AITA

2025 Volume 5 Number 1

ARTIFICIAL INTELLIGENCE THEORY AND APPLICATIONS

RESEARCH ARTICLES

REVIEW ARTICLE

https://aitajournal.com

info@aitajournal.com

ISSN: 2757-9778

ISBN: 978-605-69730-2-4

ARTIFICIAL INTELLIGENCE THEORY AND APPLICATIONS

Volume 5 – Issue 1

01.05.2025

e-ISSN: 2757-9778

Publisher: İzmir Bakırçay University

www.aitajournal.com

https://dergipark.org.tr/en/pub/aita

Journal Board

Editorial Board

International Advisory Board

International Indexes

ACADEMIC RESOURCE INDEX: RESEARCH BIBLE

INDEX COPERNICUS INTERNATIONAL

SOBIAD CITATION INDEX

ASIAN SCIENCE CITATION INDEX

INTERNATIONAL INSTITUTE OF ORGANIZED RESEARCH

EUROPUB INDEX

DIRECTORY OF RESEARCH JOURNALS INDEXING

About Journal

Artificial Intelligence Theory and Applications (AITA) provides coverage of the most significant work on principles of artificial intelligence, broadly interpreted. The scope of research we cover encompasses contributions of lasting value to any area of artificial intelligence. To be accepted, a paper must be judged to be truly outstanding in its field. AITA is interested in work in core artificial intelligence and at the boundaries, both the boundaries of sub-disciplines of artificial intelligence and the boundaries between artificial intelligence and other fields.

Scope

The best indicator of the scope of the journal is provided by the areas covered by its Editorial Board in theoretical (artificial intelligence and computing methodologies) and practical (artificial intelligence applications and applied computing) ways. These areas change from time to time, as the field evolves.

Year : 2025
Volume : 5
Issue : 1

ARTIFICIAL INTELLIGENCE THEORY AND APPLICATIONS

Table of Content **Page**

■ RESEARCH ARTICLE

Classification of Diabetic Retinopathy Disease with Deep Learning Methods

Metin Tuncel [a†], **Murat Uçar** [a]

[a] Department of Computer Engineering, İzmir Bakırçay University, İzmir, Türkiye
† tuncel155@gmail.com, corresponding author

RECEIVED DECEMBER 05, 2024
ACCEPTED APRIL 15, 2025

CITATION Tuncel, M. & Uçar, M. (2025). Classification of diabetic retinopathy disease with deep learning methods. *Artificial Intelligence Theory and Applications, 5(1*), 1-17.

Abstract

Diabetes is defined as a chronic disease caused by an increase in blood glucose levels (hyperglycemia), in which the organism is unable to make adequate use of carbohydrates, fats, and proteins due to the pancreas' inability to produce enough insulin hormone or the hormone's inability to function properly. Diabetes is the most severe chronic condition, according to a World Health Organization report. Diabetic retinopathy (DR) is a complication of type 1 diabetes. Diabetes problems can cause damage to the blood vessels in the light-sensitive tissue (retina) at the rear of the eye, resulting in DR. Diabetes is one of the top three causes of blindness, according to the International Diabetes Federation's Diabetes Atlas 10th Edition (2021). Diabetes-related blindness is mostly caused by the loss of small vessels in the retina because of chronic hyperglycemia. Approximately 25% of individuals with diabetes globally have DR of any severity. Our country has around 2 million diabetes patients, with DR accounting for 25% of the total. There are five categories of DR. These are non-proliferative diabetic retinopathy (NPDR), mild non-proliferative retinopathy, moderate non-proliferative retinopathy, severe non-proliferative retinopathy, and proliferative diabetic retinopathy (PDR), in order of severity. Using the APTOS2019 dataset, this study develops a computer-aided diagnosis system to assist doctors in making early diagnoses using convolutional deep learning (DL) models. Binary and multi-class classification was done utilizing cutting-edge models such as VGG16, InceptionResNetV2, ResNet152V2, EfficientNetB0, and MobileNetV2, which are extensively used in the literature for medical image classification. Since the amount of data in the multi-class classification in diabetic retinopathy disease images was not equal, the data were equalized utilizing data augmentation techniques in the training dataset with the Albumentations library. Among the cutting-edge models employed in binary classification, VGG16 performed best, with accuracy, precision, sensitivity, and F1-score metric values of 0.97. VGG16 was the best model employed in multi-class classification, with accuracy, precision, sensitivity, and f1-score metric values of 0.78 and 0.79, respectively.

Keywords: diabetic retinopathy, deep learning, convolutional neural networks, transfer learning

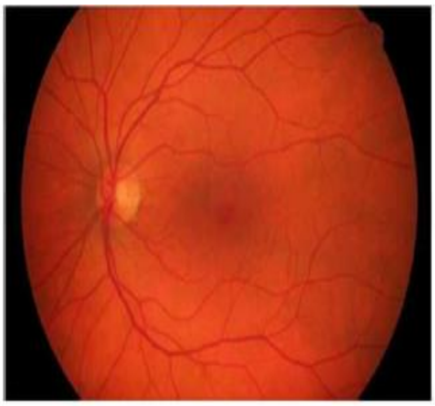
1. Introduction

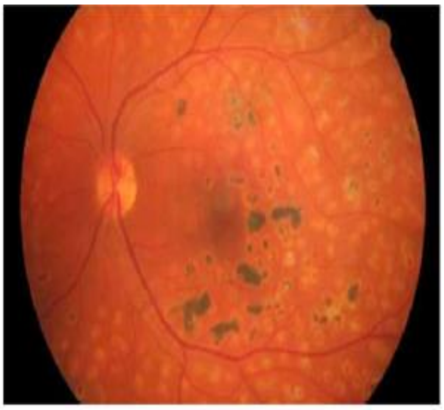
Diabetes is a rapidly increasing global health problem, with the number of people with diabetes predicted to rise to 643 million (11.3%) by 2030 and 783 million (12.2%) by 2045. [2,3]. High blood glucose is a common effect of uncontrolled diabetes. Over time, it causes serious damage to numerous systems of the body, especially the eyes, heart,

Artificial Intelligence Theory and Applications, ISSN: 2757-9778. ISBN: 978-605-69730-2-4 © 2025 İzmir Bakırçay University

kidneys, nerves, and blood vessels [2,4,5]. Diabetes is generally classified into 4 groups: Type I, Type II, gestational diabetes, and other specific types. The most common types of diabetes are Type I and Type II diabetes. The type of diabetes that occurs during pregnancy is defined as gestational diabetes, while other types are defined as high blood glucose levels that occur for many reasons and affect the pancreas [1].

Diabetic retinopathy can progress slowly or quickly, yet it can also improve on its own. However, if it worsens, it may result in partial or permanent vision loss.





a) Healthy retina image b) Retina with diabetic retinopathy

Figure 1. Sample images: a) healthy retina image, b) retina image with diabetic retinopathy.

Figure 2 shows hemorrhages, soft and hard exudates, and microaneurysms on the retina.

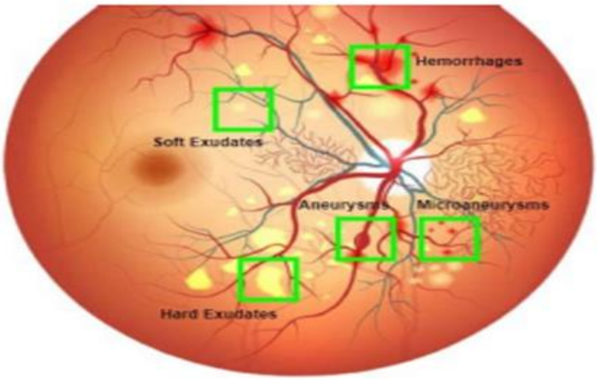


Figure 2. Fundoscopic illustration of the retina displaying microaneurysms, hemorrhages, and exudates [6]

There are 5 stages of diabetic retinopathy:[7]

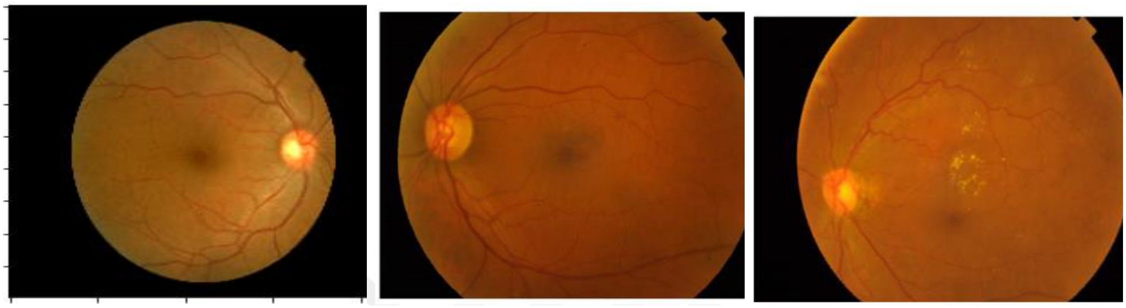
Non-proliferative diabetic retinopathy [7]: It can be seen in Figure 3a. This is the time when the sickness first appears. During this period, fluid leakage occurs in the damaged vessels, causing retinal hemorrhages. In general, the patient's vision is not affected during this period [8].

Mild non-proliferative retinopathy [7]: It can be seen in Figure 3b. It represents the first stage of diabetic retinopathy. Swellings (microaneurysms) occur in small blood vessels in the rear of the eye (retina) [9, 10].

Moderate non-proliferative retinopathy [7]: It can be seen in Figure 3c. Diabetic macular edema occurs as a result of the accumulation of blood and other fluids in the small central part of the retina (macula). Visual problems are also seen at this stage due to diabetic macular edema [9, 10].

Severe non-proliferative retinopathy [7]: It can be seen in Figure 3d. At this stage, new blood vessels and scar tissue are formed. Some or all of the blood vessels are occluded. Complete occlusion of blood vessels is called macular ischemia. As a result of this condition in the blood vessels, dark spots (flying objects) form in the visual field, and this causes blurred vision. At this stage, the possibility of visual loss is quite high [9, 10].

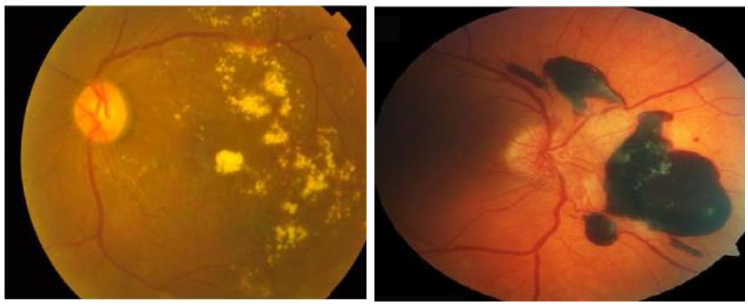
Proliferative diabetic retinopathy [7]: It can be seen in Figure 3e. This figure represents the most dangerous period of diabetic retinopathy. During this period, the blood vessels in the retinal layer are severely impaired, resulting in the formation of areas that cannot be nourished. These areas cause the development of new blood vessels. Since these vessels are very thin and fragile, they can cause sudden bleeding in the eye. The patient experiencing this process notices black spots appearing in front of the eye and moving in that direction, wherever the eye is turned. Patients with significant hemorrhage suffer visual loss and blindness [9, 10].



a) Non-proliferative diabetic retinopathy (NPDR) retinal image

b) Retinal image of mild DR [11].

c) Retinal image of moderate DR [11].



d) Retinal image of severe DR.

e) The retinal image of proliferative DR [11].

Figure 3. Sample images for the stages of diabetic retinopathy

In recent years, DL methods have been frequently used in object recognition, image classification, and segmentation in medical and ophthalmological images, and very successful results have been obtained. In particular, deep convolutional neural networks (DNNs) have been used for early detection and identification of retinal diseases such as DR, age-related macular degeneration, and glaucoma from retinal images [12,13,14,15]

In this study, image classification for DR was performed using DL methods. VGG16, InceptionResNetV2, ResNet152V2, EfficientNetB0, and MobileNetV2 transfer learning (TL) architectures from state-of-the-art models were used. Binary and multi-class classification was performed for DR. Since the amount of data in the multi-class classification was not equal, the data was equalized utilizing data augmentation techniques using the albumentations library in the training dataset.

This study contributes to the literature by demonstrating the effectiveness of deep learning and transfer learning methods in the early diagnosis of DR. The successful use of advanced models such as VGG16, InceptionResNetV2, ResNet152V2, EfficientNetB0, and MobileNetV2 in DR classification brings a significant innovation to the existing knowledge in this field. Additionally, addressing class imbalances through data augmentation techniques leads to more accurate and reliable classification results.

The rest of this paper is organized as follows. Section 2 presents the literature review. In Section 3, the proposed methods and the dataset used in the study are described in detail. Section 4 presents the results obtained in the study. The discussion and conclusion section of the paper is presented in Section 5, where a comparison with the findings of similar studies in the literature is made.

2. Literature Review

Cao et al. suggested a DL model with three major components. First, a transducer structure was incorporated into a convolutional neural network (CNN) to effectively utilize both local and long-range information. Second, disease details were collected from multiple images before self-attention was applied to improve inter-image interactions and reduce overfitting. Finally, an attention-based image transformation approach was proposed to filter information from different stages of the feature maps and adaptively capture lesion-related details. Their experiments produced a multi-class accuracy of 85.96% on the APTOS dataset and a binary class accuracy of 95.33% on the Messidor dataset, exceeding current approaches [16].

The aim of the study by Chandra et al. in 2024 was to create a CNN-based model for detecting and categorizing diabetic retinopathy using the APTOS dataset. The APTOS dataset was a large, open-access collection of fundus images that ophthalmologists analyzed for the likelihood and severity of DR. The accuracy obtained through the CNN model and the AlexNet model was 97% and 93% in quintuple classification, respectively [17].

Ohri et al. employed vision transformer-based deep learning models to classify DR diseases. The vision transformer-based DL architecture can classify fundus images into one of the DR categories by pre-training unlabeled fundus images before undertaking supervised training on labeled data. This study used DINO[VIT], MAE[VIT], and MSN[VIT] transducer-based DL architectures. The best MAE[VIT] transducer-based design achieved kappa values of 0.8341 in the low data regime and 0.9027 in the full data regime on 50 EyePACS training data using the Masked Autocoder framework [18].

Oulhadj et al. used four datasets, including EyePACS, DDR, Messidor-2, and APTOS datasets, for their study. As a preprocessing step on fundus images, they applied contrast-limiting adaptive histogram equalization and power law transformation approaches to classify DR illness using a modified capsule network-based and a fine-tuned vision transducer hybrid DL method. On the EyePACS, DDR, Messidor-2, and APTOS datasets, their efforts produced flawless test accuracy scores of 78.64%, 80.36%, 87.78%, and 88.18%, respectively [19].

In 2023, Mondal et al. combined the updated ResNet and DenseNet101 models, an enhanced version of the ResNet model for DR classification, to propose an automated ensemble DL model for better feature extraction. Two datasets, APTOS19 and DIARETDB1, were used in experiments for binary and multi-class classification. They preprocessed the images using the CLAHE approach for histogram equalization. For data augmentation, they used a GAN-based boosting strategy because of the dataset's severe class imbalance. The accuracy of the suggested approach is 86.08% for multi-class classification and 96.98% for binary classification [20].

Vijayan et al. proposed an early detection approach for automatic detection of DR that uses regression rather than multiclass classification using the convolution-based EfficientNetB0 model, one of the most advanced TL models. Better generalization was the initial advantage of the regression problem approach, and finer-grained predictions were made possible by the model's ability to give a value that falls between conventionally distinct labels. The APTOS and DDR datasets were used to test the idea. The DDR dataset had an accuracy of 84.80, whereas the APTOS dataset had an accuracy of 86.20 [21].

By merging the pyramid hierarchy of the discrete wavelet transform of the retinal fundus image with a modified capsule network and a proposed modified starting block, Oulhadj et al. discovered a novel deep hybrid model for determining the severity level of DR. Their proposed method's performance was assessed using the APTOS dataset, yielding training accuracy ratings of 97.71% and test accuracy scores of 86.54% [22].

In another study, Oulhadj et al. proposed comparing ensemble voting and five cutting-edge TL CNN models (ResNet50, DenseNet121, VGG16, InceptionV3, and Xception) for the automatic assessment of the severity of diabetic retinal disease. The results of five cutting-edge models were used by the community voting to make its choice. Using the APTOS dataset for training and testing, the suggested effort produced an accuracy of 83.63 [23].

A novel two-step convolutional DL-based approach for automatic DR detection was presented by Oulhadj et al. In the first step, known as preprocessing, the background influence was eliminated from the classification process by applying the deformable registration that covers the entire image to the retina. To identify the stage of DR, they trained four cutting-edge TL CNN models (ResNet50, InceptionV3, Xception, and DenseNet121) in the second step. The APTOS 2019 dataset was used to evaluate the proposed architecture's performance, and their model's accuracy was 85.28% [24].

Islam et al. used the APTOS 2019 dataset to automatically predict DR severity from fundus images. The state-of-the-art TL CNN model Xception was chosen as the encoder, and CLAHE was used to enhance image quality. For the binary classification of DR, the suggested model produced an AUC score of 98.50% and a test accuracy of 98.36%; for the multi-class classification using the APTOS 2019 dataset, the AUC score was 93.819% and the test accuracy was 84.364% [25].

3. Materials and Methods

This section provides detailed information about the dataset, methods, and evaluation criteria used in the study.

3.1. Dataset

APTOS2019 [11] was used in this study. Tables 1 and 2 indicate the quantity of data used for binary and multi-class classification. There are 3662 data points in this dataset. In binary and multi-class classification, 72% of the total data was used for training, 20% for testing, and 8% for validation. As a result of employing Albumentations library data augmentation methods, there are 1300 data points in each class, for a total of 6500 data points.

Table 1. Data Amounts for Binary Classification in Training, Testing, and Validation

	Train %72	**Test %20**	**Validation %8**	**Total**
0-NPDR	1300	361	144	1805
1-DR	1336	372	149	1857
Total	2636	733	293	3662

Table 2. Data Amounts for Multi-Class Classification in Training, Testing, and Validation

	Train (DataAugmented)	**Train %72**	**Test %20**	**Validation %8**	**Total**
0-NPDR	1300	1300	361	144	1805
1-MildNPDR	1300	266	74	30	370
2-ModerateNPDR	1300	719	200	80	999
3-SevereNPDR	1300	139	39	15	193
4-ProliferateDR	1300	212	59	24	295
Total	-	2636	733	293	3662
Total(DataAugmented)	6500	-	-	-	-

3.2. Method

In this study, VGG16, MobileNetV2, ResNet152V2, InceptionResNetV2, and EfficientNetB0 state-of-the-art DL models, which are frequently preferred for image classification in the field of healthcare, are used.

The Visual Geometry Group proposed the VGG16 CNN architecture. Figure 4 shows the VGG16 network architecture, which consists of 13 convolutional layers, three fully connected layers, and five pooling layers [26]. The step size is two, while the kernel size in the pooling layers is 2 × 2. In Step 1, the convolution kernel size is set to 3×3 in the convolutional layers. The rectified linear unit (ReLU) is the activation function for convolutional layers. The VGG-16 network receives an image with dimensions of 224 by 224 pixels and three channels. In the initial portion, two convolutional layers are followed by a pooling layer. Each of these convolutional layers has 64 cores and measures 224 x 224 pixels. Each of these convolutional layers has 64 cores with 224 x 224 pixels.

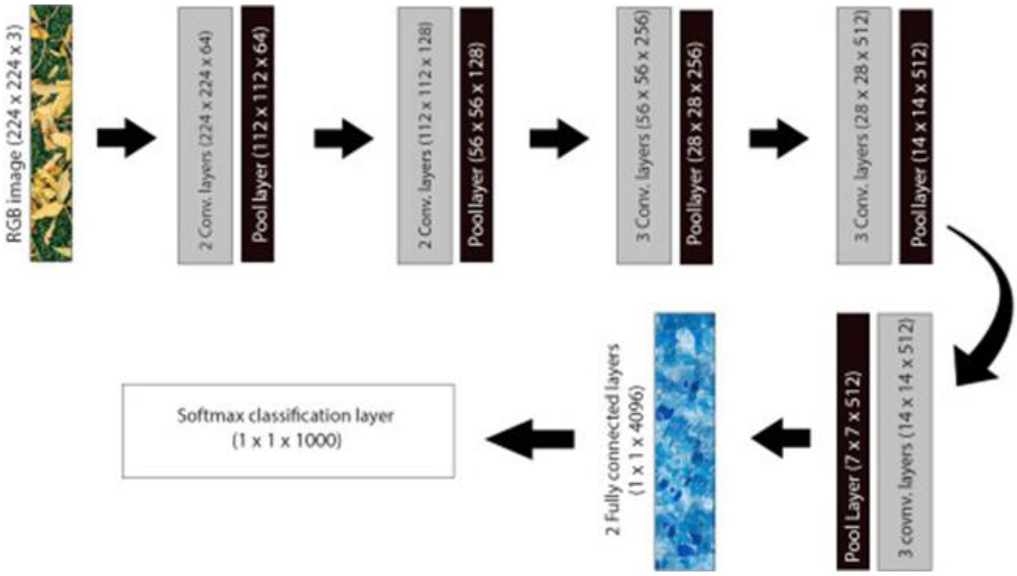


Figure 4. VGG16 Architecture Illustration [27]

The MobileNetV2 architecture shown in Figure 5 is a CNN architecture designed to work effectively on mobile devices. It is based on the inverted residual structure and enhanced with bottleneck features. This architecture improves the performance of mobile models, enabling more efficient results with lower computational requirements [28].

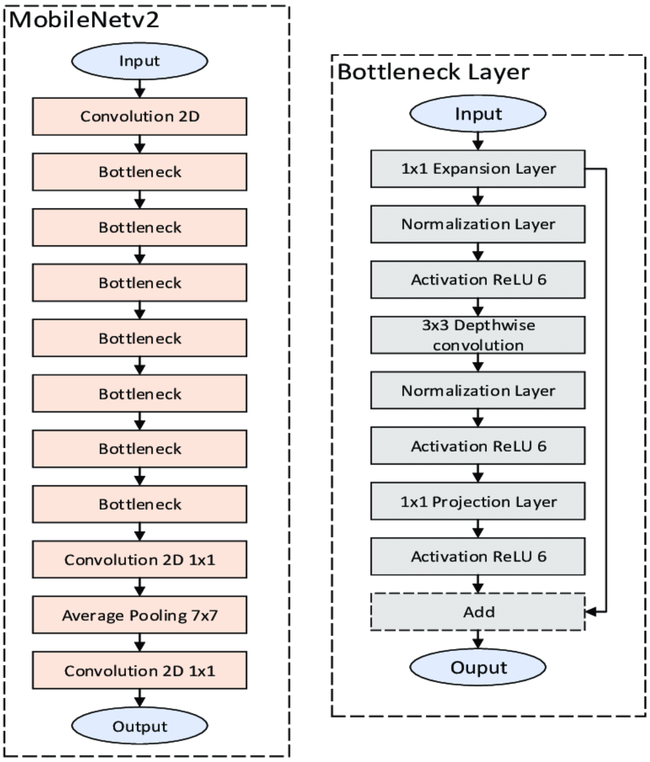


Figure 5. MobilNetV2 Architecture Illustration [29]

The ResNet152V2 architecture shown in Figure 6 is a CNN architecture used in DL. It is known for providing high accuracy, especially in image recognition tasks. The architecture includes 152 levels, improves on prior ResNet architectures, and supports the training of deeper layers utilizing residual connections [30].

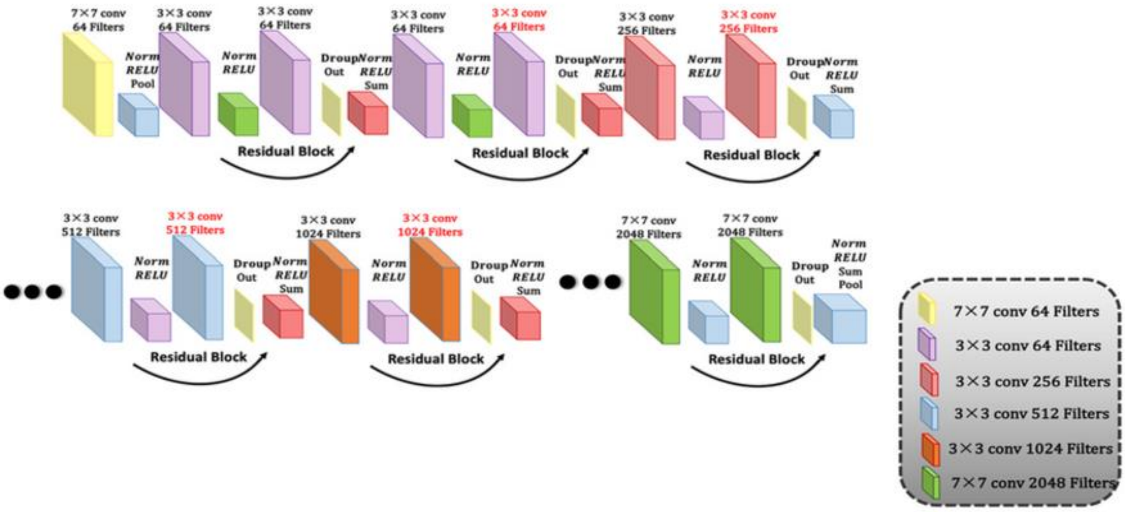


Figure 6. ResNet152V2 Architecture Illustration [31]

Figure 7 shows the Inception-ResNet-v2 architecture, which is based on the Inception architectural family. It's a CNN that combines residual connections. Training with residual connections considerably accelerates the training of initial networks [32].

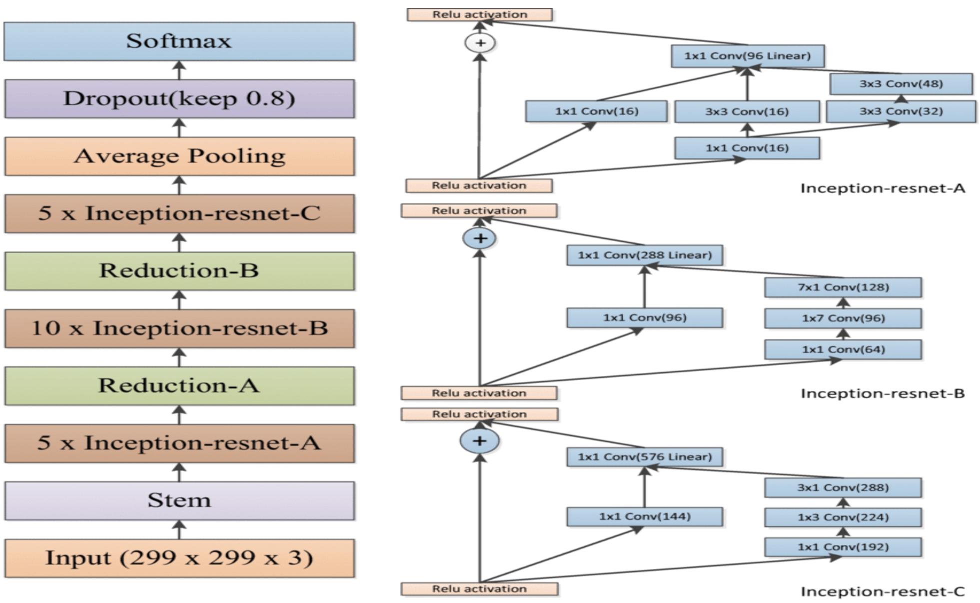


Figure 7. InceptionResNetV2 architecture [33]

The EfficientNetB0 architecture, shown in Figure 8, has a balanced design. It improves computational performance by extracting significant information from the input. EfficientNet uses a composite scaling approach that scales its layers equally in depth, width, and resolution [34].

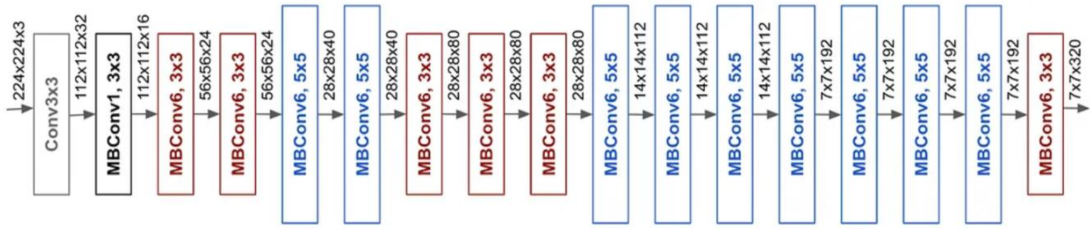


Figure 8. EfficientNetB0 architecture [35]

All models were trained by transfer learning (TL). As the initial weights of the model parameters, the weights resulting from the training on the ImageNet dataset were used. The stochastic gradient descent (SGD) method was used as the optimizer. Categorical cross-entropy was used as the loss function. The learning rate value was set as 0.00005. To reduce overfitting, the 'patience' parameter in EarlyStopping was set to 10, and training was stopped if there was no improvement in the verification loss for 10 epochs. Although the epochs were set to 1000 due to early stopping, the EfficientNetB0 model completed the binary and multi-class classification in 620 steps at most.

For data augmentation, the images were rotated horizontally and symmetrically using methods from the albumentations library. Random brightness and contrast transformation was applied in the range [-1, 1]. Random gamma correction was performed on the images. Grid distortion was applied to the image. Optical distortion in the range [-2, 2] and shift in the range [- 0.5, 0.5] were applied to the image with 50% probability. Shifting, scaling, and rotation operations were combined in the image. The images were resized to 224×224 dimensions.

3.3. Evaluation Criteria

In this study, binary and multi-class classifications were made, and the performances of the models used were evaluated using different metrics such as accuracy, precision, recall, and F1 score. These metrics given in Equations 1-4 are calculated using values such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) obtained in the confusion matrix.

$$Accuracy = \frac{(Tp + Tn)}{(Tp + Tn + Fp + Fn)} \tag{1}$$

$$Precision = \frac{Tp}{(Tp + Fp)} \tag{2}$$

$$Recall = \frac{Tp}{(Tp + Tn)} \tag{3}$$

$$F1 = 2 * \frac{precision * recall}{(precision + recall)} \tag{4}$$

4. Findings

4.1 Binary Classification Findings

The obtained results for the binary classification were listed in Table 3. The confusion matrices displaying the binary classification results were given in Figure 9.

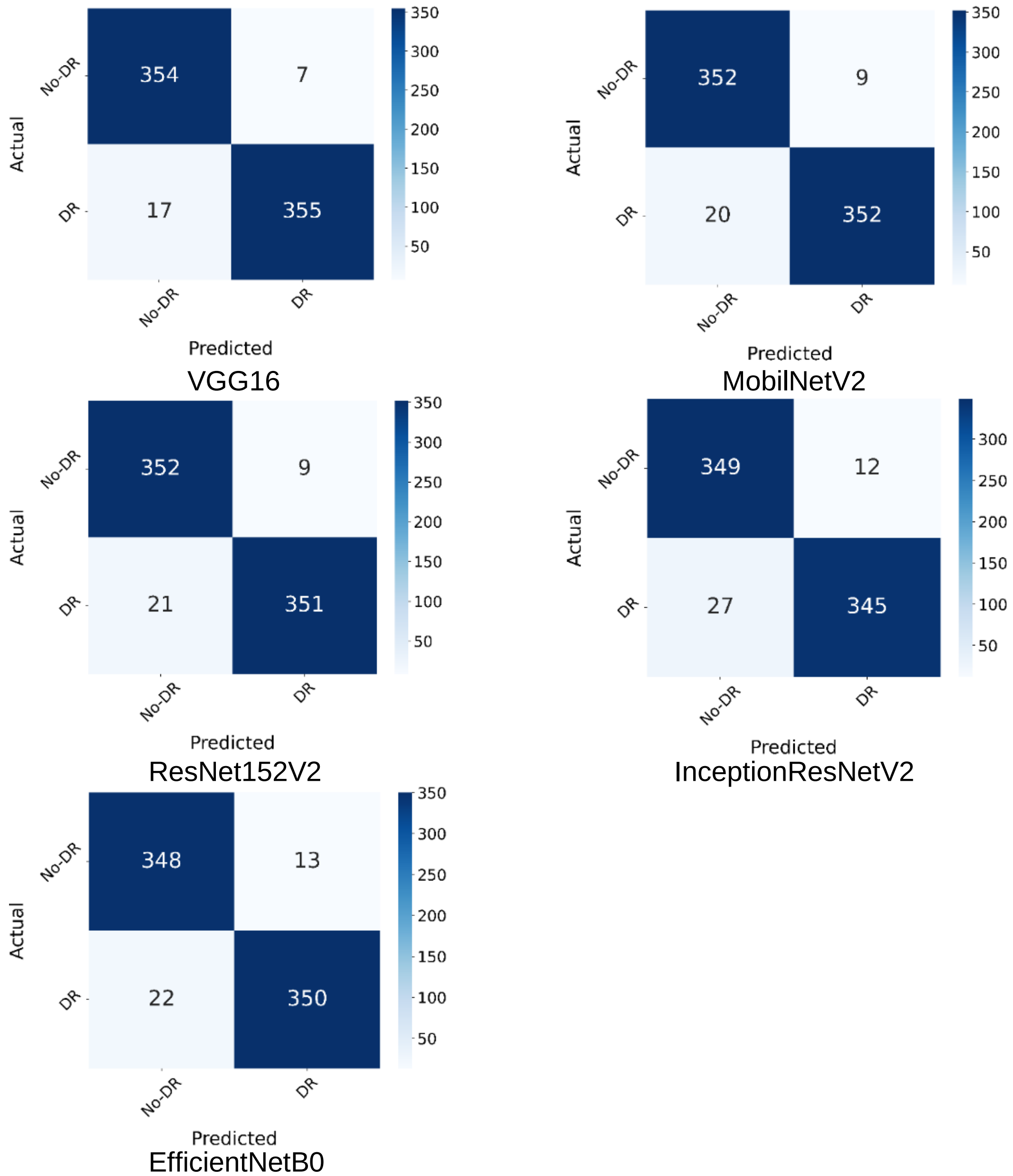


Figure 9. Confusion Matrices of Binary Classification

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Support**
VGG16	**0.97**	**0.97**	**0.97**	**0.97**	**733**
MobileNetV2	0.96	0.96	0.96	0.96	733
ResNet152V2	0.96	0.96	0.96	0.96	733
InceptionResNetV2	0.95	0.95	0.95	0.95	733
EfficientNetB0	0.95	0.95	0.95	0.95	733

Table 3. Comparison of Test Metrics for State-of-the-Art Models in Binary Classification

The best results in binary classification were obtained by the VGG16 model with an accuracy of 0.97. In the VGG16 model, according to the confusion matrix in Figure 9, 709 images were correctly classified, and 24 images were wrongly classified in the binary classification in the test data. In the No-DR class, 354 images were correctly classified and 7 incorrectly classified, while the DR class had 355 correctly classified images and 17 incorrectly classified images. The VGG16 model was followed by the MobileNetV2 and ResNet152V2 models, with an accuracy of 0.96, while the InceptionResNetV2 and EfficientNetB0 models came last with an accuracy of 0.95.

4.2 Multi-class classification findings (without data augmentation)

The multi-class classification results obtained for the InceptionResNetV2 model without data augmentation were presented in Table 4. The confusion matrix displaying the multi-class classification results of the InceptionResNetV2 model was given in Figure 10.

Table 4. Performance Metrics for Multi-Class Classification Using InceptionResNetV2 Model Without Data Augmentation

Model	Accuracy	Precision	Recall	F1-Score	Support
InceptionResNetV2 Without Data Augmentation	0.73	0.68	0.73	0.67	733

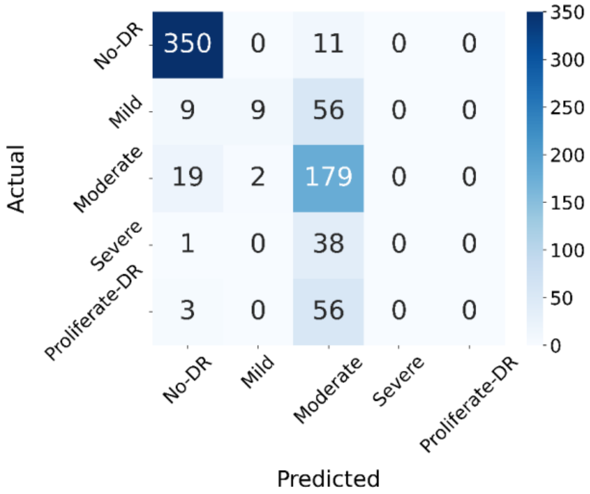
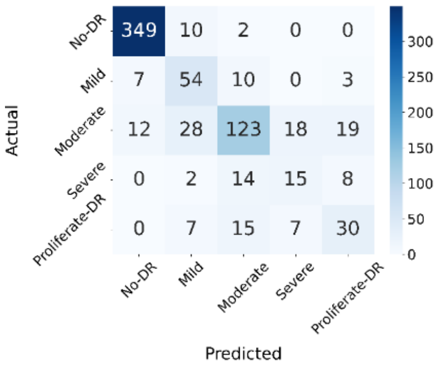


Figure 10. Confusion Matrix for Multi-Class Classification of InceptionResNetV2 Without Data Augmentation

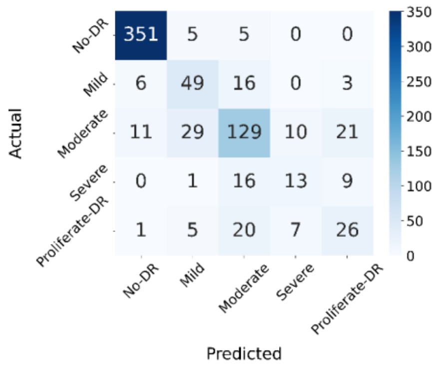
Without data augmentation, the InceptionResNetV2 model yielded multi-class accuracy and sensitivity scores of 0.73, precision of 0.68, and an F1 score of 0.67. In the test data, 538 photos were correctly classified and 195 were wrongly classified, as shown by the confusion matrix in Figure 10. The network was unable to learn sufficiently due to the lack of classification in the Severe and Proliferate DR classes; hence, Albumentations library methods were used to enrich the data.

4.3 Multi-class classification findings (with data augmentation)

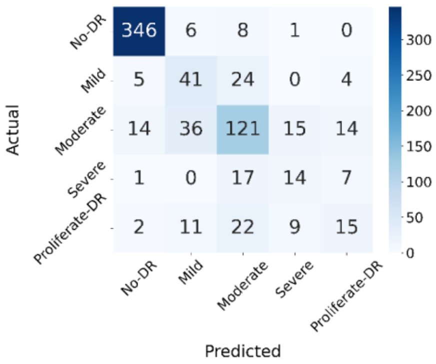
The obtained results for the multi-class classification were listed in Table 5. The confusion matrices displaying the multi-class classification results were given in Figure 11.



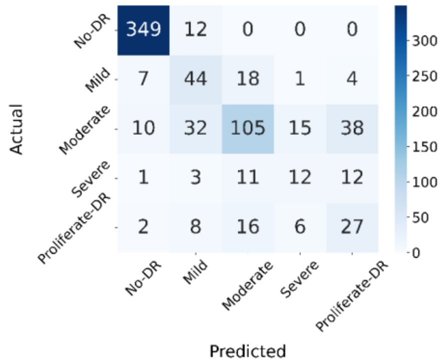
VGG16



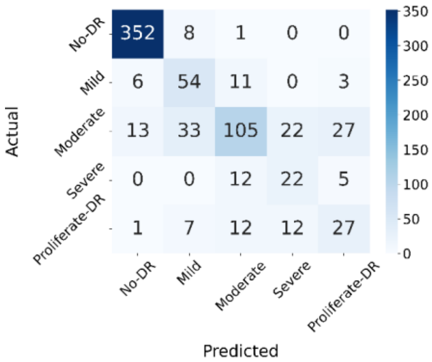
MobilNetV2



ResNet152V2



InceptionResNetV2



EfficientNetB0

Figure 11. Confusion Matrices for Multi-Class Classification

Table 5. Comparison of Performance Metrics for State-of-the-Art Models in Multi-Class Classification

Model	Accuracy	Precision	Recall	F1-Score	Support
VGG16	**0.78**	**0.79**	**0.78**	**0.78**	**733**
MobileNetV2	0.77	0.77	0.77	0.77	733
ResNet152V2	0.73	0.73	0.73	0.73	733
InceptionResNetV2	0.73	0.75	0.73	0.73	733
EfficientNetB0	0.76	0.78	0.76	0.76	733

In multi-class classification, the VGG16 model had the best accuracy value with 0.78. As can be seen from the confusion matrix in Figure 11, the VGG16 model correctly classified 571 images and incorrectly classified 162 images on the test dataset in multi-class classification. There were 349 correct and 12 incorrect images classified in the No-DR class, 54 correct and 20 incorrect images in the Mild class, 123 correct and 77 incorrect images in the Moderate class, 15 correct and 24 incorrect images in the Severe class, and 30 correct and 29 incorrect images in the Proliferate-DR class. In terms of accuracy rankings, MobileNetV2 ranks second with a value of 0.77 on the data-augmented dataset, followed by EfficientNetB0 in third place with an accuracy value of 0.76, and the ResNet152V2 and InceptionResNetV2 models in last place with an accuracy value of 0.73.

5. Discussion and Conclusion

Studies on diabetic retinopathy using the APTOS2019 dataset are displayed in Table 6. In the best model, VGG16, our suggested method for binary classification had a 97% success rate in the accuracy metric. Our method produced results that were comparable to the accuracy metric success rates of Mondal et al. (96.98%) and Kumar et al. (96.24%). Our proposed approach in multi-class classification achieved 78% success in the accuracy metric with the best model, VGG16, and remained below the success achieved in other studies. In the literature, binary and multi-class classification studies have not been conducted on ResNet152V2 and MobileNetV2 models, which are state-of-the-art models. We contributed to the literature by performing binary and multi-class classification studies on ResNet152V2 and MobileNetV2 models.

Table 6. Previous Studies on Diabetic Retinopathy Using the APTOS2019 Dataset

Authors	**Classes**	**Method/Data Set(Number of Data)**	**Approach/Algorithm**	**Metric(%)**
Proposed approach(2024)	2 5	CNN/ APTOS(3662)	VGG16**,** InceptionResNetV2, ResNet152V2, EfficientNetB0, MobileNetV2	**Accuracy,Precision,Recall, F1-Score**:**97.00**(VGG16 Binary Classification)- **Accuracy,Recall,F1-Score:78.00,Precision:79.00** (VGG16 Data Agumentation Multi Class Classificaiton)
Cao X. Et al.(2024)[16]	5	CNN,Vit / APTOS(3662)	CNN,Vit	Accuracy:85,96
Mondal et al. (2023) [36]	2 5	CNN/ APTOS(3662)	Ensemble Deep-Learning Technique(DenseNet101 ve ResNeXt)	Accuracy: 96.98-86.08 Precision: 97.00-76.00 Recall: 97.00-82.00
Oulhadj et al. (2023) [37]	5	APTOS(3662)	ViT+ CapsNet+ PLT+CLAHE	Accuracy : 88.18 Precision: 80.00 Recall : 76.00 F1-score : 78.00 Kappa score: 81.55
Vijayan et al. (2023) [21]	5	CNN/ APTOS(3662)	Efficientnet-B0	Accuracy:86.20
Oulhadj et al. (2023) [22]	5	CNN/ APTOS(3662)	CapsNet + Inception Block + DWT	Accuracy : 86.54 Kappa score : 78.77 Precision: 76.00 Recall : 70.00 F1-score: 73.00
Oulhadj et al. (2023) [23]	5	CNN/ APTOS(3662)	Transferred Learning + Voting((Xception, InceptionV3, VGG16, DenseNet121, Resnet50))	Accuracy:83.63

M. Oulhadj et al (2022) [24]	5	CNN/APTOS(3662)	Ensemble Voting(Densenet-121, Xception, Inception-v3, Resnet-50)	Accuracy:85.28 Precision:80.00 Recall:70.00 F1-Score:73.00
Islam et al. (2022)[25]	2 5	CNN/ APTOS(3662)	Supervised Contrastive Learning(Xception)	Accuracy: 98.36-84.36 Precision: 98.37-73.84 Recall: 98.36-70.51 F1-Score: 98.37-70.49 AUC: 98.50-93.82
Bodapati et al. (2022) [38]	5	Attention based CNN/ APTOS(3662)	Stacked Convolutional Auto-Encoder(VGG16, Inception, ResNet Version2 (IRV2) , Xception)	Accuracy:84.17
Zhao et al. (2022)[39]	5	CNN/ APTOS(3662)	CoTXNet	Accuracy:84.18 Kappa score:90.00
Shaik and Cherukuri (2022)[40]	5	CNN/ APTOS(3662)	Hinge Attention Networks	Accuracy:85.54
Hu et al. (2022) [41]	2 5	CNN/ APTOS(3662)	Graph Adversarial	Accuracy: 94.30-83.50
Fan et al. (2021) [42]	5	CNN/ APTOS(3662)	Multi-Scale Features (MobileNetV3)	Accuracy:85.32 Kappa score: 77.26 F1-Score:85.30 AUC:97.00
Sugeno et al. (2021) [43]	5	CNN/APTOS(3662)	Transfer Learning + EfficientNet-B3	Accuracy:84.20
Al-Antary and Arafa (2021) [44]	2 5	CNN/ APTOS(3662)	Extraction of Features + Attention	Accuracy: 98.10- 84.60 Kappa score: -89.60 AUC: 98.20 -
Kumar et al. (2021) [45]	2 5	CNN/ APTOS(3662)	VGG16+Capsule Network+Hybrid Deep Learning, DRISTI	Accuracy: 96.24-82.06

The binary classification scenario, in which the dataset's images were categorized as sick and the healthiest (npdr), yielded the best test results. Without data augmentation, the results of the multi-class classification trials, which also attempted to predict the disease's intensity, were poor. The models were biased toward classes with more examples in the learning phase since the dataset used to classify the disease by level was an unbalanced dataset. Test findings showed satisfactory success in the tests where training data sets were equalized using data augmentation approaches. In a multi-class classification, the data augmentation strategy has been demonstrated to be advantageous when treating diabetic retinal disease.

Transformer-based architectures (Vision Transformer, VitDet, Swin Transformer, etc.) can be used in future research to do classification studies for the early diagnosis of diabetic retinopathy disease. Convolution and transformation-based DL models can be used to evaluate fundus images of diabetic retinopathy patients in the hospitals, and research can be conducted to assist the ophthalmologists in making early diagnoses.

Acknowledgement

This paper is a comprehensively expanded version of the proceedings abstract presented at the IV. International Congress on Artificial Intelligence in Healthcare.

References

[1] Coşansu, G. (2015). Diyabet: Küresel bir salgın hastalık. *Okmeydanı Tıp Dergisi*, 31, 1-6.

[2] World Health Organization. (2021). Diabetes. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/diabetes

[3] International Diabetes Federation. (2021). IDF Diabetes Atlas | Tenth Edition. Retrieved from https://www.idf.org/news/240:diabetes-now-affects-one-in-10-adults-worldwide.html and https://diabetesatlas.org/atlas/tenth-edition/

[4] Grosman, B., Ilany, J., Roy, A., Kurtz, N., Wu, D., Parikh, N., Voskanyan, G., Konvalina, N., Mylonas, C., Gottlieb, R., Kaufman, F., & Cohen, O. (2016). Hybrid closed-loop insulin delivery in type 1 diabetes during supervised outpatient conditions. Journal of Diabetes Science and Technology, 10(3), 708–713. https://doi.org/10.1177/1932296816631568

[5] Swapna, G., Vinayakumar, R., & Soman, K. P. (2018). Diabetes detection using deep learning algorithms. ICT Express, 4(4), 243–246. https://doi.org/10.1016/j.icte.2018.10.005

[6] Atwany, M. Z., Sahyoun, A. H., & Yaqub, M. (2022). Deep learning techniques for diabetic retinopathy classification: A survey. IEEE Access, 10, 28642-28655. https://doi.org/10.1109/ACCESS.2022.3157632

[7] Ağca, K. (2022). Evrişimsel sinir ağları kullanarak diyabetik retinopati hastalığının tespiti [Yüksek lisans tezi, Sivas Cumhuriyet Üniversitesi]. YÖK Ulusal Tez Merkezi. https://tez.yok.gov.tr

[8] İnan, S. (2014). Diabetik retinopati ve etiyopatogenezi. Kocatepe Tıp Dergisi, 15(2), 207-217.

[9] Cunha, J. P. (2021). What are the stages of diabetic retinopathy? eMedicineHealth. Retrieved from https://www.emedicinehealth.com/what_are_the_stages_of_diabetic_retinopathy/article_em.htm

[10] Yakar, H. K. (2018). Yaşlılıkta Diyabetin Diğer Bir Yüzü: Diyabetik Retinopati Ve Düşmeler. Izmir Democracy University Health Sciences Journal, 1(2), 13-22.

[11] APTOS. (2019). Blindness detection. Kaggle. Retrieved from https://www.kaggle.com/c/aptos2019-blindness-detection/overview/aptos-2019

[12] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

[13] Lee, C. S., Baughman, D. M., & Lee, A. Y. (2017). Deep learning is effective for classifying normal versus age-related macular degeneration OCT images. Ophthalmology Retina, 1(4), 322–327. https://doi.org/10.1016/j.oret.2016.12.009

[14] Chen, X., Xu, Y., Wong, D. W. K., Wong, T. Y., & Liu, J. (2015, August). Glaucoma detection based on deep convolutional neural network. In 2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC) (pp. 715-718). IEEE.

[15] Uçar, M. (2021). Glokom Hastalığının Evrişimli Sinir Ağı Mimarileri ile Tespiti. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 23(68), 521-529.

[16] Cao, X., Lin, J., Gao, X., & Li, Z. (2024). Integrating convolution and transformer for enhanced diabetic retinopathy detection. International Journal of Bio-Inspired Computation (IJBIC), 23(4). https://doi.org/10.1504/IJBIC.2024.139257

[17] Chandra, R., Tiwari, S., Kumar, S. S., & Agarwal, S. (2024). Diabetic retinopathy prediction based on CNN and AlexNet model. In 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 382–387). IEEE. https://doi.org/10.1109/Confluence60223.2024.10463351

[18] Ohri, K., Kumar, M., & Sukheja, D. (2023). Self-supervised approach for diabetic retinopathy severity detection using vision transformer. Pattern Analysis and Applications. https://doi.org/10.1007/s13748-024-00325-0

[19] Oulhadj, M., Riffi, J., Khodriss, C., Mahraz, A. M., Yahyaouy, A., Abdellaoui, M., Andaloussi, I. B., & Tairi, H. (2024). Diabetic retinopathy prediction based on vision transformer and modified capsule network. Computers in Biology and Medicine, 175, 108523. https://doi.org/10.1016/j.compbiomed.2024.108523

[20] Mondal, S. S., Mandal, N., Singh, K. K., Singh, A., & Izonin, I. (2023). EDLDR: An ensemble deep learning technique for detection and classification of diabetic retinopathy. Diagnostics, 13(1), 124. https://doi.org/10.3390/diagnostics13010124

[21] Vijayan, M., & Venkatakrishnan, S. (2023). A regression-based approach to diabetic retinopathy diagnosis using EfficientNet. Diagnostics, 13(4), 774. https://doi.org/10.3390/diagnostics13040774.

[22] Oulhadj, M., Riffi, J., Khodriss, C., Mahraz, A. M., Bennis, A., Yahyaouy, A., ... & Tairi, H. (2023). Diabetic retinopathy prediction based on wavelet decomposition and modified capsule network. Journal of Digital Imaging, 36(4), 1739-1751. https://doi.org/10.1007/s10278-023-00813-0

[23] Oulhadj, M., Riffi, J., Khodriss, C., Mahraz, A. M., Bennis, A., Yahyaouy, A., ... & Tairi, H. (2023, January). Diabetic Retinopathy Prediction Based on Transfer Learning and Ensemble Voting. In International Conference on Digital Technologies and Applications (pp. 929-937). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-29857-8_92

[24] Oulhadj, M., Riffi, J., Chaimae, K., Mahraz, A. M., Ahmed, B., Yahyaouy, A., ... & Tairi, H. (2022). Diabetic retinopathy prediction based on deep learning and deformable registration. Multimedia Tools and Applications, 81(20), 28709-28727. https://doi.org/10.1007/s11042-022-12968-z

[25] Islam, M. R., Abdulrazak, L. F., Nahiduzzaman, M., Goni, M. O. F., Anower, M. S., Ahsan, M., ... & Kowalski, M. (2022). Applying supervised contrastive learning for the detection of diabetic retinopathy and its severity levels from fundus images. Computers in biology and medicine, 146, 105602. https://doi.org/10.1016/j.compbiomed.2022.105602

[26] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In 3rd International Conference on Learning Representations (ICLR 2015) (pp. 1–14).

[27] Sharma, P., Nayak, D. R., Balabantaray, B. K., Tanveer, M., & Nayak, R. (2024). A survey on cancer detection via convolutional neural networks: Current challenges and future directions. Neural Networks, 169, 637-659. https://doi.org/10.1016/j.neunet.2023.11.006.

[28] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520).

[29] Tragoudaras, A., Stoikos, P., Fanaras, K., Tziouvaras, A., Floros, G., Dimitriou, G., ... & Stamoulis, G. (2022). Design space exploration of a sparse mobilenetv2 using high-level synthesis and sparse matrix techniques on FPGAs. Sensors, 22(12), 4318. https://doi.org/10.3390/s22124318

[30] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[31] Vatanpour, M., & Haddadnia, J. (2024). Brain tumour segmentation of MR images based on custom attention mechanism with transfer-learning. IET Image Processing, 18(4), 886-896. https://doi.org/10.1049/ipr2.12992

[32] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1). https://doi.org/10.1609/aaai.v31i1.11231

[33] Peng, C., Liu, Y., Yuan, X., & Chen, Q. (2022). Research of image recognition method based on enhanced inception-ResNet-V2. Multimedia Tools and Applications, 81(24), 34345-34365. https://doi.org/10.1007/s11042-022-12387-0

[34] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR. http://proceedings.mlr.press/v97/tan19a.html

[35] Montalbo, F. J. P., & Alon, A. S. (2021). Empirical analysis of a fine-tuned deep convolutional model in classifying and detecting malaria parasites from blood smears. KSII Transactions on Internet and Information Systems (TIIS), 15(1), 147-165. https://doi.org/10.3837/tiis.2021.01.009

[36] Mondal, S. S., Mandal, N., Singh, K. K., Singh, A., & Izonin, I. (2022). Edldr: An ensemble deep learning technique for detection and classification of diabetic retinopathy. Diagnostics, 13(1), 124. https://doi.org/10.3390/diagnostics13010124

[37] Oulhadj, M., Riffi, J., Khodriss, C., Mahraz, A. M., Yahyaouy, A., Abdellaoui, M., ... & Tairi, H. (2024). Diabetic retinopathy prediction based on vision transformer and modified capsule network. Computers in Biology and Medicine, 175, 108523. https://doi.org/10.1016/j.compbiomed.2024.108523.

[38] Bodapati, J. D. (2022). Stacked convolutional auto-encoder representations with spatial attention for efficient diabetic retinopathy diagnosis. Multimedia Tools and Applications, 81(22), 32033-32056. https://doi.org/10.1007/s11042-022-12811-5

[39] Zhao, S., Wu, Y., Tong, M., Yao, Y., Qian, W., & Qi, S. (2022). Cot-xnet: contextual transformer with xception network for diabetic retinopathy grading. Physics in Medicine & Biology, 67(24), 245003. https://doi.org/10.1088/1361-6560/ac9fa0

[40] Shaik, N. S., & Cherukuri, T. K. (2022). Hinge attention network: A joint model for diabetic retinopathy severity grading. Applied Intelligence, 52(13), 15105-15121. https://doi.org/10.1007/s10489-021-03043-5

[41] Hu, J., Wang, H., Wang, L., & Lu, Y. (2022). Graph adversarial transfer learning for diabetic retinopathy classification. IEEE Access, 10, 119071-119083. https://doi.org/10.1109/ACCESS.2022.3220776

[42] Fan, R., Liu, Y., & Zhang, R. (2021). Multi-scale feature fusion with adaptive weighting for diabetic retinopathy severity classification. Electronics, 10(12), 1369.

[43] Sugeno, A., Ishikawa, Y., Ohshima, T., & Muramatsu, R. (2021). Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning. Computers in Biology and Medicine, 137, 104795.

[44] Al-Antary, M. T., & Arafa, Y. (2021). Multi-scale attention network for diabetic retinopathy classification. IEEE Access, 9, 54190-54200.

[45] Kumar, G., Chatterjee, S., & Chattopadhyay, C. (2021). DRISTI: a hybrid deep neural network for diabetic retinopathy diagnosis. Signal, Image and Video Processing, 15(8), 1679-1686.

List of abbreviations

CNN : Convolutional Neural Network
DL : Deep Learning
DR : Diabetic Retinopathy
NPDR : Non-Proliferative Diabetic Retinopathy
PDR : Proliferative Diabetic Retinopathy
TL : Transfer Learning

■ RESEARCH ARTICLE

Mitigating Adversarial Attacks on ECG Classification in Federated Learning via Adversarial Training

Eyüpcan Çelik[a†], **Mehmet Kemal Güllü**[a]

[a] Department of Electrical and Electronics Engineering, İzmir Bakırçay University, İzmir, Türkiye
† 6017019@bakircay.edu.tr, corresponding author

RECEIVED DECEMBER 6, 2024
ACCEPTED JANUARY 29, 2025

CITATION Çelik, E., & Güllü, M. K. (2025). Mitigating adversarial attacks on ECG classification in federated learning via adversarial training. *Artificial Intelligence Theory and Applications, 5(1*), 18-28

Abstract

Federated Learning (FL) has become an important research area in recent years, particularly when dealing with sensitive data such as healthcare information. Since healthcare data contains critical and personal information, FL provides a major advantage by enabling training on local devices without requiring data to be collected on a central server. In the analysis of healthcare data, such as electrocardiography (ECG), FL enables local processing of data while preserving privacy. However, despite its privacy benefits, FL can be vulnerable to attacks. Malicious inputs aim to degrade model accuracy, known as adversarial attacks (AA), can pose a major threat. Adversarial Training (AT) offers a defence mechanism by increasing model's robustness against such attacks. Federated Adversarial Training (FAT) extends AT into the FL environment, combining privacy advantages with enhanced resistance to adversarial inputs. In this work, we propose the use of FAT to improve both privacy and security when classifying ECG signals, ensuring robustness against AAs. This approach involves applying AT at the client level by augmenting clean ECG data with adversarial examples generated using the Projected Gradient Descent (PGD) method. A Convolutional Neural Network (CNN) architecture was employed for local training. Experiments are conducted on the MIT-BIH Arrhythmia Database (MIT-DB). For comparison, we also trained an FL model without incorporating FAT. Both models were tested on the original test data as well as on adversarially attacked versions generated using PGD, Fast Gradient Sign Method (FGSM), Carlini & Wagner (CW), and Basic Iterative Method (BIM). The results show that the FL system with FAT significantly outperforms the system without FAT in resisting AAs, with a slight compromise in performance on the original test data, thus highlighting the effectiveness of FAT in enhancing model robustness against AAs for ECG classification tasks. Code is available at *https://github.com/Skyress1/ECG-FAT-Code*.

Keywords: federated adversarial training, federated learning, adversarial attacks, ECG

1. Introduction

Machine learning enables the development of more accurate and intelligent systems by processing large datasets and provides revolutionary advances in various fields such as health [1], finance [2], and the Internet of Things [3]. However, in traditional machine learning methods, the need to collect data on a central server leads to privacy and security issues. Especially in cases where personal and sensitive data is used, these

Artificial Intelligence Theory and Applications, ISSN: 2757-9778. ISBN: 978-605-69730-2-4 © 2025 İzmir Bakırçay University

issues can expose user privacy. In response to these challenges, FL [4] is a machine learning paradigm that allows data to be processed on local devices without creating a centralized dataset. FL allows each device to train models on its local data instead of collecting and processing data from different sources. Thus, it offers significant advantages in terms of data privacy and security, especially in areas where sensitive data is used. In recent years, FL has been widely used in studies on health data to reduce privacy concerns and improve the accuracy of machine learning models.

Health data is highly sensitive, especially as data sets containing personal and biometric information. Among such data, electrocardiography (ECG) signals provide critical information about heart health by monitoring heart rhythms. While ECG data is a widely used tool for diagnosing heart conditions, the collection and processing of this data can also pose the risk of privacy violations. The high sensitivity of health data makes it imperative to ensure data privacy and security.

Protecting privacy in health data is crucial to prevent unauthorized use and sharing of individuals' personal information. While traditional centralized data processing methods require data to be collected and stored in a single location, FL reduces this risk and allows data to be processed on local devices. Thus, the protection of privacy in health data can be secured with a FL approach. Especially when it comes to highly sensitive biometric data such as ECG, the data privacy advantage provided by FL plays a critical role.

The impact of FL on health data is dramatic. FL not only ensures data privacy but also enables collaboration between different data sources. Thus, data from different hospitals or clinics can be used to train the same model without being aggregated in a centralized system. This enables improved diagnostic models for diverse patient populations.

Attacks in FL are a vulnerability that needs to be addressed in addition to the advantages offered by this technology. Especially since FL systems have a structure where model updates are made locally by each participant, malicious participants can manipulate this process. These attacks, known as AA, can degrade the accuracy and performance of the model by introducing misleading data into the model. Such attacks pose a serious threat as they can have irreversible consequences in critical areas such as health data.

One of the methods used to prevent AA is the AT approach. AT aims to make the model more resistant to attacks by using adversarial data during model training. AT provides a more robust learning process by not only improving the accuracy of the model but also its reliability. AT is used in traditional centralized learning systems. Its equivalent in FL systems is FAT.

FAT is a technique that combines the approaches of FL and AT. The FAT technique was proposed by Zizzo et al [5]. This method aims to develop models that are more resilient to AAs while maintaining the privacy advantages of FL. In terms of protecting the privacy and security of health data, FAT is considered as an important step towards developing more secure and effective machine learning models in the future.

There have been many studies on FL and ECG in the literature. Tang et al. [6] proposed a personalized FL method for ECG classification task. Manocha et al. [7] proposed a new algorithm using deep learning to classify ECG arrhythmias in a federated environment. In their proposed algorithm, they integrated a Support Vector Machine classifier with a Bi-directional Long Short-Term Memory based Auto-Encoder network. Alreshidi et al. [8] presented Fed-CL, an advanced method that combines Long Short-

Term Memory networks and Convolutional Neural Networks to accurately predict AFib utilizing FL. Çelik and Güllü [9] conducted a comparison study on server-side aggregation algorithms on Independently and Identically Distributed (IID) and Non-IID data distributions for ECG classification task.

There have also been many studies on FAT in the literature. Bondok et al. [10] used FL and AT to address privacy and security concerns in smart grids. Catak and Kuzlu [11] used FL to train a segmentation model for spectrum sensing in the presence of radar and wireless communication systems. They also used AT to combine model flexibility and local model updates into a robust global model. Luo et al. [12] proposed a new Ensemble Federated Adversarial Training (EFAT) method that enables AT to perform better in non-IID environments by extending the training data with different distortions.

In this study, we propose the use of FAT for ECG classification task to be robust against AAs while maintaining privacy and security. For this purpose, in each of the clients, the PGD [13] discarded versions of the clean data were added to the training set and the clients were made to perform AT. The original test data, PGD attacked version of the test data, FGSM [14] attacked version, CW [15] attacked version and BIM [16] attacked version of the test data were tested respectively. The results obtained are compared.

2. Materials and Methods

2.1. MIT-BIH Arrhythmia Database

The MIT-BIH Arrhythmia Database [17] (MIT-DB) was used in this study. The MIT-DB contains 48 ECG recordings of 30 minutes each. These 48 ECG recordings belong to 47 patients. The sampling frequency of all recordings is 360 Hz. The labels in the MIT-DB were edited according to the Association for the Advancement of Medical Instrumentation (AAMI) standard. The label editing is shown in Table 1. There are 5 classes in total from the AAMI standard. These are N (normal beats), S (supraventricular ectopic beats), V (ventricular ectopic beats), F (fusion beats), and Q (unclassifiable beats).

Table 1. AAMI Standards and MIT-BIH Annotation

AAMI	MIT-BIH
Normal Beat (N)	N, L, R, j, e
Supraventricular Ectopic Beat (S)	a, S, A, J
Ventricular Ectopic Beat (V)	E, V
Fusion Beat (F)	F
Unknown Beat (Q)	/, Q, f

2.2. Data Preparation and Normalization

Each ECG signal in MIT-DB was divided into 180-length windows. 180-length windows were created by taking 90 indices before and 90 indices after the beats in the ECG signals. The values in each window were normalized using Min-Max scaler to be in the range [0, 1]. The min-max scaler is given in equation (1).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

2.3. Deep Learning Architecture

In this study, Convolutional Neural Network (CNN) is used as the deep learning architecture. CNN architecture has 4 Convolutional layers, 4 MaxPool layers, 1 Flatten layer and 3 Linear layers. Convolutional layers consist of 16, 32, 64 and 128 filters respectively. They all have a kernel_size of 3 and a padding of 1. There is also a ReLU activation function at the output of each convolutional layer. In the MaxPool layers, kernel_size is set to 2. Linear layers consist of 256, 64 and 5 units respectively. The 256- and 64-unit linear layers have a ReLU activation function at the output. The 5-unit linear layer is the output layer of the model. The number of parameters of the architecture used is 410021. The architecture used in the study is given in Figure 1.

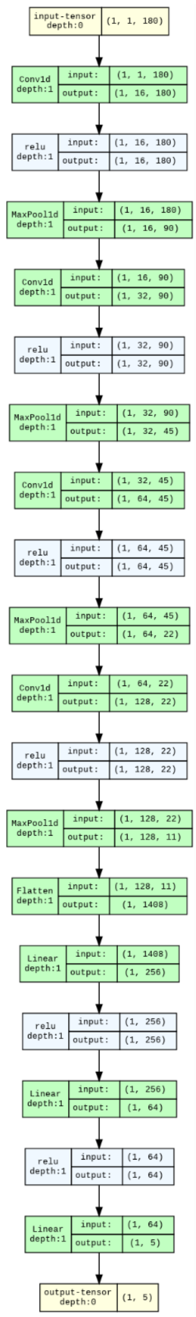


Figure 1. CNN Architecture

2.4. Algorithm

This section explains the proposed FAT system's defence mechanism against AAs in the context of ECG signal classification. The methodology is outlined in Algorithm 1, which details the local training process for a client. Initially, the client receives the global parameters. A PGD attack is then applied to the input data x_j and x_j^{adv} is obtained. Input data x_j and adversarial examples x_j^{adv} are combined to obtain $\hat{x}_j$. Original label y_j and duplicate label y_j are concatenated to obtain $\hat{y}_j$. New parameters are obtained using $\hat{x}_j$ containing both adversarial and clean samples and their labels $\hat{y}_j$. Finally, the ClientUpdate procedure is terminated by sending the new parameters to the server.

Algorithm 1. FAT for ECG signals

Input:

Client i, global parameters $\hat{\theta}$, local dataset D_i, local epoch number *E*, batch size *b*, adversarial perturbation function PGD, learning rate η

procedure ClientUpdate
    $\theta_i \leftarrow \hat{\theta}$
    for local epoch = 1, ⋯ , *E* **do**
     for mini-batch $\{x_j, y_j\}_{j=1}^{b} \sim D_i$ **do**
        $x_j^{adv} \leftarrow PGD(x_j, y_j)$
        $\hat{x}_j \leftarrow concatenate(x_j, x_j^{adv})$
        $\hat{y}_j \leftarrow concatenate(\, y_j, y_j)$
        $\theta_i \leftarrow \theta_i - \eta \nabla_{\theta_i} \ell_{CE}(\hat{x}_j, \hat{y}_j;\, \theta_i)$
     end for
    end for
    return θ_i
end procedure

After the ClientUpdate procedure is completed on all clients selected for training, the parameters sent by the clients are collected on the server. Using these parameters, the new global parameters are determined using equation (2).

$$\hat{\theta} \leftarrow \sum_{i \in S} \frac{n_i}{m}\, \theta_i \tag{2}$$

Where S is the list of clients selected in the current round, n_i is the data count of the ith client, m is the sum of the data counts of the clients selected in the round, θ_i is the local parameters sent by the ith client, $\hat{\theta}$ is the global parameters. This equation belongs to the FederatedAveraging (FedAvg) [4], which is used as the server-side aggregation method in this study.

3. Experimental Results

In this study, the experiments were conducted using the MIT-BIH Arrhythmia Database (MIT-BIH DB). The 48 ECG signals in the MIT-BIH DB were first divided into windows with length of 180 samples and normalized with the Min-Max scaler. A total of 109468 windows were obtained. Of these, 80% were used as training data and 20% were reserved as testing. The training data was distributed across 10 clients. Therefore, the training data was divided into 10 parts for 10 clients. In each round, 5 randomly selected clients participated in the training. The training continued for 10 rounds in total. In each round, the selected clients were trained locally for 5 epochs. FedAvg was used as the server-side aggregation algorithm. The study utilized two training approaches. First, the Non-FAT Model was trained using only clean data on clients. Second, the FAT Model incorporated AT by augmenting the training data with adversarial examples generated using a PGD attack. Testing was conducted at the end of each training round, using the test data in five variations: original (Clean), PGD attacked, FGSM attacked, CW attacked, and BIM attacked. For the PGD attack, epsilon was set to 8/255, alpha to 1/255 and the number of steps to 20. For the FGSM attack, the epsilon value was set to 8/255. For the CW attack parameters were set to c = 1, kappa = 0, 50 steps, and a learning rate of 0.01. Finally, the BIM attack used an epsilon of 8/255, alpha of 2/255, and 10 steps.

The CNN architecture was implemented using the Pytorch [18] library in Python. The FL environment was set up with the Flower [19] library, while the torchattacks [20] library was utilized for generating AAs (PGD, FGSM, CW, BIM). The training was performed on a system equipped with an AMD Ryzen 5 5600H processor, 16 GB of RAM and an NVDIA Geforce RTX 3050 graphics card. For the CNN architecture, the Adam optimizer was employed, and Cross Entropy Loss was used as the loss function. The results obtained are presented in tables and graphs. Table 2, 3, 4 and 5 show the Accuracy, Precision, Recall and F1 Score metrics for both FL system without FAT (Non-FAT Model) and FL system with FAT (FAT Model) across original test data (Clean) and adversarial datasets (PGD, FGSM, CW and BIM) in the 10^{th} round of training. Figure 2, 3, 4, 5 and 6 presents the variations in Accuracy, Precision, Recall and F1 Scores over 10 rounds on original test data and adversarial datasets. Note that round 0 represents the baseline results obtained using randomly initialized parameters.

Table 2. Accuracy results of 10th round for clean and adversarial data

Model	Clean (%)	PGD (%)	FGSM (%)	CW (%)	BIM (%)
Non-FAT Model	98.44	35.14	73.87	7.47	35.36
FAT Model	97.60	94.82	95.44	46.65	94.81

Table 3. Precision results of 10th round for clean and adversarial data

Model	Clean (%)	PGD (%)	FGSM (%)	CW (%)	BIM (%)
Non-FAT Model	93.95	30.12	47.05	7.27	30.15
FAT Model	90.55	81.72	83.76	26.80	81.55

Table 4. Recall results of 10th round for clean and adversarial data

Model	Clean (%)	PGD (%)	FGSM (%)	CW (%)	BIM (%)
Non-FAT Model	92.80	21.27	47.29	2.42	21.05
FAT Model	88.79	80.13	81.85	26.78	80.06

Table 5. F1 Score results of 10th round for clean and adversarial data

Model	Clean (%)	PGD (%)	FGSM (%)	CW (%)	BIM (%)
Non-FAT Model	92.94	22.61	45.46	3.20	22.55
FAT Model	89.06	79.75	81.70	24.02	79.63

Table 2, Table 3, Table 4 and Table 5 show that the Non-FAT Model outperformed the FAT Model in Accuracy, Precision, Recall and F1 Score metrics in the original (clean) test data. However, the FAT Model also achieved very high performance. On PGD, FGSM, CW and BIM attacked test data, the FAT Model outperformed the Non-FAT Model in Accuracy, Precision, Recall and F1 Score metrics.

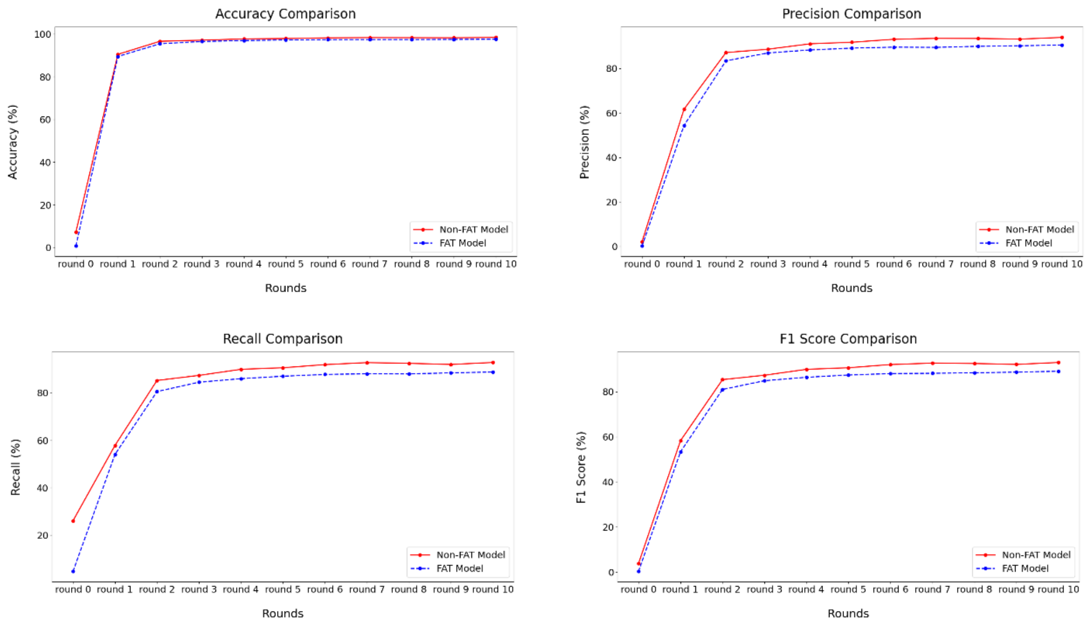


Figure 2. Clean Data Metric Results (Accuracy, Precision; Recall, F1 Score)

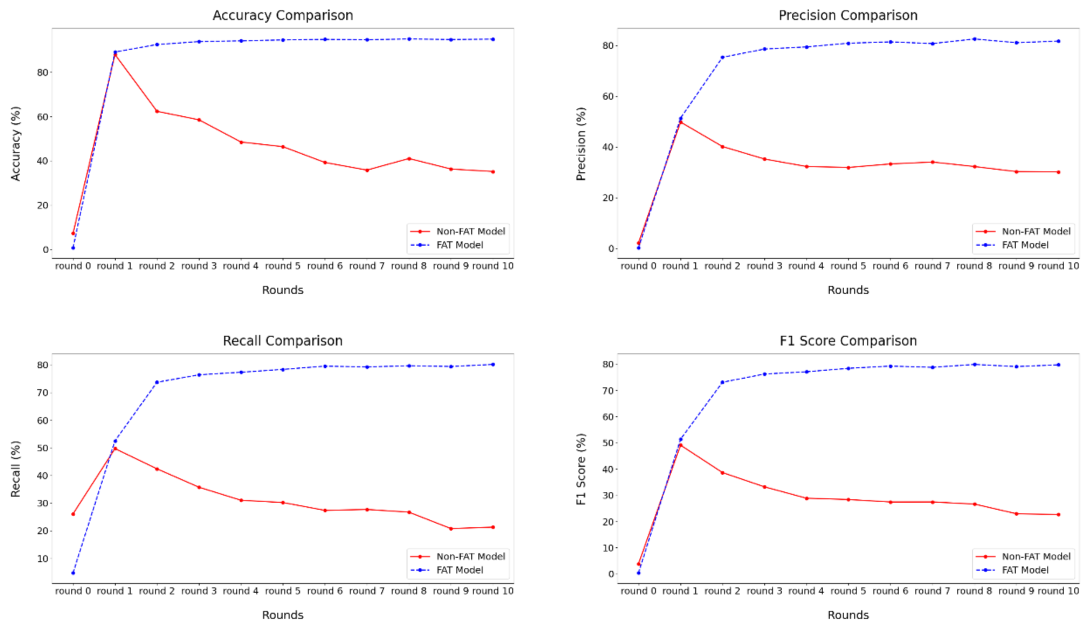


Figure 3. PGD Attacked Data Metric Results (Accuracy, Precision; Recall, F1 Score)

Figure 2 shows that the Non-FAT Model outperforms the FAT-Model across all rounds in all four metrics on the original test data. However, the performance gap between the two models is minimal. The FAT Model also achieves consistently high performance across all rounds and metrics, demonstrating its robustness even with slight compromises in comparison to the Non-FAT model.

Figure 3 shows that in round 1, the FAT and Non-FAT models exhibit comparable performance across all metrics on PGD-attacked test data, but the FAT model slightly outperforming the Non-FAT model. From round 2 onward, the FAT model performs even better, while the Non-FAT model performs very poorly against PGD-attacked data. The Non-FAT model was not able to achieve high performance on PGD-attacked data due to the lack of AT during local training and the absence of PGD-attacked instances in the dataset.

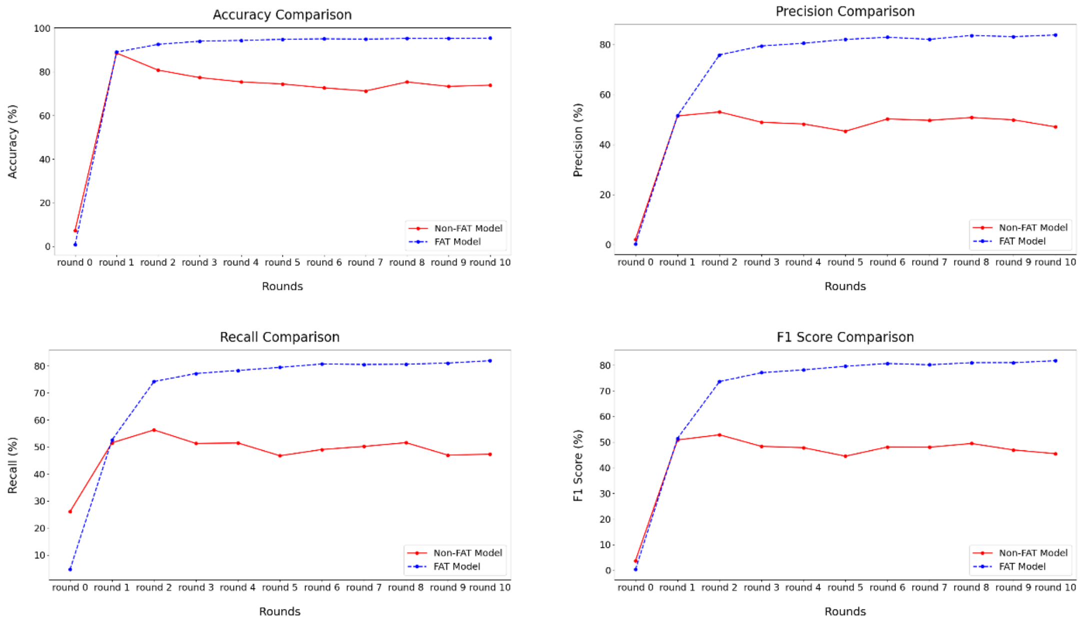


Figure 4. FGSM Attacked Data Metric Results (Accuracy, Precision; Recall, F1 Score)

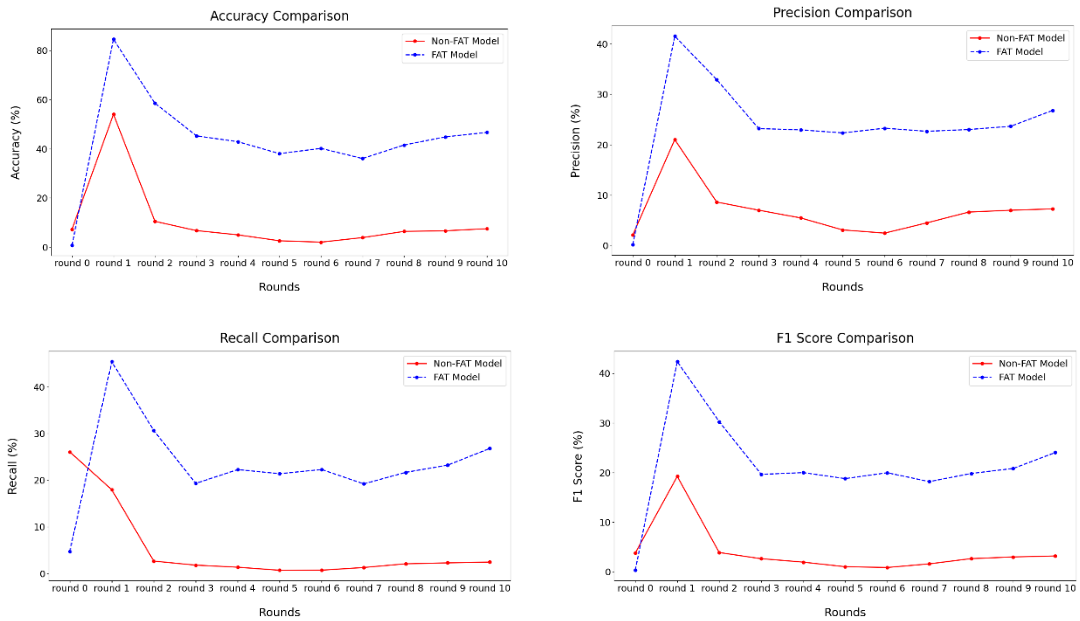


Figure 5. CW Attacked Data Metric Results (Accuracy, Precision; Recall, F1 Score)

Figure 4 illustrates that in round 1, the FAT and Non-FAT models exhibit similar performance across all four metrics on FGSM-attacked test data. However, from round

2 onward, the FAT model performs even better, while the Non-FAT model demonstrates poor performance against FGSM-attacked data. The Non-FAT model was not able to achieve high performance on FGSM-attacked data, since no AT was performed during the local training of its clients or the lack of FGSM-attacked instances in the dataset.

Figure 5 shows that the FAT model outperformed the Non-FAT model across all rounds (excluding round 0, which reflects the initial weights and is not evaluated) in all four metrics on the CW-attacked test data. The Non-FAT model achieved considerably lower results. Both models showed their highest performance in Round 1 across all metrics. However, in the following rounds, they achieved lower performance than this round. Although the FAT model outperformed the Non-FAT model, it's performance remained below the levels achieved against other types of AAs.

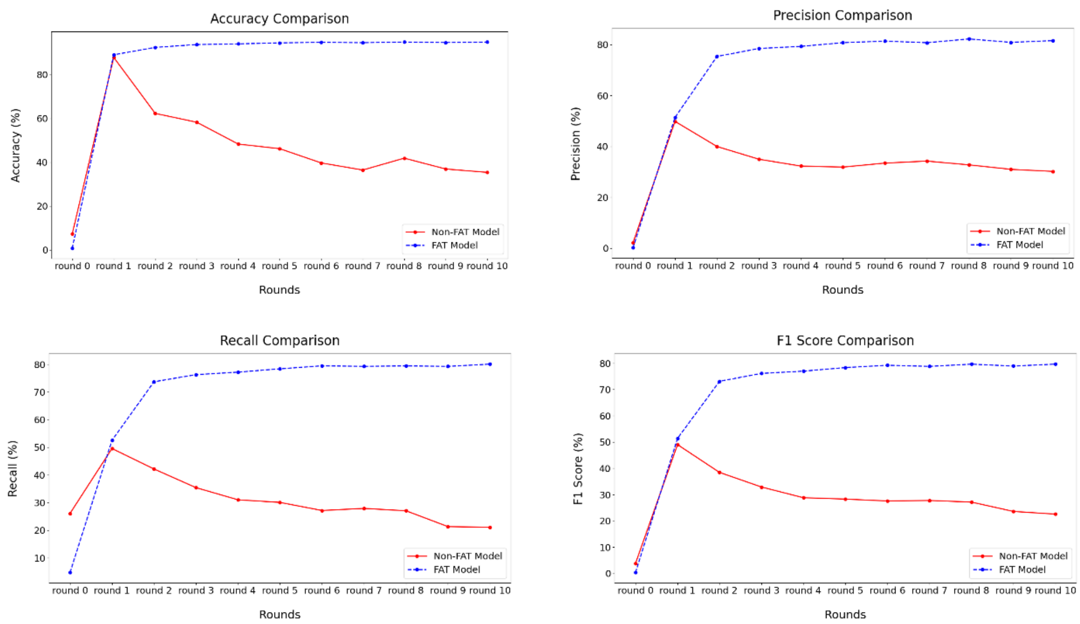


Figure 6. BIM Attacked Data Metric Results (Accuracy, Precision; Recall, F1 Score)

Figure 6 shows that in round 1, the FAT and Non-FAT Models exhibit similar performance across all four metrics on the BIM-attacked test data, with the FAT model performing slightly better. From round 2 onward, the FAT model consistently outperforms the Non-FAT model, which demonstrates very poor performance on BIM-attacked data. This underperformance of the Non-FAT model is likely due to the absence of Adversarial Training (AT) during local client training and the lack of BIM-attacked instances in the dataset.

Overall, the Non-FAT model demonstrated strong classification performance on the original data. However, it struggled with low classification performance when tested against PGD, FGSM, CW and BIM adversarial attacks. In contrast, the FAT model achieved high performance against all data except for the CW attacked data. It maintained its high performance across all datasets, except for the CW attacked data. Although its performance on CW attacks is superior to the non-FAT model, it fails to achieve the same success on CW attacks as it achieves on other adversarial data. Notably, the FAT model maintained robust performance on adversarial data with only minimal compromise in accuracy on the original data. Incorporating adversarially attacked versions of the training data during the training process, through AT, proves to be effective in enhancing the model's resilience against adversarially attacked test data.

4. Conclusion

In this study, we propose the use of FAT for ECG classification to enhance robustness against AAs while preserving privacy and security. For this purpose, in addition to clean data, adversarial examples generated using the PGD method are also used during local training on clients. The proposed framework is tested against the original test data, and adversarially attacked versions created using PGD, FGSM, CW, and BIM. Its performance was compared with that of an FL system without FAT.

The results showed that the FL system without FAT achieved high performance in Accuracy, Precision, Recall and F1 Scores on the original test data. However, it's performance dropped significantly across all four metrics for PGD, FGSM, CW and BIM attacked data. It achieved a very low performance especially against CW attacked data. In the proposed structure, i.e. the FL system using FAT, high performance is achieved in all four metrics for the original test data, PGD, FGSM and BIM attacked data. The performance on CW-attacked data, while improved compared to the Non-FAT system, was lower than for other types of adversarial attacks. When comparing the two systems, the FL system without FAT is more successful on the original test data. For the PGD, FGSM, CW and BIM attacked data, the FL system with FAT is more successful. However, the FL system with FAT is also successful on the original test data. The FL system with FAT achieves very high performance against adversarial attacked data with a little performance compromise from the original test data.

Through the FAT, ECG signal classification can achieve both enhanced privacy and security using FL, while simultaneously providing a robust defence against potential AAs. This study lays a foundation for future research exploring similar techniques with diverse types of health data.

Acknowledgement

This research has been presented in IV. International Congress on Artificial Intelligence in Health.

References

[1] Habehh, H., & Gohel, S. (2021). Machine learning in healthcare. Current Genomics, 22(4), 291–300. https://doi.org/10.2174/1389202922666210705124359

[2] Nazareth, N., & Reddy, Y. V. R. (2023). Financial applications of machine learning: A literature review. Expert Systems With Applications, 219, 119640. https://doi.org/10.1016/j.eswa.2023.119640

[3] Cui, L., Yang, S., Chen, F., Ming, Z., Lu, N., & Qin, J. (2018). A survey on application of machine learning for Internet of Things. International Journal of Machine Learning and Cybernetics, 9(8), 1399–1417. https://doi.org/10.1007/s13042-018-0834-5

[4] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. a. Y. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. International Conference on Artificial Intelligence and Statistics, 1273–1282. http://proceedings.mlr.press/v54/mcmahan17a/mcmahan17a.pdf

[5] Zizzo, G., Rawat, A., Sinn, M., & Buesser, B. (2020). FAT: Federated Adversarial Training. arXiv. https://arxiv.org/abs/2012.01791

[6] Tang, R., Luo, J., Qian, J., & Jin, J. (2021). Personalized federated learning for ECG classification based on feature alignment. Security and Communication Networks, 2021, 1–9. https://doi.org/10.1155/2021/6217601

[7] Manocha, A., Sood, S. K., & Bhatia, M. (2024). Federated learning-inspired smart ECG classification: an explainable artificial intelligence approach. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-024-20084-3

[8] Alreshidi, F. S., Alsaffar, M., Chengoden, R., & Alshammari, N. K. (2024). Fed-CL- an atrial fibrillation prediction system using ECG signals employing federated learning mechanism. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-71366-7

[9] Çelik, E., & Güllü, M. K. (2023). Comparison of federated learning strategies on ECG classification. 2023 Innovations in Intelligent Systems and Applications Conference (ASYU), 1-4. https://doi.org/10.1109/asyu58738.2023.10296796

[10] Bondok, A. H., Mahmoud, M., Badr, M. M., Fouda, M. M., Abdallah, M., & Alsabaan, M. (2023). Novel evasion attacks against adversarial training defense for smart Grid federated learning. IEEE Access, 11, 112953–112972. https://doi.org/10.1109/access.2023.3323617

[11] Catak, F. O., & Kuzlu, M. (2024). A federated adversarial learning approach for robust spectrum sensing. 2024 13th Mediterranean Conference on Embedded Computing (MECO), 1-4. https://doi.org/10.1109/meco62516.2024.10577941

[12] Luo, S., Zhu, D., Li, Z., & Wu, C. (2021). Ensemble Federated Adversarial Training with Non-IID data. arXiv. https://arxiv.org/abs/2110.14814

[13] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv. https://arxiv.org/abs/1706.06083

[14] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv. https://arxiv.org/abs/1412.6572

[15] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. 2017 IEEE Symposium on Security and Privacy (SP), 39-57. https://doi.org/10.1109/sp.2017.49

[16] Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial examples in the physical world. In Chapman and Hall/CRC eBooks (pp. 99–112). https://doi.org/10.1201/9781351251389-8

[17] Moody, G., & Mark, R. (2001). The impact of the MIT-BIH Arrhythmia Database. IEEE Engineering in Medicine and Biology Magazine, 20(3), 45–50. https://doi.org/10.1109/51.932724

[18] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., . . . Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv. https://arxiv.org/abs/1912.01703

[19] Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Li, K. H., Parcollet, T., Buarque, D. G. P. P., & Lane, N. D. (2020). Flower: a friendly federated Learning research framework. arXiv. https://arxiv.org/abs/2007.14390

[20] Kim, H. (2020). Torchattacks: a PyTorch repository for adversarial attacks. arXiv. https://arxiv.org/abs/2010.01950

■RESEARCH ARTICLE

Design of Cardiac Pacemaker Controller Based on Reinforcement Learning

Kağan Orbay [a], **Mehmet Sağbaş** [a†], **Murat Demir** [a]

[a] Department of Electrical and Electronics Engineering, İzmir Bakırçay University, İzmir, Türkiye
† sagbasm@gmail.com, corresponding author

RECEIVED DECEMBER 23, 2024
ACCEPTED APRIL 9, 2025

CITATION Orbay, K., Sağbaş, M. & Demir, M. (2025). Design of cardiac pacemaker controller based on reinforcement learning. *Artificial Intelligence Theory and Applications, 5(1)*, 29-41.

Abstract

This study investigates the derivation of PID controller parameters, commonly used for pacemaker control, using both genetic algorithm (GA) and reinforcement learning (RL) methods. We compare the PID parameters obtained by RL with those obtained by GA, a well-known and often preferred method in literature. The aim of the study is to analyze the performance of the control parameters obtained by both methods and to determine which approach is more effective in pacemaker applications. In particular, comparisons on important control criteria such as settling time, rise time and overshoot of the system will reveal the advantages and disadvantages of these methods.

Keywords: heart rhythm regulation, pacemaker control system, PID controller optimization, reinforcement learning

1. Introduction

Cardiovascular diseases, including heart attacks and arrhythmias, are among the leading causes of death worldwide [1-2]. Arrhythmias disrupt the normal electrical activity of the heart and often require medical intervention. One of the most effective solutions for regulating heart rhythm is the cardiac pacemaker, which delivers controlled electrical impulses to the heart [3]. Pacemakers continuously monitor cardiac activity and correct irregularities by providing appropriate electrical stimulation [4]. This regulation is crucial for preventing complications that can arise from untreated arrhythmias, such as stroke or heart failure. Additionally, advancements in technology have led to the development of more sophisticated pacemakers that can adapt to a patient's activity level, further improving overall cardiac health.

To improve the efficiency of pacemakers, researchers have developed advanced control strategies to optimize their performance. A key aspect of this optimization is the use of Proportional-Integral-Derivative (PID) controllers, which provide precise regulation of the heart rhythm. The PID controller has three components. The proportional (P) component provides a correction proportional to the current error, allowing a fast and accurate response to instantaneous changes in heart rate. This accuracy ensures that the heart rate is maintained at the desired level. The integral (I) component accounts for the

Artificial Intelligence Theory and Applications, ISSN: 2757-9778. ISBN: 978-605-69730-2-4 © 2025 İzmir Bakırçay University

accumulation of error over time and provides long-term corrections. This prevents the accumulation of continuous errors and helps the pacemaker maintain a more stable heart rhythm over time. Provides long-term performance improvements. The derivative (D) component analyzes the rate of change of the error and reacts quickly to instantaneous changes. This quickly compensates for sudden changes in heart rate and prevents the system from overreacting. It improves overall system performance by adapting to dynamic changes. Proper tuning of these parameters is critical to achive the desired control system behavior. However, determining optimal PID parameters remains a challenge, leading researchers to explore advanced optimization techniques such as Genetic Algorithms (GA) and Reinforcement Learning (RL) [5-6].

Several mathematical models have been proposed to describe cardiac dynamics, which are crucial for developing pacemaker control strategies. Biswas et al. (2006) modeled the cardiovascular system using a closed-loop negative unit feedback system based on transfer functions [7]. Additionally, mathematical models such as the Noble model for Purkinje fibers and the Beeler-Reuter model for ventricular myocardial cells have been widely used to simulate cardiac activity [8-9].

This study investigates the effectiveness of GA and RL in optimizing pacemaker PID controller parameters. By comparing these approaches, we aim to identify the most efficient method based on performance criteria such as settling time, rise time, and overshoot. The results provide insights into the advantages and limitations of AI-driven optimization strategies in biomedical control applications.

Traditional PID tuning methods, such as Ziegler-Nichols and Cohen-Coon, are widely used but often struggle with adaptability in dynamic physiological conditions [10-11]. To overcome this limitation, evolutionary algorithms and machine learning-based optimization techniques have been explored.

Various control techniques have been explored to improve the performance of pacemakers. Apart from classical methods, the following approaches have been utilized:

- Studies using optimization algorithms [2], [12-13]
- Embedded designs using microcontrollers and FPGAs [14-15]
- Machine learning based designs using various machine learning algorithms [16-17]
- Studies using analogue circuits [18-19]

Several studies have demonstrated the effectiveness of Genetic Algorithms (GA) in optimizing PID controllers for pacemakers. Bajpai et al. (2017) showed that GA-based tuning minimizes overshoot and improves transient response [2]. Similarly, Momani et al. (2019) examined fractional-order PID controllers tuned via GA and found improved accuracy in heart rate regulation [4]. These findings highlight GA’s ability to efficiently explore solution spaces and optimize control parameters.

However, GA has limitations [20]:

- It relies on heuristic search mechanisms that may converge to local optima.
- Its performance is highly dependent on mutation and crossover rates.
- It does not adapt well to real-time physiological variations.

In contrast, Reinforcement Learning (RL) has gained attention for adaptive control systems [21-22]. In contrast to GA, RL continuously learns from the environment, thereby improving decision-making over time [5]. Lima et al. (2023) applied RL to cardiac rhythm regulation, demonstrating its ability to dynamically adjust pacing parameters with high accuracy [16].

Despite its advantages, RL also has challenges [23-24]:

- It requires extensive training episodes to achieve convergence.
- Traditional RL methods struggle in high-dimensional continuous spaces.
- Computational complexity can be high, requiring deep RL techniques for scalability.

Although GA and RL have been studied separately, a direct comparative analysis of these methods in pacemaker control is still lacking. This study aims to bridge this gap by evaluating GA and RL in optimizing PID parameters for pacemaker applications. The key contributions are:

1. A comparative analysis of GA and RL for PID tuning in pacemakers, assessing their effectiveness in optimizing heart rhythm control.

2. A structured performance evaluation based on key control metrics (settling time, rise time, overshoot, and peak response).

3. An RL-based adaptive tuning framework, demonstrating its potential advantages over GA in reducing overshoot and improving stability.

4. A scalable optimization methodology that can be extended to other AI techniques such as Particle Swarm Optimization (PSO) and Model Predictive Control (MPC).

By integrating modern AI-driven techniques with traditional evolutionary algorithms, this study provides a novel perspective on cardiac pacemaker controller design. The findings of this study suggest that GA is more effective in achieving rapid responses, while RL offers superior long-term adaptability, making it a promising solution for real-world applications.

2. Modelling of Pacemaker

Mathematical models of the heart have been developed to facilitate understanding of cardiac function. Noble described the Purkinje fiber cell action potential in 1962 with the Noble model [8]. Beeler and Reuter introduced an electrical activity model of the ventricular myocardial cell in 1977 [9]. The mathematics of cardiac dynamics helped to design pacemaker control systems for artificial and implanted devices. Biswas et al. proposed a transfer function-based cardiovascular system mathematical model [7]. The cardiovascular system is depicted as closed loop negative unit feedback with filter and controller. Figure 1 depicts the cardiovascular closed-loop control system block diagram. Equations. (1) and (2) provide the pacemaker and heart transfer functions $G_{Pacemaker}(s)$ and $G_{heart}(s)$, for the configuration depicted in Figure 1. The closed loop system receives the real heart rate $R(s)$ and produces the target heart rate Y(s). The function of $G_K(s)$ is to serve as the controller.

$$\boldsymbol{G_{Pacemaker}(s) = \frac{\omega_{lpf}}{s+\omega_{lpf}}} \quad (1)$$

$$\boldsymbol{G_{Heart}(s) = \frac{1}{Ms^2+Bs+K}} \quad (2)$$

The cut-off frequency of the low-pass filter representing the pacemaker is ω_{lpf}, while the mass of the heart muscle is M, the viscous drag of the heart myocardial cell is B, and the torsional drag is K.

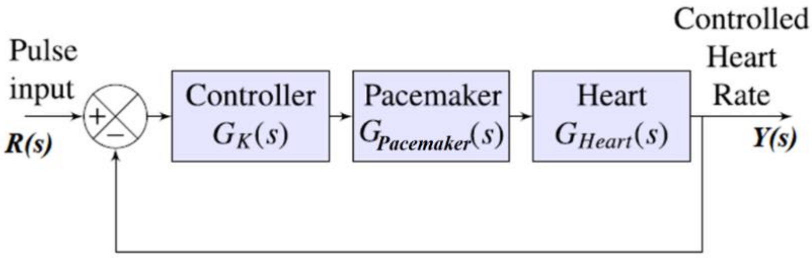


Figure 1. The block diagram of the cardiovascular system.

If the numerical values of the parameters given in Eqs. (1) and (2) are substituted in the studies in the literature, $G_P(s)$ is obtained as follows [7].

$$G_P(s) = \frac{1352}{s(s+8)(s+20.8)} \tag{3}$$

Given the closed loop system shown in Figure 1, its closed loop transfer function is as follows:

$$(s) = \frac{Y(s)}{R(s)} = \frac{G_K(s)G_P(s)}{1+G_K(s)G_P(s)} \tag{4}$$

3. PID Controllers

PID controllers are used to improve pacemaker efficiency. The PID controller improves pacemaker performance and cardiac rhythm management. The PID controller can adjust the pacemaker's output to target heart rate for accurate cardiac rhythm regulation. To respond quickly and accurately to immediate heart rate changes, the proportional (*P*) component corrects the present mistake. This accuracy keeps the heart rate at the correct level. Long-term error fixes were provided via the integral (*I*) component. Avoiding ongoing mistakes helps the pacemaker maintain a steady cardiac rhythm over time. Improves long-term performance. The derivative (*D*) component evaluates error rate and reacts swiftly to sudden changes. This swiftly adjusts for unexpected heart rate variations and minimizes overreaction. It adapts to dynamic changes to boost system performance. PID controller settings can adjust to patient activity and physiological changes. This adjustment allows the pacemaker to automatically modify heart rate based on stress or physical activity, enhancing performance.

The PID controller has three parameters: The error signal's current value determines K_p. Control action rises proportionately with mistake. The error signal's historical values determine the integral term (K_i). Long-term errors activate the integral term, boosting control action. The derivative term (K_d) predicts fault signal value. The derivative term increases control action if the mistake is rising fast. These three parameters are crucial to PID controller performance. To accomplish control system behavior, these parameters must be tuned properly depending on the application.

Equations (5) and (6) respectively provide the time-based response of the output u(t) and transfer function of the PID-controller.

$$u(t) = K_p e(t) + T_i \int_0^t e(\tau) d\tau + T_d \frac{d}{dt} e(t), \tag{5}$$

$$G_K(s) = \frac{U(s)}{E(s)} = K_p + \frac{T_i}{s} + T_d s \tag{6}$$

where E(s) and U(s) denote the Laplace transforms of the error and control signals, respectively.

4. Genetic Algorithm

Evolutionary search and optimization methods like genetic algorithms address difficult optimization issues. These algorithms replicate natural selection, crossover, and mutation to efficiently explore the solution space. Genetic algorithms analyse and choose the best potential solutions from a population. Each cycle selects the best people and assesses them using a fitness function [6, 25]. Crossover and mutation procedures expand solution space while retaining population genetic diversity. Traditional approaches fail to solve difficult and large-scale optimization issues, but this method can. Genetic algorithms' success depends on parameter values and problem-specific design.

$$\mathrm{C} = \{C_1, C_2, \dots, C_n\} \tag{7}$$

In Eq. (8), *C* represents a chromosome, and C_i represents the *i*-th gene of the chromosome.

$$f(C) = \text{Fitness Function } (C) \tag{8}$$

The fitness function in Eq. (8) determines chromosomal (solution) quality. This function represents the optimization goal. The objective is usually to maximize fitness. Selection ensures that the following generation inherits the finest population members.

$$P_i = \frac{f(C_i)}{\sum_{j=1}^{N} f(C_i)} \tag{9}$$

P_i is the probability of selection of the *i*-th individual, and $f(C_i)$ is its fitness value. Crossover creates a new person from two parental chromosomes. Single-point crossover is the most frequent crossover mechanism. Single-point crossover is a typical way to make new people from two parental chromosomes. This approach switches genes on the two chromosomes from a point. Mutations affect the value of a randomly selected gene to preserve genetic variation

5. Reinforcement Learning

Machine learning refers to algorithms that acquire knowledge from data. The domains of machine learning encompass supervised, unsupervised, and reinforcement learning. RL is extensively utilized in supervised issues because of its reward-based learning framework.

RL simply works to develop a function that produces an output based on feedback obtained from the environment utilizing data. Upon structural analysis, it is evident that it has three fundamental components: Agent, State, and Environment [26]. The learned function is referred to as policy. The policy permits the choice of the action (at) that will provide the maximum reward over time, based on the observed state [5]. The strategy may be stochastic or deterministic, contingent upon the chosen RL approach. Figure 2 depicts the overarching framework for training in RL.

This research utilizes the Q-learning algorithm, a traditional approach in RL. Q-learning is a fundamental RL method designed for systems with a discrete solution space. The Q-learning approach is especially appropriate for systems characterized by discrete action and state spaces, and in this research, it is employed to optimize the PID parameters.

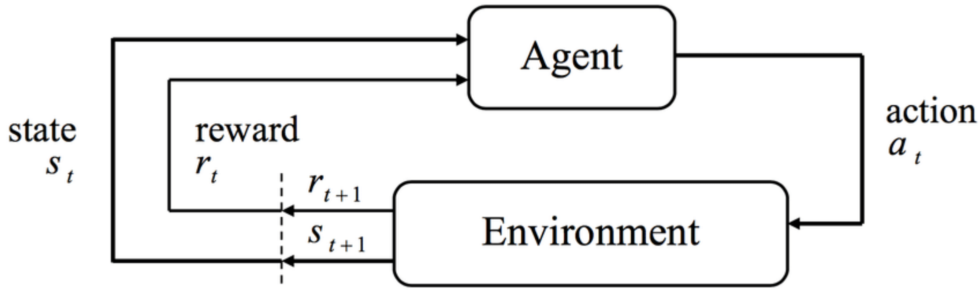


Figure 2. The reinforcement learning training schematic.

Q-learning is a model-free technique designed for RL tasks characterized by discrete state and action spaces. The agent determines the appropriate action to select in each environmental state. At each stage, the agent selects an action, obtains a reward for that action, and transitions to the subsequent state. The agent aims to select activities that optimize the cumulative reward over time. Q-learning creates a table that allocates a Q-value to each state-action combination and subsequently modifies this table to enhance learning. Acts are selected using the Epsilon-Greedy technique, whereby the agent occasionally engages in random acts for exploration and at other times opts for the action deemed optimal based on the existing Q-values. The updates consider the prospective benefits of each action, as dictated by the Bellman Equation. Consequently, the agent discerns the action that yields the maximum reward in each scenario.

6. Simulation Results and Discussion

Genetic Algorithm Optimization was first used to select the parameters of the PID controller used to control the pacemaker. The transfer functions in Equation (4) are used for the heart and pacemaker. These transfer functions model the dynamic characteristics of the heart and pacemaker systems.

The objective function of the GA is to measure the performance of the unit step response of the system, representing a heart rate of 72 bpm, by determining the parameters (K_p, K_i, K_d) of the PID controller. 72 bpm represents the healthy heart rate of an average person, and the system aims to approach this reference value with the fastest and least oscillation. By punishing changes in system response, rise time, settling time, overshoot, and peaks, the objective function tries to minimize the total error for all possible combinations of controller parameters.

The GA optimization process involves defining lower and upper bounds for the parameters of the PID controller and configuring the algorithm's operating parameters. Table 1. shows the selected parameters. We also chose a Gaussian mutation function for mutation and enabled parallel processing to speed up the calculations.

Table 1. The parameter used in simulations for GA

Parameter	Genetic Algorithm (GA)
Population Size	100
Number of Generations	200
Mutation	Gaussian Mutation
Crossover Rate	90%

To properly tune controllers and evaluate their performance, one can consider several performance criteria. The performance criteria used in this study are Integral Square Error (ISE), Integral Time Absolute Error (ITAE), Integral Time Square Error (ITSE), Integral Absolute Error (IAE), and the Discrete Time Integral Sample Based Double Square Error (dTISDSE). The following equations illustrate how these performance criteria are calculated.

$$ISE(e) = \int_0^\infty e^2(t)\,dt \tag{10a}$$

$$ITAE(e) = \int_0^\infty t|e(t)|\,dt \tag{10b}$$

$$ITSE(e) = \int_0^\infty te^2(t)\,dt \tag{10c}$$

$$IAE(e) = \int_0^\infty |e(t)|\,dt \tag{10d}$$

$$dTISDSE(e) = \sum_{k=1}^{n} k(e_k^2)^2 \tag{10e}$$

Figure 3 displays the step response of the closed-loop system resulting from GA optimization for various performance criteria. Table 2 provides the PID parameter values obtained for all performance criteria.

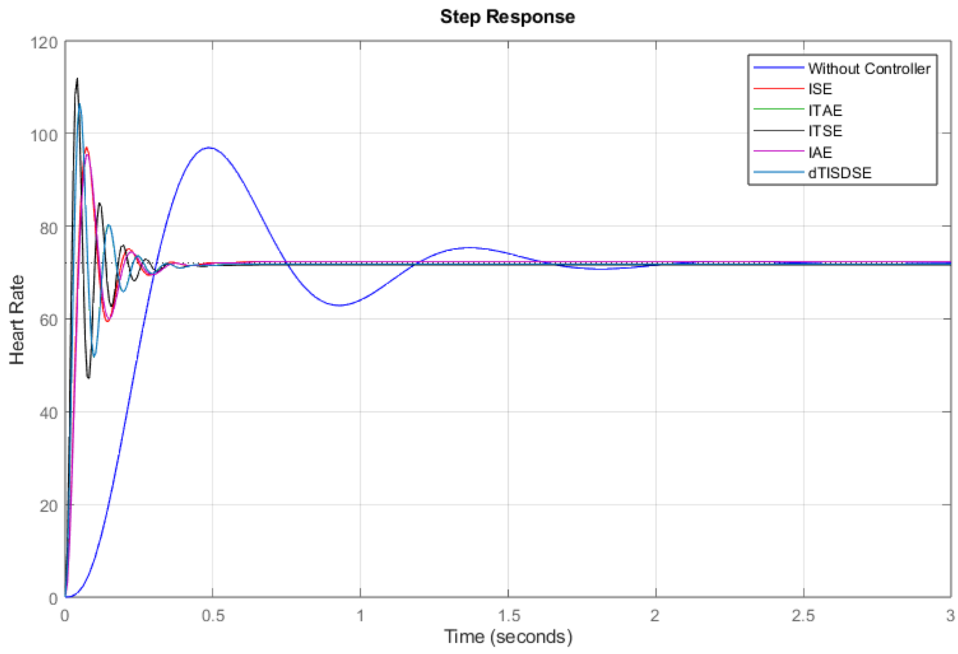


Figure 3. PID-controller responses for various performance criteria

Table 2. PID parameter values obtained using different error functions

Error Functions	K_p	K_i	K_d
ISE	4.779060	3.148517	1.487819
ITAE	6.437423	3.612167	3.090703
ITSE	7.601399	2.917295	4.871950
IAE	4.266956	2.623989	1.387273
dTISDSE	6.419923	3.476024	3.076429

Figure 4 displays the step response for ISE, yielding the best result among the used performance criteria. Table 3 shows the performance metrics for the step response obtained for ISE. Figures 3 and 4 demonstrate that the controlled system, utilizing PID parameters from the genetic algorithm, achieved the target heart rate of 72 bpm more quickly and with fewer oscillations than the uncontrolled system. As can be seen from Table 3, significant improvements are observed, especially in performance criteria such as response time, settling time, overshoot, and peak values.

Table 3. GA Optimization results for ISE performance criteria

Performance Metrics	Controlled System	Uncontrolled Closed-loop System Step Response
Rise Time (s)	0.035707	0.1908
Settling Time (s)	0.342283	1.5414
Overshoot (%)	28.820775	34.6568
Peak	92.750958	96.9529

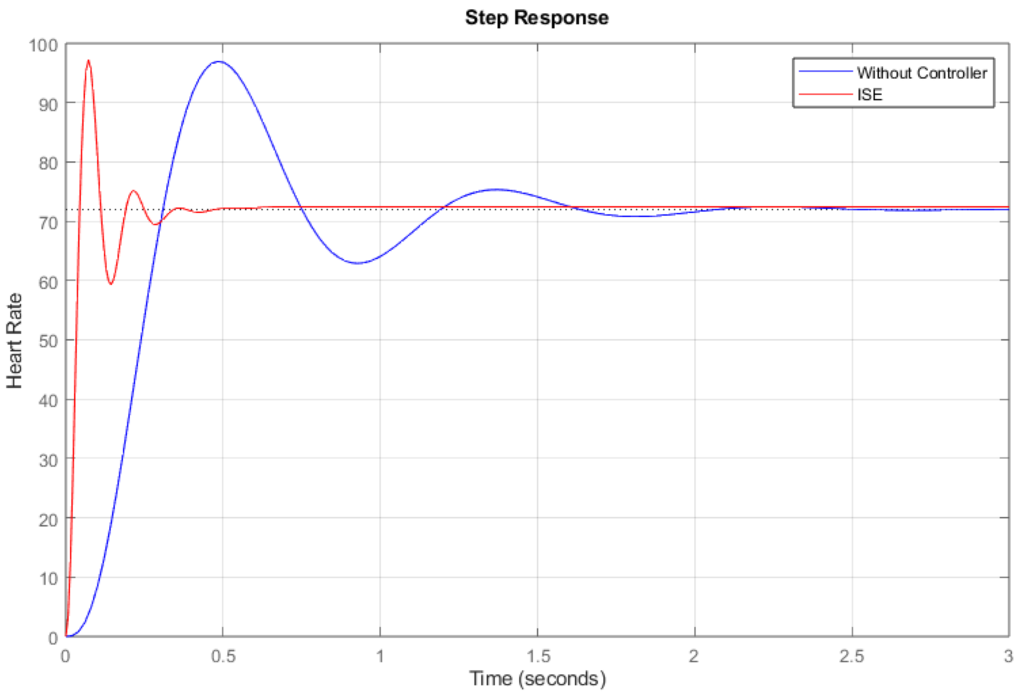


Figure 4. The step responses of the controlled system for ISE

Secondly, the PID controller parameters determined through Q-learning-based reinforcement learning (RL) are as follows: Kp = 1.3819, Ki = 0.12864, Kd = 0.26231. We use the transfer functions in Eq. (4) for heart and pacemaker dynamics. In this study, a Q-learning based PID tuning algorithm is used to find the PID parameters of the system through reinforcement learning as follows.

Algorithm: Q-learning based PID parameter optimization:
Step 1: Initialization:

- **Initialize PID Parameter Ranges:** Define the ranges for K_p, K_i, and K_d and initialize the Q-table with small random values.
- **Set Learning Parameters:** Define the learning rate (α), decay factor (γ), exploration rate (ϵ), and number of episodes.
- **Set Random Seed:** Use the rng() function to set a fixed seed value for reproducibility of simulations.

Step 2: Start the Loop (for each episode):

- **Initial State:** Choose a random combination of K_p, K_i, K_d .

For each step:

1. **Select Action:**
    - Use the Epsilon-Greedy strategy to select an action:
        - If a random number is less than ϵ\epsilonϵ, choose a random action (exploration).
        - Otherwise, select the best action based on the current Q-values (exploitation).
2. **Action Implementation:**
    - Apply the selected PID parameters (K_p, K_i, K_d) to the system.
3. **Simulate the System:**
    - Create the feedback loop according to the system's transfer function and obtain the system response (e.g., step response).
4. **Calculate Reward:**
    - Calculate the reward based on the system's performance metrics such as error, overshoot, and settling time. The reward can include penalties for these metrics.
      $$reward = -error - 0.1 * overshoot - 0.01 * settling\_time$$
5. **Update Q-Table:**
    - Update the Q-value using the Bellman equation:

$$Q(state, action) = Q(state, action) + \alpha(reward + \gamma. \max\left(Q(next_{state}, next_{action})\right) - Q(state, action))$$

6. **Transition to New State:**
    - Determine the new state based on the selected PID parameters and continue the loop for the next step.
7. **Epsilon Decay:**
    - Decrease the exploration rate (ϵ) at the end of each episode according to the decay factor (γ): $\epsilon = \max\,(0.1, \epsilon.decay\_factor)$
8. **Complete the Loop:**
    - End the episode if the error, overshoot, and settling time are within specified limits.
      $error < 0.005 \,\&\&\, overshoot < 0.05 \,\&\&\, settling\_time < 3$

Step 3: Result: Select the optimal K_p, K_i, and K_d parameters from the Q-value with the best reward.
Step 4: End of Loop:

- Terminate the algorithm when the desired performance criteria are met.

The Q-learning parameters were set in shown in Table 4. The ranges for the PID parameters utilized in the simulations were established as follows: Kp spans from 1 to 20, while both Ki and Kd range from 0.1 to 2. Each parameter was divided into 200 linearly spaced values for optimization. Additionally, a fixed seed value was used to ensure reproducibility, which was implemented using MATLAB's RNG function.

Table 4. The parameter used in simulations for RL

Parameter	Reinforcement Learning (RL)
Learning Rate	0.1
Decay factor	0.9
Exploration Rate	0.1
Number of Episodes	1000

The step response of the closed-loop system optimized with RL is illustrated in Figure 5. The optimum PID parameters obtained by reinforcement learning using the above parameters and 100 and 123 as fixed seeds are $K_p = 1.3819$, $K_i = 0.12864$, $K_d = 0.26231$ and $K_p = 1.0955$, $K_i = 0.10955$, $K_d = 0.45327$, respectively. In this case, the step response of the system is given in Figure 5. In Figure 5, Simulation I is obtained when the fixed seed is 100 and Simulation II is obtained when the fixed seed is 123.

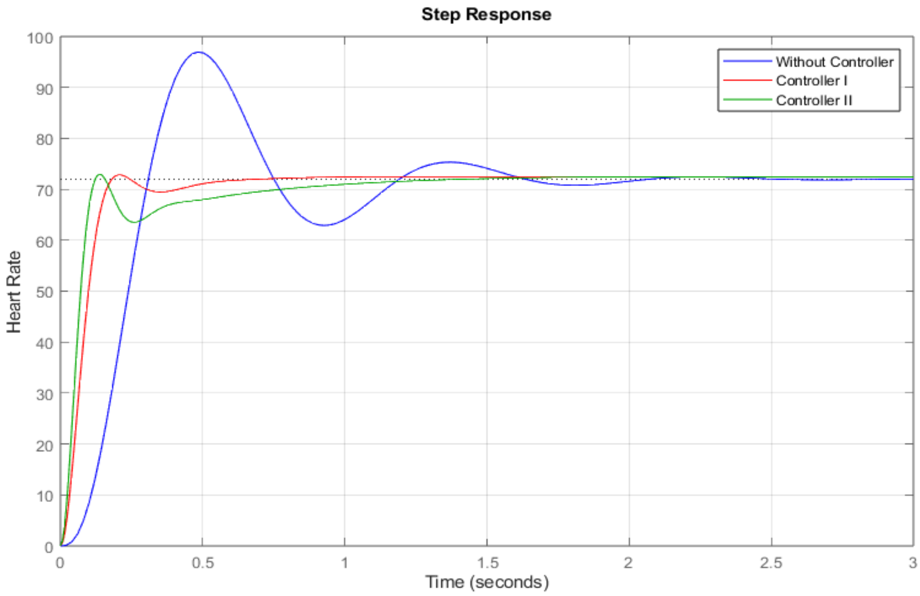


Figure 5. The simulation results of the pacemaker control system using RL.

To better understand the difference between Reinforcement Learning and Genetic Algorithm, the step response of the closed-loop system generated using the PID parameters obtained with both algorithms is shown in Figure 6. Table 5 presents a comparison of the performance metrics of GA, RL, and the uncontrolled system. As can be seen from Table 5, the step response of the pacemaker controlled with the PID controller obtained with RL.

The simulation results of the pacemaker control system in Table 5 show that the step response obtained with the control parameters optimized by both RL and GA have lower overshoot values compared to the system without controller. The overshoot value and the maximum peak value of the step response of the system controlled with PID parameters obtained by RL (1.14% and 72.83) are significantly lower than those of the system controlled with PID parameters obtained by GA (28.82% and 92.75). However, the step response of the system controlled with PID parameters generated by genetic algorithms shows much better performance for both rise time and settling time.

Tablo 5. Comparison of the performance metrics

Performance Metrics	**PID Controlled System with RL**	**PID Controlled System with GA**	**Uncontrolled System**
Rise Time (s)	0.1101	**0.035707**	0.1908
Settling Time (s)	0.4555	**0.342283**	1.5414
Overshoot (%)	**1.1413**	28.820775	34.6568
Peak (bpm)	**72.8217**	92.750958	96.9529

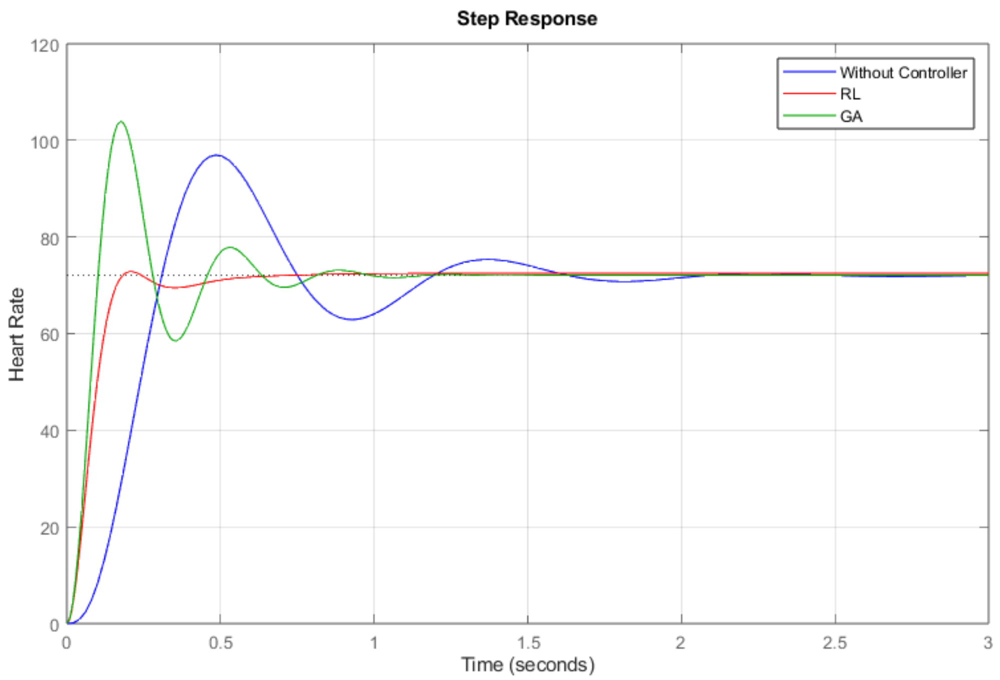


Figure 6. Comparison of the step responses

7. Conclusion

In this study, the effectiveness of Genetic Algorithm (GA) and Reinforcement Learning (RL) in optimizing PID controller parameters for pacemaker applications was investigated. The simulation results demonstrate that both methods improved the step response of the system compared to the uncontrolled closed-loop system. However, their advantages and limitations vary significantly.

Table 5 presents the performance comparison of the optimized PID parameters using both methods. The results indicate that RL-based tuning yielded significantly lower overshoot (1.14%) and peak value (72.82 bpm) compared to GA-based tuning (28.82% overshoot and 92.75 bpm peak value). This suggests that RL provides a more stable and accurate response, minimizing unwanted oscillations in heart rate regulation.

However, GA-based tuning outperformed RL in terms of rise time and settling time. The rise time for GA (0.0357 s) was significantly lower than RL (0.1101 s), and the settling time was also shorter (0.342 s for GA vs. 0.455 s for RL). This implies that GA is more effective for achieving a rapid response, which may be beneficial in scenarios requiring immediate stabilization of heart rate.

To further validate the effectiveness of these approaches, future studies could compare GA and RL with additional optimization techniques such as Particle Swarm Optimization (PSO) or Model Predictive Control (MPC). Additionally, real-time implementation and hardware validation on an actual pacemaker system would provide deeper insights into the practical feasibility of these methods.

The parameter settings for both GA and RL were carefully selected to ensure optimal performance in PID controller tuning. The following table summarizes the key parameters used in the simulations:

Acknowledgement

This research has been presented in IV. International Congress on Artificial Intelligence in Health.

This work has been supported by İzmir Bakırçay University Scientific Research Projects Coordination Unit, under grant number BBAP.2024.011. This work has been supported by İzmir Bakırçay University Scientific Research Projects Coordination Unit, under grant number BBAP.2024.011.

References

[1] Ponikowski, P., Anker, S. D., AlHabib, K. F., Cowie, M. R., Force, T. L., Hu, S., et al. (2014). Heart failure: preventing disease and death worldwide. ESC Heart Failure, 1(1), 4–25.

[2] Bajpai, S., Alam, S., Ali, M.A. (2017). Intelligent Heart Rate Controller using Fractional Order PID Controller Tuned by Genetic Algorithm for Pacemaker. International Journal of Engineering Research & Technology. 6(05), 715-720.

[3] Arunachalam, S. P., Kapa, S., Mulpuru, S.K., Friedman P.A., Tolkacheva, E.G. (2017). Intelligent Fractional-Order PID (FOPID) Heart Rate Controller for Cardiac Pacemaker, 2016 IEEE Healthcare Innovation Point-Of-Care Technologies Conference (HI-POCT), Cancun, Meksika, 2016, s. 105-108

[4] Bikki, P., Dhiraj, Y., & Kumar, R. N. (2023). Implementation of a Dual-Chamber Pacemaker for Low-Power Applications. https://doi.org/10.1109/icecct56650.2023.10179677

[5] Sutton R.S. and Barto, A.G. Reinforcement Learning: An Introduction. The MIT Press, London, 2018.

[6] Holland, J.H., Adaptation in Natural and Artificial Systems. Ann Arbor, MI: University of Michigan Press, 1975.

[7] Biswas, S.C., Das, A., Guha, P. (2006). Mathematical Model of Cardiovascular System by Transfer function Method, Calcutta Medical Journal, 4, 15-17.

[8] Noble, D., A Modification of the Hodgkin-Huxley Equations Applicable to Purkinje Fibre Action and Pace-maker Potentials, 1962 Journal of Physiology , 160, 317-352. PubMed ID: 14480151

[9] Beeler, G.W. and Reuter, H. (1977) Reconstruction of the action potential of ventricular myocardial fibres. The Journal of physiology, 268, 177-210.

[10] Ziegler, J. G., & Nichols, N. B. (1942). Optimum Settings for Automatic Controllers. Trans. ASME, 64(11), 759–768.

[11] Cohen, G. H., & Coon, G. A. (1953). Theoretical Consideration of Retarded Control. Trans. ASME, 75, 827–834.

[12] Momani, S., Batiha I.M., El-Khazali, R. (2019). Design of PIλDδ-Heart Rate Controllers for Cardiac Pacemaker, 2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Ajman, United Arab Emirates, pp. 1-5. https://doi.org/10.1109/ISSPIT47144.2019.9001785

[13] Govind K.R.A., Sekhar, R.A.: Design of a novel PID controller for cardiac pacemaker. 2014 International Conference on Advances in Green Energy (ICAGE), Thiruvananthapuram, India, pp. 82-87 (2014). https://doi.org/10.1109/ICAGE.2014.7050147

[14] Alfarhan, K.A., Mashor, M.Y., Mohd Saad, A.R., Omar, M.I. (2018). Wireless heart abnormality monitoring kit based on raspberry pi. Journal of Biomimetics, Biomaterials and Biomedical Engineering. 35, 96–108. https://doi.org/10.4028/www.scientific.net/JBBBE.35.96

[15] Srivastava, R., Kumar, B. (2022). Design of Anfis based pacemaker controller having improved transient response and its FPGA implementation. Biomedical Signal Processing and Control. https://doi.org/10.1016/j.bspc.2021.103186

[16] Lima, G.S., Savi, M.A., Bessa, W.M. (2023). Intelligent control of cardiac rhythms using artificial neural networks. Nonlinear Dynamics. 111(12), 11543–11557. https://doi.org/10.1007/s11071-023-08447-1

[17] Chen, E.Z., Wang, P., Chen, X., Chen, T., Sun, S. (2022). Pyramid Convolutional RNN for MRI Image Reconstruction. in IEEE Transactions on Medical Imaging. 41(8), 2033-2047. https://doi.org/10.1109/TMI.2022.3153849

[18] Nako, J., Psychalinos, C., & Elwakil, A. S. (2023). Minimum Active Component Count Design of a PIλDμ Controller and Its Application in a Cardiac Pacemaker System. J. Low Power Electron. Appl. https://doi.org/10.3390/jlpea13010013 (7)

[19] Yürdem, B., Aksu, M. F., & Sağbaş, M. (2024). Design of Fractional/Integer Order PID Controller Using Single DVCC and Its Cardiac Pacemaker Application. Circuits Syst Signal Process. https://doi.org/10.1007/s00034-024-02810-2 (10)

[20] Beg, A. H., & Islam, M. Z. (2016, June). Advantages and limitations of genetic algorithms for clustering records. In 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA) (pp. 2478-2483). IEEE.

[21] Chen, X., Qu, G., Tang, Y., Low, S., & Li, N. (2022). Reinforcement learning for selective key applications in power systems: Recent advances and future challenges. IEEE Transactions on Smart Grid, 13(4), 2935-2958.

[22] Zhang, T., & Mo, H. (2021). Reinforcement learning for robot research: A comprehensive review and open issues. International Journal of Advanced Robotic Systems, 18(3), 17298814211007305.

[23] Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. IEEE Signal Processing Magazine, 34(6), 26-38.

[24] Hernandez-Leal, P., Kartal, B., & Taylor, M. E. (2019). A survey and critique of multiagent deep reinforcement learning. Autonomous Agents and Multi-Agent Systems, 33(6), 750-797.

[25] Köse, B., Işıklı, İ., Sagbas, M. (2024). Estimation of Weibull Probability Distribution Parameters with Optimization Algorithms and Foça Wind Data Application. Gazi University Journal of Science, 37(3), 1236-1254. https://doi.org/10.35378/gujs.1311992

[26] Armoogum S., Li, X., (2019). Big Data Analytics and Deep Learning in Bioinformatics with Hadoop. Deep Learning and Parallel Computing Environment for Bioengineering Systems, Elsevier. pp. 17–36. https://doi.org/10.1016/B978-0-12-816718-2.00009-9

■ RESEARCH ARTICLE

AI-Enhanced Test Automation Tool for Desktop Applications

Nağme Cinel Cömertler [a†]

[a] Department of Artificial İntelligence, Ankara University, Ankara, Türkiye
[†] nagmecinel@gmail.com, corresponding author

RECEIVED DECEMBER, 26, 2024
ACCEPTED APRIL 2, 2025

CITATION Cinel Cömertler, N. (2025). AI-Enhanced test automation tool for desktop applications. *Artificial Intelligence Theory and Applications, 5(1)*, 42-50.

Abstract

Test automation is an essential part of the software testing process. This study aims to develop an AI-enhanced test automation tool for testing the user interfaces of desktop applications. The detection of each object in the graphical user interfaces of the software will be carried out using the object detection capabilities of YOLOv9 and Faster R-CNN models. The study emphasizes the importance of preprocessing steps for achieving successful outcomes in object detection processes. These preprocessing steps include image resizing, data augmentation techniques, and balancing the dataset. Additionally, the correct selection and optimization of hyperparameters (e.g., learning rate, number of epochs, network depth, and anchor box dimensions) in object detection models play a critical role in improving model performance. In this study, data analysis techniques using Python were utilized for hyperparameter optimization. Hyperparameters were evaluated and optimized based on metrics such as model accuracy, loss curves, and training time. As a result, high performance was achieved for both the test automation tool and the object detection process. This approach demonstrates the power of artificial intelligence and data analytics in test automation processes, serving as a significant example for both educational and practical applications.

Keywords: object detection, data analysis, test automation, deep learning, image classification

1. Introduction

As software systems become increasingly complex, the demand for more advanced and adaptable testing solutions continues to grow. Traditional test automation tools often face significant limitations in accurately detecting and interacting with GUI (Graphical User Interface) components. Specifically, identifying the class and location of GUI elements without access to their source code or pre-captured screenshots is considered a significant achievement in modern research [1].

This study focuses on improving object detection methods for desktop applications by leveraging cutting-edge deep learning models such as YOLOv9 and Faster R-CNN. To provide a foundational understanding of these models, an in-depth review of convolutional neural networks (CNNs) and single-stage detection architectures is conducted, shedding light on the principles underlying YOLOv9 and Faster R-CNN.

Artificial Intelligence Theory and Applications, ISSN: 2757-9778. ISBN: 978-605-69730-2-4 © 2025 İzmir Bakırçay University

Furthermore, accurate detection of graphical user interface components requires a well-prepared dataset, which involves steps such as data collection, cleaning, and preprocessing. Preprocessing techniques include image resizing, data augmentation, and dataset balancing to address class distribution issues and ensure optimal training conditions.

Subsequently, the training phase of the YOLOv9 and Faster R-CNN models focuses on optimizing key hyperparameters, including learning rate, batch size, and anchor box dimensions, to enhance detection accuracy and computational efficiency. To evaluate the performance of these models, standard metrics such as precision, recall, F1-score, and mean Average Precision (mAP) are employed. A comprehensive comparative analysis highlights the strengths and weaknesses of each model, providing valuable insights into their applicability to desktop environments.

Existing object detection methods for desktop applications often prove inadequate in handling complex and dynamic conditions. By integrating advanced deep learning techniques with rigorous data preprocessing and hyperparameter optimization, this study aims to deliver a more robust and efficient solution. The entire research process, including Literature Analysis and Synthesis, data preprocessing, model training, and performance evaluation is implemented using Python, ensuring reproducibility and efficiency.

2. Literature Analysis and Synthesis

In traditional object detection methods, techniques such as edge detection and corner detection played a fundamental role and were effective in applications with smaller data sets. However, these methods were insufficient to provide accuracy across different data sets and environments. For example, while Viola and Jones' real-time face detection method (Viola & Jones, 2001) attracted attention with its simplicity, the method was limited to static images and had scalability problems [4]. Similarly, the HOG (Histogram of Oriented Gradients) sensor was used to detect pedestrians, but could not provide consistent results in different environmental conditions [5]. Deformable Part-based Model (DPM), which was developed for the detection of more complex structures, enabled objects to be divided into parts and each part to be evaluated independently, but encountered problems such as high computational costs and low speed [6].

These limitations have accelerated the transition to deep learning-based methods. Convolutional neural networks (CNNs), in particular, have revolutionized object detection. CNN-based methods can be divided into two basic groups.

In the first stage, areas that may contain objects are determined (region proposals). In the second stage, these regions are classified in more detail and the boundary boxes are optimized. Among the methods in this category, R-CNN and Faster R-CNN have shown high performance in terms of accuracy [7]. However, the training and inference processes of these models are computationally intensive and slow.

Since object detection is performed in a single step, it is faster and suitable for real-time applications. YOLO (You Only Look Once) and SSD (Single Shot Multibox Detector) are the pioneers of this category [8]. In YOLOv4 and later, these methods have approached two-stage methods in terms of accuracy and speed [9].

In deep learning models, data preprocessing plays a crucial role in directly influencing model performance. Various methods are employed during this process to ensure the dataset is prepared for effective training. One common approach involves image resizing and normalization. For instance, Girshick demonstrated that fixing input dimensions (e.g., 224x224 pixels) in the R-CNN model improved overall accuracy [10]. Similarly, YOLO models normalize all images by scaling pixel values to the [0,1] range, ensuring consistency across diverse datasets. Additionally, data augmentation techniques, including rotation, translation, brightness adjustments, and scaling, have been widely adopted, particularly for small datasets, to enhance model performance. Such methods have been successfully implemented in models like YOLO and Faster R-CNN, as evidenced in previous studies [11].

Another essential aspect of data preprocessing is data cleaning, which addresses missing or inconsistent data within the dataset. Techniques such as imputation, deletion, or replacement with mean values are employed to ensure the integrity of the dataset and maintain statistical reliability [12]. Once these preprocessing steps are completed, the focus shifts to optimizing hyperparameters during the model training phase to maximize performance.

Hyperparameter optimization is a critical process that directly affects the accuracy and efficiency of deep learning models. For instance, systematic screening of hyperparameter ranges has shown significant improvements in models like YOLO. Studies have highlighted that parameters such as learning rate and momentum can significantly impact model accuracy, with lower learning rates yielding higher accuracy in YOLOv3 [13]. Bayesian optimization methods have also been successfully applied, as demonstrated by Johnson et al., who optimized hyperparameters in the Faster R-CNN model to achieve faster and more efficient results. Furthermore, selecting appropriate anchor box sizes tailored to the dataset has been emphasized as a key factor in enhancing performance for models such as RetinaNet and Faster R-CNN [7].

3. Data Preparation

In this study, different methods were followed to collect the data required for training. Since the designed test automation tool is intended to be used in desktop applications, GUI components in Windows and Linux operating systems were primarily collected by taking screenshots using SikuliX. Later, GUI components in web applications were obtained by web scraping methods. Web scraping was performed in Python using BeautifulSoup and Selenium libraries. The 2000 data items were collected.

Data preprocessing played an important role in preparing the data for training the model [14]. In this context, the following steps were taken:

3.1. Data Cleaning

We applied data cleaning to improve the quality and suitability of our dataset, creating a cleaner and more usable dataset for modeling. To perform the data cleaning process, Python libraries such as OpenCV, OS, Image and numpy were utilized. Through this process, corrupted/missing images, unlabeled images, duplicate images, and irrelevant images in the dataset were identified and removed, resulting in a refined dataset.

As shown in the Figure-1 below, 1500 image remaining after the cleaning process.

Figure 1. Dataset Before and After Cleaning

3.2. Data Augmentation

Data augmentation was utilized in the models we used to prevent overfitting and expand the dataset. Techniques such as rotation, flipping, cropping, brightness and contrast adjustment, zooming, and cutout were applied. These operations were implemented using the OpenCV and Albumentations libraries in Python.

This approach successfully increased the diversity of the dataset, making the model more generalizable, preventing overfitting, and effectively addressing the challenge of having a relatively small dataset by augmenting it.

After applying data augmentation techniques, the dataset was expanded to include a total of 4000 data. This process increased the diversity of the dataset and helped to improve the robustness and generalizability of the model.

3.3. Normalization

The pixel values of the dataset images we used are represented as integers in the range of 0-255. To normalize these pixel values to the range [0,1], the following formula is fundamentally used.

$$normalized_{value} = \frac{original\_value}{255}$$

This operation was performed using the OpenCV library in Python. To ensure compatibility with the YOLO model, the dimensions of the image needed to be resized to a standard input size (e.g., 416x416 pixels) therefore, resizing was also performed using the OpenCV library.

The reasons for performing normalization are as follows:

- It ensures a better scale alignment between the model weights and the input data.
- It allows optimization algorithms to work faster and more effectively.
- Large values can disrupt the learning process of the model.

3.4. Control of Labels

A total of 4000 images were automatically labeled using the Roboflow platform. The data were classified into categories such as "Button," "Checkbox," "Dropdown," and other GUI element names.

Finally, 4000 data created as a result of these processes were divided into 8 different classes and used for training YOLOv9-10 and Faster R-CNN models.

These detailed preprocessing steps played a critical role in increasing the accuracy of the models and reducing inaccurate predictions.

3.5. Division of the Data Set

The data was divided into three groups training (80%), validation (10%), and testing (10%) as shown in the Figure-1 below. It is necessary to obtain a validation data set to ensure that the model does not fit too much on the training data. This can be achieved by comparing the training progress of the model with the training and validation data sets.

This decision was made because when the data set was divided into training (70%), validation (15%), and testing (15%), lower values were obtained for metrics such as precision and recall. The validation set was used to fine-tune the hyperparameters. The test set is reserved to evaluate the overall performance of the model on previously unseen data.

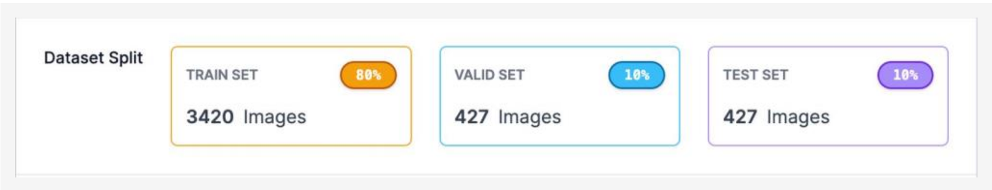


Figure 2. Train-Test-Validation Split of the Dataset

4. Model Selection and Training Process

The process of model selection is a critical step in the development of any machine learning application. For object detection tasks, selecting the right model depends on several factors, including the application requirements, available computational resources, dataset characteristics, and performance trade-offs. In this study, we evaluated two state-of-the-art object detection models, YOLOv9 and Faster R-CNN, to determine their suitability for detecting GUI elements.

4.1. YOLOv9 Training

Pre-trained weights trained on the COCO dataset were used as the starting point. This approach leveraged the model's ability to detect generic object features.

The training process was conducted using the PyTorch and YOLOv9 frameworks. Cloud-based resources, specifically Google Colab, were employed to handle the computational load.

Initial experiments set the number of epochs to 20. However, early testing revealed underfitting. Gradually, the epochs were increased to 50, balancing training duration and model performance. Other hyperparameters such as learning rate, batch size, and momentum were fine-tuned using grid search techniques.

The model was fine-tuned with a learning rate scheduler and regularization techniques to prevent overfitting.

4.2. Faster R-CNN Training

The Faster R-CNN model used pre-trained weights from the ImageNet dataset, providing a strong baseline for feature extraction. Training was performed using the torchvision library in PyTorch, which provides a robust implementation of Faster R-CNN.

Due to its complexity, training required high computational power. Experiments were conducted using local GPU resources and later scaled to cloud services when necessary. Epochs were set to 30 after evaluating overfitting and underfitting scenarios. Anchor box sizes and aspect ratios were adjusted to align with GUI element dimensions.

4.3. Comparison of YOLOv9 and Faster R-CNN Models

YOLOv9 provides significant advantages in terms of speed and real-time processing, making it suitable for applications requiring rapid inference [15]. However, its performance may degrade when detecting small or intricate objects. In contrast, Faster R-CNN excels at achieving higher accuracy in detecting complex objects but is computationally intensive, making it less ideal for real-time applications.

Table 1. Comparison of Models

Name of the Criteria	YOLOv9	Faster R-CNN
Model Type	Single-stage	Two-stage
Training Framework	PyTorch,YOLOv9	PyTorch
Pre-trained Weights Source	Collected Dataset	Collected Dataset
Optimal Number of Epochs	50	30
Key Strength	Faster inference time	High accuracy on intricate details
Computational Requirements	Moderate (Google Colab)	Moderate (Google Colab)
Evaluation Metrics	Precision and recall,	Higher mAP and F1-score
Best Use Case	Real-time applications	Detailed object detection

According to the information presented in Table 1, both models utilized custom collected datasets for fine-tuning; however, their optimal number of epochs differed significantly, with YOLOv9 requiring 50 epochs and Faster R-CNN achieving optimal performance at 30 epochs.

5. Performance Evaluation

YOLOv9 outperforms Faster R-CNN in terms of precision and F1-score, making it better suited for applications where precision is critical. However, Faster R-CNN offers slightly higher recall, which may be beneficial for tasks requiring comprehensive detection.

YOLOv9 handles categorical distinctions better, while Faster R-CNN struggles with misclassifications, especially in complex or overlapping classes. YOLOv9's higher AUC value suggests superior generalization in binary classification tasks, whereas Faster R-CNN may require further optimization to improve its ROC performance.

YOLOv9 converges faster and generalizes better, making it more suitable for time-sensitive projects with limited computational resources. Faster R-CNN, while slower to train, may benefit from additional regularization techniques to improve generalization.

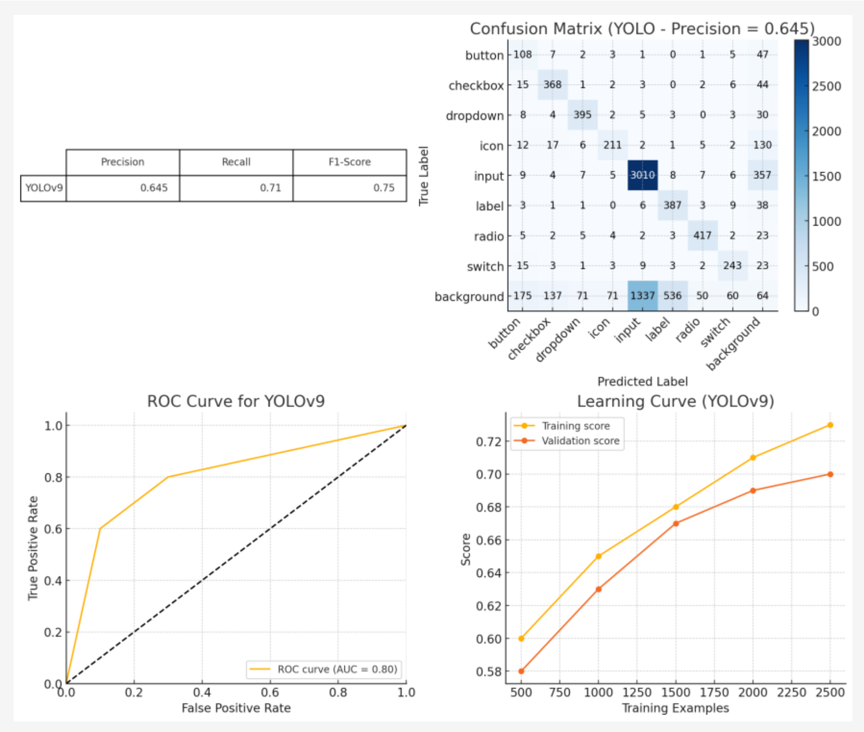


Figure 3. YOLOv9 Performance Report

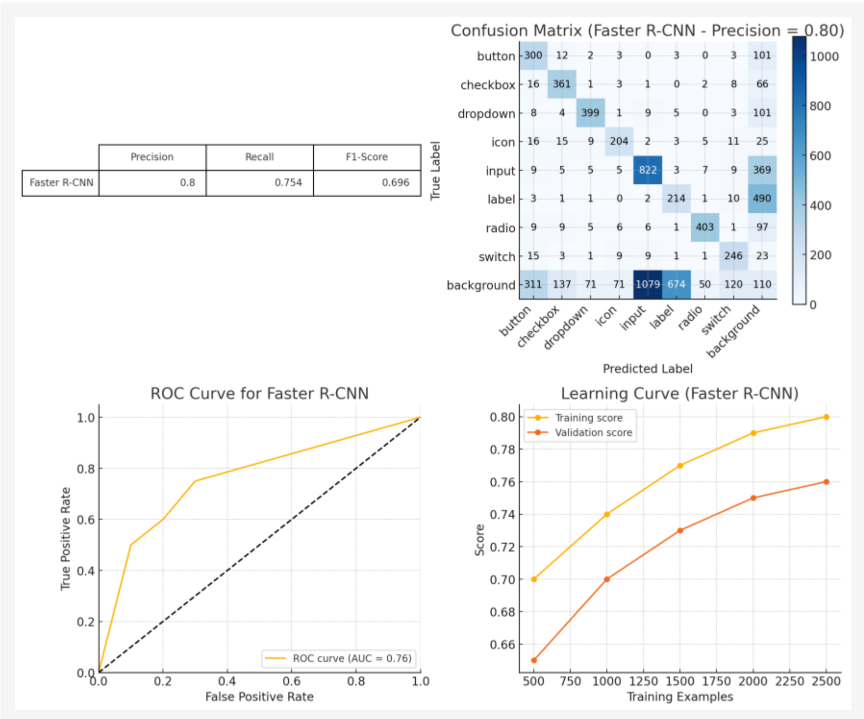


Figure 4. Faster R-CNN Performance Report

The final choice between these models should be guided by the specific application requirements. As shown in Figure-3, YOLOv9 demonstrates superior speed and moderate accuracy, making it the preferred option for tasks requiring real-time performance. On the other hand, as illustrated in Figure-4, Faster R-CNN achieves higher recall and excels in intricate detail detection, making it more suitable for applications where precision in complex scenarios is critical.

6. Conclusion and Future Work

This study underscores the growing importance of AI-powered UI element detection in the development of test automation tools. As modern software systems become increasingly complex, automating the detection and interaction with UI components has become a critical aspect of efficient and reliable software testing. Leveraging state-of-the-art object detection models such as YOLOv9 and Faster R-CNN, we explored how these approaches can enhance the accuracy and speed of UI element detection tasks.

The data collection and analysis process played a pivotal role in this study. A robust dataset of annotated GUI elements was created, involving meticulous preprocessing steps such as normalization, augmentation, and careful splitting into training, validation, and test sets. These steps ensured that the models were trained on high-quality data, which is fundamental for achieving reliable and generalizable performance.

Performance evaluation revealed clear distinctions between YOLOv9 and Faster R-CNN. YOLOv9 demonstrated its advantages in speed and computational efficiency, making it highly suitable for real-time UI element detection tasks. It offers moderate accuracy, which is sufficient for many automation scenarios, especially those requiring rapid feedback. Conversely, Faster R-CNN exhibited superior recall and detail detection, making it more suitable for scenarios that demand high precision and intricate UI interactions, albeit at the cost of computational efficiency.

The choice of model ultimately depends on the specific requirements of the application. YOLOv9 is better suited for scenarios demanding high speed, such as real-time testing or rapid prototyping, where computational resources are limited. Faster R-CNN excels in tasks requiring high recall and detailed recognition, such as testing applications with complex UI layouts or overlapping elements.

This comparative analysis highlights the trade-offs inherent in selecting the appropriate model for AI-driven test automation. The study emphasizes the importance of balancing performance metrics with application needs and resource constraints.

Future work could focus on integrating these models into real-world test automation pipelines, exploring ensemble approaches to combine the strengths of both models, and optimizing them further for domain-specific tasks. Ultimately, advancements in AI-powered UI detection have the potential to revolutionize software testing, making it faster, more accurate, and scalable for the challenges of modern software systems.

Acknowledgements

I would like to express my sincere gratitude to HAVELSAN for its valuable contributions and support to the implementation of this study. I would like to thank Mr. Kadir Herkiloğlu, Prof. Dr. Asım Egemen Yılmaz, Mr. Çağrı Şenkal, Mr. Zeynel Abidin Çelikel, Assoc. Prof. Dr. Savaş Takan and Mr. Fatih Bildirici, who made significant contributions to the process through their guidance and support. I also would like to thank İpek İşçelebi and Bora Yılmaz for their contributions to the implementation of the software; the editorial team and all individuals involved in the publication process at DergiPark for their kind support and efforts; and finally, my husband and family, who have always been by my side with their patience and support throughout this process.

References

[1] J. Chen, M. Xie, Z. Xing, C. Chen, X. Xu, L. Zhu, and G. Li, "Object detection for graphical user interface: old fashioned or deep learning or a combination" in Proc. of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2020), New York, NY, USA, 2020, pp. 1202-1214. [Online]. https://doi.org/10.1145/3368089.3409691

[2] Dwivedi, S. K., & Rawat, B. (2015, October). A review paper on data preprocessing: A critical phase in web usage mining process. In 2015 International Conference on Green Computing and Internet of Things (ICGCIoT) (pp. 506-510). IEEE.

[3] C. Zhang, T. Shi, J. Ai, and W. Tian, "Construction of GUI Elements Recognition Model for AI Testing based on Deep Learning," in Proc. 2021 8th International Conference on Dependable Systems and Their Applications (DSA), Yinchuan, China, 2021, pp. 508-515. doi: 10.1109/DSA52907.2021.00075.

[4] Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), 1, I-511-I-518.

[5] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 1, 886-893.

[6] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 32(9), 1627-1645

[7] Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 39(6), 1137–1149.

[8] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. Proceedings of the European Conference on Computer Vision (ECCV 2016), 21–37

[9] Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934

[10] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014), 580–587.

[11] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(60), 1–48.

[12] Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data (Vol. 793). John Wiley & Sons.

[13] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281–305.

[14] G. Li, G. Baechler, M. Tragut, and Y. Li, "Learning to Denoise Raw Mobile UI Layouts for Improving Datasets at Scale," in Proc. of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22), New York, NY, USA, 2022, Article 67, pp. 1-13. [Online]. https://doi.org/10.1145/3491102.3502042

[15] Huang, Jonathan et al. “Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors.” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 3296-3297.

■ RESEARCH ARTICLE

Deep Learning Based Decision Support System for Retinal Disease Classification: Diabetic Retinopathy and Macular Hole

Belinay Kabataş [a], **Emre Ölmez** [a†]

[a] Department of Biomedical Engineering, İzmir Bakırçay University, İzmir, Turkey
† emre.olmez@bakircay.edu.tr, corresponding author

RECEIVED APRIL 13, 2025
ACCEPTED APRIL 25, 2025

CITATION Kabataş, B., & Ölmez, E. (2025). Deep learning based decision support system for retinal disease classification: diabetic retinopathy and macular hole. *Artificial Intelligence Theory and Applications, 5(1)*, 51-62.

Abstract

In this study, a deep learning-based decision support system was developed to classify diabetic retinopathy (DR), macular hole (MH), and healthy cases using fundus images. A total of 1,397 fundus images, selected from the open-source Retinal Disease Classification dataset, were used in the training and testing phases. ResNet50, InceptionV3, and Xception models were trained with different hyperparameter configurations, and their performances were evaluated comparatively. Among the models, ResNet50 achieved the highest accuracy on the test set, reaching 93.79%. However, the Xception model exhibited superior robustness and stability across various hyperparameter settings, consistently delivering balanced and reliable classification performance. These findings indicate that deep learning-based approaches can be effectively utilized as clinical decision support systems for the diagnosis of retinal diseases.

Keywords: deep learning, fundus images, diabetic retinopathy, macular hole, convolutional neural networks, Resnet50, Xception, InceptionV3

---

1. Introduction

The use of artificial intelligence (AI) methods in healthcare has rapidly expanded in recent years. In particular, AI applications powered by deep learning techniques are increasingly being adopted in the medical field [1]. In this context, the early diagnosis of eye diseases is crucial for preventing permanent vision loss. Conditions such as diabetic retinopathy (DR) and macular hole (MH) can lead to severe visual impairment, especially in their advanced stages [2]. Fundus images play a critical role in the diagnosis of these diseases by enabling a detailed examination of the retinal layer, thus supporting physicians in the decision-making process [2], [3]. However, manual interpretation of fundus images is time-consuming and susceptible to human error. At this point, digital image processing and deep learning methods offer valuable assistance by serving as decision support systems for the classification of fundus images [4].

In this study, deep learning-based image processing models were trained to classify DR [5], MH [6], and healthy samples using fundus images. The Retinal Disease Classification dataset [7], an open-source collection of fundus images representing different disease

Artificial Intelligence Theory and Applications, ISSN: 2757-9778. ISBN: 978-605-69730-2-4 © 2025 İzmir Bakırçay University

categories, was used in the study. These images were processed and trained with various hyperparameter configurations using ResNet50 [8], InceptionV3 [9], and Xception [10] models, and the performance of each model was comparatively analyzed. The results indicated that ResNet50 and Xception achieved high classification accuracies of 93.70% and 92.94%, respectively, by effectively capturing distinctive features in fundus images. InceptionV3 also performed well with an accuracy of 88.70%, though slightly lower than the other two models.

This study highlights the effectiveness of deep learning-based artificial intelligence approaches in the diagnosis of retinal diseases and aims to contribute to future clinical applications. The findings support the integration of AI-powered decision support systems, particularly in the early detection of conditions such as DR and MH, where early diagnosis is critical.

2. Retinal Diseases

The retina is one of the fundamental structures responsible for the visual function of the eye. Retinal diseases can lead to serious and permanent vision loss if not diagnosed and treated in a timely manner [11]. In this section, we focus on DR and MH, two common retinal disorders that fall within the scope of this study and can cause significant visual impairment if not detected early.

2.1 Diabetic Retinopathy

Diabetic retinopathy (DR) is a serious retinal disease affecting one-third of the approximately 285 million people with diabetes worldwide. One third of these individuals also have vision-threatening symptoms of DR [12]. DR occurs in diabetic patients when the blood vessels in the retina are damaged due to high blood glucose levels [2]. Since the disease is usually asymptomatic in the initial stages, it is difficult to diagnose early and often manifests itself in later stages with symptoms such as blurred vision, dark spots in the visual field and vision loss. Therefore, early diagnosis of DR is critical to stop the progression of the disease and prevent vision loss. In the absence of early diagnosis and treatment, the disease can cause severe vision loss, up to blindness [5]. Figure 1 presents sample images labeled as DR from the Retinal Disease Classification dataset [7].

Figure 1. Fundus images labeled as DR in the Retinal Disease Classification dataset.

2.2 Macular Hole

A macular hole (MH) is a small tear or opening that occurs in the macula, the central region of the retina. Since the macula is responsible for sharp and detailed central vision, a hole in this area can significantly impair visual acuity [6]. A MH is often associated with the natural process of aging; it occurs as the structure of the vitreous fluid deteriorates and separates from the macula with age. The incidence is particularly high in individuals over 50 years of age. Early signs of the disease include distorted central vision, blurred vision and the inability to see fine details. When treatment is delayed, the damage to the macular area deepens and this can lead to permanent vision loss. Early detection of MH and appropriate intervention can preserve central vision [6], [13]. Figure 2 presents sample images labeled as MH from the Retinal Disease Classification dataset [7].

Figure 2. Fundus images labeled as MH in the Retinal Disease Classification dataset.

3. Literature Review

In recent years, deep learning-based models developed for the early diagnosis of retinal diseases—especially DR—have attracted significant interest in the research community. The potential of artificial intelligence-based systems to enhance the efficiency of clinical processes is particularly evident when faced with limited data and complex diagnostic challenges. In this section, we review some of the most influential studies on the classification and diagnosis of retinal diseases.

Kori et al. (2018) employed a CNN-based ensemble approach for the automatic grading of DR and macular edema (ME). To address the challenge of limited labeled data, the researchers utilized transfer learning by fine-tuning models that were previously trained on ImageNet, adapting them to fundus images. The final model achieved an accuracy of 83.9% for DR grading and 95.45% for ME grading. The study emphasized that the ensemble method outperformed a single CNN model and highlighted the effectiveness of transfer learning techniques [14].

Sahlsten et al. (2019) proposed a deep learning-based method for the automatic detection of DR and ME using high-resolution fundus images. Their study achieved high accuracy rates, emphasizing the potential for increased cost-effectiveness in existing screening programs [15].

Torre et al. (2020) developed a deep learning classifier aimed at improving interpretability in DR grading. Their method assigned importance scores to individual pixels or regions contributing to the final classification, enhancing transparency for clinical experts. This not only improved the diagnostic reliability but also underscored its potential for integration into clinical decision support systems [16].

Özçelik and Altan (2021) introduced a two-stage model for the early diagnosis of DR. In the first stage, two-dimensional signal processing techniques were utilized to prevent overfitting, while in the second stage, classification was performed using ESA-based transfer learning. The model was trained on 5,100 fundus images and achieved an accuracy of 97.8%. This study demonstrated the model’s speed and reliability as a diagnostic tool [5].

Aykat and Senan (2023) proposed a deep learning-based method for diagnosing retinal diseases such as DR and cataract. In their study, fundus images were enhanced using histogram equalization as preprocessing and 99% accuracy was achieved with the MobileNet-based hybrid model. These results suggest that the hybrid model outperforms similar methods in existing literature [2].

Polater and Işık (2024) conducted a study on the classification of DR severity levels using the APTOS 2019 dataset. By employing the DenseNet121 model, they achieved approximately 97% accuracy. Their findings reaffirm the superior performance of the DenseNet121 architecture and the overall efficacy of deep learning methods in DR diagnosis [17].

These studies clearly demonstrate that deep learning methods offer high accuracy and reliability in the diagnosis of DR and other retinal diseases. Validating these models on diverse datasets and across various clinical scenarios may broaden their applicability in diagnostic and treatment workflows and contribute to the development of robust clinical decision support systems.

4. Material and Method

4.1 Dataset

The Retinal Disease Classification dataset [7] used in this study is a comprehensive and open-source dataset designed for the classification of eye diseases based on retinal fundus images. It contains a total of 3,200 fundus images representing 46 distinct ocular diseases. The images were captured using three different fundus cameras—TOPCON 3D OCT-2000, Kowa VX-10, and TOPCON TRC-NW300—and each image was meticulously labeled by two senior retina specialists. The use of multiple imaging devices and expert annotations enhances both the diversity and reliability of the dataset.

The fact that the images were obtained from different devices increases the generalization capability of the deep learning models by reducing dependency on a specific device or lighting condition. Additionally, the dataset's wide range of disease classes facilitates the development of models capable of detecting multiple retinal disorders simultaneously. For the purpose of this study, three classes were selected: DR, MH, and Healthy (No Disease). These classes are commonly encountered in clinical settings and exhibit a relatively balanced distribution within the dataset, allowing for more consistent and reliable results during model training and evaluation.

From the 1,397 fundus images selected for this study, a total of 1,043 images were allocated to the training set, comprising 349 DR, 293 MH, and 401 healthy images. The remaining 354 images were used for testing, including 120 DR, 100 MH, and 134 healthy images. Accordingly, approximately 75% of the data was used for training and 25% for testing. Figure 3 illustrates representative fundus images from each of the three selected classes.

Figure 3. Sample fundus images from Retinal Disease Classification dataset. From left to right: Macular Hole, Diabetic Retinopathy, No Disease

4.1.1 Image Preprocessing

Various image preprocessing steps were applied to ensure high accuracy and generalization capability of the trained deep learning models. Fundus images were resized to 299×299 pixels to be compatible with the input layers of the deep learning models. In addition, the pixel values of the images were normalized to the range [0, 1] to facilitate the training process of the models.

In this study, data augmentation strategies were also included in the training processes. Figure 4 shows the data augmentation process used in the training scenarios where the data augmentation strategy was applied. Data augmentation aims to diversify the limited amount of training data and increase the robustness of the models against different variations. The operations in the data augmentation process were randomly applied to the images at each iteration. The applied methods include random rotation up to 30 degrees (rotation_range=30), horizontal and vertical shift up to 20% (width_shift_range=0.2, height_shift_range=0.2), shear up to 20% (shear_range=0.2), zoom in up to 20% (zoom_range=0.2), random change of brightness values within a 20% range (brightness_range=[0.8, 1.2]) and random flip on the horizontal axis (horizontal_flip=True).

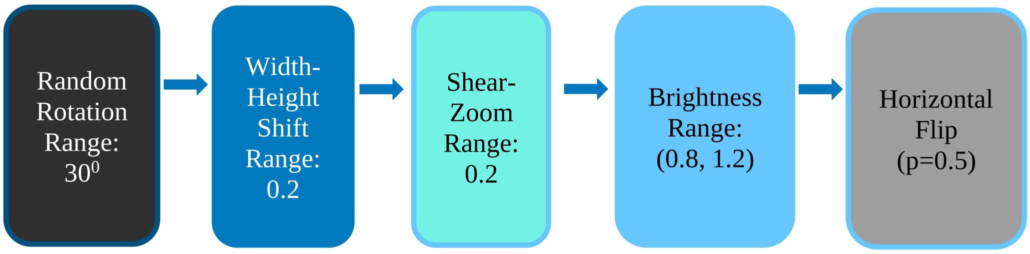


Figure 4. Data augmentation process

4.2 Deep Neural Networks

Deep learning, as one of the cornerstones of modern artificial intelligence research, has achieved significant advances in image processing and classification. In particular, convolutional neural networks (CNNs) have demonstrated remarkable success in analyzing and classifying visual data, and these models have evolved to become more efficient and effective over time [10], [18]. In this study, three different CNN-based deep learning architectures were employed for retinal image classification: ResNet50, InceptionV3, and Xception.

ResNet50, developed by He et al. in 2015, is a CNN architecture designed to address the vanishing gradient problem encountered during the training of deep neural networks [8]. It introduces **residual connections** that allow the output of a layer to be added to the input of a deeper layer, enabling more effective training of very deep architectures. ResNet50 consists of a total of 50 layers and has demonstrated high performance in complex image classification tasks.

InceptionV3 is a CNN model developed by Google that applies convolutional filters of varying sizes in parallel, allowing for the extraction of image features at multiple spatial scales [9]. This **multi-scale filtering** approach enables the model to capture visual patterns at different levels of detail while improving parameter efficiency and reducing computational cost. Thanks to this design, InceptionV3 achieves high classification accuracy and is widely adopted across various computer vision applications.

Xception is a CNN architecture designed as an enhanced version of the Inception model [10]. It utilizes depthwise separable convolutions to reduce the number of parameters and improve computational efficiency compared to traditional CNN structures. Xception has outperformed InceptionV3 on large-scale datasets such as ImageNet (ILSVRC) and JFT. Furthermore, it has proven highly effective in transfer learning scenarios, achieving strong performance across diverse classification tasks.

5. Training and Results

In this study, three different CNN based models—ResNet50, InceptionV3, and Xception—were used, and a total of 15 training scenarios were evaluated, combining three different hyperparameters (end-to-end learning, data augmentation, and learning rate) for each architecture. All models were trained using the Adam optimization algorithm for 50 epochs, with learning rates set at 0.001 and 0.0001. The detailed training and test accuracy results, along with loss values for each scenario, are summarized in Table 1. Additionally, Figure 6 illustrates the epoch-based test accuracy progression for each model throughout the training phase, while Figure 7 presents corresponding accuracy curves for the training sets. Further insights into model convergence and stability can be observed in the loss curves presented in Figures 8 and 9, showing test and training loss, respectively. Moreover, Table 2 provides a more comprehensive evaluation, detailing precision, recall, and F1-score metrics for each class and scenario, enabling deeper analysis beyond overall accuracy.

In each model architecture, the original output layers were removed, and a Global Average Pooling layer was added to the end of the base model to adapt it to the target dataset. This approach converts the output feature maps into a one-dimensional vector, helping reduce the number of parameters while preserving high-level feature representations. Following this, a Dense layer with 256 neurons and ReLU activation was added, along with 30% Dropout to prevent overfitting. Finally, a Softmax-activated output layer was used for classification into three classes: Diabetic Retinopathy, Macular Hole, and Healthy. The block diagram illustrating this modified CNN architecture is shown in Figure 5.

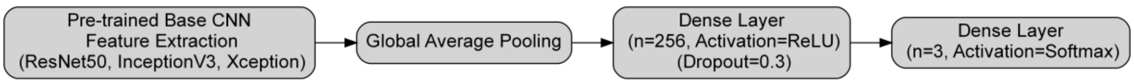


Figure 5. Modified CNN Architecture

In Table 1, when the "End to End Learning" parameter is "-", the base model weights are frozen, and only the newly added layers are trained. When the parameter is "+", training is performed end-to-end. All models were initialized with pre-trained weights from the ImageNet dataset. In scenarios where data augmentation was applied, the strategy illustrated in Figure 4 was used to evaluate the robustness of the models against image variations.

Table 1. Training and Test Results

Model	End to End Learning	Augmentation	Learning Rate	Loss	Acc	Test-Loss	Test-Acc
ResNet_1	-	-	0.0010	1.0291	0.4382	1.0527	0.4209
ResNet_2	-	+	0.0010	1.0568	0.4267	1.0506	0.4379
ResNet_3	-	-	0.0001	1.0240	0.4861	1.0628	0.4237
ResNet_4	+	-	0.0001	0.0033	0.9990	0.2751	0.9379
ResNet_5	+	+	0.0001	0.1276	0.9569	0.6505	0.8559
Inception_1	-	-	0.0010	0.0379	0.9895	0.4572	0.8814
Inception_2	-	+	0.0010	0.2624	0.8907	0.3732	0.8701
Inception_3	-	-	0.0001	0.0732	0.9818	0.3131	0.8870
Inception_4	+	-	0.0001	0.0122	0.9971	0.4580	0.8870
Inception _5	+	+	0.0001	0.0856	0.9732	0.3672	0.8729
Xception_1	-	-	0.0010	0.0179	0.9952	0.4203	0.9040
Xception_2	-	+	0.0010	0.1965	0.9271	0.2570	0.9153
Xception_3	-	-	0.0001	0.1036	0.9703	0.2301	0.9181
Xception_4	+	-	0.0001	0.0950	0.9962	0.4866	0.9294
Xception_5	+	+	0.0001	0.0203	0.9942	0.2378	0.9294

Table 2. Test set evaluation metrics

	DR			MH			No_Disesase		
Model	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
ResNet_1	0.20	0.02	0.03	0.61	0.14	0.23	0.41	0.99	0.58
ResNet_2	0.00	0.00	0.00	0.65	0.22	0.33	0.42	0.99	0.59
ResNet_3	0.50	0.03	0.05	0.64	0.14	0.23	0.41	0.99	0.58
ResNet_4	0.97	0.92	0.94	0.92	0.94	0.93	0.92	0.96	0.94
ResNet_5	1.00	0.66	0.79	0.89	0.93	0.91	0.77	0.98	0.86
Inception_1	0.96	0.78	0.86	0.85	0.86	0.86	0.85	0.99	0.91
Inception_2	0.96	0.78	0.86	0.92	0.81	0.86	0.79	0.99	0.88
Inception_3	0.93	0.83	0.88	0.91	0.83	0.87	0.85	0.98	0.91
Inception_4	0.98	0.82	0.89	0.77	0.95	0.85	0.93	0.90	0.92
Inception _5	0.99	0.81	0.89	0.89	0.84	0.87	0.79	0.96	0.86
Xception_1	0.96	0.89	0.92	0.91	0.84	0.88	0.86	0.96	0.91
Xception_2	0.93	0.90	0.92	0.92	0.87	0.89	0.90	0.96	0.93
Xception_3	0.94	0.89	0.91	0.89	0.90	0.90	0.92	0.96	0.94
Xception_4	0.95	0.95	0.95	0.98	0.84	0.90	0.89	0.98	0.93
Xception_5	0.98	0.87	0.92	0.90	0.94	0.92	0.91	0.98	0.94

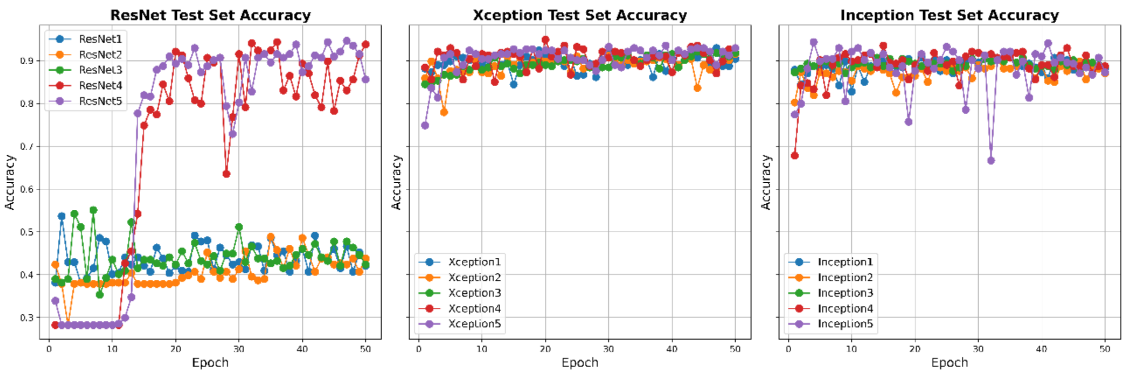


Figure 6. Test set accuracies of the models over 50 epochs

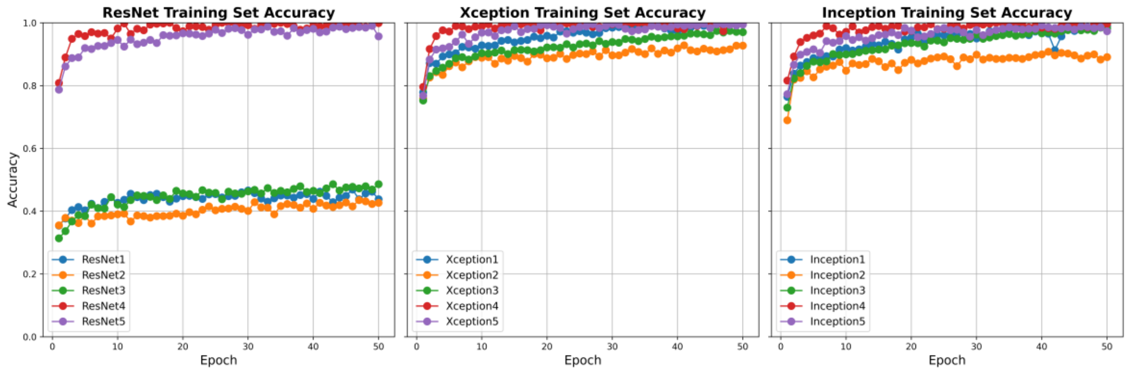


Figure 7. Training set accuracies of the models over 50 epochs

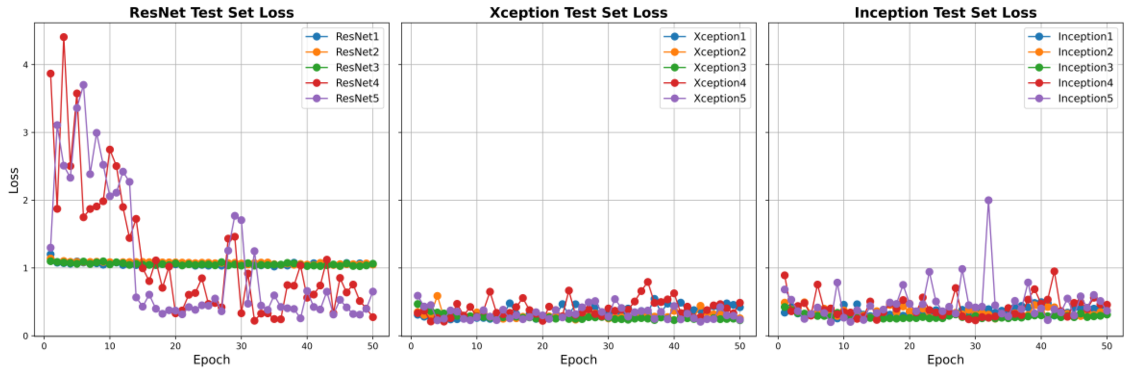


Figure 8. Test set loss of the models over 50 epochs

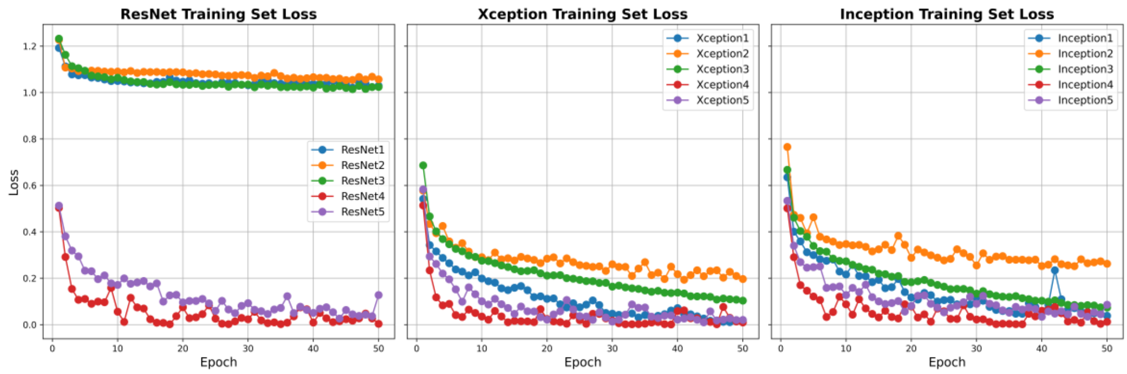


Figure 9. Training set loss of the models over 50 epochs

The results obtained from the training scenarios presented in Table 1 reveal important insights into the performances of the three convolutional neural network (CNN) architectures (ResNet50, InceptionV3, and Xception) across different hyperparameter configurations.

For the ResNet50 model, training scenarios utilizing a learning rate of 0.001 did not yield satisfactory results, especially when the base layers of the network were frozen. In these configurations, test accuracies remained notably low. However, reducing the learning rate to 0.0001 and performing end-to-end training of all layers dramatically improved the test accuracy, reaching 93.79%, which was the highest among all evaluated configurations. This significant improvement indicates that ResNet50 requires careful adjustment of learning rate and complete fine-tuning to achieve optimal performance.

The experiments conducted using the InceptionV3 architecture showed relatively consistent but limited variation in performance across different hyperparameter combinations, with accuracy typically around 88.70%. Interestingly, scenarios that involved end-to-end training and data augmentation—despite theoretical expectations of enhanced robustness—demonstrated fluctuating accuracy levels without substantial improvements. This observation implies that InceptionV3 has inherent stability in training dynamics but may have reached a saturation point in extracting discriminative features from the fundus image dataset, limiting further performance gains.

On the other hand, the Xception architecture consistently demonstrated high and stable performance across all evaluated scenarios. It not only achieved higher average test accuracy compared to ResNet50 and InceptionV3 but also showed robust generalization and sensitivity to hyperparameter changes. Particularly, training the base layers with a lower learning rate notably enhanced its performance, underscoring the adaptability and robustness of the Xception architecture.

Beyond accuracy, additional metrics provided in Table 2 (precision, recall, and F1-score) offer deeper insight into the classification performance across each disease category (Diabetic Retinopathy, Macular Hole, and Healthy). Analysis of these metrics further highlights the superiority of Xception. Across all classes, Xception consistently outperformed the other architectures, achieving precision, recall, and F1-scores frequently exceeding 0.90 in configurations involving end-to-end learning and data augmentation. These findings reinforce Xception's suitability for accurate multi-class classification tasks in ophthalmological applications.

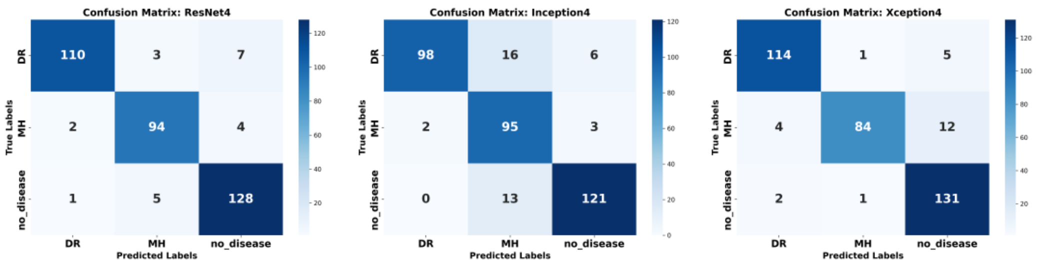


Figure 10. Confusion matrices of the models with high classification performance

Figure 10 presents the confusion matrices of the best-performing models trained with identical hyperparameter configurations. When evaluating class-wise performance based solely on the confusion matrices, the Xception model demonstrates the highest number of correct classifications for the DR class. Xception produced fewer misclassifications in this class compared to the other two models, indicating greater reliability in distinguishing DR cases. For the MH class, the most successful results were achieved by the InceptionV3 model, which attained the highest number of correct predictions, distinguishing itself from the other architectures. Although ResNet50 performed comparably to InceptionV3 in the MH class, it was observed that Xception made noticeably more classification errors in this category. In the no_disease class, Xception once again stood out, delivering the highest number of correct classifications and exhibiting strong performance in identifying healthy individuals. In contrast, InceptionV3 produced more misclassifications in this class, frequently confusing it with MH. In summary, the confusion matrices indicate that while Xception achieved superior results in the DR and no_disease classes, InceptionV3 was the most effective in classifying MH cases. ResNet50, on the other hand, displayed a balanced performance across all classes with relatively low misclassification rates.

In contrast, ResNet50 exhibited significant performance variability between different scenarios. While scenarios involving frozen base layers and higher learning rates resulted in poor precision and recall scores, the architecture successfully recovered in configurations that involved end-to-end training with a reduced learning rate. This pattern suggests that ResNet50 requires a carefully controlled training environment to mitigate overfitting and achieve its full potential.

InceptionV3, while maintaining moderate stability, demonstrated competitive yet generally lower overall performance compared to Xception. Particularly, its precision, recall, and F1-scores for the Macular Hole and Healthy classes were commendable, but it was less consistent across the Diabetic Retinopathy class, highlighting a limitation in effectively distinguishing between visually challenging classes.

The training dynamics presented in Figures 6 and 7 further illustrate differences in learning behaviors among the architectures. Notably, Xception exhibited smoother and more consistent progression in both training and test accuracies throughout the epochs. Additionally, the loss trajectories shown in Figures 8 and 9 complement these observations, demonstrating more stable loss convergence for Xception, whereas ResNet50 and InceptionV3 experienced noticeable fluctuations, indicative of overfitting tendencies, particularly when training base layers were frozen or when higher learning rates were employed.

Collectively, these detailed analyses underscore the robustness, stability, and superior generalization capabilities of the Xception model. Its consistently high performance across various metrics and configurations makes it particularly suitable as the basis for reliable and effective clinical decision support systems in the early diagnosis of retinal diseases.

6. Conclusion and Future Work

This study provides a comparative analysis of the classification performance of deep learning-based models for the diagnosis of retinal diseases such as diabetic retinopathy (DR) and macular hole (MH). Three different CNN architectures—ResNet50, InceptionV3, and Xception—were trained and evaluated under various hyperparameter configurations. According to the findings, ResNet50 achieved the highest test accuracy (93.79%) among the models used in the study. However, as illustrated in Figure 6, the Xception model demonstrated more stable performance across different training scenarios. InceptionV3, while exhibiting more consistent performance than ResNet50, achieved lower accuracy than Xception. The results also indicate that both InceptionV3 and Xception models converged more rapidly during the early training stages compared to ResNet50, achieving high accuracy values in a shorter time frame, as depicted in Figure 6.

These results highlight the strong potential of deep learning methods in ophthalmologic image analysis and diagnostic decision support systems. Such systems, developed as alternatives to time-consuming and expert-dependent manual evaluations, can expedite clinical decision-making, reduce the workload on healthcare professionals, and enable earlier intervention for patients at risk of vision loss.

For future work, we plan to evaluate the current models on larger and more imbalanced datasets involving multi-class disease classification. Additionally, we aim to integrate explainable AI (XAI) techniques such as Grad-CAM and LIME to improve the interpretability of the models' decision-making processes. Exploring lightweight and optimized architectures (e.g., MobileNet, EfficientNet) suitable for real-time applications will also be a crucial step toward integration into portable medical devices. Furthermore, training models using multi-center, multi-device fundus image datasets is expected to improve the generalizability and reliability of decision support systems.

In conclusion, this study highlights the effectiveness of deep learning-based models in classifying retinal diseases using fundus images, providing a foundation for future research toward more reliable and interpretable clinical decision support systems in eye care.

Acknowledgement

We would like to thank Izmir Bakırçay University, The Center for Artificial Intelligence Studies and Research in Healthcare, for the resources and support provided for the implementation of the study

References

[1] B. Vatansever, H. Aydın, and A. Çetinkaya, "Genetik Algoritma Yaklaşımıyla Öznitelik Seçimi Kullanılarak Makine Öğrenmesi Algoritmaları ile Kalp Hastalığı Tahmini," *Journal of Scientific Technology and Engineering Research*, Nov. 2021, doi: 10.53525/jster.1005934.

[2] Ş. Aykat *et al.*, "Derin Öğrenme Kullanılarak Fundus Görüntülerinden Katarakt ve Diyabetik Retinopati Tespiti Detection of Cataract and Diabetic Retinopathy from Fundus Images Using Deep Learning," 2023.

[3] Q. Wei *et al.*, "Development and Validation of an Automatic Ultrawide-Field Fundus Imaging Enhancement System for Facilitating Clinical Diagnosis: A Cross-sectional Multicenter Study," *Engineering*, Oct. 2024, doi: 10.1016/j.eng.2024.05.006.

[4] K. Jin and J. Ye, "Artificial intelligence and deep learning in ophthalmology: Current status and future perspectives," Nov. 01, 2022, *Elsevier Inc.* doi: 10.1016/j.aopr.2022.100078.

[5] Y. B. Özçelik and A. Altan, "Diyabetik Retinopati Teşhisi için Fundus Görüntülerinin Derin Öğrenme Tabanlı Sınıflandırılması," *European Journal of Science and Technology*, Dec. 2021, doi: 10.31590/ejosat.1011806.

[6] E. Yaşar, N. Erol, M. D. Bilgeç, and A. İ. Çakmak, "Coexistence of peripheral retinal diseases with macular hole," *Turk J Ophthalmol*, vol. 49, no. 4, pp. 209–212, Aug. 2019, doi: 10.4274/tjo.galenos.2019.06706.

[7] S. Pachade *et al.*, "Retinal fundus multi-disease image dataset (Rfmid): A dataset for multi-disease detection research," *Data (Basel)*, vol. 6, no. 2, pp. 1–14, Feb. 2021, doi: 10.3390/data6020014.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385

[9] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Dec. 2016, pp. 2818–2826. doi: 10.1109/CVPR.2016.308.

[10] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Institute of Electrical and Electronics Engineers Inc., Nov. 2017, pp. 1800–1807. doi: 10.1109/CVPR.2017.195.

[11] F. S. Sorrentino *et al.*, "Novel Approaches for Early Detection of Retinal Diseases Using Artificial Intelligence," Jul. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/jpm14070690.

[12] R. Lee, T. Y. Wong, and C. Sabanayagam, "Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss," Dec. 01, 2015, *BioMed Central Ltd*. doi: 10.1186/s40662-015-0026-2.

[13] D. Mikhail *et al.*, "The role of artificial intelligence in macular hole management: A scoping review," Jan. 01, 2024, *Elsevier Inc.* doi: 10.1016/j.survophthal.2024.09.003.

[14] A. Kori, S. S. Chennamsetty, M. S. K. P., and V. Alex, “Ensemble of Convolutional Neural Networks for Automatic Grading of Diabetic Retinopathy and Macular Edema,” Sep. 2018, [Online]. Available: http://arxiv.org/abs/1809.04228

[15] J. Sahlsten *et al.*, “Deep Learning Fundus Image Analysis for Diabetic Retinopathy and Macular Edema Grading,” *Sci Rep*, vol. 9, no. 1, Dec. 2019, doi: 10.1038/s41598-019-47181-w.

[16] J. de la Torre, A. Valls, and D. Puig, “A deep learning interpretable classifier for diabetic retinopathy disease grading,” *Neurocomputing*, vol. 396, pp. 465–476, Jul. 2020, doi: 10.1016/j.neucom.2018.07.102.

[17] S. N. Polater *et al.*, “Diyabetik Retinopati Tespiti İçin Derin Öğrenmeye Dayalı Sınıflandırma Deep Learning-Based Classification for Diabetic Retinopathy Detection,” 2024.

[18] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” May 27, 2015, *Nature Publishing Group*. doi: 10.1038/nature14539.

■ REVIEW ARTICLE

A New Era in Diabetes Management: Generative Artificial Intelligence

Meleknur Göktaş [a†], **Tuğba Bilgehan** [b]

[a] Dr. Hulusi Alataş Elmadağ State Hospital , Ankara, Turkey
[b] Department of Nursing, Ankara Yıldırım Beyazıt University Ankara, Turkey.
[†] meleknur.gkts@gmail.com, corresponding author.

RECEIVED DECEMBER 06, 2024
ACCEPTED MARCH 18, 2025

CITATION Göktaş, M., & Bilgehan, T. (2024). A new era in diabetes management: generative artificial intelligence. Artificial Intelligence Theory and Applications*, 5(1*), 63-81.

Abstract

Diabetes mellitus (DM) is a rapidly increasing global health issue that requires effective self-management to prevent complications and improve quality of life. In recent years, advancements in generative artificial intelligence (GenAI) have created new opportunities to support DM self-management by providing personalized care solutions. This study is designed as a systematic review. Numerous studies in the literature have examined the contributions of GenAI models to DM self-management, and reviewing these studies is essential to provide a general framework on this topic. The primary aim of this study is to systematically examine research that utilizes GenAI in DM management. This systematic review was conducted in accordance with PRISMA guidelines. A comprehensive literature search was carried out between February and October 2024 across PubMed, Scopus, Web of Science, Google Scholar, Ulakbim, Türk Medline, and national databases. Using the keywords "diabetes," "generative artificial intelligence," and "diabetes self-management," studies published between 2018 and 2024 were identified. A total of 19 studies that met the inclusion criteria were analyzed in terms of the GenAI models used, application areas, and reported outcomes. Among the reviewed studies, GPT-based models were predominant, appearing in 53% of the research. In addition, models such as GAN, LSTM, WaveNet, GRU, Markov-Bayes, Google Bard, and Mobiguide were also utilized. Moreover, the findings of this study highlight that GenAI-based systems are widely adopted in DM self-management and possess significant potential to facilitate this process. These systems not only provide information but also incorporate advanced support mechanisms that enhance patient monitoring and clinical decision-making processes. GenAI has made notable contributions to DM care, particularly by developing personalized care plans, offering tailored dietary and exercise recommendations, generating educational materials, predicting blood glucose (BG) levels, providing individualized guidance, and supporting clinical workflows. As GenAI continues to evolve and adapt to the specific contexts and demands of the medical field, its role in DM care is expected to become increasingly prominent. However, several challenges have been reported, including concerns over data security, privacy, misinformation generation, and suboptimal performance in detecting critical conditions such as hypoglycemia. Addressing these ethical, technical, and security-related limitations requires further research and technological advancements. Future studies should prioritize enhancing the reliability, usability, and diagnostic accuracy of GenAI applications to ensure their seamless integration into clinical practice.

Keywords: generative artificial intelligence, diabetes self-management, diabetes

Artificial Intelligence Theory and Applications, ISSN: 2757-9778. ISBN: 978-605-69730-2-4 © 2025 İzmir Bakırçay University

1. Introduction

Diabetes Mellitus (DM), a major contributor to multiple health issues, has emerged as an escalating public health problem, significantly impacting a substantial portion of the global population [1, 2]. The prevalence of DM is estimated to rise to 643 million by 2030 and 783 million by 2045 [2, 3]. The primary objectives of DM management are to prevent or mitigate complications and preserve an optimal quality of life. Thus, effective long-term self-management is essential for individuals living with DM as a chronic condition [4]. Individuals with diabetes face an increased risk of developing various complications if they fail to maintain optimal blood glucose (BG) control [5]. The progression of chronic complications not only diminishes quality of life but also has detrimental physical, psychological, and social effects on individuals, while substantially escalating the economic burden of DM on healthcare systems [3,6]. Maintaining optimal glycemic control is a cornerstone of DM management to mitigate the risk of acute and chronic complications. Nevertheless, through comprehensive DM management and strict metabolic control, the onset of these serious complications can be delayed, or in some cases, entirely prevented [1, 3, 5, 7]. Many factors that affect successful DM self-management are modifiable and practical. Key behaviors, including healthy eating, regular physical activity, BG monitoring, adherence to prescribed medication regimens, developing healthy coping strategies, and problem-solving skills, are fundamental components of effective DM self-management [8]. Attaining optimal glycemic control in DM necessitates active participation in self-management efforts, which not only enhances the effectiveness of DM care but also fosters patient empowerment [1].

Each person is unique, and individuals with diabetes may have different preferences, values, and goals for their self-management. Therefore, creating a personalized management plan is essential. Such plans should consider various factors, including the individual’s age, cognitive abilities, work or school schedule, health beliefs, support systems, dietary habits, physical activity level, social situation, financial concerns, cultural factors, and literacy [9]. Furthermore, a comprehensive management approach should integrate factors such as DM history (including duration, complications, and current medication regimen), comorbidities, health priorities, other medical conditions, patient care preferences, and life expectancy [9, 10]. To facilitate this process, a DM care team plays an essential role in supporting individuals with diabetes. However, due to constraints such as time limitations, financial barriers, or other challenges, individuals with diabetes may encounter difficulties in regularly consulting a DM educator [11]. In these cases, a variety of strategies and techniques should be utilized to support self-management efforts. Technology can play a pivotal role in this context, facilitating daily DM self-management activities, including blood glucose (BG) monitoring, physical activity, healthy eating, medication adherence, complication monitoring, and problem-solving. Although the use of technology to support DM self-management is not a new concept, the diversity of technological strategies has expanded as individuals have become more tech-savvy, devices have become more accessible, and new technologies have emerged [12].

Recent advancements in artificial intelligence (AI) and machine learning (ML) techniques have become indispensable in addressing the complexities of DM management, empowering both patients and healthcare professionals in their daily management of DM [13]. AI refers to a set of techniques that enable computers to simulate human intelligence, encompassing ML as a subset. Often referred to as machine intelligence, AI involves the capacity of computer systems to learn from data inputs or historical information. The term 'AI' is commonly used to describe scenarios where a machine mimics cognitive functions of the human brain, such as learning and problem-solving

[14]. Within the healthcare sector, ML models have been effectively applied and are widely recognized as potent tools that enable computers to learn from data [15].

ML is a subset of AI that encompasses techniques enabling machines to enhance their performance in tasks through experience, and includes deep learning (DL). DL, in turn, is a subset of ML that utilizes neural networks, enabling a machine to autonomously train itself to perform tasks. The hierarchical evolution of these technologies can be summarized as follows: AI, ML, and DL [16]. ML models leverage extensively pre-trained data to generate accurate and relevant outputs. In ML, users must define and supply algorithms with sufficient information to generate accurate predictions. In contrast, DL algorithms utilize artificial neural network architectures to autonomously process data, allowing them to learn and generate accurate predictions based on high-level features extracted from the data [17].

ML technology, referred to as GenAI, can generate new data based on the training dataset. Generative models produce data that closely resemble the original dataset. A distinctive feature of GenAI is its capability to perform unsupervised learning, meaning it can identify patterns from data without the need for explicitly labeled examples. Some GenAI models learn how real-world data is distributed and subsequently generate new datasets that are statistically similar to the original dataset (Figure 1) [18, 19].

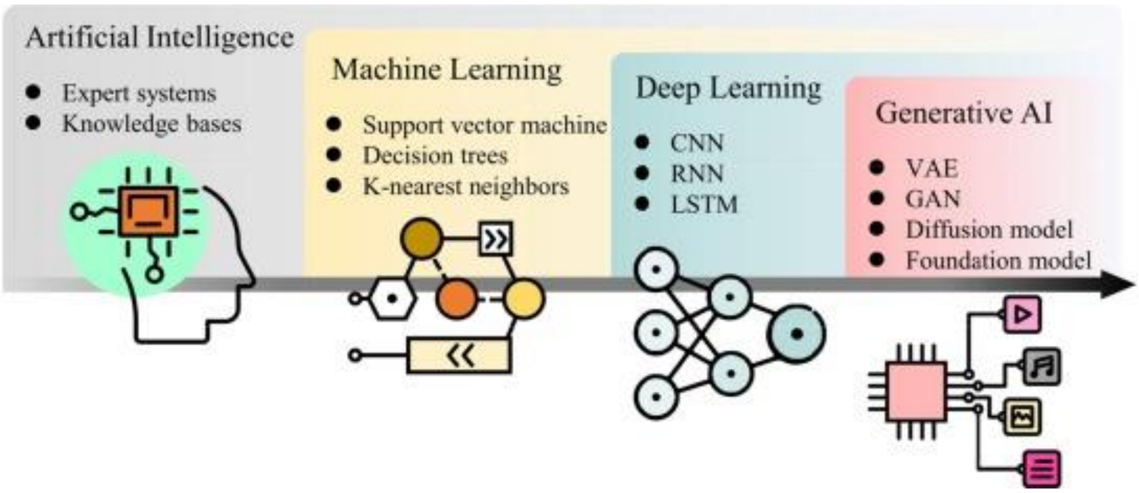


Figure 1: The development of AI and GenAI [19].

GenAI encompasses various models, including Generative Pre-trained Transformers (GPTs), Generative Adversarial Networks (GANs), Bayesian Networks (BNs), Artificial Neural Networks (ANNs), and Large Language Models (LLMs) (Figure 2) [20, 21, 22, 23].

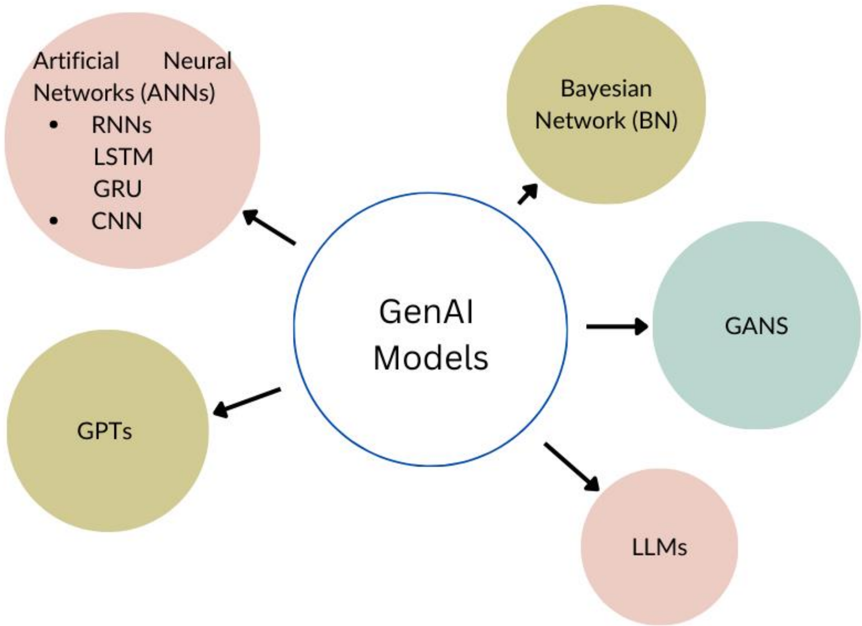


Figure 2: Key Techniques and Models in GenAI [20, 21, 22, 23].

GenAI models are emerging as promising tools within the context of healthcare. These models can analyze an individual's genetic profile, lifestyle, and medical history to provide tailored predictions about treatment options. By considering factors that may influence an individual's response to medication, GenAI can aid in optimizing therapeutic efficacy and improving individual outcomes [24, 13, 25, 26]. By generating personalized scenarios and responses tailored to individual needs, GenAI can effectively address the unique needs of individuals [26]. The predictions generated by GenAI guide the adjustment and optimization of treatments, thereby enabling the provision of more personalized care. Furthermore, by addressing individuals' health concerns and anxieties, GenAI offers supportive responses that help patients feel more reassured and less isolated throughout their healthcare journey [27].

GenAI also demonstrates significant potential in the field of patient education. It can generate personalized educational materials tailored to patients' specific conditions, symptoms, or inquiries. For instance, GenAI can provide individuals with diabetes information on BG management, nutrition, exercise, and medication adherence. Through interactive dialogues, patients can pose questions and receive answers that enhance their understanding of their conditions. This feature is particularly valuable for patients who may feel hesitant or embarrassed to directly ask specific questions of healthcare professionals. Furthermore, GenAI can simplify complex medical concepts by generating visual materials, such as diagrams or infographics. For example, it can illustrate how a specific medication works within the body to enhance patient comprehension [27]. Furthermore, patients with different levels of education and health literacy can improve their health literacy through GenAI-generated content tailored to various reading levels [28]. For instance, GenAI can deliver medication adherence reminders through email or text messages and provide explanations regarding the importance of adhering to prescribed treatment plans. Moreover, to enhance accessibility for individuals whose primary language is not English, GenAI can generate educational materials in multiple languages [29].

By analyzing patient-specific data, GenAI can predict health outcomes, identify potential risks, and recommend personalized treatment plans. Its capacity to extract insights from data presents significant potential for enhancing patient care processes [27]. Acting as a tool for patient engagement, education, and personalized interventions, GenAI provides a promising opportunity to improve DM management and transform healthcare delivery.

A review of the literature indicates that the majority of studies on GenAI have been published in the past three years, highlighting its rapidly growing use in DM management. For example, recent studies have shown that ChatGPT can simplify health information for individuals with diabetes [30], predict disease risks [31], and support patients in achieving positive outcomes by aiding them in managing lifestyle behaviors [32]. Furthermore, GenAI's ability to exhibit human-like empathy is highly regarded by users, as it delivers responses that are comparable to those provided by physicians, thereby fostering a sense of trust and rapport among patients [33].

GenAI demonstrates substantial potential in monitoring treatment adherence in individuals with diabetes and is increasingly recognized as a transformative tool for the future of DM management and care.

2. Materials-Methods

2.1. Purpose, Significance, and Scope of the Study

Purpose and Type of the Study

This research was designed as a descriptive systematic review. Numerous studies in the existing literature have examined the contributions of GenAI models to DM self-management, and reviewing these studies is essential for providing a comprehensive framework on this topic. The primary aim of this study is to systematically examine research that utilizes GenAI in DM management.

Research Question

What methods, applications, and outcomes have been reported in the current literature regarding the use of GenAI in DM self-management?

2.2. Method and Analysis of the Study

Studies on the use of GenAI in DM self-management, published between 2018 and 2024, were reviewed through searches conducted between February and October 2024 in the Scopus, Web of Science, PubMed, Ulakbim, Türk Medline, and Google Scholar databases. The searches were conducted using the keywords 'diabetes,' 'generative artificial intelligence,' and 'diabetes self-management,' either individually or in various combinations.

In the study selection process, relevant keywords were used to search the databases in line with the research question. After applying the inclusion and exclusion criteria, the remaining studies were retrieved, and their titles and abstracts were screened. Studies that did not correspond to the research question were excluded, while the full texts of the remaining studies were evaluated in detail. Ultimately, those meeting the eligibility criteria were included in the final set of studies for this systematic review.

All stages of the study—including article identification, screening, and selection—were conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic

Reviews and Meta-Analyses) guidelines [34]. The PRISMA flow diagram, which outlines the study selection process, has been presented in Figure 3.

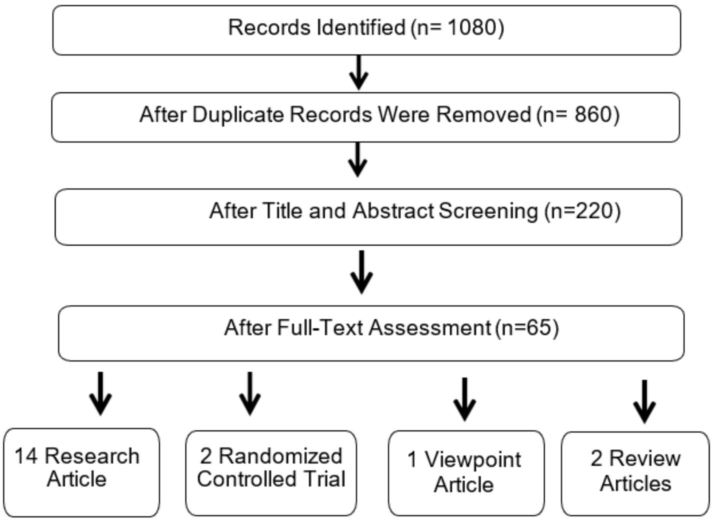


Figure 3. Flow Diagram of the Study

2.3 Inclusion and Exclusion Criteria

Inclusion Criteria:

- Studies addressing the use of GenAI in DM self-management.
- Articles published in English or Turkish.
- Research published between 2018 and 2024.
- Original research articles (including observational studies, experimental studies, clinical studies, and reviews).
- Studies with full-text availability.
- Studies with sufficient methodological quality to be included in a systematic review in accordance with PRISMA guidelines.

Exclusion Criteria:

- Studies related to DM or GenAI that do not focus on DM management.
- Abstracts, conference proceedings, commentaries, book chapters, editorial articles, and theses.
- Articles requiring paid access.

2.4. Data Collection

In the initial stage of the review process, an evaluation form outlining the inclusion criteria was developed. Based on this form, database searches were systematically conducted. As a result of the screening process, a total of 19 studies published between 2018 and 2024, focusing on the use of GenAI in DM management, were identified. All studies that met the inclusion criteria (n = 19) have been included in the review.

In the second stage of the review, a structured checklist was developed, comprising the following components: study title, study type and design, sample group and size, year of publication, and study outcomes. In accordance with this checklist, the titles and abstracts of all relevant articles identified through the database searches were independently reviewed by the author. No other individuals participated in the data collection phase of this study.

3. RESULTS

3.1. Characteristics of the Included Studies

As a result of the review, a total of 19 studies conducted in Turkey between 2018 and 2024 focusing on the use of GenAI in DM management were identified. Among these, 14 (73%) were original research articles, 2 (11%) were review articles, 1 (5%) was a viewpoint article, and 2 (11%) were randomized controlled trials.

3.2. Study Designs and Sample Characteristics

The characteristics of the included studies are presented in detail across separate tables: research articles in Table 1, randomized controlled trials in Table 2, the viewpoint article in Table 3, and review articles in Table 4.

Table 1. Descriptive Characteristics of the Research Articles on the Application of GenAI in DM Self-Management (n=14)

Title of the Study	Type of Study	Theory/Model Used	Sample Group	Study Reference	Study Findings
Blood glucose prediction with deep neural networks using weighted decision level fusion.	Research Article	A fusion of Long Short-Term Memory (LSTM), WaveNet, and Gated Recurrent Units (GRU) architectures.	The study sample is derived from the expanded OhioT1DM dataset, which comprises the BG history of 12 individuals with diabetes over an eight-week period and encompasses 19 distinct data types, including administered insulin doses and physiological sensor readings.	Dudukcu et al., 2021 [35]	The fusion performance of "LSTM + WaveNet + GRU" yielded more successful results in BG prediction. The predicted values are planned to be used for insulin dose calculations, with future development as a mobile application.
Blood glucose prediction for type 1 DM using generative adversarial networks	Research Article	A novel DL model using a modified GAN architecture for predicting BG levels in individuals with type 1 diabetes (T1DM).	BG data from 12 individuals with T1DM over an eight-week period.	Zhu et al., 2020 [36]	In this study, a novel DL model is proposed to predict future BG levels based on past continuous glucose monitoring measurements, meal intake, and insulin administration. When compared to the RNN (Recurrent Neural Network) prediction model, the GAN model demonstrated better validation performance and a smaller RMSE (Root Mean Square Error) for most of the contributors during the training process.

Using an optimized generative model to infer the progression of complications in type 2 diabetes patients	Research Article	Markov Jump Process and Bayesian network	Longitudinal EHRs of 9,298 individuals with type 2 diabetes (T2DM) or prediabetes (2005–2016, China).	Wang et al., 2022 [37]	The findings of this study demonstrated that the system could predict 55.3% of individual complications and 31.8% of complication patterns of progressive T2DM at an early stage, allowing for appropriate management that could potentially delay or prevent these complications.
The Future of Patient Education: AI-Driven Guide for Type 2 Diabetes	Research Article	OpenAI's ChatGPT	70 T2DM-related questions, each asked three times.	Hernandez et al., 2023 [38]	98.5% of responses aligned with care standards, outperforming traditional online search engines, with minimal inappropriate responses (1.5%), underscoring the need for continuous AI improvements.
An AI Dietitian for Type 2 Diabetes Mellitus Management Based on Large Language and Image Recognition Models: Preclinical Concept Validation Study	Research Article	Creating an AI based nutritionist program using advanced language and image recognition models using ChatGPT and GPT 4.0	206 individuals with T2DM and 26 endocrinologists	Sun et al., 2023 [39]	Positive feedback from dietitians and accurate food recognition, enabling personalized meal analysis and dietary guidance. The model developed at the end of this study can identify ingredients from images of a patient's meal and provide nutritional guidance and diet recommendations.
Building Trustworthy Generative Artificial Intelligence for Diabetes Care and Limb Preservation: A Medical Knowledge Extraction Case	Research Article	OpenAI's ChatGPT-4 with a RAG architecture.	NIH Diabetes Self-Management Education Standards knowledge base, 295 articles, 175 questions.	Mashatian et al., 2024 [40]	RAG model effectively delivers reliable medical information for self-education and emphasizes the importance of content validation and prompt engineering.
A Study on the Development of a Chatbot Using Generative AI to Provide Diets for Diabetic Patients	Research Article	OpenAI's ChatGPT	Data from 10 dietary guidelines, adapted for personalized seasonal diet plans.	Lee et al., 2024 [41]	Facilitates personalized diets and supports elderly health management with enhanced services and datasets.
Comparative evaluation of generative artificial intelligence systems for patient queries on age-related macular degeneration and diabetic macular edema	Research Article	ChatGPT-3.5, ChatGPT-4, Google Bard	22 patient queries from 68 anti-VEGF-treated individuals.	Posa et al., 2024 [42]	This study compared the effectiveness of three GenAI systems – Chat GPT-3.5, Chat GPT-4, and Google Bard – in providing clear and concise answers to patient questions about diabetic macular edema. GPT-4 was deemed most effective for patient communication due to its clear and simple language.

Artificial intelligence chatbots for the nutrition management of diabetes and the metabolic syndrome	Research Article	GPT-3.5-turbo0301	The total number of requests is 63, with 9 requests entered for each of the 7 conditions.	Naja et al., 2024 [43]	While ChatGPT provided human-like responses, significant gaps were identified, emphasizing that it cannot replace dietitian expertise.
Appropriateness of Artificial Intelligence Chatbots in Diabetic Foot Ulcer Management	Research Article	ChatGPT (OpenAI) GPT-4 (OpenAI) GPT-4 Turbo (OpenAI), GoogleBard (Google LLC), BingAI Balanced-mode (Microsoft Corp.), Perplexity (Perplexity AI) and Claude-2'den (Anthropic)	42 clinical questions on diabetic foot ulcers.	Shiraishi et al., 2024 [44]	Chatbots showed 91.2% accuracy but inconsistent evidence levels. Claude-2 had the highest reference accuracy; ChatGPT had the lowest. Variability and hallucinations highlight the need for cautious clinical use.
Assessment of the information provided by ChatGPT regarding exercise for patients with type 2 diabetes: a pilot study	Research Article	ChatGPT (V.4.0)	14 common patient exercise questions reviewed by two DM care specialists.	Chung and Chang, 2024 [45]	ChatGPT can serve as supplementary educational material but may provide incomplete answers for certain exercise-related questions.
Evaluation of ChatGPT-4 Performance in Answering Patients' Questions About the Management of Type 2 Diabetes	Research Article	ChatGPT-4	24 patient questions	Gokbulut et al., 2024 [46]	It has been observed that, while answering a series of questions related to the pharmacological management of T2DM, no inaccurate information was identified, and responses were highly consistent and reliable. However, readability levels varied, with many responses being classified as difficult to read.
Using Generative AI to Improve the Performance and Interpretability of Rule-Based Diagnosis of Type 2 Diabetes Mellitus	Research Article	GPT	Dataset of 768 instances with eight predictors and one outcome class.	Kopitar et al., 2024 [47]	This study, which explores the combination of association rule mining with GPT-based advanced natural language processing for classifying non-insulin-dependent DM, demonstrates that ChatGPT is effective in predicting diabetic and non-diabetic conditions. However, further research is required to enhance diagnostic accuracy in DM classification.
MobiGuide: guiding clinicians and chronic patients anytime, anywhere.	Research Article	MobiGuide	10 atrial fibrillation patients in Italy and 20 gestational individuals with diabetes in Spain	Peleg et al., 2022 [48]	Higher adherence rates and improved health outcomes were observed, including better glycemic control, with enhanced clinician engagement and patient quality of life.

Table 2. Characteristics of Randomized Controlled Trials on the Use of Generative AI in Diabetes Management (n=2)

Title of the Study	Type of Study	Theory/Model Used	Sample Group	Study Reference	Study Findings
Use of Voice-Based Conversational Artificial Intelligence for Basal Insulin Prescription Management Among Patients With Type 2 Diabetes	Randomized Controlled Trial	Alexa	32 adults with T2DM requiring initiation or adjustment of basal insulin therapy	Nayak et al., 2023 [49]	In this randomized clinical trial of a voice-based conversational AI application for autonomous basal insulin management in adults with T2DM, participants in the AI group demonstrated significantly greater improvements in the time required to achieve the optimal insulin dose, insulin adherence, glycemic control, and DM-related emotional distress compared to those in the standard care group.
Decoding Type 2 Diabetes Through Point-of-Care Testing, Cloud-based Monitoring, and Generative Augmented Retrieval Model-driven Virtual Diabetes Education: A Comprehensive Approach to Glycemic Control	Randomized Controlled Trial	AI-Powered Metabolic Coach Designed to Provide Personalized Recommendations	The study sample consists of 100 individuals aged between 18 and 65 who have been diagnosed with T2DM.	Shaikh et al., 2024 [50]	Participants in the intervention group, who received guidance from an AI-powered metabolic coach designed to provide personalized recommendations, demonstrated superior outcomes in HbA1c improvement, plasma glucose control, and related parameters compared to the control group.

Table 3. Characteristics of the Viewpoint Article on the Use of GenAI in DM Management (n=1)

Title of the Study	Type of Study	Theory/Model Used	Study Reference	Study Findings
ChatGPT in diabetes care: An overview of the evolution and potential of generative artificial intelligence model like ChatGPT in augmenting clinical and patient outcomes in the management of diabetes.	Viewpoint Article	ChatGPT	Dey, 2023 [51]	It can provide personalized care, where individualized treatment plans, glucose monitoring, and medication reminders are generated based on personal patient data. However, ethical considerations and data security must be carefully addressed, and any obtained information should be verified by healthcare professionals.

Table 4. Characteristics of the Review Articles on the Use of GenAI in DM Management (n=2)

Title of the Study	**Type of Study**	**Theory/Model Used**	**Study Reference**	**Study Findings**
The Future of Diabetes Care: Navigating with Generative Language Models	Review	OpenAI's GPT-3	Khan, 2023 [52]	This review concluded that generative language models can facilitate personalized care by creating customized treatment plans, glucose monitoring, and medication reminders based on individual patient data. However, ethical concerns and data security should be carefully considered, and any recommendations should be validated by healthcare professionals.
Potential and Pitfalls of ChatGPT and Natural-Language Artificial Intelligence Models for Diabetes Education	Review	OpenAI's GPT	Sng et al., 2023 [53]	This review found that ChatGPT performs well in generating comprehensible and generally accurate responses to questions related to DM self-management and education. The application of large language models has the potential to alleviate some of the burden on individuals with diabetes, allowing those with adequate knowledge of their condition to focus on more complex self-management and educational tasks. However, it is important to acknowledge that ChatGPT is constrained by the datasets on which it was trained. These limitations may lead to errors, such as difficulties in distinguishing between different types of insulin or recognizing variations in BG measurement units. The review emphasizes that healthcare providers should exercise due diligence when assessing AI chatbots for clinical care enhancement and patient guidance, ensuring a thorough understanding of both the strengths and limitations of these models.

Tables 1 through 4 reveal that the most commonly used GenAI models were GPT (53%), followed by Google Bard (6%), WaveNet (3.12%), Mobiguide (3.12%), BingAI Balanced-mode (Microsoft Corp.) (3.12%), Perplexity AI (3.12%), Claude-2 (Anthropic) (3.12%), Alexa (3.12%), AI-Powered Metabolic Coach (3.12%), as well as GAN (3.12%), LSTM (3.12%), GRU (3.12%), Markov Jump Process (3.12%), RAG (3.12%), and Bayesian Network (3.12%).

Thematic analysis of the included studies revealed the distribution of GenAI functions in DM self-management, which is presented in Table 5. When examining the areas in which GenAI models used in the studies included in the scope of this systematic review can be applied to DM self-management, it becomes evident that many models can serve common purposes. Additionally, this comparison highlights which model has previously been utilized for specific self-management actions.

Table 5. The Role of GenAI Applications in DM Self-Management Domains

Self-Management Action	Included Studies
Personalized recommendations (e.g., nutrition, exercise, BG management) [1]	Shaikh et al., 2024 [50], Peleg et al., 2022 [48], Hernandez et al., 2023 [38], Sun et al., 2023 [39], Lee et al., 2024 [41], Naja et al., 2024 [43], Shiraishi et al., 2024 [44], Chung and Chang, 2024 [45], Gokbulut et al., 2024 [46], Nayak et al., 2023 [49], Dey, 2023 [51], Khan, 2023 [52], Sng et al., 2023 [53]
Provision of DM education and information	Shaikh et al., 2024 [50], Hernandez et al., 2023 [38], Sun et al., 2023 [39], Mashatian et al., 2024 [40], Posa et al., 2024 [42], Shiraishi et al., 2024 [44], Chung and Chang, 2024 [45], Gokbulut et al., 2024 [46], Nayak et al., 2023 [49], Dey, 2023 [51], Khan, 2023 [52], Sng et al., 2023 [53]
Early detection of complications and complication alerts	Wang et al., 2022 [37], Mashatian et al., 2024 [40], Posa et al., 2024 [42], Shiraishi et al., 2024 [44], Chung and Chang, 2024 [45], Gokbulut et al., 2024 [46], Dey, 2023 [51], Sng et al., 2023 [53]
Emotional support and stress management	Mashatian et al., 2024 [40], Nayak et al., 2023 [49], Dey, 2023 [51]
Insulin and medication management	Shaikh et al., 2024 [50], Mashatian et al., 2024 [40], Gokbulut et al., 2024 [46], Nayak et al., 2023 [49], Dey, 2023 [51], Sng et al., 2023 [53]
Continuous glucose monitoring, BG prediction [2], and personalized analysis	Dudukcu et al., 2021 [35], Zhu et al., 2020 [36], Mashatian et al., 2024 [40], Dey, 2023 [51]
Tracking DM progression and providing lifestyle recommendations	Peleg et al., 2022 [48], Wang et al., 2022 [37], Dey, 2023 [51], Sng et al., 2023 [53]
Communication with healthcare professionals/ decision support for healthcare providers	Peleg et al., 2022 [48], Khan, 2023 [52]
Prediction of future risks and early diagnosis	Peleg et al., 2022 [48], Wang et al., 2022 [37], Posa et al., 2024 [42], Shiraishi et al., 2024 [44], Chung and Chang, 2024 [45], Gokbulut et al., 2024 [46], Kopitar et al., 2024 [47], Dey, 2023 [51], Sng et al., 2023 [53]
Improvements in HbA1C and BG levels	Shaikh et al., 2024 [50], Lee et al., 2024 [41], Hernandez et al., 2023 [38], Sun et al., 2023 [39], Mashatian et al., 2024 [40], Gokbulut et al., 2024 [46], Nayak et al., 2023 [49], Dey, 2023 [51]
Alerts for necessary actions	Peleg et al., 2022, [48], Hernandez et al., 2023 [38], Shaikh et al., 2024 [50], Posa et al., 2024 [42], Shiraishi et al., 2024 [44], Chung and Chang, 2024 [45], Gokbulut et al., 2024 [46], Nayak et al., 2023 [49], Dey, 2023 [51], Sng et al., 2023 [53]

[1] Assessing dietary habits, physical activity, BG levels, and medication adherence based on individual characteristics, climate conditions, and variations in caloric intake.
[2] Predicting future BG levels based on historical continuous glucose monitoring (CGM) measurements, meal intake, and insulin administration.

Table 5 presents a detailed summary of various DM self-management actions and the corresponding studies that contribute to these domains. Key areas frequently explored include personalized recommendations (e.g., nutrition, exercise, BG management) and DM education, underscoring the extensive application of GenAI-powered systems to improve daily self-management practices. Furthermore, key aspects that directly impact clinical management—such as early detection and alert systems for complications, insulin and medication management, continuous glucose monitoring, and BG prediction—have been extensively studied. These findings suggest that GenAI-based systems provide more than simple information and guidance, offering an advanced support framework incorporating in-depth data analysis and patient monitoring. Moreover, the increasing focus on communication with healthcare professionals, decision support, risk prediction, and improvements in HbA1C and BG levels demonstrates the evolving role of AI-driven solutions in DM management. Such systems are anticipated to enhance both patient adherence to treatment and the optimization of clinical decision-making by healthcare professionals. In conclusion, the table highlights the diverse applications of GenAI in DM management, offering valuable insights into how these technologies can support both patients' self-management and healthcare professionals' monitoring strategies. The expansion of GenAI's role in DM self-management, alongside its integration into clinical practice, represents a promising area for future research.

According to the analysis of the studies included in this review, GenAI applications demonstrate considerable potential advantages in DM self-management. However, the studies also highlight several limitations of these applications, as presented in Table 6.

Table 6: Potential Benefits and Limitations of GenAI in DM Self-Management

Advantages	Disadvantages
Personalized guidance	Collection and processing of patients' personal health data
Real-time support	Concerns regarding data security and privacy
Integration of evidence-based best practices	Resistance to the adoption of new Technologies
High accuracy in predicting future BG levels	Sometimes insufficient in detecting a small number of hypoglycemia events
Providing reliable medical information for self-education and self-management of DM	Lack of user trust in the system
Proven improvements in DM self-management	Requirement for comprehensive training and support to use the system effectively
High accuracy of outputs as a result of being trained with accurate books and data sources	Challenges due to insufficient technological infrastructure, especially in low-resource healthcare settings
Ability to offer personalized solutions to complex problems	Possibility of providing misinformation

The studies highlight that GenAI plays a significant role in promoting active patient involvement in self-management processes. Key functionalities such as offering personalized guidance, delivering real-time support, and integrating evidence-based clinical guidelines are central to this enhanced participation. Furthermore, the ability of GenAI to predict future BG levels accurately and provide access to reliable medical information has made DM management more proactive and informed. The high accuracy of GenAI model outputs, ensured by training on high-quality and comprehensive datasets, further strengthens its effectiveness. Additionally, the robust capability of these systems to generate personalized solutions for complex challenges greatly enhances the overall effectiveness of DM self-management. Collectively, these factors contribute to scientifically validated improvements in DM self-care practices.

Foremost among the concerns are data security and privacy, especially with regard to the collection and processing of patients' personal health information. Additionally, challenges such as resistance from healthcare professionals and patients to adopting new technologies, low user trust, and the need for extensive training and technical support complicate the implementation of GenAI solutions. Furthermore, certain models have been found inadequate in detecting critical conditions, such as hypoglycemia, while limited technological infrastructure in certain regions restricts the practical application of GenAI. The potential for GenAI to generate misinformation is also recognized as a significant risk factor. Consequently, while GenAI applications present substantial contributions to DM self-management, critical areas still require development to ensure these systems are more reliable, user-friendly, and ethically compliant (Table 6).

4. Discussion

This review has examined examples of GenAI applications that have had the potential to facilitate and enhance DM management. The analyzed studies highlight the use of GenAI in DM care across various areas, including BG control strategies, detection of hypoglycemia and hyperglycemia, insulin bolus calculators, decision support systems, risk assessment, patient personalization, meal and exercise tracking, error detection, and lifestyle support [48, 49, 50, 52, 53]. DL and GenAI are advancing towards ushering in a new era in DM management. By providing personalized recommendations, simplifying monitoring and follow-up processes, and assisting in error prevention, GenAI significantly contributes to DM self-management [54, 55].

In a randomized controlled trial that has been conducted by Shaikh et al. [50], the effectiveness of an AI-powered metabolic coach designed to provide personalized recommendations has been evaluated over a 12-week period. The study has assessed the impact of the metabolic coach on various glycemic parameters, including HbA1c levels, plasma glucose, glycemic variability (which has been measured using the glucose management indicator score), and predicted postprandial glucose levels. The trial has included 100 individuals aged 18–65 years who have been diagnosed with T2DM and have been willing to utilize digital technology for health monitoring. In this study, in particular, the observed improvement in postprandial glucose regulation has demonstrated that the AI-driven metabolic coach has effectively guided individuals in managing postprandial glucose fluctuations and maintaining stable glycemic control throughout the day. Therefore, the study findings have indicated substantial improvements in both short-term and long-term glycemic control. As a result, participants with T2DM in the intervention group have demonstrated superior outcomes compared to the control group, including significant reductions in HbA1c levels, lower plasma glucose concentrations, and notable decreases in postprandial glucose levels. This AI-driven metabolic coach, which has employed a holistic approach, presents a comprehensive strategy for addressing multiple aspects of metabolic health in DM management. This study has underscored the potential of AI-driven interventions to provide an integrative approach to DM management. The utilization of AI has had the potential to drive significant advancements in personalized care by targeting multiple facets of metabolic health. Consequently, the positive outcomes that have been observed in this study highlight the transformative potential of AI in enhancing the metabolic health of individuals with diabetes through the delivery of personalized healthcare solutions. These findings have provided strong evidence supporting the integration of AI technologies into DM management strategies.

Similarly, in a study conducted by Zhu et al. [36], a novel DL model has been developed utilizing a modified Generative Adversarial Network (GAN) architecture to predict future BG levels in individuals with T1DM, based on historical continuous glucose monitoring (CGM) measurements, meal intake, and insulin delivery. To train the model, BG-related data have been collected over eight weeks from 12 individuals with T1DM. The dataset included BG levels recorded every five minutes via CGM, insulin delivery data from insulin pumps, self-reported events (such as meals, work, sleep, psychological stress, and physical exercise) through a smartphone application, and physical activity data captured by a sensor band. The developed model has been found to provide appropriate treatment recommendations regardless of prediction error, demonstrating high clinical accuracy. For individuals with T1DM, maintaining BG within the target range is essential to prevent periods of hypoglycemia and hyperglycemia, which can lead to severe complications. Accurate BG prediction can reduce this risk and facilitate early interventions to improve DM management. However, DM management remains challenging due to the complex nature of glucose metabolism and the wide range of lifestyle factors that can affect it. For this reason, the DL model developed by Zhu et al. [36], which incorporates personalized data to predict future BG levels, holds promise as an innovative tool for advancing DM management.

In a study conducted by Dudukcu et al. [35], a fusion model has been developed using the extended OhioT1DM dataset, which includes historical BG data from 12 individuals with diabetes. The model has combined Long Short-Term Memory (LSTM), WaveNet, and Gated Recurrent Units (GRU) architectures, incorporating decision-level fusion of these models. The study has demonstrated that the fusion model incorporating 'LSTM + WaveNet + GRU' architecture has achieved superior performance in BG prediction. Dudukcu et al. [35] plan to utilize the prediction values generated by the fusion model in calculating the required insulin doses. If these efforts prove successful, the developed system is planned to be converted into a mobile application. This would provide individuals with diabetes access to more accurate BG prediction and insulin dosage guidance through a GenAI-powered DM management tool, offering the added convenience of mobile use. In another study by Peleg et al. [48], the Mobiguide application has been tested with 10 atrial fibrillation patients in Italy and 20 gestational individuals with diabetes in Spain. Additionally, Mashatian et al. [40] have developed an AI-based question-answering model using a Retrieval-Augmented Generation (RAG) architecture to address inquiries related to DM and diabetic foot care. In this study, Pinecone has been utilized as a vector database alongside GPT-4, developed by OpenAI. The NIH National Standards for Diabetes Self-Management Education have served as the foundation for training the model. A total of 58 keywords have been used to select 295 articles, and the model has been tested with 175 questions covering various topics. According to the results, the RAG model is considered a promising tool for delivering reliable medical information to the public for self-education and self-management in the field of DM.

In a study that has been conducted by Lee et al. [41], a chatbot using GenAI has been developed to provide dietary recommendations for individuals with diabetes. The chatbot has been trained using an additional dataset that has been produced expressly for the system, and it is based on OpenAI's ChatGPT model. This approach allows patients to access personalized diet plans that have been tailored to their physical needs, including options that have considered seasonal changes. When compared with existing applications, the system has demonstrated superior capabilities in managing the health status of elderly individuals by incorporating additional data sources and offering a wider range of services. As a result, through this system, a new chatbot has been made accessible, which has focused on dietary guidance, has assumed the role of an expert, and has performed precise caloric calculations to help individuals with diabetes manage their health effectively.

In another study, Wang et al. [37] have developed a generative Markov-Bayes-based model, which has been based on previous GenAI models. In this study, longitudinal electronic health records from 9,298 individuals with T2DM or prediabetes, which have been collected from a large regional healthcare delivery network in China between 2005 and 2016, have been utilized to generate 5,000 synthetic disease trajectories. The findings have shown that 55.3% of individual complications and 31.8% of complication patterns associated with progressive T2DM can be predicted early and appropriately managed, potentially delaying or preventing them through lifestyle modifications that reduce the risk of DM development or progression.

However, there are several limitations associated with the use of GenAI. These include the production of inaccurate or fabricated content, reliance on unreliable information sources, and the provision of incorrect answers to user queries. In addition, various practical challenges exist within clinical applications [56, 57]. These findings indicate that GenAI stands out for its role in facilitating DM self-management and delivering information on DM care, while highlighting the need for further improvement in critical areas such as emotional support and medication adherence. As GenAI continues to adapt and develop in response to the unique settings and requirements of the medical field, it is expected to play a more significant role in DM care [50]. Based on evidence-based models that examine perceived usefulness and ease of use—key factors influencing the adoption of new technologies—these elements appear to be critical for the successful integration of GenAI into clinical practice [58].

5. Conclusion

The findings of this systematic review demonstrate that GenAI technologies have gained increasing significance in DM self-management, offering powerful tools that support patients in managing their own care processes. The majority of the reviewed studies indicate that GenAI applications are being effectively utilized in various areas, such as providing personalized recommendations, delivering DM education, early detection of complications, and offering emotional support. GenAI-powered solutions have been shown to make significant contributions in critical aspects of DM management, including BG prediction, dietary management, exercise recommendations, insulin dose optimization, and the generation of patient education materials. However, the limitations highlighted in the literature are also noteworthy. Issues such as data security and privacy, ethical concerns, the risk of misinformation, challenges in predicting critical situations like hypoglycemia, and limited access to technology in low-resource healthcare settings have been identified as areas requiring further development for GenAI applications. Additionally, hesitations regarding technology adoption among healthcare professionals and patients, as well as the need for technical support and comprehensive training to ensure effective use of these systems, pose challenges to their implementation.

Overall, the results of this review highlight the innovative solutions and potential benefits offered by GenAI in DM self-management, while also emphasizing the need for improvement to make these technologies more reliable, transparent, user-friendly, and ethically sound. In this context, there is a need for the development of larger datasets, increased interpretability of GenAI models, and more comprehensive evaluations of these systems. Future research is recommended to focus on developing strategies that enhance data security, accuracy, interpretability, and user satisfaction to strengthen the integration of GenAI applications into clinical practice.

Acknowledgement

This paper is a comprehensively expanded version of the proceedings abstract presented at the IV. International Congress on Artificial Intelligence in Healthcare.

References

[1] The Society of Endocrinology and Metabolism of Turkey. (2024). *Guidelines for the diagnosis, treatment, and monitoring of diabetes mellitus and its complications-2024.* https://file.temd.org.tr/Uploads/publications/guides/documents/diabetesmellitus2024.pdf

[2] Republic of Türkiye Ministry of Health. (2023). *Türkiye diabetes program 2023 – 2027.* https://hsgm.saglik.gov.tr/depo/birimler/saglikli-beslenme-ve-hareketli-hayat-db/Dokumanlar/Programlar/Turkiye-Diyabet-Programi.pdf

[3] Magliano, D. J., Boyko, E. J. IDF Diabetes Atlas 10th edition scientific committee. (2021). *Idf diabetes atlas*. (10th ed.). International Diabetes Federation. https://pubmed.ncbi.nlm.nih.gov/35914061/

[4] World Health Organization. (2003). *Adherence to long-term therapies: evidence for action*. https://iris.who.int/handle/10665/42682

[5] American Diabetes Association. (2023). *Standards of care in diabetes-2023.* https://www.portailvasculaire.fr/sites/default/files/docs/2023_ada_diabete_standards_of_care_in_diabetes_diab_care.pdf

[6] Harding, J. L., Pavkov, M. E., Magliano, D. J., Shaw, J. E., & Gregg, E. W. (2019). Global trends in diabetes complications: a review of current evidence. *Diabetologia*, *62*, 3-16. https://doi.org/10.1007/s00125-018-4711-2

[7] American Diabetes Association. (2024). 2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes-2024. *Diabetes Care, 47*(1), 20-42. https://doi.org/10.2337/dc24-S002

[8] American Association of Diabetes Educators. (2009). AADE Guidelines for the Practice of Diabetes Self-Management Education and Training (DSME/T). *The Diabetes Educator, 35*(3), 85-107. https://doi.org/10.1177/0145721709352436

[9] Williams, D. M., Jones, H., & Stephens, J. W. (2022). Personalized type 2 diabetes management: an update on recent advances and recommendations. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, 281-295. https://doi.org/10.2147/DMSO.S331654

[10] Khor, X. Y., Pappachan, J. M., & Jeeyavudeen, M. S. (2023). Individualized diabetes care: Lessons from the real-world experience. *World Journal of Clinical Cases*, *11*(13), 2890. https://doi.org/10.12998/wjcc.v11.i13.2890

[11] Song, M., Choe, M. A., Kim, K. S., Yi, M. S., Lee, I., Kim, J., Lee, M., Cho, M.Y., & Shim, Y. S. (2009). An evaluation of Web-based education as an alternative to group lectures for diabetes self-management. *Nursing & health sciences*, *11*(3), 277-284. https://doi.org/10.1111/j.1442-2018.2009.00458.x

[12] Hunt, C. W. (2015). Technology and diabetes self-management: An integrative review. *World journal of diabetes*, *6*(2), 225–233. https://doi.org/10.4239/wjd.v6.i2.225

[13] Sheng, B., Pushpanathan, K., Guan, Z., Lim, Q. H., Lim, Z. W., Yew, S. M. E., Goh, J. H. L., Bee, Y. M., Sabanayagam, C., Sevdalis, N., Lim, C. C., Lim, C. T., Shaw, J., Jia, W., Ekinci, E. I., Simó, R., Lim, L. L., Li, H., & Tham, Y. C. (2024). Artificial intelligence for diabetes care: current and future prospects. *The Lancet. Diabetes & Endocrinology*, *12*(8), 569–595. https://doi.org/10.1016/S2213-8587(24)00154-2

[14] Goel, A.K., & Davies, J. (2020). Artificial Intelligence. In R. J. Sternberg (Ed.), *The Cambridge Handbook of Intelligence* (602-625). Cambridge: Cambridge University Press.

[15] Cahn, A., Akirov, A., & Raz, I. (2018). Digital health technology and diabetes management. *Journal of diabetes*, *10*(1), 10–17. https://doi.org/10.1111/1753-0407.12606

[16] Aggarwal, M., & Murty, M. N. (2021). Deep learning. *Machine Learning in Social Networks: Embedding Nodes, Edges, Communities, and Graphs* (35-66). Springer Singapore. https://doi.org/10.1007/978-981-33-4022-0_3

[17] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, *2*(4). https://doi.org/10.1136/svn-2017-000101

[18] García-Peñalvo, F. J., & Vázquez-Ingelmo, A. (2023). What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in Generative AI. *International Journal of Interactive Multimedia and Artificial Intelligence, 8*(4), 7-15. https://doi.org/10.9781/ijimai.2023.07.006

[19] Liu, S., Chen, J., Feng, Y., Xie, Z., Pan, T., & Xie, J. (2024). Generative artificial intelligence and data augmentation for prognostic and health management: Taxonomy, progress, and prospects. *Expert Systems with Applications,* 255, 124511. https://doi.org/10.1016/j.eswa.2024.124511

[20] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems,* 27.

[21] Hernandez-Matamoros, A., Fujita, H., & Perez-Meana, H. (2020). A novel approach to create synthetic biomedical signals using BiRNN. *Information Sciences, 541*, 218-241. https://doi.org/10.1016/j.ins.2020.06.019

[22] Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q. L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122-1136. https://doi.org/10.1109/JAS.2023.123618

[23] Balasubramaniam, S., Chirchi, V., Kadry, S., Agoramoorthy, M., Gururama, S. P., Satheesh, K. K., & Sivakumar, T. A. (2024). The road ahead: Emerging trends, unresolved issues, and concluding remarks in generative AI—A comprehensive review. *International Journal of Intelligent Systems*, *2024*. https://doi.org/10.1155/2024/4013195

[24] Howell, M. D. (2024). Generative artificial intelligence, patient safety and healthcare quality: a review. *BMJ Quality & Safety*, *33*(11), 748-754. https://doi.org/10.1136/bmjqs-2023-016690

[25] Kline, A., Wang, H., Li, Y., Dennis, S., Hutch, M., Xu, Z., Wang, F., Cheng, F., & Luo, Y. (2022). Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, *5*(1), 171. https://doi.org/10.1038/s41746-022-00712-8

[26] Yang, K., Ji, S., Zhang, T., Xie, Q., & Ananiadou, S. (2023). On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. arXiv preprint

[27] Reddy, S. (2024). Generative AI in healthcare: An implementation science informed translational path on application, integration and governance. *Implementation Science*, *19*(1), 27. https://doi.org/10.1186/s13012-024-01357-9

[28] Gabrielson, A. T., Odisho, A. Y., & Canes, D. (2023). Harnessing generative artificial intelligence to improve efficiency among urologists: welcome ChatGPT. *Journal of Urology*, *209*(5), 827-829.

[29] Hiba, A., Al-Qudheeby, M., Kheyami, Z. A., Khalil, R., Khamees, N., Hijjawi, O., Sallam, M., & Barakat, M. (2024). Cross-linguistic evaluation of generative AI models for diabetes and endocrine queries. *Jordan Medical Journal*, *58*(4). https://doi.org/10.35516/jmj.v58i4.3369

[30] Ayre, J., Mac, O., McCaffery, K., McKay, B. R., Liu, M., Shi, Y., Rezvan, A., & Dunn, A. G. (2024). New frontiers in health literacy: using ChatGPT to simplify health information for people in the community. *Journal of General Internal Medicine*, *39*(4), 573-577. https://doi.org/10.1007/s11606-023-08469-w

[31] Chen, A., Chen, D. O., & Tian, L. (2024). Benchmarking the symptom-checking capabilities of ChatGPT for a broad range of diseases. *Journal of the American Medical Informatics Association*, *31*(9), 2084-2088. https://doi.org/10.1093/jamia/ocad245

[32] Huang, R. S. T., Lu, K. J. Q., Meaney, C., Kemppainen, J., Punnett, A., & Leung, F. H. (2023). Assessment of resident and AI chatbot performance on the University of Toronto family medicine residency progress test: comparative study. *JMIR Medical Education*, *9*, e50514. https://doi.org/10.2196/50514

[33] Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, *183*(6), 589-596. https://doi.org/10.1001/jamainternmed.2023.1838

[34] Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & the PRISMA Group. (2009). Reprint—preferred reporting items for systematic reviews and metaanalyses: The PRISMA Statement. *Physical Therapy, 89*(9), 873-880.

[35] Dudukcu, H. V., Taskiran, M., & Yildirim, T. (2021). Blood glucose prediction with deep neural networks using weighted decision level fusion. *Biocybernetics and Biomedical Engineering*, *41*(3), 1208-1223. https://doi.org/10.1016/j.bbe.2021.08.007

[36] Zhu, T., Yao, X., Li, K., Herrero, P., & Georgiou, P. (2020). Blood glucose prediction for type 1 diabetes using generative adversarial networks. In *CEUR workshop proceedings* (Vol. 2675, pp. 90-94).

[37] Wang, X., Lin, Y., Xiong, Y., Zhang, S., He, Y., He, Y., Plasek, J. M., Zhou, L., Bates, D. W., & Tang, C. (2022). Using an optimized generative model to infer the progression of complications in type 2 diabetes patients. *BMC Medical Informatics and Decision Making*, *22*(1), 174.

[38] Hernandez, C. A., Gonzalez, A. E. V., Polianovskaia, A., Sanchez, R. A., Arce, V. M., Mustafa, A., Vypritskaya, E., Gutierrez, O. P., Bashir, M., & Sedeh, A. E. (2023). The future of patient education: AI-driven guide for type 2 diabetes. *Cureus*, *15*(11).

[39] Sun, H., Zhang, K., Lan, W., Gu, Q., Jiang, G., Yang, X., Qin, W., & Han, D. (2023). An AI dietitian for type 2 diabetes mellitus management based on large language and image recognition models: preclinical concept validation study. *Journal of medical Internet research*, *25*, e51300.

[40] Mashatian, S., Armstrong, D. G., Ritter, A., Robbins, J., Aziz, S., Alenabi, I., Huo, M., Anand, T., & Tavakolian, K. (2024). Building trustworthy generative artificial intelligence for diabetes care and limb preservation: a medical knowledge extraction case. *Journal of Diabetes Science and Technology*, 19322968241253568. https://doi.org/10.1177/19322968241253568

[41] Lee, H. E., Jun Woo, C. H. O. I., & Kang, M. S. (2024). A Study on the development of a chatbot using generative AI to provide diets for diabetic patients. *Korean Journal of Artificial Intelligence*, *12*(3), 25-31. https://doi.org/10.24225/kjai.2024.12.3.25

[42] Posa, K. N., Peralta, G. E. Q., Hidalgo, I. F., Prat-Oriol, B., & Abreu, R. (2024). Comparative evaluation of generative artificial intelligence systems for patient queries on age-related macular degeneration and diabetic macular edema. *Investigative Ophthalmology & Visual Science*, *65*(7), 2326-2326.

[43] Naja, F., Taktouk, M., Matbouli, D., Khaleel, S., Maher, A., Uzun, B., Alameddine, M., & Nasreddine, L. (2024). Artificial intelligence chatbots for the nutrition management of diabetes and the metabolic syndrome. *European Journal of Clinical Nutrition*, *78*(10), 887-896. https://doi.org/10.1038/s41430-024-01476-y

[44] Shiraishi, M., Lee, H., Kanayama, K., Moriwaki, Y., & Okazaki, M. (2024). Appropriateness of artificial intelligence chatbots in diabetic foot ulcer management. *The international journal of lower extremity wounds*, 15347346241236811. https://doi.org/10.1177/15347346241236811

[45] Chung, S. M., & Chang, M. C. (2024). Assessment of the information provided by ChatGPT regarding exercise for patients with type 2 diabetes: a pilot study. *BMJ Health & Care Informatics*, *31*(1), e101006. https://doi.org/10.1136/bmjhci-2023-101006

[46] Gokbulut, P., Kuskonmaz, S. M., Onder, C. E., Taskaldiran, I., & Koc, G. (2024). Evaluation of ChatGPT-4 performance in answering patients' questions about the management of type 2 diabetes. *Medical Bulletin of Sisli Etfal Hospital*, *58*(4). https://doi.org/10.14744/SEMB.2024.23697

[47] Kopitar, L., Fister Jr, I., & Stiglic, G. (2024). Using generative AI to improve the performance and interpretability of rule-based diagnosis of type 2 diabetes mellitus. *Information*, *15*(3), 162. https://doi.org/10.3390/info15030162

[48] Peleg, M., Shahar, Y., & Quaglini, S. (2022). MobiGuide: guiding clinicians and chronic patients anytime, anywhere. Communications of the ACM, 65(4), 74-79. https://doi.org/10.1145/3511596

[49] Nayak, A., Vakili, S., Nayak, K., Nikolov, M., Chiu, M., Sosseinheimer, P., Talamentes, S., Testa, S., Palanisamy, S., Giri, V., Schulman, K., & Schulman, K. (2023). Use of voice-based conversational artificial intelligence for basal insulin prescription management among patients with type 2 diabetes: a randomized clinical trial. *JAMA network open*, *6*(12), e2340232-e2340232. doi:10.1001/jamanetworkopen.2023.40232

[50] Shaikh, A., Baluni, S., Malpani, N., Lodha, P., & Meena, A. (2024). Decoding type 2 diabetes through point-of-care testing, cloud-based monitoring, and generative augmented retrieval model-driven virtual diabetes education: A comprehensive approach to glycemic control. *International Journal of Diabetes and Technology*, *3*(1), 25-31. https://doi.org/10.4103/ijdt.ijdt_5_24

[51] Dey, A. K. (2023). ChatGPT in diabetes care: An overview of the evolution and potential of generative artificial intelligence model like ChatGPT in augmenting clinical and patient outcomes in the management of diabetes. *International Journal of Diabetes and Technology*, *2*(2), 66-72. https://doi.org/10.4103/ijdt.ijdt_31_23

[52] Khan, S. (2023). The Future of Diabetes Care: Navigating with generative language models. *Indus Journal of Medical and Health Sciences*, *1*(1), 109-116. https://induspublishers.com/IJMHS/article/view/58

[53] Sng, G. G. R., Tung, J. Y. M., Lim, D. Y. Z., & Bee, Y. M. (2023). Potential and pitfalls of ChatGPT and natural-language artificial intelligence models for diabetes education. *Diabetes Care*, *46*(5), 103-105. https://doi.org/10.2337/dc23-0197

[54] Thyde, D. N., Mohebbi, A., Bengtsson, H., Jensen, M. L., & Mørup, M. (2021). Machine learning-based adherence detection of type 2 diabetes patients on once-daily basal insulin injections. Journal of diabetes science and technology, 15(1), 98-108. https://doi.org/10.1177/1932296820912411

[55] Cordeiro, R., Karimian, N., & Park, Y. (2021). Hyperglycemia identification using ECG in deep learning era. Sensors, 21(18), 6263. https://doi.org/10.3390/s21186263

[56] Megahed, F. M., Chen, Y. J., Ferris, J. A., Knoth, S., & Jones-Farmer, L. A. (2024). How generative AI models such as ChatGPT can be (mis) used in SPC practice, education, and research? An exploratory study. Quality Engineering, 36(2), 287-315. https://doi.org/10.1080/08982112.2023.2206479

[57] Kanitz, R., Gonzalez, K., Briker, R., & Straatmann, T. (2023). Augmenting organizational change and strategy activities: Leveraging generative artificial intelligence. *The Journal of Applied Behavioral Science, 59*(3), 345-363. https://doi.org/10.1177/00218863231168974

[58] Stamler, J., Vaccaro, O., Neaton, J. D., Wentworth, D. (1993). Diabetes, other risk factors, and 12-yr cardiovascular mortality for men screened in the multiple risk factor intervention trial. *Diabetes care*, *16*(2), 434-444. https://doi.org/10.2337/diacare.16.2.434