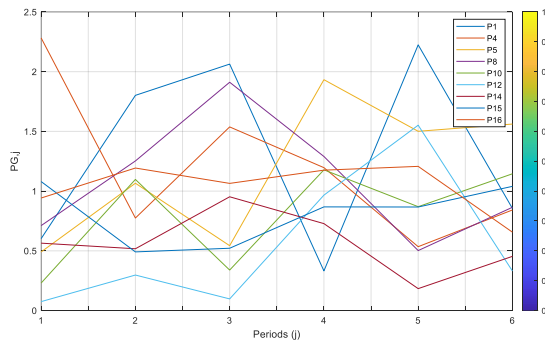


Figure 10. Variation of the values of active power generated in each period (GSA - Case 1)

The water volume values in the reservoirs at the end of the sub-periods and the error rates within acceptable limits for the solution in the 7th study, which has the best total fuel cost among the 50 solutions made with GSA, are given in Table 6. The $V_m^{\text{end,solution}}$ value in the table represents the calculated value of the volume of water remaining in the reservoir of the

system are given, respectively. Firstly, the values of active and reactive power generated by the generation units, total fuel cost values (*TFC*), total transmission line losses (*TTLL*), and their durations for the run (Run: 7) where the best solution is obtained for the GSA algorithm given in Table 4 are shown in Table 5. With GSA, the best solution was obtained in run 7 with 149279.809206 R, and the worst solution was obtained in run 1 with 151626.969393 R. The average time for each solution was 279.944503 s, while the best solution took 280.653264 s. For the run (Run:7), where the best solution values given in the table are obtained, the variation of the active power values generated in each period is shown in Figure 10.

m^{th} hydraulic unit in the last period in the optimal solution of the test problem.

TABLE VI. RESERVOIR WATER VALUES OF THE BEST SOLUTION OBTAINED WITH GSA (CASE 1)

	Reservoir water amount (acre-ft)			
	V_{10}	V_{12}	V_{14}	V_{16}
V_{start}	50000	45000	46600	40000
$V_m^{\text{end,solution}}$	48019.192239	46607.093781	40597.795222	50594.163206
V_{end}	48000	46600	40600	50600
%ErrorV	0.039984	0.015223	0.005430	0.011535
%TotalErrorV	0.0722			

Secondly, the values of active and reactive power generated by the generation units, total fuel cost values (*TFC*), total transmission line losses (*TTLL*), and durations of the run (Run: 24) where the best solution is obtained for the DE algorithm given in Table 4 are shown in Table 7.

TABLE VII. VALUES FOR THE BEST SOLUTION FOR DE (CASE-1)

Generation unit no (n,m)		Period (j)					
		1	2	3	4	5	6
1	$P_{GS,1j}$	2.191521	2.068250	1.593236	1.368718	2.822614	1.720313
	$Q_{GS,1j}$	0.771368	0.755874	0.755874	1.019166	1.314003	0.467084
4	$P_{GS,4j}$	0.549319	0.538031	1.120449	1.770410	1.570902	0.524182
5	$P_{GS,5j}$	1.241671	0.400000	1.233339	0.687416	1.186104	1.022992
8	$P_{GS,8j}$	0.500000	1.356888	1.239920	0.842645	0.874138	0.811322
10	$P_{GH,10j}$	0.538195	1.094782	0.873570	1.029616	0.545843	0.786761
12	$P_{GH,12j}$	0.123876	0.137408	0.927461	0.548743	0.944911	0.671698
14	$P_{GH,14j}$	0.256920	0.909208	0.389392	1.117681	0.177734	0.547895
15	$P_{GS,15j}$	0.565038	0.741216	0.504363	1.229466	0.527297	0.699885
16	$P_{GH,16j}$	0.987006	1.253919	1.115424	1.090656	0.802402	0.971872
Ploss (pu)		0.153546	0.199702	0.247154	0.285351	0.301945	0.15692
TFC (R)		148235.166212					
TTLL (pu)		1.344620					
Time (s)		228.469464					

For the run (Run: 24), where the numerical values given in the table are obtained, the variation of the active power values generated in each period is shown in Figure 11.

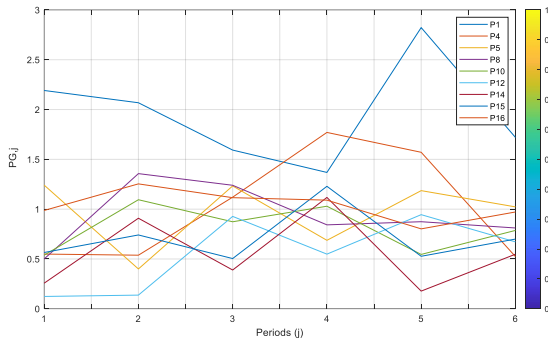


Figure 11. Variation of the values of active power generated in each period (DE - Case 1)

The best solution with DE occurred in run 24 with 148235.166212 R , and the worst solution occurred in run 44 with 151546.908949 R . The average time for each solution was 232.055713 s, while the best solution, 24, took 228.469464 s. The water volume values and error rates in the reservoir at the end of the sub-time periods for the solution in the 24th study, which has the best total fuel cost among the 50 solutions made with DE, are given in Table 8. Thirdly, the active and reactive power values, total fuel cost values (TFC), total transmission line losses ($TTLL$), and durations for the study (Run: 27) where the best solution is obtained for the ADE algorithm given in Table 4 are shown in Table 9.

Out of the 50 solutions obtained with the ADE algorithm, the best fuel cost solution was obtained in run 27 with

147839.995227 R , and the worst solution was obtained in run 29 with 152929.663208 R . The average time for each solution was 254.302013 seconds, while the best solution in run 27 took 244.675648 seconds.

TABLE VIII. RESERVOIR WATER VALUES OF THE BEST SOLUTION OBTAINED WITH DE (CASE 1)

	Reservoir water amount (acre-ft)			
	V_{10}	V_{12}	V_{14}	V_{16}
V_{start}	50000	45000	46600	40000
$V_{end, solution}$	48000.889136	46600.949646	40598.645451	50600.487163
V_{end}	48000	46600	40600	50600
%ErrorV	0.001852	0.002038	0.003336	0.000963
%TotalErrorV	0.0082			

For the run (Run: 27), where the numerical values given in the table are obtained, the variation of the active power values generated in each period is shown in Figure 12.

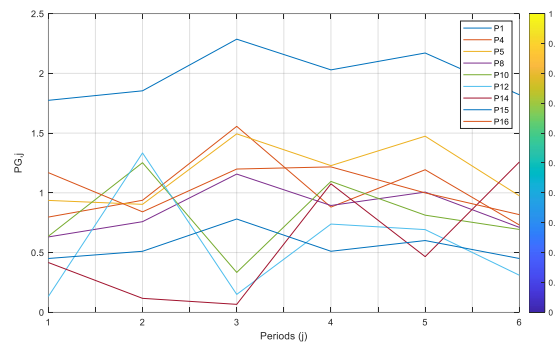


Figure 12. Variation of the values of active power generated in each period (ADE - Case 1)

TABLE IX. VALUES OF THE BEST SOLUTION FOR ADE (CASE-1)

Generation unit no (n,m)		Period (j)					
		1	2	3	4	5	6
1	$P_{GS,1j}$	1.774286	1.854101	2.286389	2.029399	2.170580	1.820789
	$Q_{GS,1j}$	0.694538	0.727651	1.023229	1.033674	1.125214	0.538163
4	$P_{GS,4j}$	0.796903	0.937555	1.556399	0.880511	1.193176	0.730586
5	$P_{GS,5j}$	0.936526	0.903555	1.493806	1.227610	1.474010	0.974627
8	$P_{GS,8j}$	0.630608	0.758486	1.157036	0.893858	1.006261	0.710798
10	$P_{GH,10j}$	0.636121	1.252304	0.333587	1.095247	0.813018	0.693263
12	$P_{GH,12j}$	0.131896	1.333722	0.149566	0.738431	0.692042	0.309951
14	$P_{GH,14j}$	0.416561	0.116569	0.066265	1.075311	0.465786	1.258104
15	$P_{GS,15j}$	0.450193	0.510470	0.780742	0.510678	0.600481	0.450132
16	$P_{GH,16j}$	1.169090	0.840846	1.199116	1.217093	0.999479	0.817257
Ploss (pu)		0.142184	0.207608	0.272906	0.268138	0.264833	0.165507
TFC (R)		147839.995227					
TTLL (pu)		1.321178					
Time (s)		244.675648					

For Case 1, the comparison of the results obtained in this study in terms of total fuel cost (TFC) with the results of different meta-heuristic algorithms previously published in the literature is given in Table 10.

TABLE X. LITERATURE COMPARISON (CASE 1)

	GA [10]	GSA	DE	ADE
TFC (R)	148767.660	149279.809	148235.166	147839.995

The table shows that the best solution is obtained with the ADE algorithm. The comparisons in the table were made over fuel costs regardless of the amount of water discharged or not discharged by the hydraulic units in the system. The

comparisons in the table were made over fuel costs irrespective of the amount of water discharged or not discharged by the hydraulic units in the system. The water volume values and error rates in the reservoir at the end of the sub-time periods for the solution in the 27th study, which has the best total fuel cost among the 50 solutions made with ADE, are given in Table 11.

B. CASE-2: STHCP WITH NON-CONVEX THERMAL FUEL COST FUNCTION

This case is designed to contribute a sample test problem for non-convex lossy STHCP with fixed head to the literature. Because the B matrix is usually used in the literature to solve such problems, however, as shown in Figure 4, the losses of

the example power system can be solved with load flow when the states of the generation units at the buses, loads, and R , X , and B values of the transmission lines are known. Therefore, in this case, to contribute to the literature, the non-convex fuel cost functions of Figure 4 are taken as in equation (3). The values are used for the coefficients in equation (3), and this problem is solved 50 times each by GSA, DE, and ADE algorithms, respectively. First, the statistical analysis of the aggregated results obtained from the solutions of the three different algorithms is given in Table 12.

TABLE XI. RESERVOIR WATER VALUES OF THE BEST SOLUTION OBTAINED WITH ADE (CASE 1)

	Reservoir water amount (acre-ft)			
	V_{10}	V_{12}	V_{14}	V_{16}
V_{start}	50000	45000	46600	40000
$V_{m}^{end, solution}$	47999.020605	46599.072418	40599.173708	50598.980142
V_{end}	48000	46600	40600	50600
%ErrorV	0.002040	0.001991	0.002035	0.002016
%TotalErrorV	0.0081			

TABLE XII. VALUES FOR 50 SOLUTIONS (CASE 2)

	Solution Methods		
	GSA	DE	ADE
The best TFC (R)	177414.249557 (Run: 2)	171627.154727 (Run: 47)	156316.715581 (Run: 47)
The worst TFC (R)	193928.660732 (Run: 50)	195154.228166 (Run: 4)	175161.020942 (Run: 46)
The mean TFC (R)	187202.809936	181854.837533	163483.43463
Standard deviation	3832.031012	5006.662745	4700.180592
Total time (s)	13949.699205	11402.936344	12946.616234
The mean time (s)	278.993984	228.058727	258.932325

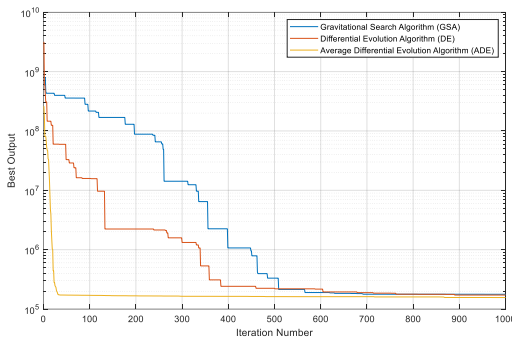


Figure 13. Variation of the fitness functions of the best solutions according to iterations (Case 2)

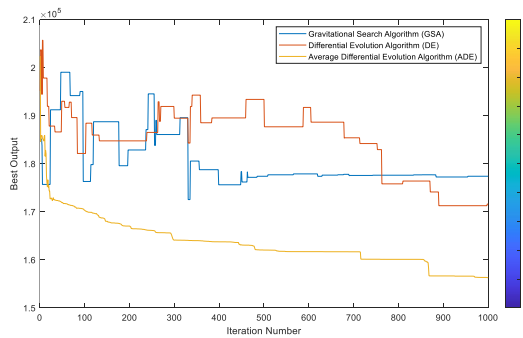


Figure 14. Variation of TFC of the best solutions according to iterations (Case 2)

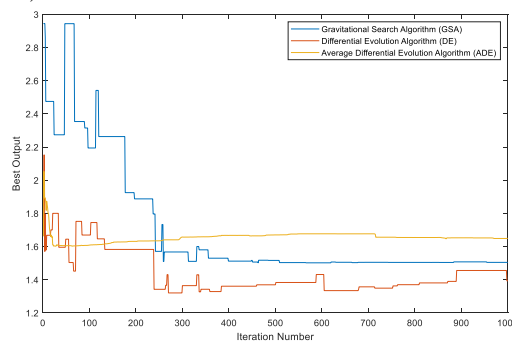


Figure 15. Variation of total transmission line losses (TTLL) according to iterations for the example system (Case 2)

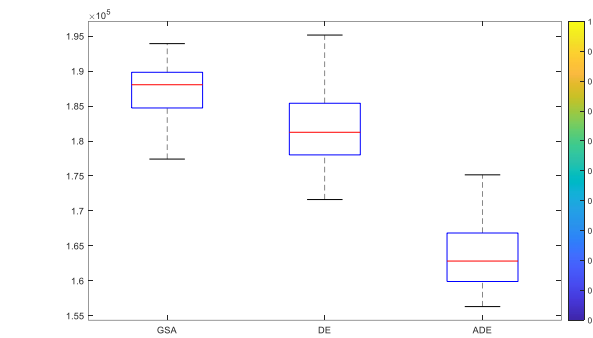


Figure 16. Box plots for 50 solutions (Case 2)

Table 12 shows that the ADE algorithm is remarkably superior to the other algorithms for the best, worst, and mean solution values when the solution values are analyzed 50 times each for this case, where non-convex fuel cost functions are used. Valve point effects are also taken into account. Considering the solution times, the DE algorithm provides an acceptable solution proposal in a shorter time than the other algorithms. For case 2, the graphs obtained for the solutions with all three algorithms are given below for comparison. For each algorithm, the logarithmic variation of the fitness functions concerning iteration is shown in Figure 13, the variation of TFC is shown in Figure 14, the variation of total transmission line losses (TTLL) in pu is shown in Figure 15, and the box plots of the 50 solutions given in table 15 are shown in Figure 16.

When the change of the fitness functions according to the algorithms given in Figure 13 is analyzed, it is seen that the ADE algorithm converges faster than the other two algorithms in this case as well. Also, when the change in TFC of ADE compared to the other algorithms is analyzed in Figure 14, it can be said that it monotonically decreases continuously after the first 20th iterations and reaches the optimum value by resetting the penalty functions at the 870th iteration. The monotonically continuous decreases in the TFC changes of the other two algorithms occur only after approximately the 900th iteration. Looking at the

transmission line losses given in Figure 15, it can be seen that ADE, again, the algorithm that converges fastest to the optimal value, makes minor adjustments after the 700th iteration. On the other hand, of the other two algorithms, GSA converges at about the 630th iteration, whereas DE continues to make adjustments until the last iteration. The box plots in Figure 16 show the visualization of the statistical values in Table 12, and it can be stated that ADE captures the best values.

TABLE XIII. VALUES FOR THE BEST SOLUTION FOR GSA (CASE 2)

Generation unit no (n,m)		Period (j)					
		1	2	3	4	5	6
1	$P_{GS,1j}$	0.427296	0.323931	2.242475	1.792951	0.328232	0.543327
	$Q_{GS,1j}$	0.613009	0.756310	1.145564	1.030746	1.207636	0.566284
4	$P_{GS,4j}$	1.039219	0.671167	2.400021	1.739655	0.573841	2.500000
5	$P_{GS,5j}$	1.625803	0.520303	0.590747	1.206818	1.840584	0.632617
8	$P_{GS,8j}$	1.367482	1.728999	0.523720	1.364441	0.774934	0.731281
10	$P_{GH,10j}$	0.462019	0.493085	1.022084	0.669538	1.270047	0.889547
12	$P_{GH,12j}$	0.098420	1.578938	0.010959	0.324641	0.748655	0.552449
14	$P_{GH,14j}$	0.313522	1.311731	0.962809	0.176430	0.283474	0.344727
15	$P_{GS,15j}$	0.474695	1.100372	0.621935	1.059844	2.316393	0.837272
16	$P_{GH,16j}$	1.131955	0.815768	0.694223	1.329163	1.349343	0.771285
Ploss (pu)		0.140411	0.244294	0.318973	0.263481	0.335503	0.202505
TFC (R)		177414.249557					
TTLL (pu)		1.505169					
Time (s)		289.714972					

TABLE XIV. VALUES FOR THE BEST SOLUTION FOR DE (CASE-2)

Generation unit no (n,m)		Period (j)					
		1	2	3	4	5	6
1	$P_{GS,1j}$	0.804705	1.997618	2.020782	2.383630	2.983880	0.537479
	$Q_{GS,1j}$	0.651131	0.725297	0.936389	1.048838	1.345280	0.487256
4	$P_{GS,4j}$	1.755174	0.903577	0.503391	0.581640	1.485320	0.513693
5	$P_{GS,5j}$	0.419269	0.480630	1.087418	0.400000	0.460585	0.939872
8	$P_{GS,8j}$	0.500000	1.681917	1.796214	1.713031	0.529823	2.000000
10	$P_{GH,10j}$	0.692746	0.556634	0.977754	0.893860	0.822674	0.927008
12	$P_{GH,12j}$	0.398535	0.082468	0.491910	1.337532	0.404853	0.640892
14	$P_{GH,14j}$	0.335638	0.853500	0.309287	0.794631	0.395915	0.708327
15	$P_{GS,15j}$	1.111508	0.918296	0.547912	1.039262	1.075420	0.450000
16	$P_{GH,16j}$	0.944397	1.016227	1.281899	0.536860	1.306330	1.061765
Ploss (pu)		0.161972	0.190867	0.266567	0.280446	0.3148	0.179036
TFC (R)		171627.154727					
TTLL (pu)		1.393689					
Time (s)		218.800278					

For Case 2, the best solution with GSA was obtained in run 2 with 177414.249557 R, and the worst solution was obtained in run 50 with 193928.660732 R. The mean time for each solution was 278.993984 seconds, while the best solution took 289.714972 seconds. For the run (Run:2), where the best solution values given in the table are obtained, the variation of the active power values generated in each period is shown in Figure 17.

Secondly, for Case 2, the values of active and reactive power generated by the generation units, total fuel cost values (TFC), total transmission line losses (TTLL), and durations for the run (Run: 47) where the best solution is obtained for the DE algorithm are given in Table 14.

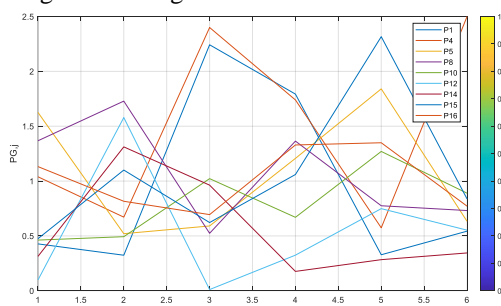


Figure 17. Variation of active power values generated in each period (GSA - Case 2)

In this section, the optimal results obtained by the GSA, DE, and ADE algorithms for the solution of the non-convex fuel cost example test system, where valve point effects are also considered, are given respectively. Firstly, the values of active and reactive power generated by the generation units, total fuel cost values (TFC), total transmission line losses (TTLL), and durations for the run (Run: 2) where the best solution is obtained for GSA are given in Table 13.

The water volume values in the reservoirs at the end of the sub-time periods of the solution with the best total fuel cost among the 50 solutions made with GSA, and the error rates within acceptable limits are given in Table 15.

For case 2, the best solution with DE occurred in run 47 with 171627.154727 R, and the worst was in run 4 with 195154.228166 R. The mean time for each solution was 228.058727 s, while the best solution, 47, took 218.800278 s. The water volume values and error rates in the reservoir at the end of the sub-time periods for the solution in the 47th run, which has the best total fuel cost among the 50 solutions with DE, are given in Table 16.

TABLE XV. RESERVOIR WATER VALUES OF THE BEST SOLUTION OBTAINED WITH GSA (CASE 2)

	Reservoir water amount (acre-ft)			
	V_{10}	V_{12}	V_{14}	V_{16}
V_{start}	5000	45000	46600	40000
$V_{end, solution}$	48000.626009	46601.291171	40598.999362	50601.038430
V_{end}	48000	46600	40600	50600
%ErrorV	0.001304	0.002771	0.002465	0.002052
%TotalErrorV	0.0086			

For the study in which the numerical values given in Table 15 were obtained (Run: 47), the variation of the active power values generated in each period is shown in Figure 18.

Out of the 50 solutions obtained with the ADE algorithm, the solution with the best fuel cost value was obtained in run 47 with 156316.715581 *R*, and the worst solution was obtained in run 46 with 175161.020942 *R*. The mean time for each solution was 258.932325 s, while the best solution in run 47 took 265.273997 s.

For the run where the numerical values given in the table are obtained (Run: 47), the variation of the active power values generated in each period is shown in Figure 19.

Thirdly, the active and reactive power values, total fuel cost values (*TFC*), total transmission line losses (*TTLL*), and durations produced by the generation units belonging to the run (Run: 47) in which the best solution is obtained for the ADE algorithm given in Table 17.

TABLE XVI. RESERVOIR WATER VALUES OF THE BEST SOLUTION OBTAINED WITH DE (CASE 2)

	Reservoir water amount (acre-ft)			
	V_{10}	V_{12}	V_{14}	V_{16}
V_{start}	50000	45000	46600	40000
$V_{m}^{end, solution}$	47997.095585	46599.555628	40602.106815	50601.246730
V_{end}	48000	46600	40600	50600
%ErrorV	0.006051	0.000954	0.005189	0.002464
%TotalErrorV	0.0147			

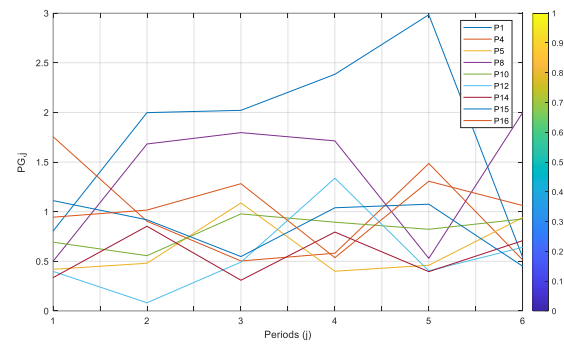


Figure 18. Variation of active power values generated in each period (DE - Case 2)

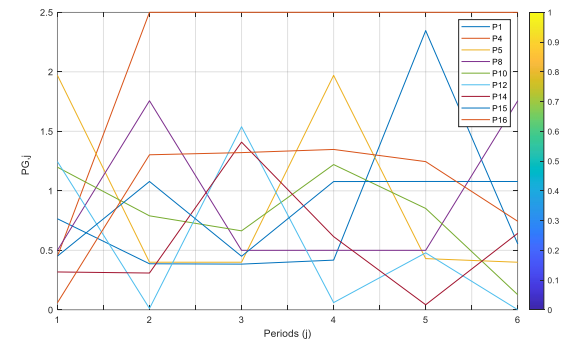


Figure 19. Variation of active power values generated in each period (ADE - Case 2)

TABLE XVII. VALUES OF THE BEST SOLUTION FOR ADE (CASE-2)

Generation unit no (n,m)		Period (j)					
		1	2	3	4	5	6
1	$P_{GS,1j}$	0.765391	0.387043	0.384259	0.417756	2.345753	0.553943
	$Q_{GS,1j}$	0.656796	0.789343	1.303577	1.146303	1.352644	0.588291
4	$P_{GS,4j}$	0.450000	2.499846	2.500000	2.500000	2.499962	2.499494
5	$P_{GS,5j}$	1.970791	0.400169	0.400052	1.969904	0.430456	0.400003
8	$P_{GS,8j}$	0.500019	1.756696	0.500028	0.500195	0.500000	1.756686
10	$P_{GH,10j}$	1.198536	0.789944	0.663727	1.220638	0.852020	0.127014
12	$P_{GH,12j}$	1.245557	0.011604	1.538311	0.061015	0.478980	0.000950
14	$P_{GH,14j}$	0.318356	0.309334	1.409196	0.615241	0.042858	0.643276
15	$P_{GS,15j}$	0.451172	1.078534	0.450091	1.077901	1.078717	1.078377
16	$P_{GH,16j}$	0.059350	1.302802	1.321014	1.347606	1.245407	0.743823
Ploss (pu)		0.159172	0.235972	0.416678	0.310256	0.324153	0.203566
TFC (R)		156316.715581					
TTLL (pu)		1.649799					
Time (s)		265.273997					

The water volume values and error rates in the reservoir at the end of the sub-time periods for the solution in the 47th run, which has the best total fuel cost among the 50 solutions with DE, are given in Table 18.

TABLE XVIII. RESERVOIR WATER VALUES OF THE BEST SOLUTION OBTAINED WITH ADE (CASE-2)

	Reservoir water amount (acre-ft)			
	V_{10}	V_{12}	V_{14}	V_{16}
V_{start}	50000	45000	46600	40000
$V_{m}^{end, solution}$	47999.033619	46599.619931	40599.122758	50599.072088
V_{end}	48000	46600	40600	50600
%ErrorV	0.002013	0.000816	0.002161	0.001834
%TotalErrorV	0.0068			

The comparison of the results obtained in this study for Case 2 with the results of different heuristic algorithms previously published in the literature is given in Table 19.

When the table is examined, it is seen that the ADE algorithm achieved the best result with 156316.715 *R* in terms of TFC. The comparisons in the table are based on total fuel

costs regardless of the amount of water discharged or not discharged by the hydraulic units in the system. The comparisons are made for acceptable tolerance values for the feasible solution to the problem in all studies.

TABLE XIX. LITERATURE COMPARISON (CASE-2)

	IGSA-1 [26]	IGSA-2 [26]	IGSA-3 [26]
TFC (R)	166516.440	165259.461	164762.279
	GSA	DE	ADE
	177414.249	171627.154	156316.715

VI. RESULTS AND CONCLUSION

To our knowledge, ADE, one of the newly developed meta-heuristic algorithms for solving STHCP with a constant drop, is applied for the first time in this study for two cases (convex and non-convex cases). The same problems are solved with both GSA and DE metaheuristics to evaluate the performance of ADE on fundamental issues. For each case, the results obtained from the solutions of the three algorithms (GSA,

DE, and ADE) are compared with the values in the literature and with each other.

For the optimal solution of the fixed head STHCP, transmission line losses are calculated using the Newton-Raphson load flow method in three meta-heuristic algorithms (GSA, DE, and ADE). The study solved problems 50 times each with all three algorithms. The best TFC values in the solution of the example test system in Case 1 were 149342.548 R for GSA, 148184.488 R for DE, and 147743.228 R for ADE. ADE, which was applied to this type of problem for the first time, achieved a better result of approximately 441.26 R than the classical DE in terms of TFC values. When compared in terms of solution times, the ADE algorithm reached the solution in shorter times than the GSA and DE algorithms. Similarly, the best TFC values in the solution of the example test system in Case 2 were obtained as 177414.249 R for GSA, 171627.154 R for DE, and 156316.715 R for ADE. In this problem, when compared to TFC values, ADE achieved a better result than classical DE. However, the DE algorithm solved the issue faster than the solution times. In the optimal solutions of GSA, DE, and ADE algorithms for both cases in the study, the amount of water required to be spent by the hydraulic production units of the test systems was within the maximum error tolerance value of 0.2%.

The penalty function method is adopted in this study. That is, penalty functions provide the problem's constraints in the algorithms. In the optimal solution, this method causes an increase in the number of iterations depending on the number of constraints and thus increases the solution time. This is the disadvantage of the penalty method.

In this study, it has been successfully demonstrated that the STHCP with a lossy fixed head, which is one of the optimization problems with many constraints that has a considerable place in the literature and is of great importance in electrical engineering, can be solved with ADE, one of the newly developed meta-heuristic algorithms.

ACKNOWLEDGMENT

This research has been supported by Kütahya Dumlupınar University Scientific Research Projects Coordination Office under grant number #2021-09.

We would also like to thank Kütahya Dumlupınar University Intelligent Systems Design Application and Research Center (ASTAM) for providing some basic research facilities.

CONFLICT OF INTEREST

The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editor, and the reviewers. The publisher does not guarantee or endorse any statements, claims, performances, or results.

REFERENCES

- [1] S. A. Papazis, G. C. Bakos, "Generalized Model of Economic Dispatch Optimization as an Educational Tool for Management of Energy Systems," *Advances in Electrical and Computer Engineering*, vol.21, no.2, pp.75-86, 2021. doi:10.4316/aecce.2021.02009
- [2] C. Yaşar, S. Özyön, "A Modified Incremental Gravitational Search Algorithm for Short-Term Hydrothermal Scheduling with Variable Head," *Engineering Applications of Artificial Intelligence*, vol.95 (103845), pp.1-17, 2020. doi:10.1016/j.engappai.2020.103845
- [3] M. Basu, "Hopfield Neural Networks for Optimal Scheduling of Fixed Head Hydrothermal Power Systems," *Electric Power Systems Research*, vol.64, no.1, pp.11-15, 2023. doi:10.1016/S0378-7796(02)00118-9
- [4] C. E. Zoumas, A. G. Bakirtzis, J. B. Theocharis, V. Petridis, "A Genetic Algorithm Solution Approach to the Hydrothermal Coordination Problem," *IEEE Transactions on Power Systems*, vol.19, no.2, pp.1356-1364, 2004. doi:10.1109/TPWRS.2004.825896
- [5] J. Sasikala, M. Ramaswamy, "Optimal Gamma based Fixed Head Hydrothermal Scheduling using Genetic Algorithm," *Expert Systems with Applications*, vol.37, no.4, pp.3352-3357, 2009. doi:10.1016/j.eswa.2009.10.015
- [6] V. S. Kumar, M. R. Mohan, "A Genetic Algorithm Solution to the Optimal Short-term Hydrothermal Scheduling," *International Journal of Electrical Power & Energy Systems*, vol.33, no.4, pp.827-835, 2010. doi:10.1016/j.ijepes.2010.11.008
- [7] S. Özyön, C. Yaşar, Y. Aslan, H. Temurtaş, "Solution to Environmental Economic Power Dispatch Problems in Hydrothermal Power Systems by Using Genetic Algorithm," in *6th International Conference on Electrical and Electronics Engineering (ELECO'09)*, Bursa, TÜRKİYE, 2009, pp.387-392.
- [8] S. Özyön, C. Yaşar, "Gravitational Search Algorithm Applied to Fixed Head Hydrothermal Power System with Transmission Line Security Constraints," *Energy*, vol.155, pp.392-407, 2018. doi:10.1016/j.energy.2018.04.172
- [9] M. S. Fakhra, S. A. R. Kashif, S. Liaquat, A. Rasool, S. Padmanaban, M. A. Iqbal, M. A. Baig, B. Khan, "Implementation of APSO and Improved APSO on Non-cascaded and Cascaded Short-term Hydrothermal Scheduling," *IEEE Access*, vol.9, pp.77784-77797, 2021. doi:10.1109/access.2021.3083528
- [10] S. Fadıl, C. Yaşar, "A Pseudo Spot Price Algorithm Applied to Short-term Hydrothermal Scheduling Problem," *Electric Power Components and Systems*, vol.29, no.11, pp.112-119, 2010. doi:10.1080/153250001753239202
- [11] A. E. Nezhad, S. Jowkar, T. T. Sabour, E. Rahimi, F. Ghanavati, F. Esmailnezhad, "A Short-term Wind-Hydrothermal Operational Framework in the Presence of Pumped-Hydro Storage," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol.8 (100577), pp.1-16, 2024. doi:10.1016/j.prime.2024.100577
- [12] M. Basu, "A Simulated Annealing-based Goal-Attainment Method for Economic Emission Load Dispatch of Fixed Head Hydrothermal Power Systems," *International Journal of Electrical Power & Energy Systems*, vol.27, no.2, pp.147-153, 2004. doi:10.1016/j.ijepes.2004.09.004
- [13] M. Basu, "Economic Environmental Dispatch of Fixed Head Hydrothermal Power Systems using Nondominated Sorting Genetic Algorithm-II," *Applied Soft Computing*, vol.11, no.3, pp.3046-3055, 2010. doi:10.1016/j.asoc.2010.12.005
- [14] M. Basu, "Artificial Immune System for Fixed Head Hydrothermal Power System," *Energy*, vol.36, no.1, pp.606-612, 2010. doi:10.1016/j.energy.2010.09.057
- [15] T. T. Nguyen, D. N. Vo, A. V. Truong, "Cuckoo Search Algorithm for Short-term Hydrothermal Scheduling," *Applied Energy*, vol.132, no.1, pp.276-287, 2014. doi:10.1016/j.apenergy.2014.07.017
- [16] T. T. Nguyen, D. N. Vo, "Modified Cuckoo Search Algorithm for Short-term Hydrothermal Scheduling," *International Journal of Electrical Power & Energy Systems*, vol.65, pp.271-281, 2014. doi:10.1016/j.ijepes.2014.10.004
- [17] N. Narang, J. S. Dhillon, D. P. Kothari, "Scheduling Short-term Hydrothermal Generation using Predator-prey Optimization Technique," *Applied Soft Computing*, vol.21, pp.298-308, 2014. doi:10.1016/j.asoc.2014.03.029
- [18] A. Rasoulzadeh-akhijahani, B. Mohammadi-ivatloo, "Short-term Hydrothermal Generation Scheduling by a Modified Dynamic Neighborhood Learning based Particle Swarm Optimization," *International Journal of Electrical Power & Energy Systems*, vol.67, pp.350-367, 2014. doi:10.1016/j.ijepes.2014.12.011
- [19] N. Gouthamkumar, V. Sharma, R. Naresh, "Disruption-based Gravitational Search Algorithm for Short-term Hydrothermal

- Scheduling,” *Expert Systems with Applications*, vol.42, no.20, pp.7000-7011, 2015. doi:10.1016/j.eswa.2015.05.017
- [20] G. Chen, M. Gao, Z. Zhang, S. Li, “Hybridization of Chaotic Grey Wolf Optimizer and Dragonfly Algorithm for Short-term Hydrothermal Scheduling,” *IEEE Access*, vol.8, pp.142996-143020, 2020. doi:10.1109/access.2020.3014114
- [21] M.A. Almubaidin, A.N. Ahmed, L.M. Sidek, K.A.H. AL-Assifeh, A. El-Shafie, “Deriving Optimal Operation Rule for Reservoir System Using Enhanced Optimization Algorithms. *Water Resources Management*, vol.38, no.4, pp.1207-1223, 2024. doi:10.1007/s11269-010-9712-6
- [22] M.S. Fakhar, S. Liaquat, S.A.R. Kashif, A. Rasool, M. Khizer, M.A. Iqbal, S. Padmanaban, “Conventional and metaheuristic optimization algorithms for solving short term hydrothermal scheduling problem: A review,” *IEEE Access*, vol.9, pp.25993-26025, 2021. Doi:10.1109/access.2021.3055292
- [23] S. Özyön, “Optimal Short-term Operation of Pumped-storage Power Plants with Differential Evolution Algorithm,” *Energy*, vol.194(116866), pp.1-13, 2019. doi:10.1016/j.energy.2019.116866
- [24] A.J. Wood, B.F. Wollenberg, G.B. Sheble, *Power Generation Operation and Control*, IEEE & Wiley, 2013, p.656.
- [25] D. P. Kothari, J. S. Dhillon, *Power System Optimization*, PHI Learning Private Limited, 2007, p.732.
- [26] S. Özyön, “Optimal Aktif Güç Dağıtımı için Karşıt Öğrenme Tabanlı Diferansiyel Gelişim Algoritması.” *Uludağ Üniversitesi Mühendislik Fakültesi Dergisi*, vol.25, no.1, pp.231-246, 2020. doi:10.17482/uumfd.635957
- [27] B. Durmuş, “Optimal Components Selection for Active Filter Design with Average Differential Evolution Algorithm,” *AEU - International Journal of Electronics and Communications*, vol.94, pp.293-302, 2018. doi:10.1016/j.aeue.2018.07.021
- [28] B. Durmuş, H. Temurtaş, S. Özyön, “The Design of Multiple Feedback Topology Chebyshev Low-pass Active Filter with Average Differential Evolution Algorithm,” *Neural Computing & Applications*, vol.32, no.22, pp.17097-17113, 2020. doi:10.1007/s00521-020-04922-7
- [29] S. Özyön, “The Solution of the Short-term Hydrothermal Coordination Problem by Improved Incremental Gravitational Search Algorithm,” *Electrical-Electronics Engineering*, Ph.D. Thesis, Kütahya Dumlupınar University, Institute of Science and Technology, 2008.

BIOGRAPHIES



S. Özyön was born in Ayaş, Turkey, in 1981. He received a B.Sc. degree in electrical electronics engineering from Dumlupınar University, Kütahya, Turkey, in 2005 and an M.Sc. degree from the Department of Electrical Electronics Engineering, Dumlupınar University, Kütahya, Turkey, in 2009. He works

as an Associate Professor in the Department of Electrical Electronics Engineering at Kütahya Dumlupınar University. His research areas include power systems analysis, economical operation of power systems, power distribution systems, renewable energy systems, and optimization techniques.



H. Temurtaş was born in Mersin, Turkey, in 1967. He received a B.Sc. degree in electrical electronics engineering from Middle East Technical University, Ankara, Turkey, in 1993; an M.Sc. degree from the Department of Electrical Electronics Engineering, Dumlupınar University, Kütahya, Turkey, in 1996; PhD degree in Electrical and Electronic Engineering

from Sakarya University, Sakarya, Turkey in 2004. He works as an Associate Professor in the Department of Computer Engineering at Kütahya Dumlupınar University. His research areas include programming languages, control algorithms, and optimization techniques.



B. Durmuş was born in Pazaryeri, Turkey, in 1978. He received his B.Sc. and M.Sc. degrees in 2000 and 2003 from the Department of Electrical Electronics Engineering, Dumlupınar University, Kütahya, Turkey, respectively. He received a Ph.D. degree from Sakarya University. He works as an Associate Professor

in the Department of Electrical Electronics Engineering at Kütahya

Dumlupınar University. His areas of research include optimization techniques and computational intelligence.



C. Yaşar was born in Kütahya, Turkey, in 1958. He received the B.Sc.E.E. degree from the Yıldız Technical University, the M.Sc. E.E degree from the Anadolu University and Ph.D. degree from the Osmangazi University, Turkey, in 1980, 1988 and 1999, respectively. He works as a Professor at the Electrical and Electronics Engineering Department of Dumlupınar

University, Turkey. His research interests include analysis of power systems, economical operation of power systems, power distribution systems, power system protection, and renewable energy systems.

Publication Ethics

The journal publishes original papers in the extensive field of Electrical-electronics and Computer engineering. To that end, it is essential that all who participate in producing the journal conduct themselves as authors, reviewers, editors, and publishers in accord with the highest level of professional ethics and standards. Plagiarism or self-plagiarism constitutes unethical scientific behavior and is never acceptable.

By submitting a manuscript to this journal, each author explicitly confirms that the manuscript meets the highest ethical standards for authors and coauthors

The undersigned hereby assign(s) to *Balkan Journal of Electrical & Computer Engineering* (BAJECE) copyright ownership in the above Paper, effective if and when the Paper is accepted for publication by BAJECE and to the extent transferable under applicable national law. This assignment gives BAJECE the right to register copyright to the Paper in its name as claimant and to publish the Paper in any print or electronic medium.

Authors, or their employers in the case of works made for hire, retain the following rights:

1. All proprietary rights other than copyright, including patent rights.
2. The right to make and distribute copies of the Paper for internal purposes.
3. The right to use the material for lecture or classroom purposes.
4. The right to prepare derivative publications based on the Paper, including books or book chapters, journal papers, and magazine articles, provided that publication of a derivative work occurs subsequent to the official date of publication by BAJECE.
5. The right to post an author-prepared version or an official version (preferred version) of the published paper on an internal or external server controlled exclusively by the author/employer, provided that (a) such posting is noncommercial in nature and the paper is made available to users without charge; (b) a copyright notice and full citation appear with the paper, and (c) a link to BAJECE's official online version of the abstract is provided using the DOI (Document Object Identifier) link.



ISSN: 2147- 284X
Year: June 2025
Volume: 13
Issue: 2

CONTENTS

Research Article	Eren Gündüzvar, Abdulsamet Kayık, Mehmet Ali Altuncu; Alzheimer's Disease Diagnosis in MRI Images Using Transfer Learning Methods: Evaluation of Different Model Performances, ...	119–127
Research Article	İsmail Kırbaş, Ahmet Çifci; Leveraging SHAP for Interpretable Diabetes Prediction: A Study of Machine Learning Models on the Pima Indians Diabetes Dataset,	128–139
Research Article	Süleyman Dal, Necmettin Sezgin; Heart Attack Classification with a Machine Learning Approach Based on the Random Forest Algorithm,	140–147
Research Article	Şilan Fidan Vural, Nida Kumbasar; Comparison of VT-based and CNN-based Models on Teeth Segmentation,	148–156
Research Article	Cemanur Aydinalp, Gülşah Yıldız Altıntaş; Breast Cancer Detectability and Tumor Differentiation based on Microwave Dielectric Property Changes with Reverse Time Migration,	157–163
Research Article	Aykut Satıcı; Control Through Contact using Mixture of Deep Neural-Net Experts,	164–173
Research Article	Ahmet Hamdi Özkurt, Emrah Aydemir, Yasin Sönmez; Large Language Models vs. Human Interpretation: Which is More Accurate in Text Classification?,	174–182
Research Article	Hadjer Brioua, Havvanur Siyambaş, Durmuş Özkan Şahin; Phishing E-mail Detection with Machine Learning and Deep Learning: Improving Classification Performance with Proposed New F.,	183–193
Research Article	Emre İrtəm, Nesli Erdoğan; Fingerprint Generation for DNN Training: A Case Study in Fingerprint Classification,	194–202
Research Article	Emrah Aslan, Yıldırım Özüpak; Performance Comparison of Deep Learning Models in Brain Tumor Classification,	203–209
Research Article	Vedat Yılmaz; A Bibliometric Analysis on Cybersecurity Using VOSviewer: An Evaluation for Public Security,	210–218
Research Article	Abdulkadir Gozuoglu; Intelligent Modular Energy Hub: Advanced Optimization of Second-Life Lithium-Based Batteries for Sustainable Power Utilization,	219–229
Research Article	Serdar Özyön, Hasan Temurtaş, Burhanettin Durmuş, Celal Yaşar; Application of Average Differential Evolution Algorithm to Lossy Fixed Head Short-Term Hydrothermal Coord. Pr.,	230–242

BALKAN JOURNAL OF ELECTRICAL & COMPUTER ENGINEERING

(An International Peer Reviewed, Indexed and Open Access Journal)

Contact

Batman University
Department of Electrical-Electronics Engineering
Bati Raman Campus Batman-Turkey

Web: <https://dergipark.org.tr/en/pub/bajece>
<http://www.bajece.com>
e-mail: bajece@hotmail.com

