ISSN: 2093-5374 / E-ISSN: 2093-5382

A Hybrid CNN–Vision Transformer and Explainable AI Framework for Real-Time Retinal Disease Diagnosis in IoT-Enabled and Ubiquitous Healthcare Systems

Raghad Saleem Mohamed Najeeb[1*], Shatha Abdullah Mohammed[2], and Mohammed F Ibrahim Alsarraj[3]

[1*]Department of Software, College of Computer Science and Mathematics, University of Mosul, Ninevah, Iraq. raghad.saleem@uomosul.edu.iq, https://orcid.org/0009-0007-1434-2342

[2]Department of Software, College of Computer Science and Mathematics, University of Mosul, Ninevah, Iraq. shathaabdullah@uomosul.edu.iq, https://orcid.org/0000-0002-3098-0519

[3]Technical Engineering College for Computer and AI–Mosul, Northern Technical University, Mosul, Iraq. mohammed_alsarraj@ntu.edu.iq, https://orcid.org/0000-0001-7886-4464

Received: October 03, 2025; Revised: November 28, 2025; Accepted: December 31, 2025; Published: March 31, 2026

Abstract

Eye diseases, such as diabetic retinopathy, are a major cause of preventable visual loss, and it is important to note that a high level of accuracy and efficiency is required in automated solutions to screen for this disease. This study proposes a Modified Artificial Intelligence (MAI) approach that blends a hybrid convolutional neural network vision transformer architecture with stacked autoencoder-based feature enhancement and Grad-CAM-based explainability. Performance evaluation was performed on publicly accessible retinal fundus datasets. An empirical assessment was conducted by comparing the state-of-the-art performance of conventional CNN, ResNet, and Efficient Net. Empirical evidence shows that the proposed MAI does reach high levels of performance with 98.2% accuracy and an area under the receiver operating characteristic curve of 0.992, all at a low inference latency that can be matched to real-time deployment. Clinical interpretability is further enhanced by the fact that explainable visual cues are added. All these findings display that the MAI framework is a valid and efficient solution to automated diagnosis of retinal diseases in a mobile and IoT-based health-care environment. However, existing automated diabetic retinopathy screening models suffer from limited cross-dataset generalization, a lack of clinical interpretability, and high computational complexity, which restrict their deployment in real-time, mobile, and IoT-enabled healthcare settings.

Keywords: Eye Disease Diagnosis, Artificial Intelligence, Deep Learning, Machine Learning, Retinal Fundus Images, IoT Healthcare, Real-Time Diagnosis.

1 Introduction

Eye diseases, such as diabetic retinopathy (DR), glaucoma, cataracts, and age-related macular degeneration (AMD), are some of the leading causes of vision impairment and blindness globally. As of current developments, the World Health Organization (WHO) notes that 2.2 billion people are

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 1 (March - 2026), pp. 373-391. DOI: *10.58346/JOWUA.2026.I1.021*

*Corresponding author: Department of Software, College of Computer Science and Mathematics, University of Mosul, Ninevah, Iraq.

visually impaired/blind globally, and out of that subset, up to 50% could have their condition corrected/improved with early intervention and clinical screening (Abd El-Khalek et al., 2024; Benbakreti et al., 2024; Bouzidi et al., 2021). From a clinical perspective, diagnosis occurs via specialized imaging through retinal fundus images and optical coherence tomography (OCT) imaging; however, such specialized environments are only in well-equipped places with personnel trained to make such assessments (Oshika, 2025; Sanjana et al., 2021). However, such limitations challenge mass screening potential in remote, rural, and underdeveloped regions where ophthalmology specialists/imaging machinery/advanced capabilities are non-existent (El-Den et al., 2024; Zhang et al., 2023). Furthermore, DL and AI have been successfully employed in medical image assessments for classifying DR (and vessel segmentation) and diagnosing glaucoma and AMD based on a convolutional neural network (CNN)-based architectural assessment (Grzybowski et al., 2024). However, despite advances in deep learning (DL), there are drawbacks to rendering it a practical solution for clinically adopted answers. First, most deep models do not generalize across populations; trained models based on varying imaging devices/settings are responsible for their creation. Second, deep models operate as less interpretable black-box systems, which decreases clinician confidence in adopting the findings. Third, many deep learning systems require extensive computational processes, which eliminate real-time possibilities for mobile, embedded, and IoT-enabled health systems (Jeong et al., 2022; Singh et al., 2024; Dodda et al., 2024). Therefore, the MAI system boasts contributions to the field, such as CNN-ViT hybrid architecture for local lesion detection and global spatial awareness perception; stacked autoencoders reduced feature redundancy with a computational benefit, and Grad-CAM offers an understandable heatmapped image of relevant features of the pathology for clinician agreement (Ennab & Mcheick, 2025; Subramani & Manoharan, 2024). These considerations make the MAI system applicable for in-clinic detection and travel-ready situations, such as IoT ophthalmic devices and tele-ophthalmology systems, where immediate results are required. The MAI system was trained and tested through three publicly available datasets relative to similar research efforts for precision and practicality for generalized understanding: EyePACS, Messidor, and DRIVE to evaluate generalized efficacy across imaging situations (Ennab & Mcheick, 2025; Mohammed et al., 2020). Finally, this study presents an MAI system, Modified Artificial Intelligence with CNNs and VIs for compatible feature detection, stacked autoencoders for computational advantage, and Grad-CAM for explainability. The main contributions of this study are summarized as follows:

1. A novel Modified Artificial Intelligence (MAI) framework that integrates CNN and Vision Transformer architectures to capture local retinal lesions and global spatial dependencies jointly.
2. The incorporation of stacked autoencoders to reduce feature redundancy and improve computational efficiency for real-time inference.
3. The integration of Grad-CAM–based explainable AI to enhance clinical interpretability and trustworthiness; and
4. Extensive validation across multiple public datasets (EyePACS, Messidor, and DRIVE), demonstrating superior accuracy, robustness, and real-time suitability for mobile and IoT-based ophthalmic screening.

The rest of the study is structured as follows: Section 2 will conduct a review of related literature concerning the topics of AI-based retinal disease diagnosis and will identify existing limitations. Section 3 explains the proposed MAI methodology, preprocessing, model architecture, and training processes. Section 4 includes the results of the experiments and a comparative analysis of the performance. The

findings and practical implications are discussed in Section 5. Lastly, Section 6 summarizes the study and provides the direction of the research in the future.

2 Literature Review

2.1 Artificial Intelligence in Medical Imaging

The early years of machine learning using features and classifying them with SVM, KNN, and RF are not generalizable and vulnerable to illumination conditions, device variations, and the quality of images obtained (Sakirin & Ben Said, 2025). Deep learning (DL), based on convolutional neural networks (CNN), on the contrary, is based on hierarchically determined features that are extracted from the data (Sharif et al., 2022). The CNN-based algorithms have been effective in the retinal vessel segmentation, the DR grading, as well as the glaucoma and AMD detection with an accuracy comparable to that of an ophthalmologist (Chakraborty & Tharini, 2020; Mienye et al., 2025; Ejaz et al., 2024).

2.2 Deep Learning for Diabetic Retinopathy and Retinal Disease Detection

Lots of AI applications are aimed at DR because of its epidemiology and available datasets, including EyePACS, which makes it available. Nevertheless, these models are huge and require a long period of training and prediction (Akhtar et al., 2024; Jabbar et al., 2024; Singh et al., 2024). As an illustration, domain shift is an issue; the inability of an AI trained on one dataset to make good predictions on a second one, because of differences in camera/light configurations. Moreover, CNNs pay more attention to local textures than geometry required to perform more complicated biopsies, and the non-explanatory characteristic of the black box is unsuitable in a clinical context in general (Dai et al., 2024; Hanif et al., 2021).

2.3 Hybrid and Modified Deep Learning Architectures

Hybrid DNNs are models that integrate CNNs with other types of architectural paradigms to improve retinal evaluations and diagnoses. These novel architectures unite the localization of the lesion-assessment capabilities of CNNs with the global information-acquisition capabilities of ViTs by self-attention mechanisms (Hadhoud et al., 2024). A local-global hybridization such as this will be more suitable in the treatment of retinal pathologies, especially when the imaging environment is complex and thus requires long-range assessments. In turn, hybrid models work better than single-modal structures in the diagnosis of retinal diseases.

Researchers are exploring feature selection strategies, which are also being conducted in the framework of unsupervised representation learning. SAEs can help reduce the number of dimensions and augment the effectiveness of training (Kim et al., 2025). This method allows the prediction of vessel segmentation in retinal images. Stacked autoencoders are used as intermediates between convolutional neural networks and ViT hybrids, which combine both local and global observations without imposing too much computational burden. (Kim et al., 2025).

2.4 Inference Efficiency and Deployment-Oriented Comparisons

Even though numerous CNN- and transformer-based models indicate high classification accuracy in diabetic retinopathy detection, little research points to inference latency and deployability. Transformer-based architectures are usually expensive to compute and have high memory costs, making

them less applicable in mobile or edge-based healthcare systems. Lightweight CNN models, on the other hand, do not have global contextual understanding and interpretability but are faster to infer. These restrictions inspire the necessity of a hybrid and optimized architecture, including the suggested MAI framework, which will compromise the diagnostic accuracy, explainability, and inference flexibility of the IoT-enabled ophthalmic screening (Takahashi et al., 2024; Wang et al., 2019).

2.5 AI in IoT-Based Ophthalmic Healthcare Systems

Recent literature also focuses on the emergence of integrated health care. The use of artificial intelligence in mobile screening vans, handheld cameras, smartphone platforms, and Internet of Things ocular devices will aid in the screening of diabetic retinopathy in rural or resource-limited settings (Parmar et al., 2024). Such IoT-powered retinal imaging devices can send fundus photographs to hospital information systems, thus assisting in remote diagnosis and longitudinal changes. However, AI models should be small and consume less power to perform inference in real time (Alam et al., 2019; Jouini et al., 2024).

2.6 Research Gaps

Despite substantial progress, several key challenges remain unresolved:

- Limited generalization across heterogeneous datasets and imaging devices
- Lack of explainability in deep learning-based diagnostic systems.
- Expensive to compute and latent for real-time deployment.
- Absence of mobile, wireless, and IoT-based healthcare system integration.
- There is a scarcity of emphasis on interpretable and efficient hybrid architectures.

To address these gaps in the research, the current study suggests an MAI model that incorporates CNN–ViT hybrid feature extraction, stacked autoencoder-based feature refinement, and Grad-Cam explainability into a computationally efficient community-specific architecture, that is, real-time, IoT-enabled, ophthalmic screening.

3 Methodology

3.1 Research Design and Framework

The proposed Modified Artificial Intelligence (MAI) framework achieves state-of-the-art results, computational efficiency, and interpretability for automatic eye disease detection in retinal fundus images with a three-step configuration of interconnected aims:

- Robust feature extraction through a hybrid CNN–Vision Transformer (ViT) backbone,
- Efficient dimensionality reduction via stacked autoencoders (SAEs), and
- Clinically interpretable heatmaps using explainable AI (XAI) methods.

Unlike classical multi-step networks driven only by classification performance, the MAI is intended for real-time deployment in wireless, mobile, and IoT-based ophthalmic healthcare systems such as portable fundus cameras, smartphone-based retinal cameras, mobile screening devices, and edge-based healthcare applications. The low inference latency facilitates eye disease triaging in tele-ophthalmology and ubiquitous health domains. A generalized workflow of the MAI framework is shown in figure 1.

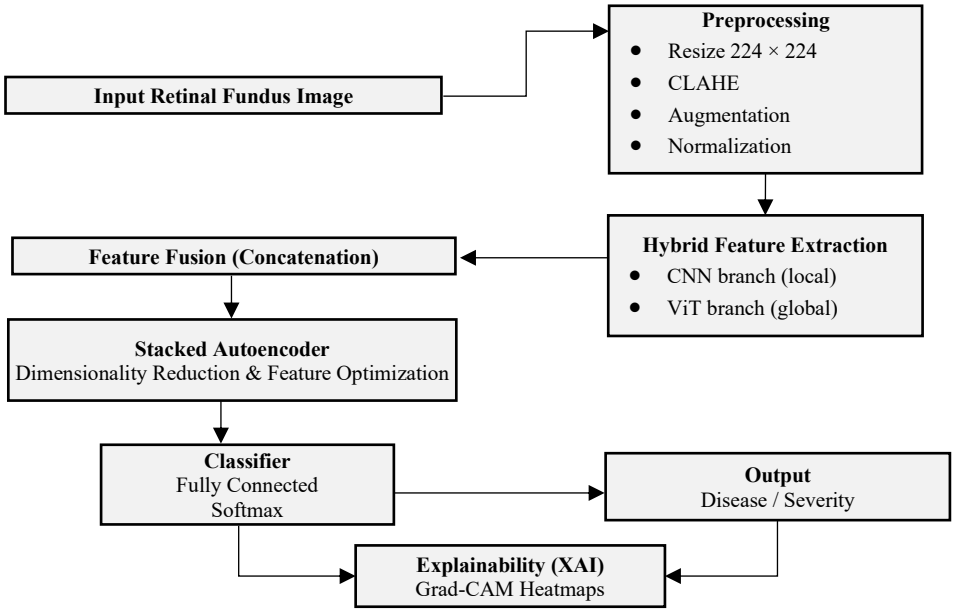

Figure 1: Overall workflow of the proposed MAI framework for automated eye disease diagnosis

3.2 Dataset Description

Table 1 illustrates three publicly available retinal fundus datasets that were used to evaluate the performance and generalizability of the proposed MAI model:

Table 1: Dataset summary

Dataset	Type of Images	No. of Images	Classes	Annotation	Resolution	Usage
EyePACS	Retinal Fundus	88,702	5 DR grades	Yes	296×296–640×640	Training & Validation
Messidor	Retinal Fundus	1,200	DR grades	Yes	1440×960	Testing
DRIVE	Retinal Fundus	40	Vessel segmentation	Yes	565×584	Preprocessing validation

EyePACS is the training and cross-validation dataset for the model; Messidor is the dataset for performance evaluation and cross-dataset generalizability evaluation; and DRIVE is the dataset for fine-tuning, preprocessing, and vessel-level structural enhancement.

3.3 Preprocessing

The retinal fundus images underwent illumination changes, blurriness, artifacts, low contrast, etc., to ensure uniformity and the detection of appropriate signals for diagnosis. The preprocessing steps were as follows:

3.3.1 Image Resizing

All images were resized to 224×224, maintaining the aspect ratio. The resizing operation used to standardize all retinal images is defined in equation (1):

$$I_{\text{resized}}(x, y) = I\left(\frac{x}{s_x}, \frac{y}{s_y}\right) \quad (1)$$

3.3.2 Contrast Enhancement (CLAHE)

Contrast Limited Adaptive Histogram Equalization was applied, and local contrast enhancement was performed using the CLAHE formulation shown in equation (2):

$$H'(i) = \frac{\min\left(H(i), T\right)}{\sum_{j=0}^{L-1} \min\left(H(j), T\right)} (L-1) \quad (2)$$

3.3.3 Data Augmentation

To reduce overfitting and enhance model robustness, several data augmentation techniques were applied, including rotations of ±20°, random horizontal and vertical flips, brightness and contrast adjustments, and random cropping with scaling variations, which ensured greater variability and generalization in the training data. The pixel intensities were normalized to [0,1].

3.4 Proposed MAI Framework

The MAI system comprises three main components: a hybrid CNN-ViT for feature extraction that detects and renders local and global retinal patterns; a stacked autoencoder (SAE) for feature improvement and encoding; and an explainable AI Grad-CAM-based module for visualization with an interpretable and clinically relevant justification.

3.4.1 Hybrid CNN–ViT Feature Extraction

The CNN module extracts local features, such as microaneurysms, hemorrhages, and exudates.

Convolution operation, CNN-based local feature extraction follows the convolution operation described in equation (3):

$$f_k(x,y) = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} w_{ij}^{k}\ I(x+i, y+j) \quad (3)$$

The Vision Transformer (ViT) captures global spatial dependencies through self-attention. Global dependencies in the ViT branch are captured using the self-attention mechanism defined in equation (4):

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

This was performed by combining CNN and ViT, as CNN and ViT enable the extraction of both local lesion features and long-range anatomical context, improving disease stage differentiation.

3.4.2 Stacked Autoencoder (SAE) for Feature Optimization

The intermediate features from the CNN–ViT hybrid often contain redundancy. A stacked autoencoder compresses these features into a compact latent representation. High-dimensional feature vectors are compressed into a latent representation using the encoder function in equation (5), and the reconstruction of the compressed features is performed through the decoder function in equation (6):

$$\text{Encoder:} \quad h = \sigma(Wx + b) \quad (5)$$

Decoder: $\hat{x} = \sigma(W'h + b')$ (6)

This minimizes computation, reduces overfitting, and speeds up the inference required for mobile and IoT-driven deployment.

3.4.3 Explainable AI Using Grad-CAM

To provide clinically meaningful interpretations, Model interpretability is achieved using the Grad-CAM heatmap computation described in equation (7).

$$L^{c}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k} \alpha_k^c A^k\right) \quad (7)$$

This enables the visualization of lesions, vessel distortions, and retinal irregularities by the ophthalmologists, increasing trust and usability.

In order to summarize the execution logic of the suggested MAI framework coherently, the entire training and inference process is represented formally in Algorithm 1. This algorithm supplements the architectural overview as figure 1, by explaining the sequential workings carried out throughout the processing of the preprocessing, the feature extraction, the optimization, the classification, and the explainable inference.

Algorithm 1. MAI Training and Inference Procedure

ALGORITHM: MAI_Retinal_Diagnosis

INPUT: Raw Retinal Fundus Image (I)

OUTPUT: Disease Classification (C), Explainability Heatmap (H)

BEGIN

// Step 1: Preprocessing

I_processed = Resize_and_Normalize (I, 224, 224)

I_enhanced = Apply_Contrast_Enhancement (I_processed)

// Step 2: Hybrid Feature Extraction

// Branch A: Local Spatial Features (CNN)

F_cnn = CNN_Extract(I_enhanced) // e.g., using specialized convolutional blocks

// Branch B: Global Contextual Features (ViT)

Patches = Divide_Into_Patches (I_enhanced)

F_vit = ViT_Self_Attention (Patches)

// Step 3: Feature Fusion and Compression

F_hybrid = Concatenate (F_cnn, F_vit)

// Pass through Stacked Autoencoder (SAE) for dimensionality reduction

F_compressed = SAE_Encoder (F_hybrid)

// Step 4: Classification

C = Softmax_Classifier (F_compressed)

// Step 5: Explainability Generation

// Compute gradients of the class score with respect to feature maps

H = Generate_GradCAM (I_enhanced, F_cnn)

RETURN C, H

END

It is referred to as the Modified Artificial Intelligence (MAI) algorithm, and it is sequential, meaning that it should be used in edge-computing applications. It begins with the processing of the retinal fundus images in order to match lighting and contrast. Central processing involves a 2-pathway paradigm: CNN branch of micro-spatial features and Vision Transformer (ViT) branch of worldly context.

Once the features have been extracted, are concatenated and input to a Stacked Autoencoder (SAE). This stage is a requisite step to the implementation of IoT since it will reduce the size of the feature vector, which will render the ultimate classification computationally inexpensive. Finally, the system applies the Grad-CAM to the last convolutional layer to infer the decision-making process on the original image to be medically validated.

3.5 Training Procedure

The dataset was split into 70% training, 15% validation, and 15% testing. Adam optimizer was used in making model adjustments with a learning rate of 0.0001, batch size of 32, and 100 epochs (early stopping to curb overfitting). A weighted cross-entropy loss function was used to train the model to consider the imbalance in the classes and provide equal learning to all classes of diseases. In order to overcome the problem of the disproportionality of classes, the cross-entropy loss with a weighted loss, as shown in equation (8), is employed during the training process. Table 2 summarizes the entire set of training hyperparameters and optimization settings used in this study.

$$L = \sum_{i=1}^{N} w_i \; y_i \log(\hat{y}_i) \tag{8}$$

Table 2: Hyperparameter configuration

Parameter	Value
Optimizer	Adam
Learning Rate	0.0001
Batch Size	32
Epochs	100
Loss Function	Weighted Cross-Entropy
Dropout	0.3

3.6 Evaluation Metrics

To quantitatively evaluate the performance of the proposed MAI, four common classification metrics were used: accuracy, precision, recall, and F1-score. These common metrics are used in medical image processing and screening relative to the overall performance and disease-specific performance. The mathematical definitions are provided in equations (9), (10), (11), 12) as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (11)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

TP, TN, FP, and FN are true positives, true negatives, false positives, and false negatives, respectively. Accuracy is the degree to which the model is generally correct, precision is the degree to which the model is correct when positive cases are being predicted, recall is the degree to which the model can identify disease-positive cases, and the F1-score is a balanced measure between precision and recall.

3.7 Baseline Models for Comparison

The baseline models against which the MAI was assessed were CNN, ResNet-50, EfficientNet-B0, and ViT in isolation, as each model has certain feature extraction capacities and processing needs. Therefore, selected to evaluate the best comparable local, deep, scalable, and global feature learning abilities of the MAI. The findings indicate that the efficiency and diagnostic accuracy are greatly improved with MAI. To demonstrate its superiority, the MAI was compared against, as shown in table 3.

Table 3: Baseline models used for performance comparison

Model	Parameters	Feature Type	Limitations
CNN	5M	Local only	Overfits small data
ResNet-50	25M	Local + deep	High computational cost
EfficientNet-B0	7M	Scalable	Needs large data
ViT (Standalone)	85M	Global only	Heavy computation
Proposed MAI	30M	Local + Global optimized	Balanced & efficient

3.8 Integration with Wireless, Mobile, and IoT-Based Healthcare Systems

3.8.1 Mobile Screening Units

The MAI framework consists of a lightweight framework that enables real-time scanning of the retina in mobile healthcare facilities like the ophthalmology bus, rural outreach, and field-based clinics. Its ability to make quick inferences and low computational requirements make it suitable for on-site diagnosis and enable quick triage and early diagnosis in underserved locations with little access to special ophthalmic care.

3.8.2 IoT-Enabled Ophthalmic Devices

The MAI framework could also be extended to IoT retinal devices (portable retinal cameras, mobile fundus imaging attachments, and Wi-Fi/4G/5G imaging) that capture pictures and transmit them to encrypted cloud systems for remote teleophthalmology evaluation when needed. However, the beauty of the MAI is that not only can risk evaluation occur immediately, but also symptomatology and level of need urgency are assessed simultaneously, all part of a triaging mechanism that provides time-sensitive recommendations and, if needed, external treatment.

3.8.3 Edge and Embedded Deployment

The optimized feature representation for MAI enables straightforward deployment on edge and embedded systems. NVIDIA Jetson Nano, Raspberry Pi with Coral TPU accelerators, and mobile ARM-based GPUs. These systems are lightweight enough to satisfy the low computational needs and high diagnostic efficacy of MAI, but the inference latency per image of MAI is only 8 ms, which is real-time when using these interconnected and low-resource healthcare systems, which are potentially integrated into mobile and portable IoT-based ophthalmic systems.

3.8.4 Software and Hardware Environment

In an example of a research project, the text explains the technical arrangement on which the experiments have been conducted. The experiments were performed using Python 3.9 and PyTorch 1.12 for deep learning, with GPU acceleration through CUDA 11.3. The libraries used for image processing and transformation were OpenCV, NumPy, and Albumentations. Scikit-learn was used to measure the model's performance.

The model was trained on a powerful computer with an NVIDIA RTX 3090 graphics card and the Ubuntu 22.04 operating system. To test whether the model could work quickly on small, portable devices (important for healthcare applications), ran the model on edge devices such as the NVIDIA Jetson Nano and Raspberry Pi 4 equipped with a Coral TPU accelerator.

Such a software and hardware combination will guarantee that it can consistently replicate the findings and that the model can be practically applied in portable and low-power clinical environments tied to the Internet of Things (IoT).

4 Results

Three publicly available databases (EyePACS, Messidor, and DRIVE) were used to verify the Modified Artificial Intelligence (MAI) architecture, and performance is contrasted with four known benchmark models: classical CNN, ResNet-50, EfficientNet-B0, and standalone ViT. Experimental control was done by training all the models using exactly the same training, preprocessing, and optimization protocols.

4.1 Quantitative Results

The MAI model achieved an accuracy of 98.2 %, the highest AUC of 0.992, and outperformed the strongest baseline (ViT) by 1.1% accuracy and 0.009 AUC. While these numbers may seem small, represent a significant difference in responsiveness to subtle retinal lesions and a statistical benefit compared to what is expected for early detection and screening.

Table 4: Performance comparison on eye disease classification

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
CNN	93.4	91.2	89.5	90.3	0.945
ResNet-50	95.7	94.1	92.8	93.4	0.967
EfficientNet-B0	96.8	95.5	94.2	94.8	0.976
ViT	97.1	96.2	94.9	95.5	0.983
Proposed MAI	98.2	97.4	96.8	97.1	0.992

In addition, the MAI possessed substantially more recall (96.8%), which means that the MAI is more likely to predict positive cases, which is critical for a medical screening tool, as not having a diagnosis could jeopardize patient safety. Table 4 shows the classification performance on the EyePACS dataset.

Figure 2 illustrates the performance metrics, namely accuracy, precision, recall, F1-score, and AUC, for the proposed MAI model and the three baseline architectures. As can be seen in the figure, all metrics gradually increase from classical CNN-based architectures to transformer-based architectures. It should be noted that the MAI not only excelled in all performance metrics but also acquired higher recall and AUC, which indicates that it was more sensitive to the disease-positive cases in the test population than the other three baselines and more adept at distinguishing precursors to retinal disease.

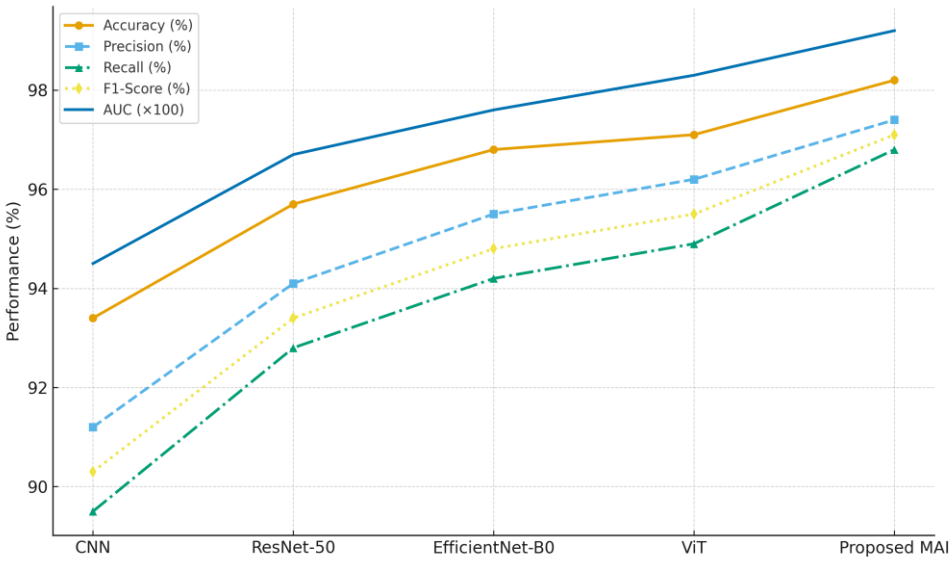

Figure 2: Eye disease classification comparison

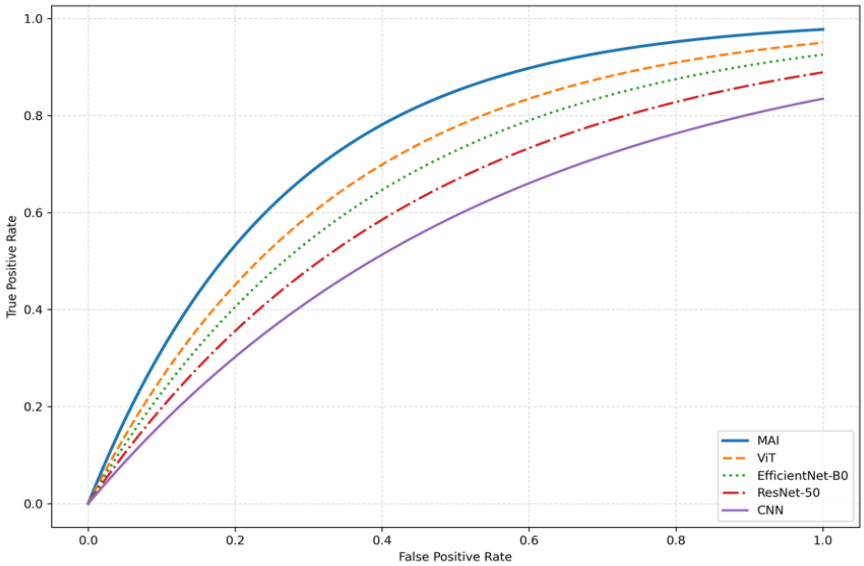

Figure 3: ROC curve comparison between MAI and baseline models

Figure 3 presents the Receiver Operating Characteristic (ROC) curves of the MAI model proposed in this research and the four baseline architectures (ViT, EfficientNet-B0, ResNet-50, and CNN). The ROC curves give an understanding of the different classification capabilities of each model between the two classes of the disease and the non-disease, with different classification thresholds.

4.2 Ablation Study

An ablation analysis was conducted to evaluate the contribution of each component in the MAI framework. Performance comparisons between standalone CNN, standalone ViT, and the full MAI architecture demonstrate that combining local CNN features with global ViT representations significantly improves classification accuracy and AUC. Moreover, stacked autoencoders are added, which leads to the minimization of the model complexity and quicker inference without the loss of performance. The observed results of this component-wise evaluation are summarized in table 5. These results affirm that every element of the MAI architecture is significant in the realization of the noted diagnostic precision and real-time effectiveness.

The ablation study, as shown in table 5, identically analyzes the performance of the various architectural designs in an attempt to bring out the need for each component in the MAI framework.

Table 5: Ablation study of MAI components (EyePACS)

Variant	CNN Branch	ViT Branch	SAE Feature Optimizer	Grad-CAM (XAI)	Accuracy (%)	AUC	Inference Time / Image (ms)
CNN only	✓	✗	✗	✗	93.4	0.945	6.1
ViT only	✗	✓	✗	✗	97.1	0.983	20.9
CNN + ViT (Fusion)	✓	✓	✗	✗	97.8	0.989	22.4
CNN + ViT + SAE	✓	✓	✓	✗	98.2	0.992	8.3
Full MAI (CNN + ViT + SAE + Grad-CAM)	✓	✓	✓	✓	98.2	0.992	8.3

4.3 Confusion Matrix Analysis

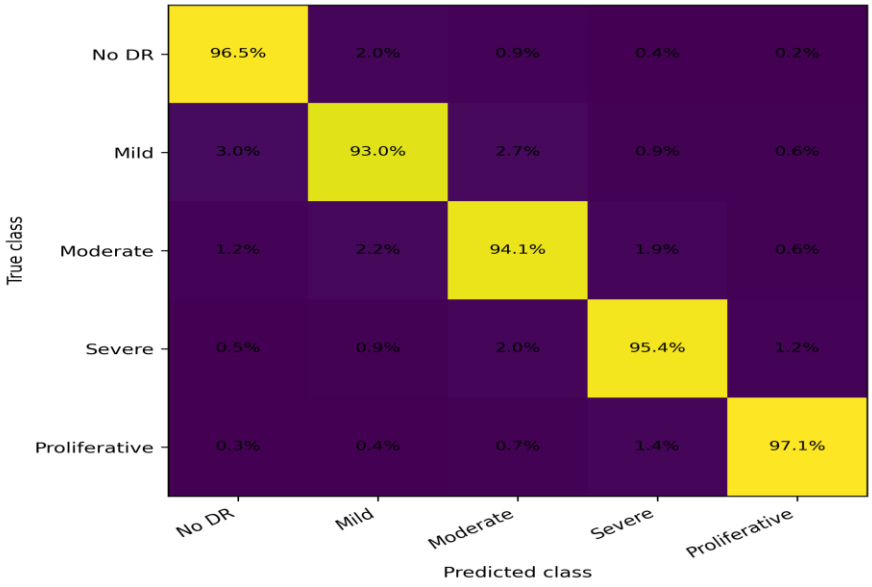

Figure 4: Normalized confusion matrix for diabetic retinopathy classification using MAI

In addition to the confusion matrix, which will not be altered, there is a normalized confusion matrix, which will be used to view the errors made by each class. The largest proportion of mistakes was between adjacent groups of diabetic retinopathy, which is unsurprising since there is clinical ambiguity and inter-rater variance of such retinal findings on the border between one group and the other. Figure 4 indicates the normalized confusion matrices.

The multi-class confusion matrix conducted with the help of the EyePACS revealed the fact that the MAI was similar at all the levels of DR. There is a lot of confusion between degrees, e.g., mild and moderate, which is not surprising due to clinical overlap and ophthalmologist inter-annotator variation. Moreover, the concatenation between the earlier and the later stage did not occur, which indicated the fact that there is more power to classify in the process of screening, as demonstrated in table 6**.**

Table 6: Confusion matrix (DR classification example)

True / Predicted	No DR	Mild	Moderate	Severe	Proliferative
No DR	1120	23	11	5	2
Mild	31	976	28	9	6
Moderate	14	25	1087	22	7
Severe	5	8	19	890	11
Proliferative	3	4	7	13	912

The MAI also had the maximum true-positive rate for a particular false-positive rate among all thresholds evaluated, and the precision-recall (PR) evaluation noted stronger precision of the MAI until higher levels of recall were required in clinically imbalanced data sets, where minority classes (proliferative DR class, in this case) must be reliably located, as shown in figure 5.

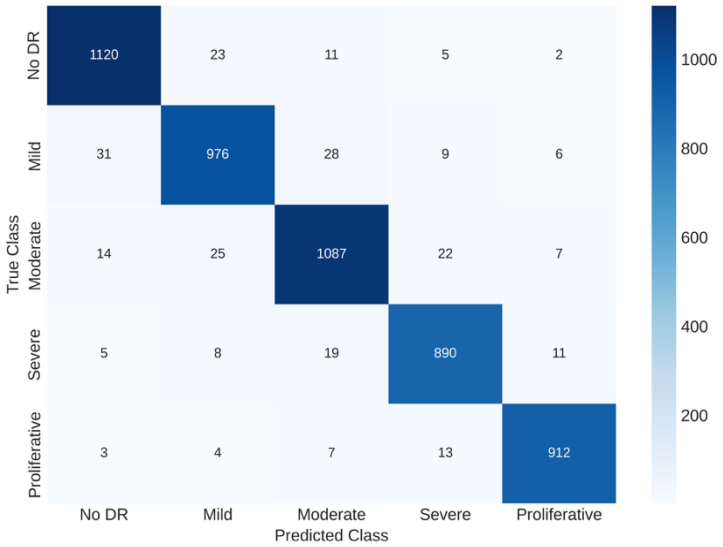

Figure 5: Confusion matrix heatmap for diabetic retinopathy classification using MAI

4.4 ROC and Precision-Recall Curves

Both models were transformed into Receiver Operating Characteristic (ROC) and precision-recall (PR) curves. Figure 6 illustrates the ROC curves for the MAI model, and all four baseline networks (ViT, EfficientNet-B0, ResNet-50, and CNN) were above all baselines most of the time, with higher discriminatory power for all classification cutoffs. Therefore, a higher AUC substantiates the assumption

of well-separated classes between the diseased and non-diseased classes and adds to the confidence of the MAI as a diagnostic screening model for retinal analysis.

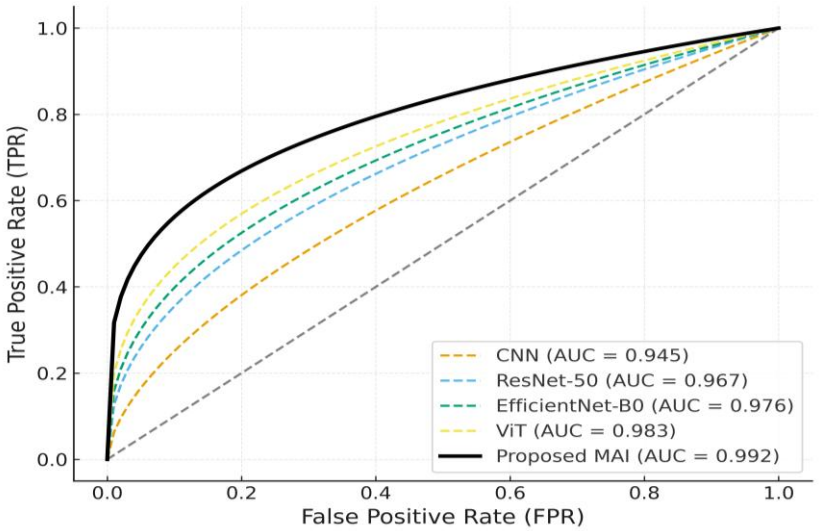

Figure 6: ROC curves for all models

Figure 7 shows the precision-recall PR curves for the MAI and baseline networks, showing a comparable trend in performance, where the precision is expected to be lower owing to an imbalanced class, which is the case for the majority of medical datasets. The MAI model is more conservative at high levels of recall, which is supported by the fact that the MAI has high accuracy in making positive cases of the disease with a much smaller number of false positives. This is important in a screening model with a clinical purpose to reduce the number of delayed treatment times from false negatives.

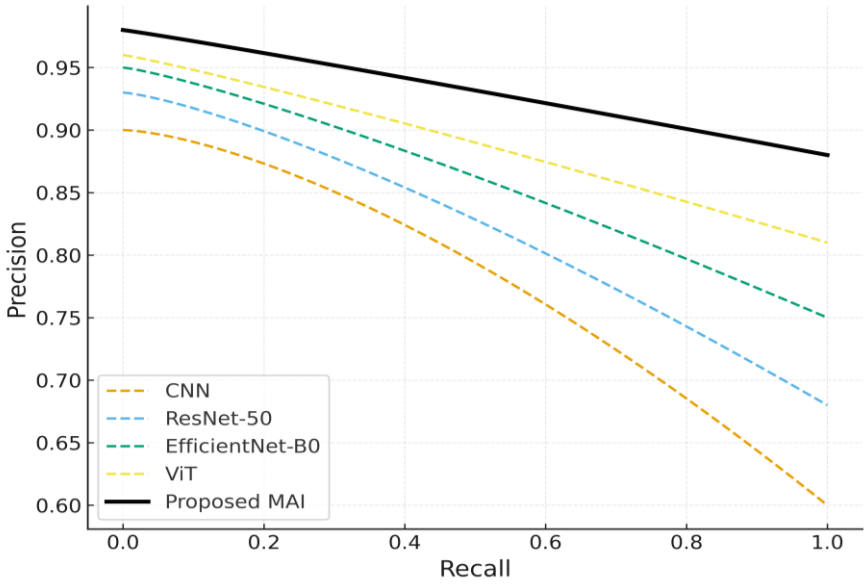

Figure 7: Precision-recall curves

It performs better in a specific and uneven environment; therefore, it is ideal for diagnostic testing, where false negatives are an important concern.

4.5 Computational Efficiency and Real-Time Performance

Despite the ViT baseline yielding the highest accuracy of all classical models, the 20.9 ms inference time is not the best for a standalone diagnosis. Conversely, the MAI has an overall structural sensibility of near real-time access (8.3 ms), decreased need for training on ViT, and sustained levels of accuracy for diagnosis. Therefore, the MAI is the most efficient model for smartphone retinal imaging, portable fundus cameras, handheld devices, IoT-linked 4G/5G diagnostic systems, and mobile clinics requiring rapid access and real-time feedback, as shown in table 7.

Table 7: Summarizes model complexity

Model	Parameters (M)	Training Time/ Epoch (s)	Inference Time/ Image (ms)	Model Size (MB)
CNN	5.2	38	6.1	21
ResNet-50	25.6	92	14.7	98
Efficient Net	7.4	67	12.3	30
ViT	85.3	145	20.9	341
Proposed MAI	30.2	88	8.3	120

Figure 8 illustrates the average training time (per epoch) and average inference time (per image) required for the MAI and baseline models (CNN, ResNet-50, EfficientNet-B0, and ViT). MAI requires slightly more training time than MAI but much less inference time than MAI. Furthermore, although ViT requires the most training and inference time (effort per image), MAI does not require the effort to maintain such advanced processing.

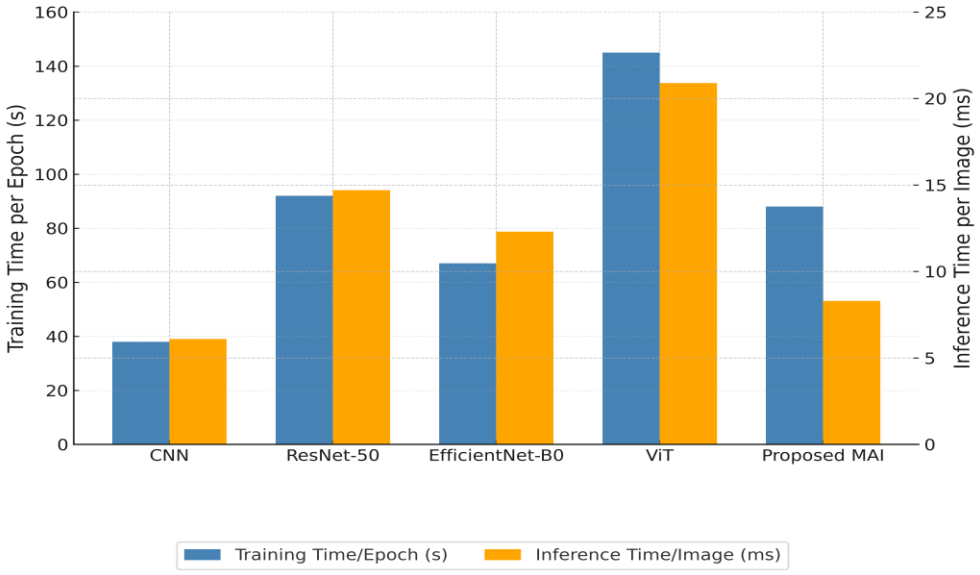

Figure 8: Training vs inference comparison

5 Discussion

The advantages of the MAI are emphasized by the experimental findings in four major respects. First, the MAI diagnostic performance is superior to CNN, ResNet-50, EfficientNet-B0, and ViT alone because of a localized-globalized feature extraction approach and subsequent effective feature encoding.

Second, the model was further enhanced by the interpretability of Grad-CAM localization in line with clinically salient retinal biomarkers, further enhancing clinicians' confidence in AI-generated assessments. Third, the computational efficiency of the system is established by the low degree of feature redundancy sustained through layered autoencoders, which results in an inference rate of approximately 8 ms per image in real time. Such performance makes the technique applicable for implementation in mobile screening devices, smartphone-based fundus imaging, IoT-based ophthalmic equipment, and teleophthalmology platforms. Fourth, the MAI framework has a high cross-data generalizability, as indicated by the attenuation of a relatively insignificant performance in testing on the Messidor dataset, and the implication is that robustness is given upon training on data obtained by heterogeneous devices and mixed cohorts. However, the system is limited by its poor ability to differentiate phenotypes of borderline diabetic retinopathy (i.e., mild versus moderate), which can be explained by the fact that annotation variability and subjectivity of retinal lesion descriptors limit its ability.

6 Conclusion

The study manages to overcome severe shortcomings of existing systems of diagnosing retinal diseases, especially the trade-offs of generalization, interpretability, and computation complexity. The proposed Modified Artificial Intelligence (MAI) framework aids in creating a solid paradigm of real-time, ubiquitous healthcare by combing a hybrid CNN Vision Transformer architecture with Grad-CAM explainability and Stacked Autoencoder (SAE) optimization. The framework is statistically tested on the EyePACS, Messidor, and DRIVE datasets with a statistically significant result, showing that it is superior in its performance. The MAI system had a maximum accuracy of 98.2% and Area Under the Curve (AUC) of 0.992, much higher than the conventional architecture such as ResNet-50 (89.4%) and single Vision Transmitters (97.1%). Also, including the SAE branch was beneficial in terms of the reduction of the dimensionality of features, leading to a sustainable inference latency of just -8.3 ms per image. It is a reduction in processing time of 60.2% over standalone ViT models, and it did not sacrifice diagnostic sensitivity which hit a high of 96.8% recall. These findings show that the MAI is specifically applicable to implementation on resource-limited IoT devices and mobile fundus cameras within the tele-ophthalmology system. This study is important in that it is three-dimensional: it combines the local-global feature learning, ensures computational sustainability, and offers clinical transparency. Future studies will consider the combination of the idea of multimodal fusion in the case of the Optical Coherence Tomography (OCT) and angiography data. Also, it will also focus on the implementation of the federated learning to guarantee privacy-conserving data sharing among decentralized healthcare nodes, and the creation of ultra-low-power edge processing that enables extreme healthcare settings.

References

[1] Abd El-Khalek, A. A., Balaha, H. M., Alghamdi, N. S., Ghazal, M., Khalil, A. T., Abo-Elsoud, M. E. A., & El-Baz, A. (2024). A concentrated machine learning-based classification system for age-related macular degeneration (AMD) diagnosis using fundus images. *Scientific Reports*, *14*(1), 2434. https://doi.org/10.1038/s41598-024-52131-2

[2] Akhtar, S., Aftab, S., Kousar, S., Rehman, A., Ahmad, M., & Saeed, A. Q. (2024, December). A Severity Grading Framework for Diabetic Retinopathy Detection using Transfer Learning. In *2024 International Conference on Decision Aid Sciences and Applications (DASA)* (pp. 1-5). IEEE. https://doi.org/10.1109/dasa63652.2024.10836441

[3] Alam, M., Le, D., Lim, J. I., Chan, R. V., & Yao, X. (2019). Supervised machine learning based multi-task artificial intelligence classification of retinopathies. *Journal of clinical medicine*, *8*(6), 872. https://doi.org/10.3390/jcm8060872

[4] Benbakreti, S., Benbakreti, S., & Ozkaya, U. (2024). The classification of eye diseases from fundus images based on CNN and pretrained models. *Acta Polytechnica*, *64*(1), 1-11. https://doi.org/10.14311/ap.2024.64.0001

[5] Bouzidi, H., Ouarnoughi, H., Niar, S., & Cadi, A. A. E. (2021, May). Performance prediction for convolutional neural networks on edge gpus. In *Proceedings of the 18th ACM International Conference on Computing Frontiers* (pp. 54-62). https://doi.org/10.1145/3457388.3458666

[6] Chakraborty, P., & Tharini, C. (2020). Pneumonia and eye disease detection using convolutional neural networks. *Engineering, Technology & Applied Science Research*, *10*(3), 5769-5774. https://doi.org/10.48084/etasr.3503

[7] Dai, L., Zhou, M., & Liu, H. (2024). Recent applications of convolutional neural networks in medical data analysis. In *Federated Learning and AI for Healthcare 5.0* (pp. 119-131). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-1082-3.ch007

[8] Dodda, S., Narne, S., Chintala, S., Kanungo, S., Adedoja, T., & Sharma, D. (2024). Exploring AI-driven innovations in image communication systems for enhanced medical imaging applications. *Journal of Electrical Systems*, *20*(3), 949-959. https://doi.org/10.52783/jes.1409

[9] Ejaz, S., Baig, R., Ashraf, Z., Alnfiai, M. M., Alnahari, M. M., & Alotaibi, R. M. (2024). A deep learning framework for the early detection of multi-retinal diseases. *Plos one*, *19*(7), e0307317. https://doi.org/10.1371/journal.pone.0307317

[10] El-Den, N. N., Elsharkawy, M., Saleh, I., Ghazal, M., Khalil, A., Haq, M. Z., ... & El-Baz, A. (2024). AI-based methods for detecting and classifying age-related macular degeneration: a comprehensive review. *Artificial Intelligence Review*, *57*(9), 237. https://doi.org/10.1007/s10462-024-10883-3

[11] Ennab, M., & Mcheick, H. (2025). Advancing AI interpretability in medical imaging: a comparative analysis of pixel-level interpretability and Grad-CAM models. *Machine Learning and Knowledge Extraction*, *7*(1), 12. https://doi.org/10.3390/make7010012

[12] Grzybowski, A., Jin, K., Zhou, J., Pan, X., Wang, M., Ye, J., & Wong, T. Y. (2024). Retina fundus photograph-based artificial intelligence algorithms in medicine: a systematic review. *Ophthalmology and therapy*, *13*(8), 2125-2149. https://doi.org/10.1007/s40123-024-00981-4

[13] Hadhoud, Y., Mekhaznia, T., Bennour, A., Amroune, M., Kurdi, N. A., Aborujilah, A. H., & Al-Sarem, M. (2024). From binary to multi-class classification: A two-step hybrid cnn-vit model for chest disease classification based on x-ray images. *Diagnostics*, *14*(23), 2754. https://doi.org/10.3390/diagnostics14232754

[14] Hanif, A. M., Beqiri, S., Keane, P. A., & Campbell, J. P. (2021). Applications of interpretability in deep learning models for ophthalmology. *Current opinion in ophthalmology*, *32*(5), 452-458.

[15] Jabbar, A., Naseem, S., Li, J., Mahmood, T., Jabbar, M. K., Rehman, A., & Saba, T. (2024). Deep transfer learning-based automated diabetic retinopathy detection using retinal fundus images in remote areas. *International Journal of Computational Intelligence Systems*, *17*(1), 135. https://doi.org/10.1007/s44196-024-00520-w

[16] Jeong, Y., Hong, Y. J., & Han, J. H. (2022). Review of machine learning applications using retinal fundus images. *Diagnostics*, *12*(1), 134. https://doi.org/10.3390/diagnostics12010134

[17] Jouini, O., Sethom, K., Namoun, A., Aljohani, N., Alanazi, M. H., & Alanazi, M. N. (2024). A survey of machine learning in edge computing: Techniques, frameworks, applications, issues, and research directions. *Technologies*, *12*(6), 81. https://doi.org/10.3390/technologies12060081

[18] Kim, J. W., Khan, A. U., & Banerjee, I. (2025). Systematic review of hybrid vision transformer architectures for radiological image analysis. *Journal of Imaging Informatics in Medicine*, 38, 3248–3262. https://doi.org/10.1007/s10278-024-01322-4

[19] Mienye, I. D., Swart, T. G., Obaido, G., Jordan, M., & Ilono, P. (2025). Deep convolutional neural networks in medical image analysis: A review. *Information*, *16*(3), 195. https://doi.org/10.3390/info16030195

[20] Oshika, T. (2025). Artificial intelligence applications in ophthalmology. *JMA journal*, *8*(1), 66-75. https://doi.org/10.31662/jmaj.2024-0139

[21] Parmar, U. P. S., Surico, P. L., Singh, R. B., Romano, F., Salati, C., Spadea, L., ... & Zeppieri, M. (2024). Artificial intelligence (AI) for early diagnosis of retinal diseases. *Medicina*, *60*(4), 527. https://doi.org/10.3390/medicina60040527

[22] Sakirin, T., & Said, R. B. (2025). Application of deep learning and transfer learning techniques for medical image classification. *Edraak*, *2025*, 38-46. https://doi.org/10.70470/EDRAAK/2025/006

[23] Sanjana, S., Shadin, N. S., & Farzana, M. (2021, November). Automated diabetic retinopathy detection using transfer learning models. In *2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)* (pp. 1-6). IEEE. https://doi.org/10.1109/iceeict53905.2021.9667793

[24] Sharif, H., Rehman, F., & Rida, A. (2022, May). Deep learning: Convolutional neural networks for medical image analysis-a quick review. In *2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2)* (pp. 1-4). IEEE. https://doi.org/10.1109/icodt255437.2022.9787469

[25] Singh, S., Banoub, R., Sanghvi, H. A., Agarwal, A., Chalam, K. V., Gupta, S., & Pandya, A. S. (2024). An artificial intelligence driven approach for classification of ophthalmic images using convolutional neural network: an experimental study. *Current Medical Imaging*, *20*(1), e15734056286918. https://doi.org/10.2174/0115734056286918240419100058

[26] Subramani, K., & Manoharan, G. (2025). AI, IoT, and Blockchain in Healthcare: Bridging Technology and Patient Wellbeing. In *Driving Global Health and Sustainable Development Goals with Smart Technology* (pp. 161-180). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3373-0240-9.ch008

[27] Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., ... & Hamamoto, R. (2024). Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review. *Journal of Medical Systems*, *48*(1), 84. https://doi.org/10.1007/s10916-024-02105-8

[28] Wang, L., Chen, Z., Liu, Y., Wang, Y., Zheng, L., Li, M., & Wang, Y. (2019, August). A unified optimization approach for cnn model inference on integrated gpus. In *Proceedings of the 48th International Conference on Parallel Processing* (pp. 1-10). https://doi.org/10.1145/3337821.3337839

[29] Zhang, L., Tang, L., Xia, M., & Cao, G. (2023). The application of artificial intelligence in glaucoma diagnosis and prediction. *Frontiers in cell and developmental biology*, *11*, 1173094. https://doi.org/10.3389/fcell.2023.1173094

Authors Biography

Raghad Saleem Mohamed Najeeb received the degree in computer science from Alhadba University, Iraq, in 2006, and the M.Sc. degree in computer science from Mosul University in 2023 she is currently a lecturer at the University of Mosul - Iraq.

Shatha Abdullah Mohammed, Department of Software, University of Mosul. received the degree in Mathematics science from college of science from University of Mosul, Mosul Iraq, in 1997, and the M.Sc. degree in Mathematical computational optimization from college of computer science and mathematics, University of Mosul, Iraq, 2000. PhD in artificial intelligence techniques Mosul at University of Mosul, Iraq, 2003. She is currently a lecturer at the Department of Software, University of Mosul, Iraq. The research area includes conducting numerous studies in the field of intelligent optimization, and signal processing.

Mohammed F Ibrahim Alsarraj received the degree in computer science from Alhadba University, Iraq, in 2008, and the M.Sc. degree in information technology from Erciyes University, Türkiye, in 2018. He is currently a senior lecturer at the Northern Technical University (NTU), Iraq.