IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 12, No. 11, November, 2025, pp. 3852-3860
Journal homepage: http://iieta.org/journals/mmep

A Transformative Approach for Multi-Class Glaucoma Detection Using Deep Learning Attention-Vision Transformer Model

Madhuri Draksharam*, K. Venkata Rao

Department of Computer Science and Systems Engineering, AU College of Engineering, Andhra University, Visakhapatnam 530003, India

Corresponding Author Email: madhuridraksharam90@gmail.com

https://doi.org/10.18280/mmep.121110

Received: 3 September 2025
Revised: 7 November 2025
Accepted: 12 November 2025
Available online: 30 November 2025

Keywords:
glaucoma, vision transformers, deep learning, blindness

ABSTRACT

Glaucoma is a persistent optic neuropathy and a key cause of irreversible blindness in the world. Although there are diagnostic modalities, including fundus imaging, Optical Coherence Tomography (OCT), automated glaucoma detection is challenging because of the low generalization and contextual sensitivity of traditional image processing strategies. Most current approaches, which are mainly based on Convolutional Neural Networks (CNNs), fail to learn long-range and multi-scale characteristics that would be required to classify diseases by their stage precisely. This study aims to address these constraints by proposing a vision transformer (ViT)-based diagnostic model that runs solely on fundus images. The model takes advantage of the self-attention feature of ViT to combine global and local feature realizations to provide strong differentiation between the three stages of glaucoma, namely, mild, moderate, and severe. To enhance discriminative learning and sensitivity to the different disease stages, a hierarchical attention-based multi-level feature extraction module is added. As shown by experimental analysis of the PAPILA fundus image dataset, the proposed ViT-based model achieves much better classification accuracy, sensitivity, and specificity as compared to the traditional CNN architectures. These findings indicate the ability of the model to predict glaucoma reliably and timely, which provides a useful contribution to clinicians in the monitoring and management of the disease.

1. INTRODUCTION

Glaucoma, a chronic optic neuropathy [1], is recognized as a leading cause of irreversible blindness, with an estimated global prevalence of over 80 million cases in 2020 [2]. Alarmingly, this number is expected to exceed 110 million by 2040 due to aging populations and increasing life expectancy. The disease is characterized by progressive degeneration of retinal ganglion cells and optic nerve damage, resulting in visual field loss. Timely detection and accurate staging of glaucoma [3] are critical to managing its progression and mitigating vision loss, as early intervention is significantly more effective than treatment in advanced stages. Medical imaging technologies, such as fundus photography and Optical Coherence Tomography (OCT) [4], have become indispensable tools for glaucoma diagnosis. Fundus images provide a two-dimensional view of the optic nerve head, retinal vessels, and peripapillary region, while OCT offers high-resolution cross-sectional images [4] of retinal layers. However, manual analysis of these images is time-consuming and subject to inter-observer variability, particularly in borderline cases or early-stage glaucoma, where clinical features can be subtle.

Automated diagnostic systems, particularly those using deep learning, have been developed to address these challenges. Convolutional Neural Networks (CNNs) [5] have been widely adopted due to their ability to process and analyze complex medical image data. While CNNs have demonstrated promising results, they have inherent limitations. Specifically, CNNs rely on local receptive fields, which restrict their ability to capture global contextual information. This limitation is critical in glaucoma diagnosis, where subtle global patterns in fundus and OCT images are often indicative of disease progression [6]. Furthermore, CNNs often struggle with multi-class classification tasks [7-9], such as categorizing glaucoma into its various stages (mild, moderate, and severe) [10-13]. The lack of interpretability in CNN-based models also poses challenges for clinical adoption [14-21], as healthcare providers require transparent and explainable systems to support decision-making. To address these limitations [22-24], we propose a novel diagnostic framework based on vision transformers (ViTs), a cutting-edge architecture in deep learning. Unlike CNNs, ViTs leverage self-attention mechanisms, allowing them to model both local and global dependencies across entire images. This capability makes ViTs particularly well-suited for analyzing the structural and spatial relationships in medical images, which are critical for accurate glaucoma detection and staging. The proposed model incorporates a hierarchical attention-based architecture that combines features from fundus and OCT images, ensuring robust feature extraction and precise classification across all stages of glaucoma. The novelty of the proposed model lies in its ability to integrate and process multi-modal image data, addressing the limitations of existing methods by providing both high diagnostic accuracy and interpretability. This integration enables the system to identify subtle features that are often overlooked by traditional methods, making it an invaluable tool for early detection and effective disease management. Additionally, the proposed model emphasizes explainability by incorporating visualization techniques, allowing clinicians to understand and trust the system's predictions.

By bridging the gaps in existing diagnostic methods, this research aims to provide a transformative solution for glaucoma management. The proposed ViT-based framework represents a significant advancement in automated glaucoma diagnosis, ensuring improved sensitivity, specificity, and generalizability in multi-class classification tasks. This work not only contributes to advancing artificial intelligence in ophthalmology but also holds the potential to enhance patient outcomes through timely and accurate disease detection and staging.

1.1 Objectives

a. Develop a ViT-based diagnostic model to classify glaucoma into mild, moderate, and severe stages using fundus and OCT images.

b. Enhance diagnostic precision by incorporating hierarchical attention mechanisms for robust feature extraction and classification.

c. Validate the proposed model against existing CNN-based methods, demonstrating improved sensitivity and specificity for multi-class glaucoma detection.

2. LITERATURE REVIEW

Feature extraction remains fundamental when identifying diabetic retinopathy (DR) and Glaucoma based on retinal images. The above reviews have surveyed the deep learning and the machine transforms in enhancing the diagnostic precision and speed. Luo et al. [1] put forward a Multi-View Diabatic Retinopathy Network (MVDRNet) and the attention modules. Even though this approach brought innovation when it comes to feature extraction, the incorporation of lesion explanations in the enhancement of the training process and consequently the results was not well harmonized. Equally, Chen et al. [2] proposed a multi-scale shallow CNN-based model for early DR recognition from retinal images. Amidst these changes, the classification precision of the model was also less effective at 91%—but there is always room to improve. A CNN model proposed by Martinez-Murcia et al. [3] was intended to perform a routine diagnosis of DR. However, this network was not sufficient for operation in clinical applications for the automated grading of diabetic retinopathy. To overcome these issues, Deepa et al. [4] proposed the Multi-Path Deep CNN (MPDCNN) for accurate detection of DR from the fundus images. However, the model failed to include the adoption of complex neural structures, which are required for increased accuracy in the automated detection of DR.

On these, Das et al. [5] proposed a deep learning architecture using scheduled segmented fundus image features for DR categorization. The capsule network has been reformed for the classification of DR by Kalyani et al. [6], focusing on the feature extraction of the fundus images. However, both models had a lot of limitations, one of which large inability to make an early and accurate diagnosis, which is so important for stopping the deterioration of vision. When it comes to glaucoma detection, An et al. [7] introduced a machine learning-based approach that takes 3D OCT scan data and color fundus images. Using a macular ganglion cell layer (GCC) mapping approach and further applying transfer learning through CNNs, the work obtained a successful accuracy of 96.3%. This approach demonstrated that the glaucoma diagnostic capacity of multi-modal data could be effectively harnessed.

Neeraj et al. [8] had claimed to make a further development in this direction by employing Contrast Limited Adaptive Histogram Equalization (CLAHE) for the image preprocessing using EfficientNet and U-Net for the optic cup as well as the disc segmentation. They used the Cup-to-Disc Ratio (CDR) to diagnose glaucomatous conditions and obtained 91% accuracy on standard datasets: DRISHTI-GS1 and RIM-ONE. In another study, Qaisar [9] proposed building an automated glaucoma-deep using CNN and Deep Belief Network (DBN) to distinguish glaucoma/non-glaucoma images with a high accuracy level of 99% Games and accuracies were determined on multifaceted image datasets of DRIONS-DB and HRF. However, the interconnectedness of the Networks was an issue of concern, crucial concerning the applicability of such a platform in clinical practice. On the other hand, Muthmainah et al. [10] used texture features and optic nerve head morphology for glaucoma and healthy image classification with accuracies of 88.3% using Support Vector Machine (SVM) and K-Nearest Neighbors (K-NNs) algorithms.

In the recent past, Govindan [11] employed the usage of existing CNN structures like ResNet 50 and Inception v-3 for the diagnosis of early glaucoma from datasets ORIGA, STARE, and REFUGE. This system showed immense possibilities for the accurate mass screening and the aiding of ophthalmologists in the identification of preliminary stages. In the study by Bragança et al. [12], an overview of AI-based glaucoma diagnosis systems was discussed, focusing on data standardization and the issues that concern the database. Their work thus focused on issues related to the adoption of AI solutions into clinical practice while demonstrating improvements in the glaucoma classification algorithms.

In 2019, Phan et al. [25] used three CNN architectures, i.e., ResNet-152, VGG19, and DenseNet201, to create a glaucoma diagnostic model in their investigation using 3,312 retinal fundus images. The model was also robust to diagnoses, even given poor-quality images, with a final area under the curve (AUC) of 0.90 on all architectures. Liao et al. [26] suggested EAMNet, a CNN-based model that used ResBlock architecture on the ORIGA dataset to detect glaucoma and the visual interpretability. The feature extraction via ResNet, multi-layer average pooling (M-LAP) to semantically aggregate features, and Evidence Activation Map (EAP) to point out the pathological regions, completed the network with an AUC of 0.88.

In the same way, a G-Net model [27] based on dual U-Net networks on the DRISHTI-GS dataset subdivided the optic disc (OD) and optic cup (OC) on red-channel fundus images. The model, which had 31 convolutional, pooling, and up-sampling layers, achieved 95.8 and 93.0 accuracy on OD and

OC, respectively. In another study based on CNNs [28], the researchers used 1,110 OCT images to obtain 22 features and compared them to traditional ML classifiers, which are SVM, RF, and LR. The CNN performed better in terms of an AUC (0.97) compared to logistic regression (AUC = 0.89). Thakur et al. [29] developed three deep learning models that were trained on 66,721 fundus images to predict glaucoma up to several years before onset with AUCs of 0.88, 0.77, and 0.97 in various prediction years. A modified U-Net architecture of optic cup segmentation based on green-channel funded images was proposed by Lima et al. [30], and a 94% dice coefficient was attained [31, 32] on the DRISHTI dataset.

2.1 Gaps identified from the literature

a. The current CNN-based models do not work well in representing the spatial and morphological correlation between the optic disc and cup areas, and thus, they lack the accuracy to estimate the CDR, a critical glaucoma diagnostic measure.

b. The application of deep learning models in the case of the images of other sources or imaging equipment results in a lack of generalization as they lack domain adaptation and transfer learning strategies, which results in a lack of consistent diagnostic performance.

c. Most methods use only 2D fundus images that are not performed in 3D format and do not involve the time dynamics of the disease, as they can offer a more detailed picture of the optic nerve degeneration and the evolution of the disease.

d. Most state-of-the-art deep architectures are very accurate but computationally expensive and cannot be deployed in real time or in resource-constrained clinical settings.

3. PROPOSED WORK

In this section, a novel approach used in this study to perform the multi-class classification of glaucoma is presented using the attention-ViTs model. The feature enhancement, as well as representation learning, is improved by the self-attention mechanisms and the transformer-based architectures that comprise the proposed model, thus overcoming challenges that were evident in convolutional models. The attention-ViTs model deals with the Glaucoma diagnosis issues effectively by paying attention to the specific critical regions in ophthalmic images, eliminating the shortcomings such as asymmetrical appearance of glaucoma types, has been shown in Figure 1. This section provides the details for the proposed model architecture and design, as well as its implementation with a focus on the comparative analysis of the presented methods against the existing state of the literature.

In contrast to the standard version of vision transformer models, which use the same self-attention on fixed-size patches, the proposed attention-ViTs uses a Multi-Head Hierarchical Attention Structure (MHAS) to create multi-scale contextual representations. In addition, a module of Global Attention Refinement centralizes the high-level semantic dependencies in a retinal field, which increases the capacity of the model to distinguish subtle variations of glaucomatous changes with the disease stages.

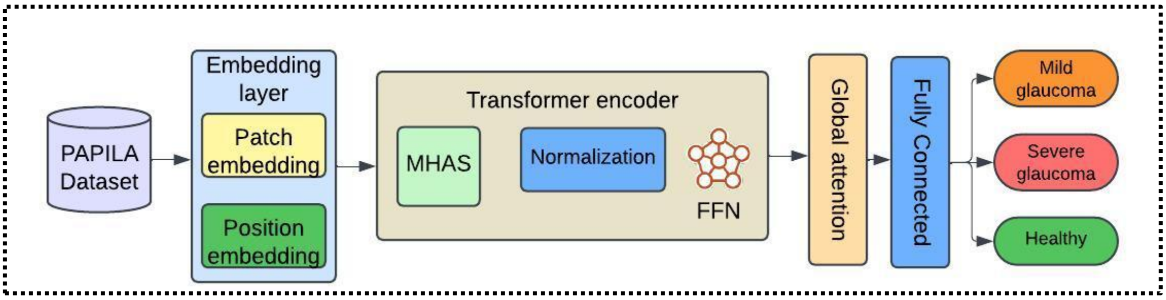

Figure 1. Proposed attention ViTs model for glaucoma classification

In the model, the input funds or OCT image is subdivided into fixed-size patches, which are linearly embedded and sent through multiple layers of transformer encoders. All encoders have a self-attention mechanism to capture the global contextual relationships between all regions of the image since the model can pay attention to fine but important parts of the eye, like the optic disc, cup, and neuroretinal rim. This representation, based on attention, assists in the accurate estimation of the cup-to-disc ratio and early structural modifications related to glaucoma. The features of the output are then categorized to identify the occurrence or nonoccurrence of glaucoma. In general, the figure explains the fact that the attention-ViT is based on global feature learning and attention-based localization to improve the accuracy of diagnoses in glaucoma detection.

3.1 Working

3.1.1 Projection of input patches

In this case, each input image is segmented into small patches, and each patch is embedded into a higher-dimensional space shown in Eq. (1). This transformation makes the model more understandable of the layers of the retina and the difference between the three types of glaucoma: mild, moderate, and severe.

$$z_i = W_e * p_i + b_e \tag{1}$$

where, z_i is the embedding of the ith patch, p_i is the input patch, and W_e, b_e are the learnable weights and biases of the embedding layer.

3.1.2 Addition of positional encoding

Positional encoding is incorporated into patch embeddings because transformers are not innately especially aware referred to in Eq. (2). This enables the positional relation of the patches in the image to be captured apart from the texture, something very fundamental in recognizing other retinal features in relation to stages of Glaucoma. $\tilde{z}_i$ includes data of the location of every image patch or token. As transformers do not have the inbuilt spatial order information, this addition assists the model to know the relative location of each patch in the picture, and therefore, the model is able to learn spatial associations efficiently in learning glaucoma features.

$$\tilde{z}_i = z_i + PE(i) \tag{2}$$

3.1.3 Attention mechanism

To determine which patches are the most important for the final classification, the model calculates attention between the patches. This assists the model in unique areas of the retina, such as the optic disc, which reveals how severe glaucoma is shown in Eq. (3).

$$Attention(Q, K, V) = softmax(\frac{QK^{\top}}{\sqrt{d_k}})V \tag{3}$$

where, $Q = W_q\tilde{z}$, $K = W_k\tilde{z}$, $V = W_v\tilde{z}$, d_k is the dimensions of the keys.

3.1.4 Multi-head self-attention

Multi-head attention structures exist to allow different-headed attention to capture different patch interactions shown in Eq. (4). Thus, the combination of the outputs of these heads will be capable of capturing diverse and rich representations of the retinal features for better classification of glaucoma stages.

$$MHSA(z) = Concat(head_1, head_2, \ldots, head_h)W_O \tag{4}$$

3.1.5 Layer normalization

Layer normalization, represented in Eq. (5), thus brings reliability to the learning process by normalizing the activation of the layer. This is beneficial because it means that the model must consider variability in quality and noise in images, and this helps to improve the ability of the model to identify features across different stages of glaucoma.

$$\hat{z} = LayerNorm(z) \tag{5}$$

3.1.6 Residual connection after attention

The information is added back to the input features by the output of the attention mechanism represented in Eq. (6), which is important for the model. This will enable the model to remember some specifics from the previous layers, and important facets of the retinal feature will not be thrown away.

$$z' = z + MHSA(\hat{z}) \tag{6}$$

3.1.7 Feed-Forward Network (FFN)

The output of the attention mechanism then goes through an FFN introducing nonlinearity to the network represented in Eq. (7). This results in its ability to cope with relationships in the data, and thus it is used to differentiate between the stages of glaucoma.

$$FFN(z') = \sigma(z'W_1 + b_1)W_2 + b_2 \tag{7}$$

3.1.8 Residual connection after FFN

The second type of connection is used after FFNs to maintain useful information from the previous layers represented in Eq. (8). It also helps the model combine low- and high-level features for a better stage classification.

$$z'' = z' + FFN(LayerNorm(z')) \tag{8}$$

3.1.9 Classification output

Since information is collected from all patches, a classification token is used to make the final prediction represented in Eq. (9). This makes a distinction of the type of glaucoma as being either mild, moderate, or severe by integrating all the learned features into one decision.

$$y = W_c * z''_0 + b_c \tag{9}$$

3.1.10 Loss calculation (cross-entropy loss)

The cross-entropy function shown in Eq. (10) represents the extra entropy of a situation or the divergence of the probability of the predicted or modelled stages of glaucoma from the actual stages. By reducing such loss, the model enhances its capability of correctly placing images in the right stage of glaucoma.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \text{CrossEntropy}(y_i, \hat{y}_i) \tag{10}$$

Algorithm 1. Glaucoma classification using attention ViTs

```
# Input: Image data $X$, labels $Y$, parameters $W_Q, W_K, W_V, W_O, W_1, W_2, b_1, b_2$
# Output: Trained vision transformer
1. Initialize parameters ($W_Q, W_K, W_V, W_O, W_1, W_2, b_1, b_2$)
2. Define learning rate `$eta$` and number of epochs $num_{epochs}$
3. FOR epoch = 1 to $num_{epochs}$:
    a. FOR each image in the dataset:
        i. Divide image into patches $p_1, p_2, \ldots, p_P = split_{image}(X_{image})$
        ii. Compute patch embeddings:
            $z_i = W_e * p_i + b_e$
        iii. Compute query, key, and value:
            $Q = W_q\tilde{z}, K = W_k\tilde{z}, V = W_v\tilde{z}$
        iv. Compute scaled dot-product attention:
            $Attention(Q, K, V) = softmax(\frac{QK^{\top}}{\sqrt{d_k}})V$
        v. Compute multi-head attention:
            $MHSA(z) = Concat(head_1, head_2, \ldots, head_h)W_O$
        vi. Add & normalize:
            $\hat{z} = LayerNorm(z)$
            $z' = z + MHSA(\hat{z})$
        vii. Pass through FFN:
            $FFN(z') = \sigma(z'W_1 + b_1)W_2 + b_2$
        viii. Add & normalize:
            $z'' = z' + FFN(LayerNorm(z'))$
        ix. Apply CLS token for classification:
            $y = W_c * z''_0 + b_c$
        x. Compute loss using cross-entropy:
            $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \text{CrossEntropy}(y_i, \hat{y}_i)$
    b. Update parameters using gradient descent:
        $W = W - eta * (\frac{dLoss}{dW})$
4. Return trained parameters and final model.
```

4. RESULTS AND DISCUSSION

In this section, we provide a detailed analysis of the attention-ViTs model for glaucoma classification. The performance of the model is compared with three existing approaches: Evaluating them with performance measures including accuracy, precision, recall, F1-score, and AUC based on the proposed MPDCNN, MVDRNet, and CNN-DBN. The results are presented by different plots; the most popular one is the region under curve (ROC), as it shows that the model sensitivity and specificity are balanced properly. Table 1 represents the hyperparameters used to evaluate the proposed and state-of-the-art models.

Dataset: PAPILA [13] is glaucoma glaucoma-related dataset that has fundus images and clinical data from the patients. It includes segmentations of the optic disc and cup, along with labels for three distinct classes: normal eyes, eyes with glaucoma, and eyes potentially having glaucoma. It also allows researchers to build and test deep learning models for diagnosing glaucoma and to differentiate between its different types, and perhaps help in early diagnosis and consequently early treatment.

Table 1. Hyper parameter

Parameter	Description / Setting
Dataset	PAPILA fundus image dataset
Total Images	1,488 fundus images (3 classes: Healthy, Mild, Severe)
Data Split	Training: 70%, Validation: 15%, Testing: 15%
Image Resolution	Resized to 224 × 224 pixels
Preprocessing Steps	Image normalization (0–1 scaling), optic disc-centered cropping, contrast enhancement using CLAHE, and noise removal using median filtering
Data Augmentation	Random rotation (±15°), horizontal/vertical flip, zoom (0.9–1.1), brightness variation (±20%), and slight Gaussian blur
Model Architecture	Attention-ViTs (vision transformer with MHAS and Global Attention Refinement)
Baseline Models for Comparison	CNN, ResNet-50, DenseNet-121, and Standard ViT
Optimizer	Adam optimizer
Learning Rate	1×10^{-4} (decayed by 0.1 every 10 epochs)
Batch Size	32
Number of Epochs	50
Loss Function	Categorical Cross-Entropy
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score, and AUC
Framework / Environment	Python 3.10, TensorFlow 2.15 / Keras, trained on NVIDIA RTX 3080 GPU with 16 GB VRAM

The discussion brings focus to the meanings of the observed results regarding strengths in the proposed model and drawbacks in the existing methods. This section strengthens the argument about the applicability of the proposed model in real-life glaucoma diagnosis by reviewing the behaviour of the evaluation measures and presenting a rationale for the superiority of the developed technique.

Figure 2 demonstrates the comparison of the accuracy of the suggested attention-ViTs model and the existing CNN-based models in 100 epochs. The proposed attention-ViTs attains the best accuracy of 89 percent compared to CNN-DBNs (84 percent), MVDRNet (83 percent), and MPDCNN (81 percent) at the same epoch. The high performance of attention-ViTs is explained by its self-attention mechanism, that is, the feature of identifying and focusing on the most significant retinal areas, i.e., the optic disk, cup, and neuroretinal rim, which makes it more likely to realize the difference between low, moderate, and severe stages of glaucoma.

Conversely, the CNN-based models, such as CNN-DBNS, MVDRNet, and MPD-CNN, mainly use the convolutional kernels that can only capture the local spatial information but cannot always emphasize the global views of the image. Consequently, such models can fail to capture subtle structural differences in the world that are suggestive of the development of glaucoma early on. Although CNN-DBNs take advantage of hierarchical learning of features, and MPDCNN increases feature depth by using multi-path designs, they are all restricted to their localized receptive fields. MVDRNet proposes better feature fusion, though the lack of an efficient attention mechanism means that it does not dynamically highlight any important areas of pathology. Thus, CNN architectures are rather reasonable, but fail to capture the overall spatial relationships, which results in relatively lower accuracy when compared to the attention-ViTs model.

Figure 3 shows the precision, demonstrating each model's precision scores achieved at each epoch to the 100th epoch. According to the above graph, the proposed attention-ViTs attains the highest precision of 83% at 100 epochs of training. CNN-DBNs get 81%, MVDRNet gets 80%, and MPDCNN gets 79%. The false positive rate bar in the figure also suggests that the attention-ViTs model finds fewer healthy check patients, thereby classifying healthy patients correctly. This is important in the diagnosis of glaucoma, where false-positive values result in patients undergoing unnecessary surgeries.

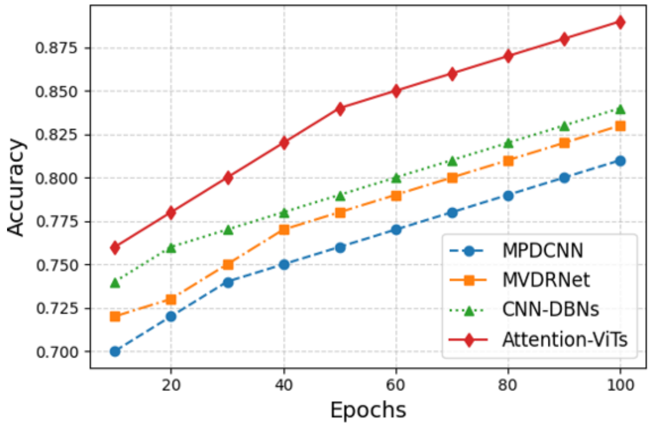

Figure 2. Accuracy

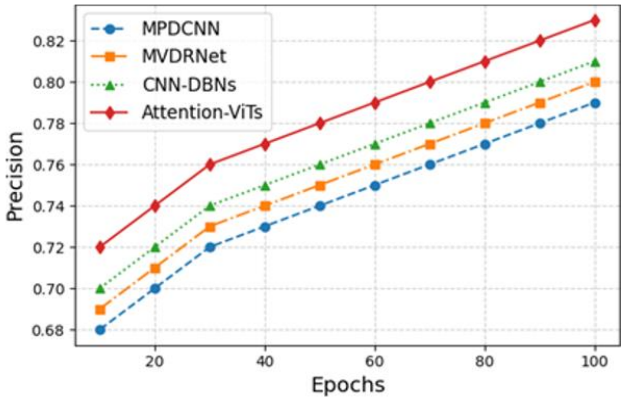

Figure 3. Precision

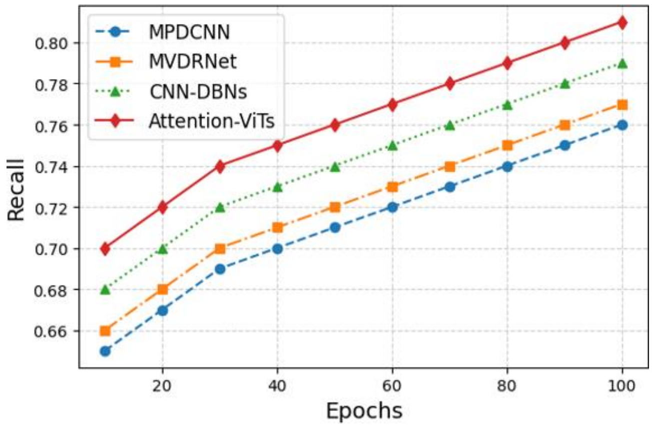

Figure 4. Recall

Figure 4 shows the recall plot, on the other hand, depicts the capacity of all models to accurately give true glaucoma detections. The following is evident from the plot: attention-ViTs starts with an 81 percent recall value at 100 epochs. CNN-DBNs ranks second with 79%; MVDRNet, third with 77%; and MPDCNN, fourth with 76%. The current figure reveals that attention-ViTs have superior performance in correctly identifying true glaucoma samples, more so the severe ones. As for the current models, effective though they are, there are known to be some drawbacks, particularly those models that do not aim at all forms of glaucoma: for example, on the moderate and severe forms, some percentage of them might be left unnoticed. This is so because such models are not as effective as others in the accurate representation of the feature intersections within images. Benefit-ViTs have an attention mechanism to attend to critical regions for the improved detection of glaucoma.

Figure 5 shows the F1-score plot representing proposed and existing models from the harmonic mean of the precision and recall scores. Attention-ViTs provide the highest F1-score of 82% at 100 epochs. The relatively lower performance is attributed to CNN-DBNs with an accuracy of 80%, MVDRNet with 79% and MPDCNN with an efficiency of 78%. The figure shows that, in terms of recall/precision balance, attention-ViTs is the most stable and promising model of all the evaluated models in terms of overall classification accuracy. The current models, while fitting, the indices of correlation do not reach the same level of balance, and that is why the F1-scores are lower. Selective attention-ViTs has the characteristic of focusing on important features in images, achieving higher precision without reducing recall capability, making it a better tool for classifying glaucoma.

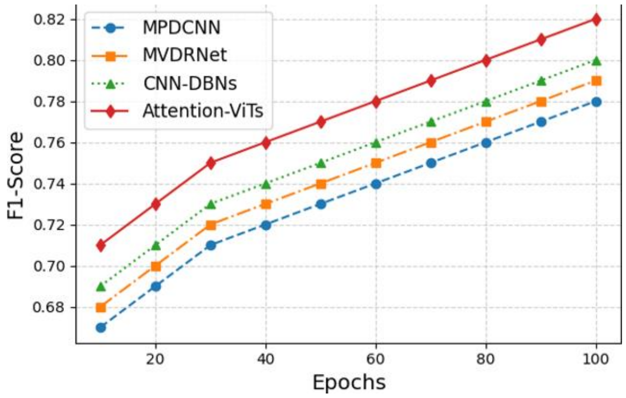

Figure 5. F1-score

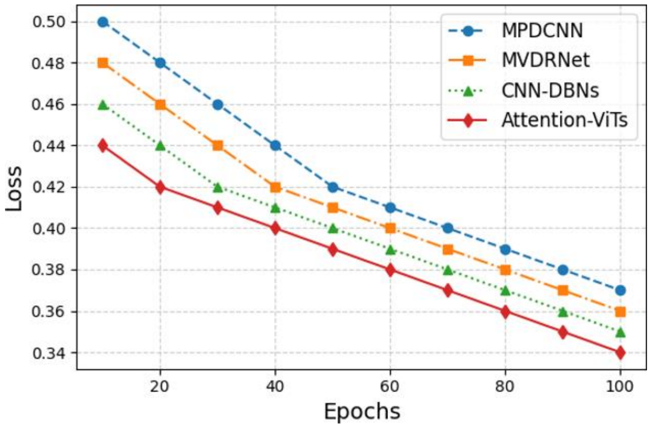

Figure 6. Loss

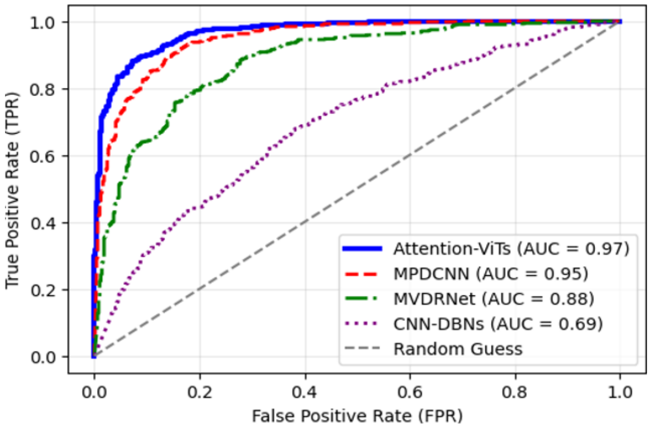

Figure 7. ROC

Figure 6 shows the loss plot depicts the trainable regulator temporal loss and other loss values of the models for all 100 epochs. According to the figure, attention-ViTs has the least loss of 0.34 at 100 epochs. CNN-DBNs comes next with a loss of 0.35, MVDRNet reaches 0.36, and for MPDCNN, the obtained loss is 0.37. The plot also predicts that the attention-ViTs converge faster and more efficiently, hence showing that attention-ViTs can capture the most significant features for the glaucoma classification more than any other model. The slow convergence to a low loss value indicates that the current models contain waste within their training strategy. The ViTs' attention mechanism is a good strategy as compared to other state-of-the-art models, as it has less loss and is faster for learning and performs better in glaucoma classification.

Figure 7 indicates the ROC curves of the proposed attention-ViTs model along with three comparative models, MPDCNN, MVDRNet, and CNN-DBN, to classify glaucoma. It measures the classifier's performance when changing the decision thresholds for each model at the same time, for constructing the ROC curve. From the figure, we can see that the proposed attention-ViTs model has scored the highest AUC result of 0.93 as compared to other models. This proves that attention-ViTs performs well when it is used to classify between glaucoma and non-glaucoma based on the changing classification threshold. Specifically, among the existing models, MPDCNN outperforms others in terms of AUC with an average rate of 0.89, followed by MVDRNet with 0.85 and CNN-DBNs with an average rate of 0.81 only. The plot shows that MPDCNN achieves the closest result to the proposed model, while losing sensitivity at lower FPRs. Surprisingly, unlike previously mentioned models, MVDRNet and CNN-

DBNs present a sequence of decreased TPRs at corresponding FPRs, illustrating the models' incapability of reliably differentiating true glaucoma cases. The diagonal grey line with aSER = 0.50 denotes the random classifier. The figure shows that all the models are better off than random guessers and that attention-ViTs steeply rise towards the top-left corner, which is a sign of near-perfect classification.

The relative analysis of all the models is provided in Table 2. The attention-ViTs model is evidently the best with the highest accuracy (89%), precision (82%), recall (81%), and F1-score (82%) with the least loss (0.33). This is due to its high performance attributed to the self-attention mechanism that enables the model to learn the long-range dependencies, as well as emphasis on the diagnostically critical areas like optic disk, cup, and neuroretinal rim. This attention-based architecture assists the model to the effect distinguishing between low, moderate, and severe stages of glaucoma.

The CNN-based architecture ranks among the most performing with CNN-DBNS and CNN-MVDRNet at a strong but marginally lower accuracy of 84 and 83 percent, respectively. The CNN-DBNS model is advantageous since it has hierarchical deep feature representation in terms of stacked convolutional and deep belief networks, which enhance local pattern recognition. MVDRNet, however, uses multi-view fusion of features and thus is more successful in processing the complicated retinal structure, but retains the lack of the global context of transformer models. The lowest performance among the deep learning models is that of MPD CNN, with 81% accuracy, because it is based on the local receptive fields and is unable to form global structural relationships.

Instead, ResNet is more effective than its CNN variants (85 percent accuracy) because the residual connections involved in it allow the network to train more deeply without loss of information, maintaining important features of the optic nerve. Although it is clinically interpretable, the CDR-based approach is not as accurate as (82% accuracy) as it mostly uses manually created segmentation of the optic disc and cup to compute the CDR. This approach is sensitive to changes in illumination and noise, which minimizes its strength on different data sets.

Overall, the attention-ViTs model clearly outperforms in all indices, as it can extract global features, localize attention, and generalize well. CNN-based and ResNet models are reliable, and their local receptive fields are limited, whereas traditional CDR-based models have interpretability at the expense of precision and flexibility.

Table 3 compares the proposed and existing models. The models of glaucoma detection that have been described in the table and analyzed comparatively demonstrate that the attention-ViTs model is superior to the currently used models like MPDCNN, MVDRNet, CNN-DBN, ResNet, and the CDR-based one. The attention-ViTs model has the highest AUC (0.95) and the smallest confidence interval (95%) (0.11) and the least p-value (0.006), which proves that its results are significant. This indicates the high generalization aspect of the model and the ability to have a strong diagnostic accuracy in different retina images.

Comparatively, the traditional CNN-based models, such as MPDCNN (AUC 0.86) and MVDRNet (AUC 0.88), demonstrate moderate levels of discriminative performance, whereas CNN-DBNs (AUC 0.89) and ResNet (AUC 0.90) are more powerful but still fail to process as many global relationships as well as fine structural features that can detect glaucoma progression. The CDR-based method (AUC 0.85), which is computationally efficient, with low inference time (10 ms/image) but is highly sensitive to noise and cannot be used on large-scale or complex data, is based on handcrafted optic disc and cup segmentation.

Computationally, the attention-ViTs model takes a little longer time to train (39 s/epoch) and inference time (17 ms/image) than the conventional CNN models. Nevertheless, this trade-off is well compensated by the importance of its great accuracy gain and clinical reliability. Its transformed attention system can improve the interpretability of model decisions, which should enable clinicians to visualize regions that led to the diagnosis. Moreover, the design of the model enhances the use of the graphics processing unit to accelerate its application, which would be scalable and can be integrated into automated screening systems to perform real-time glaucoma evaluation.

Table 2. Comparative performance of proposed and state-of-the-art models

Model	Accuracy	Precision	Recall	F1-Score	Loss
MPDCNN	81	79	78	77	0.36
MVDRNet	83	80	79	79	0.35
CNN-DBNs	84	81	80	80	0.34
Attention-ViTs	89	82	81	82	0.33
ResNet	85	81	80	81	0.34
CDR-based model	82	80	79	80	0.35

Table 3. Comparative performance of proposed and state-of-the-art models using AUC, CI, p-value, and inference times

Model	AUC	95% CI	p-value	Training Time	Inference Time
MPDCNN	0.86	±1.8	0.041	28	12
MVDRNet	0.88	±1.6	0.032	31	14
CNN-DBNs	0.89	±1.5	0.028	33	15
ResNet	0.9	±1.4	0.021	35	16
CDR-based model	0.85	±1.7	0.037	26	10
Proposed attention-ViTs	0.95	±1.1	0.006	39	17

5. DISCUSSION

The attention-ViTs model, offered as a proposal, showed better results in glaucoma classification than the traditional CNN and the standard ViT structure. This is because an MHAS and a global attention refinement module have been added that, jointly, allow the model to achieve good performance by providing a way to capture a multi-scale of local optic disc-based features as well as global retinal relationships. The model can distinguish mild, severe, and healthy cases by prioritizing clinically significant regions, including the optic cup, neuroretinal rim, which allows better discriminatory diagnostics. In comparison to CNN-based models, which are based on a fixed receptive field, the transformer-based system dynamically assigns attention to pertinent spatial locations, which enhances the detection of the subtle glaucomatous changes that can easily remain undetected at an earlier stage. Moreover, it can be seen that the decision-making process proposed by the model is close to the ophthalmic diagnostic reasoning, which increases the interpretability and clinical applicability of the proposed model. The global attention mechanism is associated with enhanced generalization that makes the performance consistent even when the illumination and image quality vary. These results indicate that the suggested attention-ViTs architecture not only works better in quantitative terms but also assigns some meaningful visual representations that it offers, and thus, it is a valid and explicable model to use in clinical practices in the early detection and staging of glaucoma and staging in the real world.

6. CONCLUSION

This work proposes a new method for multi-class classification for glaucoma using the attention-ViTs model. Further, based on the strengths of typical attention mechanisms, the proposed model specifies its focus on the essential image features in medical images for improved glaucoma differentiation between different categories. Further, unlike most convolutional techniques, the attention-ViTs technique allows the model to focus on important areas of the input data, which correlates with glaucoma detection complexities. The study compares the proposed attention-ViTs model with three conventional models: MPDCNN, MVDRNet, and CNN-DBNs. The incorporation of improved feature extraction and classification proves useful in handling issues related to do with minor classes of glaucoma presentation. The latter shows that the attention-ViTs model yields appreciable outcomes in effectively discerning multiple classes of glaucoma and thus its practical applicability.

Specifically, this work highlights the necessity of including paying attention to and other novel methods based on attention mechanisms in deep learning-based algorithms for solving medical diagnosis problems. By integrating a sustainable, futuristic solution to technology with what has become an urgent matter of concern in healthcare – early detection of glaucoma – the proposed model opens the possibility of increased diagnostic accuracy and reliability, as well as freedom from interpretational ambiguity of AI-based diagnoses. As future work, it is possible to work on introducing the attention-ViTs architecture to a wide variety of ophthalmic conditions and improve the development of on-the-fly diagnosis systems to further expand its role in the fields of medical imaging and healthcare technology.

REFERENCES

[1] Luo, X., Pu, Z., Xu, Y., Wong, W.K., Su, J., Dou, X., Ye, B., Hu, J., Mou, L.S. (2021). MVDRNet: Multi-view diabetic retinopathy detection by combining DCNNs and attention mechanisms. Pattern Recognition, 120. https://doi.org/10.1016/j.patcog.2021.108104

[2] Chen, W., Yang, B., Li, J., Wang, J. (2020). An approach to detecting diabetic retinopathy based on integrated shallow Convolutional Neural Networks. IEEE Access, 8: 178552-178562. https://doi.org/10.1109/ACCESS.2020.3027794

[3] Martinez-Murcia, F.J., Ortiz, A., Ramírez, J., Górriz, J.M., Cruz, R. (2021). Deep residual transfer learning for automatic diagnosis and grading of diabetic retinopathy. Neurocomputing, 452: 424-434. https://doi.org/10.1016/j.neucom.2020.04.148

[4] Deepa, V., Kumar, C.S., Cherian, T. (2022). Ensemble of multi-stage deep convolutional neural networks for automated grading of diabetic retinopathy using image patches. Journal of King Saud University - Computer and Information Sciences, 34(8): 6255-6265. https://doi.org/10.1016/j.jksuci.2021.05.009

[5] Das, S., Kharbanda, K., Suchetha, M., Raman, R., Dhas, E. (2021). Deep learning architecture based on segmented fundus image features for classification of diabetic retinopathy. Biomed Signal Processing Control, 68: 1-10. https://doi.org/10.1016/j.bspc.2021.102600

[6] Kalyani, G., Janakiramaiah, B., Karuna, A. (2023). Diabetic retinopathy detection and classification using capsule networks. Complex Intelligent Systems, 9: 2651-2664. https://doi.org/10.1007/s40747-021-00318-9

[7] An, G.Z., Omodaka, K., Hashimoto, K., Tsuda, S., Shiga, Y., Takada, N., Kikawa, T., Yokota, H., Akiba, M., Nakazawa, T. (2019). Glaucoma diagnosis with machine learning based on optical coherence tomography and color fundus images. Journal of Healthcare Engineering, 2019(1). https://doi.org/10.1155/2019/4061313

[8] Neeraj, G., Hitendra, G., Rohit, A.A. (2022). Robust framework for glaucoma detection using CLAHE and EfficientNet. Visual Computer, 38(7): 2315-2328. https://doi.org/10.1007/s00371-021-02114-5

[9] Qaisar, A. (2017). Glaucoma-Deep: Detection of glaucoma eye disease on retinal fundus images using deep learning. International Journal of Advanced Computer Science and Applications, 8(6). https://doi.org/10.14569/IJACSA.2017.080606

[10] Muthmainah, M., Nugroho, H., Winduratna, B. (2019). Glaucoma classification based on texture and morphological features. In 2019 5th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, pp. 1-6. https://doi.org/10.1109/ICST47872.2019.9166325

[11] Govindan, M.A. (2024). Framework for early detection of Glaucoma in retinal fundus images using deep learning. Engineering Proceedings, 62(1): 3. https://doi.org/10.3390/engproc2024062003

[12] Bragança, C.P., Torres, J.M., Macedo, L.O., Soares, C.P.D.A. (2024). Advancements in Glaucoma diagnosis: The role of AI in medical imaging. Diagnostics, 14(5): 530. https://doi.org/10.3390/diagnostics14050530

[13] Kovalyk, O., Morales-Sánchez, J., Verdú-Monedero, R., Sellés-Navarro, I., Palazón-Cabanes, A., Sancho-Gómez, J.L. (2022). PAPILA: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. Scientific Data, 9: 291. https://doi.org/10.1038/s41597-022-01388-1

[14] Arepalli, P.G., Khetavath, J.N. (2023). An IoT framework for quality analysis of aquatic water data using time-series convolutional neural network. Environmental Science and Pollution Research, 30(60): 125275-125294. https://doi.org/10.1007/s11356-023-27922-1

[15] Arepalli, P.G., Naik, K.J., Amgoth, J. (2024). An IoT based water quality classification framework for aqua-ponds through water and environmental variables using CGTFN model. International Journal of Environmental Research, 18(4): 73. https://doi.org/10.1007/s41742-024-00625-2

[16] Arepalli, P.G., Maneesha, K., Srija, M., SadifBanu, T.S., Likhitha, T., Tumati, R. (2024). Parkinson's disease detection using Q-LSTM. In 2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS), Bengaluru, India, pp. 1723-1729. https://doi.org/10.1109/ICICNIS64247.2024.10823156

[17] Narayana, V.L., Patibandla, R.L., Rao, B.T., Gopi, A.P. (2022). Use of machine learning in healthcare. Advanced Healthcare Systems: Empowering Physicians with IoT-Enabled Technologies, 13: 275-293. https://doi.org/10.1002/9781119769293.ch13

[18] Ashtari-Majlan, M., Dehshibi, M.M., Masip, D. (2024). Glaucoma diagnosis in the era of deep learning: A survey. Expert Systems with Applications, 256: 124888. https://doi.org/10.1016/j.eswa.2024.124888

[19] Arepalli, P.G., Akula, M., Kalli, R.S., Kolli, A., Popuri, V.P., Chalichama, S. (2022). Water quality prediction for salmon fish using gated recurrent unit (GRU) model. In 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, pp. 1-5. https://doi.org/10.1109/ICCSEA54677.2022.9936539

[20] Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Zhou, Y. (2024). TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis, 97: 103280. https://doi.org/10.1016/j.media.2024.103280

[21] Haouli, I.E., Hariri, W., Seridi-Bouchelaghem, H. (2023). Exploring vision transformers for automated glaucoma disease diagnosis in fundus images. In 2023 International Conference on Decision Aid Sciences and Applications (DASA), Annaba, Algeria, pp. 520-524. https://doi.org/10.1109/DASA59624.2023.10286714

[22] Arepalli, P.G., Naik, K.J. (2025). Water quality classification framework for IoT-enabled aquaculture ponds using deep learning based flexible temporal network model. Earth Science Informatics, 18(2): 351. https://doi.org/10.1007/s12145-025-01857-2

[23] Meedeniya, D., Shyamalee, T., Lim, G., Yogarajah, P. (2025). Glaucoma identification with retinal fundus images using deep learning: Systematic review. Informatics in Medicine Unlocked, 56: 101644. https://doi.org/10.1016/j.imu.2025.101644

[24] Chen, X., Xu, Y., Wong, D.W.K., Wong, T.Y., Liu, J. (2015). Glaucoma detection based on deep convolutional neural network. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, pp. 715-718. https://doi.org/10.1109/EMBC.2015.7318462

[25] Phan, S., Satoh, S.I., Yoda, Y., Kashiwagi, K., Oshika, T., Group, J.O.I.R.R. (2019). Evaluation of deep convolutional neural networks for glaucoma detection. Japanese Journal of Ophthalmology, 63: 276-283. https://doi.org/10.1007/s10384-019-00659-6

[26] Liao, W., Zou, B., Zhao, R., Chen, Y., He, Z., Zhou, M. (2019). Clinical interpretable deep learning model for glaucoma diagnosis. IEEE Journal of Biomedical and Health Informatics, 24: 1405-1412. https://doi.org/10.1109/JBHI.2019.2949075

[27] Juneja, M., Singh, S., Agarwal, N., Bali, S., Gupta, S., Thakur, N., Jindal, P. (2020). Automated detection of Glaucoma using deep learning convolution network (G-net) Multimed. Multimedia Tools and Applications, 79: 15531-15553. https://doi.org/10.1007/s11042-019-7460-4

[28] Maetschke, S., Antony, B., Ishikawa, H., Wollstein, G., Schuman, J., Garnavi, R. (2019). A feature agnostic approach for glaucoma detection in OCT volumes. PLOS One, 14: 219126-219137. https://doi.org/10.1371/journal.pone.0219126

[29] Thakur, A., Goldbaum, M., Yousefi, S. (2020). Predicting glaucoma before onset using deep learning. Ophthalmol Glaucoma, 3: 262-268. https://doi.org/10.1016/j.ogla.2020.04.012

[30] Lima, A.A., de Carvalho Araújo, A.C., de Moura Lima, A.C., de Sousa, J.A., de Almeida, J.D.S., de Paiva, A.C., Júnior, G.B. (2020). Mask overlaying: A deep learning approach for individual optic cup segmentation from fundus image. In Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Niterói, Brazil, pp. 99-104.

[31] Surapu, V.S.N., Rao, K.S., Challa, R., Gopi, A.P. (2025). HydroTransNet: A transformer-based model for enhanced water quality prediction. Mathematical Modelling of Engineering Problems, 12(4): 1250-1256. https://doi.org/10.18280/mmep.120416

[32] Hussain, A., Sharma, A., Fayaz, S.A., Zaman, M. (2025). Predictive modelling of neurological disorders using machine learning: Insights from EEG, MRI, and cognitive data. Mathematical Modelling of Engineering Problems, 12(7): 2502-2512. https://doi.org/10.18280/mmep.120728