

Helformer Transformer-Based Multimodal Framework For Early Retinal Disease Detection Using Fundus And OCT Images

Patel Vandanababen Gopalbhai^{1*} and Dr. Jayesh N. Modi²

¹ Research Scholar, Department of Computer Science, Hemchandracharya North Gujarat University, Patan, India

² PHD Computer Science, Department of computer Science, Hemchandracharya North Gujarat University, Patan, India

Abstract

Retinal eye disease is a wide range of medical conditions that impacts the eye structure and function, resulting in visual impairments or blindness. In order to avoid these kinds of severe issues, an early eye disease detection mechanism should be developed for detecting abnormalities in the eyes. However, the images of the retina and fundus may be impacted due to the pixel smudge, noise, blur, or inadequate lighting at the particular disease portion, which reduces the accuracy of early disease diagnosis. To overcome these several limitations, a Helformer transformer-based classifier model had been proposed for detecting the eye disease earlier based on the input image. Initially, the input image is collected from both fundus and OCT for automated eye disease prediction. The fundus image are preprocessed using Weighted Wavelet Visual Perception Fusion (WWPF) for contrast enhancement and Conditional Filter Network (CFNet) for denoising the fundus image, along with that the OCT image are preprocessed using Adaptive Deep Residual Network (AdResNet) for removing the noise. Then the preprocessed image of fundus and OCT image are fed into the Gradient-weighted Class Activation Mapping (Grad-CAM) for highlighting the disease region of eye. After region detection, the fundus image are segmented using Multiscale Feature Interaction Network (MFI-Net) for segmenting the blood vessels in eye region and Semantic Graph Attention Network (SGANet) is used in OCT image for segmenting the retinal layer of the eye. To extract the features, Graph-Based Radiomics Feature Extraction Algorithm (GBRF) for fundus image and fully dense fusion neural network (FDFNN) for OCT images, which extracts the most relevant features from the image. The extracted features from both the fundus and OCT image are fused together with a fusion process. The fused feature vector can be passed to a Helformer transformer-based classifier for predicting the eye disease effectively. Finally, the output can be described as Normal, Diabetic Retinopathy, Diabetic Macular Edema and Drusen based on the input image. For eye disease classification, the proposed approach achieves a high accuracy of 98.57%, high precision of 97.05% and low error value of 1.42% respectively. These proposed model automatically learns and extracts seasonal patterns from the time-series data, which differentiate against from existing techniques.

Keywords: OCT; Fundus; CFNet; Ad Res Net; Grad-CAM; MFI-Net.

1. INTRODUCTION

Retina is the only portion of the eye at which blood vessels can be observed without invasive procedures, and thus is a key location for early detection of ocular and systemic disease. Retinal diseases leading to structural and functional abnormalities are on a rise and represent a significant threat of irreversible loss of sight or blindness when not treated [1]. In order to address these increasing challenges, the last two decades have seen extraordinary improvements in digital imaging technologies that have revolutionised fundus imaging and improved clinical diagnostic potential [2]. Retina is one of the essential organs for vision by functioning as a light sensor that transforms the incident light into neural signals that are sent to the visual cortex of the brain where they are interpreted and object recognition occurs [3]. However, if the ocular disease goes undiagnosed, progressive retinal damage can result in permanent loss of vision. However, the low public awareness, high consultation cost and shortage of ophthalmologists worldwide have limited the diagnosis of early diabetic retinopathy at

an early stage, and thus there is a strong demand for automated solutions for screening [4]. Although digital image acquisitions continue to lag behind those from conventional imaging techniques in terms of ultra-high resolution, current systems can provide images with sufficiently high quality to satisfy most clinical needs [5]. Among the most popular diagnostic modalities, OCT and colour fundus photography (CFP) has its own complementary advantages. However, OCT is a more precise method of measuring retinal thickness for early disease detection but expensive and less widely available, CFP is a relatively inexpensive tool that can be used for fundus photography to help ophthalmologists diagnose glaucoma, diabetic retinopathy, cataracts, and other potentially vision-threatening diseases [6].

Artificial Intelligence (AI) is an important tool for detection of eye diseases, through the improved imaging and prediction modelling [7]. Standard machine learning (ML) algorithms such as Fuzzy C-Means (FCM), Random Forest (RF), Support Vector Machine (SVM), Gaussian Mixture Model (GMM), K-Nearest Neighbour (KNN) have been applied in several applications, but they suffer from high computational complexity with high dimensional data, initialization sensitivity and the difficulty in finding optimal parameters [8]. These shortcomings are further increased by high-dimensional and large datasets, class imbalance and robustness to outliers [9]. Furthermore, the occurrence of spurious features and categorical bias usually results in poor predictions and biased results [10]. In this regard, Deep Learning approaches have recently gained popularity for overcoming these shortcomings as they have shown better performance in complex data management and increasing the accuracy of eye disease prediction.

DL is a subset of AI which is based on the learning from huge amounts of data using multi-layered neural networks. Typical DL architectures consist of Convolutional Neural Networks (CNN), Fully Connected Networks (FCN), Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN) [11]. These models generally work better than traditional Machine Learning (ML) methods, but also have a number of challenges. However, DL models are not well suited for complex spatial hierarchies or temporal dependencies [12]. In many applications, training is slow and resource intensive which complicates deployment [13]. Further, the vanishing and exploding gradient problem restricts the learning of long-range dependencies [14]. Also, high-dimensional inputs introduce computational overhead and degrade the performance [15]. To overcome these problems, the proposed multimodal deep learning framework introduces fundus and OCT image fusion, which allows the model to learn the vascular patterns on both the surface and the depth information from the retina layers. Through the use of complementary modes across modalities, the framework minimizes the effect of noisy high-dimensional input and enhances representation learning, which in turn leads to the robustness against gradient-related issues. The joint feature space can achieve faster convergence and better generalisation, which makes the model more efficient and robust compared with traditional single-modality DL methods for predicting eye diseases accurately.

Major contribution of the work is given below.

- A deep learning multimodal framework is proposed that combines fundus and OCT images for the purpose of a comprehensive eye disease detection.
- Weighted Wavelet Visual Perception Fusion (WWPF) is used for the contrast enhancement and visual quality improvement of fundus images for better feature extraction.
- Conditional Philtre Network (CFNet) is applied to filter the noise from the fundus image and retain the vital retinal structures.
- Adaptive Deep Residual Network (AdResNet) is used to remove unwanted regions and enhance image quality of OCT images.

- To enable interpretability for the model, a Grad-CAM is used to visualise the most impacted retinal area in both fundus and OCT images.
- Implemented a novel segmentation pipeline in which MFI-Net is used for blood vessel segmentation of fundus and SGANet for retinal layer segmentation of OCT images.
- Combined Graph-Based Radiomics Feature Extraction Algorithm (GBRF) for structural feature extraction of fundus image with Fully Dense Fusion Neural Network (FDFNN) for deep feature extraction of OCT image provides richer multimodal feature representation than single-methods.
- Implemented a Helformer transformer-based classifier over fused features capturing long-range dependencies and cross-modal correlations better.

The article's sections are organized as follows: Section 2 contains several research articles on ML and DL methods for detecting the eye disease. Section 3 provides a description of the proposed method. The outcomes and performance criteria of the proposed strategy are covered in Section 4. Section 5 offers a summary of the entire framework.

2. LITERATURE SURVEY

Eye disease using fundus and retinal images can be identified in a number of ways. An outline of some of the current detection methods can be seen below.

Menaouer et al., [16] have demonstrated Hybrid DL technique for diabetic retinopathy classification. The established technique employed two VGG network models (VGG16 and VGG19) and the deep CNN approach to identify and categorize diabetic retinopathy based on the visual risk associated with the degree of retinal ischemia. In fact, comprehending the images and their context in relation to the categories was a key component of DR classification. This research were carried out using a large scale of retinal images for accurate eye detection. The outcomes of the experiment, which was conducted on 5584 images from a collection of internet datasets, and these technique contains the accuracy value of 90.90%, F1 score of 94% and Recall of 95% respectively.

Chavan and Pete [17] have introduced the multilevel glowworm swarm optimized convolutional neural network (MGSCNN) for multiclass disease prediction on retinal images. In this framework, the images are collected from the RFMiD dataset for multi-disease classification on eye. Two phases were created for the suggested system: preprocessing and categorization. The images were first scaled, smoothed, and normalized in order to get them ready for preprocessing. To categorize an image as normal or abnormal, the pre-processed images were sent into the MGSCNN classifier, which covers 39 distinct illness categories. Glowworm Swarm Optimizer (GSO) were allows to simultaneously and optimally identify the CNN classifier's structure and hyperparameters. This method has an exceptional accuracy of 95.09%.

Latif et al., [18] have created the Optic Disk localization and Glaucoma Diagnosis Network (ODGNet) for automatically predicting the OD localization and Glaucoma classification using fundus images. For efficient OD localization from the fundus images, a visual saliency map combined with shallow CNN was employed in the first phase and the second stage involves diagnosing glaucoma using pre-trained models based on transfer learning. To evaluate the transfer learning models such as AlexNet, ResNet, and VGGNet, Five public retinal datasets (ORIGA, DRIONS-DB, HRF, RIM-ONE and DR-HAGIS) were used to differentiate among normal and glaucoma. The experimental findings of this research show that the suggested ODGNet assessed on ORIGA for glaucoma diagnosis achieved the accuracy, specificity, sensitivity, and area under the curve of 95.75, 94.90, 94.75, and 97.85%, respectively.

Al-Kahtani et al., [19] have suggested the optimization based Multi-Retinal Disease with Transfer Learning Aided Detection on fundus image. These methods are distinguished between normal and glaucomatous images, which are tested on five publicly available retinal datasets, including ORIGA, HRF, DRIONS-DB, DR-HAGIS, and RIM-ONE. The primary goal of this technique was to utilize the fundus images to identify and categorize various retinal disorders. The Wiener filtering (WF) technique was employed in the DMBOTL-MRDD technique to remove noise. Moreover, feature vector extraction was carried out using ShuffleNetv2 lightweight CNN model. Additionally, the ShuffleNetv2 model's optimal hyperparameter selection was done using the DMBO approach. The Multi-Layer AutoEncoder was used to identify a number of retinal conditions, including glaucoma, AMD, and DR. These model outperforms the accuracy value of 97.12%, Recall value of 88.96% and the precision value of 90.04%.

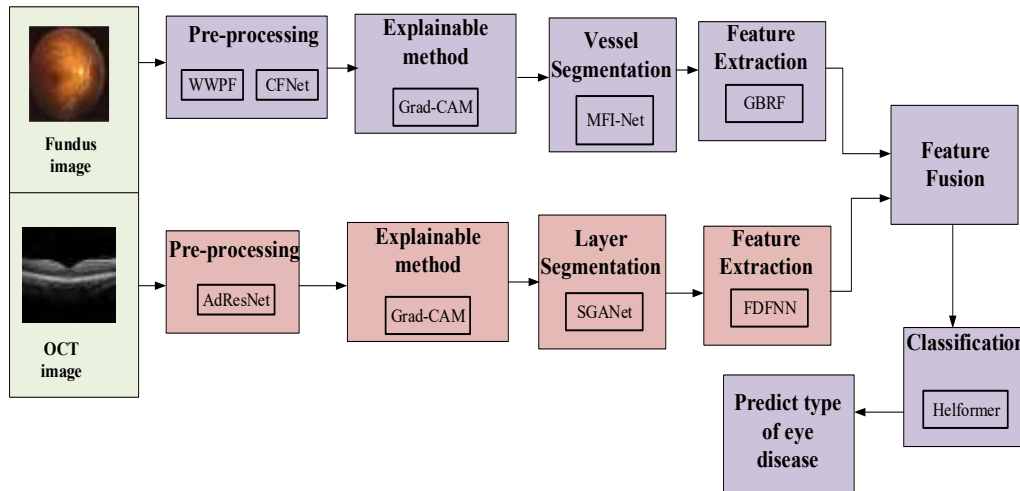
Al-Antary and Arafa [20] have developed the Multi-Scale Attention Network (MSA-Net) for the classification of Diabetic Retinopathy. The suggested technique uses a two datasets such as EyePACS and APTOS for implementing the classification performance on DR. The image of retina was embedded in high-level representation space using the encoder network in the suggested method, where the representation was enhanced by combining mid and high-level features. These model was trained using the cross-entropy loss in a traditional way to classify the DR severity level. The model's ability to discriminate between images of unhealthy retina was enhanced by this surrogate task. These technique shows the accuracy value of 84.6%, Specificity of 90.5% and Sensitivity of 91% respectively.

2.1. Research Gap

According to recent assessments, many modern methods for eye disease detection are still very limited. A prevalent shortcoming is the dependency on a small number of images from existing datasets, which limits the representativeness of models, and reduces the reliability of the diagnostic. In addition, training times for these methods are often very long, thus making them unsuitable for large-scale fundus screening [16]. However, poor generalisation ability, slow convergence, and high computational cost make the optimization of such techniques difficult. For instance, although ODGNet has proven to be effective in some cases, it requires a lot of training time and cannot be used in real-time clinical applications [18]. However, the DMBOTL-MRDD method has the limitation of relying on a very small training set, which limits its generalisation ability over different populations, and needs considerable parameter tuning for proper training [19]. Moreover, in more complex diagnostic applications, manual feature extraction based algorithms tend to underestimate the variability of retinal abnormalities, which in turn results in poor performance and limited adaptability [20]. To address these gaps, this work presents a Helformer transformer-based classifier that exploits multimodal fundus and OCT image fusion to obtain accurate, generalised, and computationally efficient retinal disease classification.

3. PROPOSED METHODOLOGY

According to statistics from the World Health Organization, 2.2 billion people have a near-or distance vision impairment. In addition, AI has proven to be an innovative factor in medical imaging, providing automatic evaluations of imaging modalities. However, there has been limited exploration into interventional research on multi-label disease prediction and predictive models of disease progression. Thus, the proposed model based on the multi-modal deep learning framework for eye disease classification has been developed and the framework is illustrated in figure 1.

Figure 1. Multi-Modal DL Framework for Eye Disease Classification.

The proposed architecture presents a multimodal DL framework for classification of eye diseases by fusing fundus and OCT images. In this framework, the fundus images are first sharpened using Weighted Wavelet Visual Perception Fusion (WWPF), then denoised by a Conditional Filter Network (CFNet) to increase image sharpness. Preprocessing is done through an Adaptive Deep Residual Network (AdResNet) to remove unwanted regions while retaining clinically relevant structures for OCT images. To highlight the disease-affected areas, Grad-CAM is used on both modalities. For segmentation, modality-specific networks are used: MFI-Net is used to segment blood vessels in fundus images, and SGANet is used to delineate the layers of the retina in OCT images. Once segmented, feature extraction is performed using a variety of strategies for each modality. A Graph-Based Radiomics Feature Extraction Algorithm (GBRF) is used to extract deep structural features from fundus images while a Fully Dense Fusion Neural Network (FDFNN) is used to extract comprehensive features from OCT images. In addition, the features extracted from both modalities are fused by a feature fusion process to produce an integrated multimodal representation. Finally, a Helformer transformer-based classifier is used to take the fused feature vector and effectively model long-range dependencies within the multimodal features to predict eye conditions. The model is trained to categorise the retinal status into four classes: Normal, DR, DME and Drusen. The upcoming below sections elaborate the techniques used in the proposed framework.

3.1. Preprocessing of Fundus Image

Preprocessing is employed to improve the digital image quality prior to machine learning or deep learning algorithms processing and analyzing it further in fundus image. In this process, a WWPF technique is utilized for enhancing the contrast level of images and Conditional Filter Network (CFNet) for removing the unwanted pixels from the images.

3.1.1. WWPF

In order to generate a high-quality fundus image, traditional fusion techniques use numerous feature maps to combine many upgraded versions, to produce many features [21]. There are significant differences in the algorithm's resilience and the feature map's effectiveness. In contrast to these fusion approaches, utilizing a WWPF technique for integrating the frequency components of low and high at various scales to provide a

better fundus image. The multi-resolution, multi-scale, and multi-temporal properties of the wavelet transform have been extensively employed in image fusion, denoising, and compression. These can be expressed as in eqn (1).

$$W_{\varphi'}(k, m, n) = \frac{\sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x, y) \phi'_{k, m, n}(x, y)}{\sqrt{M \times N}}, k \in \{V, D, H\} \quad (1)$$

Where, I represents the input image, $W_{\varphi'}(k, m, n)$ represents the detail coefficient, $W_{\varphi}(k, m, n)$ indicates the approximate coefficient, $M \times N$ is the size, $\phi(x, y)$ represents the wavelet function of two dimension, $\varphi(x, y)$ denotes the Gaussian function, V, D and H denotes the Vertical, Diagonal and Horizontal components of high frequency.

While the finer details are found in the high-frequency components, the low-frequency components convey the image's main information and structural components. The decomposed low and high-frequency components are upsampled using the Wavelet Inverse Transform (WIT) to provide a better fundus image. These can be shown as in eqn (2).

$$I_E = \sum_l U_d (WIT_l(I_{Alf}^l, I_{Vlf}^l, I_{Hlf}^l, I_{Dlf}^l)) \quad (2)$$

Where, l denotes the layer in wavelet pyramid, I_E represents the enhanced fundus image, U_d indicates the upsampling operator, Alf is the approximate low frequency, Vlf denotes the low frequency of the vertical samples, Hlf and Dlf are the horizontal and diagonal low frequency of an image. One metric to determine an image's level of detail is the average gradient. These can be expressed as in eqn (3), (4), (5) and (6).

$$\nabla G = \frac{\sum_{i=1}^{M-1} \sum_{j=1}^{N-1} \sqrt{(dx(i, j))^2 + (dy(i, j))^2}}{(M-1)(N-1)} \quad (3)$$

$$\lambda_V = 1 + \frac{\nabla G_{VH}}{\nabla G_{VH} + \nabla G_{HH} + \nabla G_{DH}} \quad (4)$$

$$\lambda_H = 1 + \frac{\nabla G_{HH}}{\nabla G_{VH} + \nabla G_{HH} + \nabla G_{DH}} \quad (5)$$

$$\lambda_D = 1 + \frac{\nabla G_{DH}}{\nabla G_{VH} + \nabla G_{HH} + \nabla G_{DH}} \quad (6)$$

Where, ∇G denotes the average gradient and $\lambda_V, \lambda_H, \lambda_D$ are the vertical, horizontal and diagonal components of reconstructed image. Then the gradient calculation for weighted wavelet fusion and weight calculation can be expressed as in eqn (7).

$$I_{FE} = \sum_l U_d (WIT_l(\lambda_V I_{Alf}^l, \lambda_H I_{Vlf}^l, \lambda_H I_{Hlf}^l, \lambda_D I_{Dlf}^l)) \quad (7)$$

Where, I_{FE} represents the final enhanced fundus image.

3.1.2. Conditional Filter Network

The CFNet is built with an exclusive conditional filter that dynamically deduces the kernels of the local windows determined by the analysis of the noise map and image features [22]. These may be written as in eqn (8).

$$x = CFNet(\theta|y) + y \quad (8)$$

Where, y is the input of the image and x is the noise-free version.

A conditional filter block is a combination of one conditional filter and two convolutional layers. Between the input and the output, there is a skip connection between the image features. The conditional filter uses the noise and corresponding image features to provide an estimate of the denoising kernels within the local windows. These may be written as in eqn (9).

$$\hat{F}_{img}(p, c) = \sum_{q \in N(p, c)} \mu(q, c) \cdot \gamma(q, c) \cdot F_{img}(q, c) \quad (9)$$

Where, F_{img} is the input features of the image, $\hat{F}_{img}$ is the output. c is the index of channel, $N(p, c)$ is the local window, p, q are the spatial positions, and m, g are the respective kernels.

3.1.3. GRAD-CAM

GradCAM is an innovative approach to understanding the ConvNet and deep learning in general [23]. It creates a heatmap that highlights exact regions of the input image that were responsible for the greatest effect on the model's output. Additionally, this heatmap is applied to the input image, which allows the model's prediction to be visually recognized. The purpose of this method is to generate a preliminary localization map by utilizing the gradients of the model's output in relation to the feature mappings in a convolutional layer. Using the compiled and rescaled image of interest, the input image is classified using the compiled features from the model. GradCAM heatmaps, which are content relevance focused maps over the image, are created using this as one of the inputs. GradCAM highlights the regions of interest on the affected region of eye which helps to identify the disease of eye regions. The average pooling of gradient calculation can be shown as in eqn (10).

$$a_k^c = \frac{1}{z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{k_{ij}}} \quad (10)$$

Where, a_k^c denotes the weight, y^c represents the class score and A_k is the feature map. The heatmap for the Grad-CAM can be computed as in eqn (11).

$$L_{Grad-CAM}^c = ReLU(\sum_k a_k^c A_k) \quad (11)$$

Where, $L_{Grad-CAM}^c$ denotes the heatmap of the Grad-CAM and $ReLU$ represents the Rectified linear unit.

3.1.4. Vessel Segmentation

Retinal vessel segmentation on fundus images is a critical and vital step in the detection and assessment of the risk of ocular and microvascular condition and involves the division of the area of blood vessels in the fundus image with the help of MFI-Net.

i) MFI-Net

The MFI-Net is encoder-decoder based and has a U-shaped structure comprising of PSE and C2F modules [24]. The adaptation to multiple vessel widths of the networks through the use of multiscale information extraction and attention learning by the PSE module. These network can pay adaptive attention to the vessels of different width because the use of multiscale information extraction and attention learning by the PSE module. The goal of the C2F module is to conserve features of the vessels during decoding and reduce the semantic gap between the encoder and the decoder. Deep supervision and feature fusion are used on the decoder blocks output to make the segmentation more accurate.

As the U-Net has connections linking features that are skipped across the encoder and decoder, they cannot overcome the semantic gap between these two modules. Under this change, the C2F module improves the decoder in 3 phases. Prior to passing through a convolutional block and the PSE module, the up-sampled feature maps are concatenated with the encoder feature maps to create F_{deco} . After a convolutional layer with batch normalization, the up-sampled feature maps are offset by the encoder-generated feature maps to generate the residual feature maps, or F_{res} . Once F_{deco} and F_{res} are concatenated, the result of the C2F module is then eventually available after being transferred through the PSE module and a convolutional block. The semantic aspects that the encoder and decoder conflict on due to F_{res} can be found through the C2F module. These attributes are then concatenated to the decoding process to add value to the information. The loss of segmentation is due to the fact that the MFI-Net results will provide segmentation at three scales. They can be represented as in eqn (12).

$$L = \sum_{i=1}^3 \lambda_i \times L_{BCE}(Y_i, Y_i') \quad (12)$$

In which, L_{BCE} is the binary cross-entropy, Y_i is the result of the segmentation, $Y_i Y_i'$ means the ground truth, and λ_i is the weighting parameter. The results of all

decoders are passed through a convolutional layer using batch normalization followed by an upsampling to recover the input size.

3.1.5. Feature Extraction Of Fundus Image

For the feature extraction phase, utilizing the technique of GBRF for extracting the relevant information from the segmented image.

i) Graph-Based Radiomics Feature Extraction Algorithm

Radiomics features are often divided into groups according to their characteristics and the features of the image [25]. This technique introduces a new category within radiomics beyond the previously listed ones by employing a graph-based technique to extract the vascular network from fundus images which demonstrate blood vessels.

In radiomics, feature extraction refers to computations in which feature descriptors are employed to quantify attributes of the ROI, such as pixel spacing, gray-level intensities, gray-level histogram bins, and so on. The formula of GBRF can be expressed as in eqn (13) and (14).

$$G = (V, E) \quad (13)$$

$$E \subseteq \{(u, v) | u, v \in V\} \quad (14)$$

Where, E represents the edges, V denotes the node and G denote the Graph. T-points are used to compare the allocated SEs of class 1 for bifurcation and 0 for non-bifurcation. The coordinates of the center pixel of the SE at every position in the skeletonized image are recovered when the SE fits, that is, if each pixel in the SE is connected to the corresponding pixel of the neighborhood under the SE at that location. The feature fragment to be retrieved from the grayscale image is then centered on these coordinates and the feature fragment is then binarized and segmented. It employs a two-class categorization strategy in this work: one class comprises overlaps, noisy fragments, and non-bifurcations, while the other class is group bifurcations.

3.2. Preprocessing of OCT Image

For the process of cleaning the unwanted noises to improve the quality of the Optical Coherence Tomography (OCT), the AdResNet is employed.

3.2.1. AdResNet

AdResNet is an image denoising model that uses residual learning and deep CNN networks [26]. The research goes on to provide a step-by-step explanation of the procedure, supported by mathematical formulas. It can be shown in eqn (15).

$$y = x + v \quad (15)$$

Where, v is the addition of noise in an image, y is the noisy image, and x is the noise removal image.

To normalize the output of each layer, batch normalization re-centers and re-scales the activations. The explanation can be expressed as in eqn (16).

$$\hat{F}_l(X) = \frac{F_l(X) - \mu}{\sqrt{\sigma^2 + \epsilon}} \quad (16)$$

Where, ϵ is a tiny constant, μ is the mean and σ^2 is the variance of mini-batch during training.

After completing normalization, a scaling and shifting procedure is expressed by eqn (17).

$$\hat{F}_l(X) = \gamma \hat{F}_l(X) + \beta \quad (17)$$

Where, γ and β are learned scaling and shifting parameters.

When the MSE is smaller, the denoised image is more like the original. These can be explained as in eqn (18).

$$MSE = \frac{1}{m.n} \sum_{i=1}^m \sum_{j=1}^n (I_{original}(i, j) - I_{denoised}(i, j))^2 \quad (18)$$

Where, $I_{original}(i, j)$ and $I_{denoised}(i, j)$ are the pixel values at position (i, j) , m and n are the image's dimensions.

3.2.2. Layer Segmentation

For automatically recognizing and defining the borders of various retinal layers inside an optical coherence tomography (OCT), the Semantic Graph Attention Network (SGANet) is employed and explained in the below section.

i) Semantic Graph Attention Network

A standard FCN is used as a backbone and the 3D UNet is employed to implement the network and validate the quadrant-segmentation network [27]. Although the traditional FCNs are not especially concerned with the spatial relationships among two or more objects that are being segmented, this method relies on the SGA system to explain the anatomical topology of the retina. To train the discriminative voxel-wise feature embeddings that can distinguish between fine-grained retinal demarcation, the SGA specifically uses two crucial components. The former is a set of global-context attention (GCA) modules for more effectively attending to segmentation that combine the local details with semantic feature embeddings, and the latter is a graph-based attention (GBA) module that reduces confusion with distinguishing between adjacent regions that have similar appearances. These may be written as in eqn (19).

$$X_{class} = \sigma(\varphi(X; W_C)); X_{hidden} = \sigma(\varphi(X; W_M)) \quad (19)$$

Where, X_{class} denotes the possibility of X for the specific class, $\varphi(.)$ denotes the convolution operation, C is the channel, M denotes the feature map size, X_{hidden} indicates the hidden representation of X , W represents the weight matrix and $\sigma(.)$ is the activation function.

By combining the voxel-wise hidden representations, one may easily determine the first prototype representations of various anatomical objectives. These can be derived as in eqn (20).

$$F^0 = \tilde{X}_{class}^T \times \tilde{X}_{hidden} \quad (20)$$

Where, F^0 represents the target of anatomy,

Regarding the adjacency matrix given to the original F^0 , GCNs can be further implemented by explicitly showing the spatial relationship among various targets of the anatomy. This will enhance the semantic data that may be stored in each prototype representation. In particular, forward path of every layer in GCN can be represented as in eqn (21).

$$F^i = \sigma(AF^{i-1}W_i) \quad (21)$$

Where, A denotes the adjacency matrix and $\sigma(.)$ represents the activation function. To update the segmentation encoder's feature embedding, the output of GCN can be shown as in eqn (22).

$$\tilde{X}^* = \sigma_{softmax}(\tilde{X}_{class}) \times \tilde{F} \quad (22)$$

Where, $\tilde{X}^*$ denotes the reshaped size, $\tilde{F}$ represents the GCN's output and $\sigma_{softmax}(.)$ indicates the softmax normalization. Finally, each global-context attention module's output characteristics may be described as eqn (23).

$$X_{gca} = X + \sigma_{sigmoid}(F'' \cdot X) \quad (23)$$

Where, $\cdot$ denotes the multiplication in channel-wise, F'' represents the weighted vector, X_{gca} is the global-context attention module and $\sigma_{sigmoid}$ denotes the sigmoid function.

3.2.3. Feature Extraction Of OCT Image

To identify the structural features and textual patterns of the OCT image, the Fully Dense Fusion Neural Network (FDFNN) is employed and illustrated as a brief in the below section.

i) Fully Dense Fusion Neural Network

The FDFNN architecture consist of two fully connected layers with 60 million free parameters, three pooling layers, and five convolution layers based on the AlexNet architecture[28]. In the proposed method FDFNN was utilized for feature extractor model for the OCT images. The CNN is optimized by reducing the difference between the real and output values. Gradient-based adjustments are often employed in CNNs. Based on the stochastic optimization technique, Adam optimization was applied to work effectively with large parameters.

One of the hyperparameters in CNNs that significantly affects the network's training is the learning rate. The network is more affected by input data at higher learning rates, whereas the long-term learning is accomplished on the network at lower learning rates. Therefore, learning rate could be the most important hyperparameter to consider while building the neural network. The feature maps from the convolution layers progressively appear towards the conclusion of the FDFNN training process. In this method, all of the FDFNN features were retrieved for classification by activating the final fully connected layer. After these process, the both OCT and fundus image features can be fused for performing the classification task.

3.3. Classification

In this proposed framework, the Helformer based classifier is employed for classifying the fused features of OCT and the fundus based on the eye diseases.

3.3.1. Helformer

The Helformer uses a single encoder component instead of two parts. This Helformer encoder structure consists of a dense layer, an LSTM component, residual connections, an attention mechanism, and a sequence of breakdown blocks [29]. Without losing the prediction accuracy, uses a single Transformer design structure lowers computing resource consumption and model complexity memory bottlenecks. Optuna is used to further optimize the model, adjusting hyperparameters like learning rate, dropout rate, and attention head count to ensure optimal performance. Furthermore, the neural network model incorporates the exponential smoothing coefficients directly, for allowing the improvement with additional parameters using the same model optimizer. Holt-Winters smoothing is used in the decomposition block to identify important parameters. These can be shown in eqn (24), (25) and (26).

$$L_t = \alpha \frac{X_t}{S_t} + (1 - \alpha)L_{t-1} \quad (24)$$

$$S_{t+m} = \gamma \frac{Y_t}{L_{t+1}} + (1 - \gamma)S_t \quad (25)$$

$$Y_t = \frac{X_t}{S_t L_t} \quad (26)$$

Where, S_t denotes the seasonality, L_t is the level, X_t represents the datapoint, m denotes the weighted mean, t is the time point and $t - 1$ represents the previous time point, α and γ are the local parameters.

In order to make the sequential data more appropriate for the attention mechanism, it has been adjusted to remove seasonal changes and unnecessary information. Specifically, self-attention configuration of the transformer model can be expressed as in eqn (27), (28), (29).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (27)$$

$$Multihead(Q, K, V) = Concat(head_1, \dots, head_h)W^O \quad (28)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (29)$$

Where, Q is the query matrix, K is the key matrix, V stands for the value matrix and d is the hidden dimension of keys. The pseudocode for the identification of retinal disease using fundus and OCT image can be shown below.

Algorithm 1: Pseudo code for Eye disease prediction using fundus and OCT image
Input: Fundus and OCT image based eye disease prediction (AT) and (GV) # Preprocessing of Fundus image { VN=WWPF(AT) // remove noise in the fundus images using eqn (6) DB= CFNet (VN)// enhance the clarity of the fundus image using eqn (8) } #Preprocessing of OCT image { QC=AdResNet (GV)// Denoising of OCT image using eqn (17) } #Explainable Method { SX=GradCAM (DB), (QC)// Highlights the affected region of eye in both fundus and OCT image using eqn (10) } # Vessel Segmentation { HQ= MFI-Net (SX)// Segment the blood vessel from the fundus image using eqn (11) } #Layer Segmentation { WZ= SGA Net (SX)// Segments the layer in OCT image using eqn (22) } # Feature Extraction { JI= GBRF (HQ)// Extracting the features of fundus image using eqn (13) LO= FDFNN (WZ)// Extracts the most relevant features from OCT image } # Classification RE= Helformer(JI, LO))// Classify the image to predict the eye disease using eqn (28) If Class=0 Normal Else Class=1 Diabetic Retinopathy If Else Class=2 Diabetic Macular Edema Else If Class=3 Drusen Output: Accurately detect the eye disease from the images

END

4. RESULT AND DISCUSSION

The proposed Helformer-based multimodal framework fuses the fundus and OCT images for precise eye disease classification. Grad-CAM was included in the preprocessing step to mark disease-affected regions to boost the quality of the subsequent segmentation and feature extraction. The combined features from both modalities were then categorised by Helformer model. Experimental results show that the proposed framework achieves better performance than the existing approaches in terms of accuracy, robustness, and generalisation. The system was developed in Python 3.8 on a system containing an Intel Core i5 CPU, 16 GB RAM and NVIDIA GeForce GTX 1650 GPU. Table 1 illustrates the Modelling parameters for simulations performed in the experiments.

Table 1. Simulation parameter for WWPF, CFNet, Grad-CAM, MFI-Net, GBRF, SGANet, FDFNN and Helformer

Parameter	Methods	Values
batch_size	WWPF	16
optimizer		Adam
Learning Rate		3e-5
epochs		50
batch_size	CFNet	32
learning_rate		1e-3
optimizer		Adam
epochs		80
batch_size	Grad-CAM	32
learning_rate		1e-3
optimizer		Adam
epochs		50
batch_size	MFI-Net	16
learning_rate		1e-3
optimizer		Adam
epochs		120
batch_size	GBRF	16
learning_rate		1e-4
optimizer		Adam
epochs		100
batch_size	SGANet	16
optimizer		Adam
Learning Rate		1e-4
epochs		150
batch_size	FDFNN	32
optimizer		Adam
Learning Rate		1e-3
epochs		80
batch_size		8

optimizer	Helformer	Adam
Learning Rate		1e-4
epochs		100

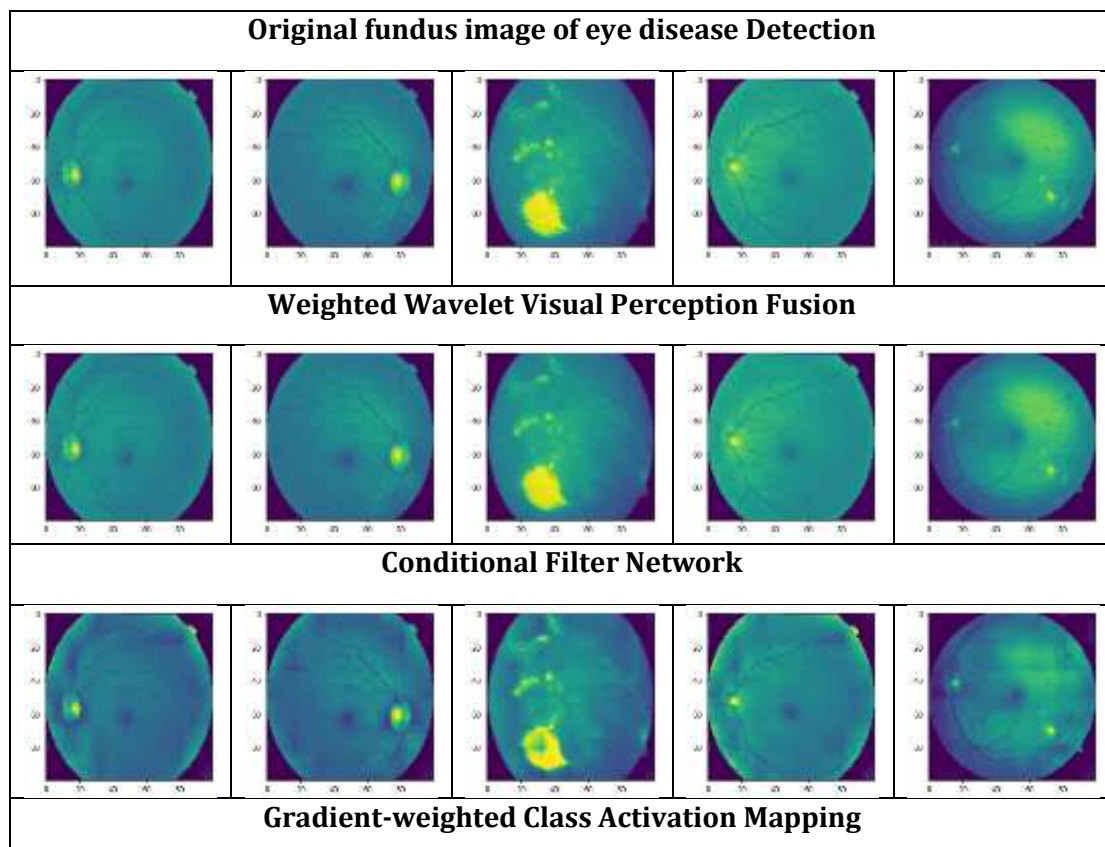
The simulation parameters forWWPF, CFNet, Grad-CAM, MFI-Net, GBRF, SGANet, FDFNN and Helformer are listed in Table 1, which includesbatch_size, learning_rate, optimizer and epochs.

4.1.Dataset Description

In this dataset [30], the 1000 fundus images, which are divided into 39 groups, were obtained from the Joint Shantou International Eye Centre (JSIEC), which is located in Shantou, Guangdong province, China. The 209,494 fundus images that will be utilized to train, validate, and test a deep learning system are far larger than these images.

In this dataset [31],the high-resolution cross sections of the retinas of living patients are obtained using an imaging technique called retinal optical coherence tomography. An estimated 30 million OCT scans are performed annually, and interpreting and analyzing the data takes a lot of effort. These two dataset fused together for classifying the eye disease accurately. There are 1,400samples in the data of eye disease prediction, out of which 273 (20%) are used for testing and 1,127 (80%) are used for training.

Figure 2.Sample Images for the detection of eye disease in fundus image.



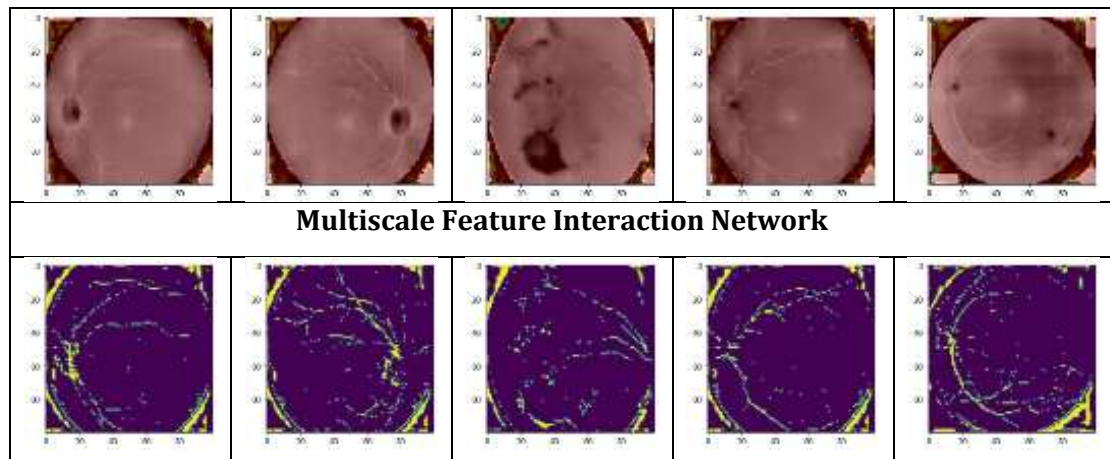
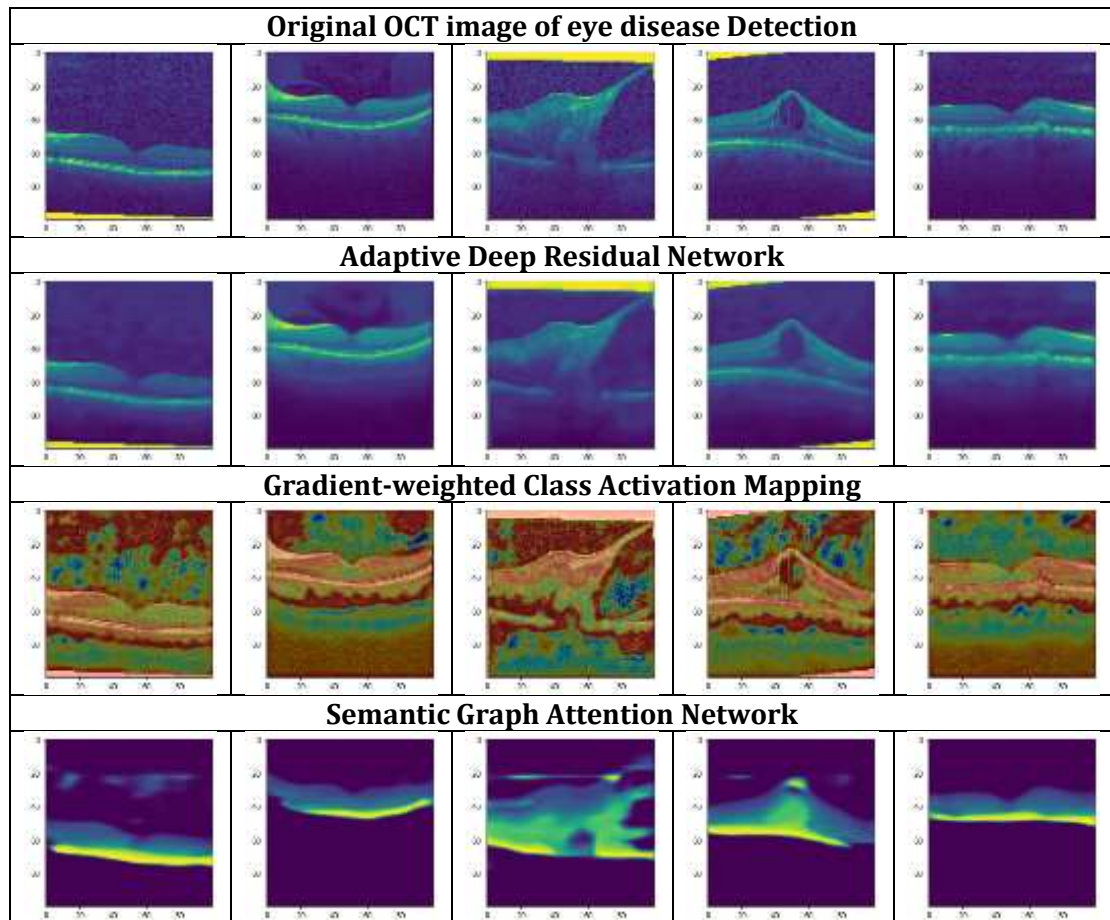


Figure 2 describes the samples of fundus image for eye disease prediction. The original fundus image of eye are displayed in the first row. The second row contains a WWPF for improving the quality of images, which is based on gradients. The third row shows the CFNet for denoising of fundus image. Then the GradCAM were shown by the fourth row, which highlights the affected region of eye. Finally the last row describes the output of the Multiscale Feature Interaction Network to segment the blood vessels efficiently.

Figure 3. Sample Images for Eye disease detection using OCT image.



The OCT image samples used to forecast eye diseases which are shown in Figure 3. The first row shows the original image of the OCT eye disease prediction. A gradient-based Adaptive Deep Residual Network is used in the second row to eliminate the noise. The GradCAM can highlight the eye's afflicted area, which is seen in the third row. The output of the SGANet to effectively segment the retinal layer of eye is finally described in the last row.

Figure 4. AUC and ROC plots for the proposed techniques.

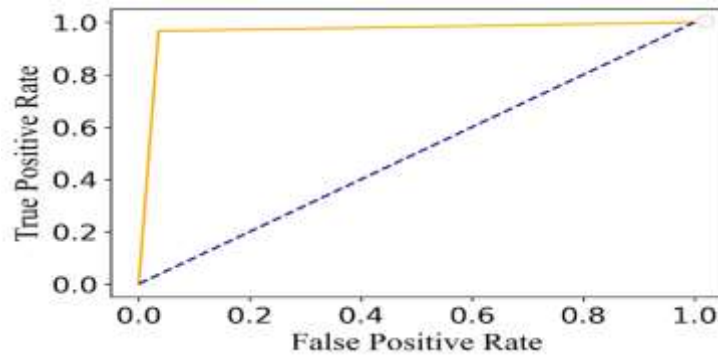
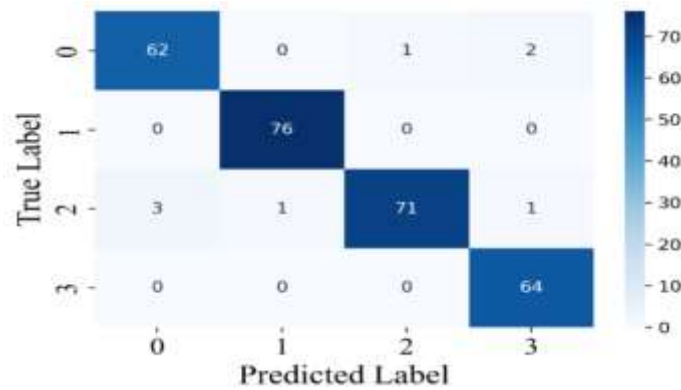


Figure 4 displays the AUC and ROC curves for the proposed methods. The ROC curve, which shows the True Positive Rate against the False Positive Rate, illustrates the model's capacity for class discrimination. The blue dashed line shows the random guessing baseline, while the orange curve shows the classifier's performance. A curve in the upper-left corner indicates improved model accuracy.

Figure 5. Confusion matrix of the proposed Helformer transformer-based classifier.



The confusion matrix for the proposed approach is shown in Figure 5. The effectiveness of the model is evaluated by the analysis of four distinct classes generated based on eye disease. Class 0 (Normal) attained 62 true predictions as per the actual label, Class 1 (Diabetic Retinopathy) got 76, Class 2 (Diabetic Macular Edema) attained 71 and class 3 (Drusen) attained 64. This results that the proposed model can accurately classify the disease for retinal disease prediction.

4.2. Comparison Analysis for Classification

In this comparison analysis, the proposed technique of Helformer transformer-based classifier will be compared with other existing techniques such as DMBOTL-MRDD [19],

ODGNet [18], MGSCNN [17] and Hybrid DL [16], in order to predict the effectiveness of the proposed model. The performance metrics used in these evaluation which includes Accuracy, Precision, Recall, Specificity, Negative Positive Rate(NPV), F1-score, Error, FDR, False Negative Rate (FNR), FPR and FOR, the graph of these metrics can be shown below.

Figure 6. Comparison of Helformer Accuracy among different Algorithms.

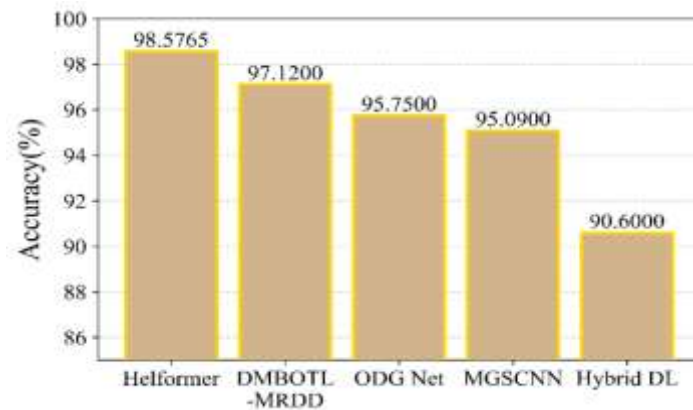


Figure 7. Precision value for Helformer compared to other existing methods.

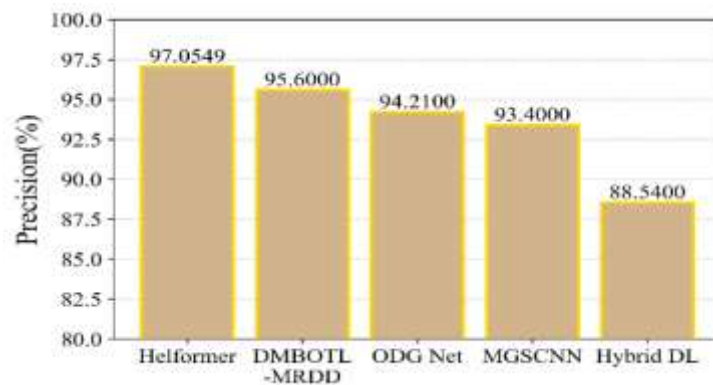
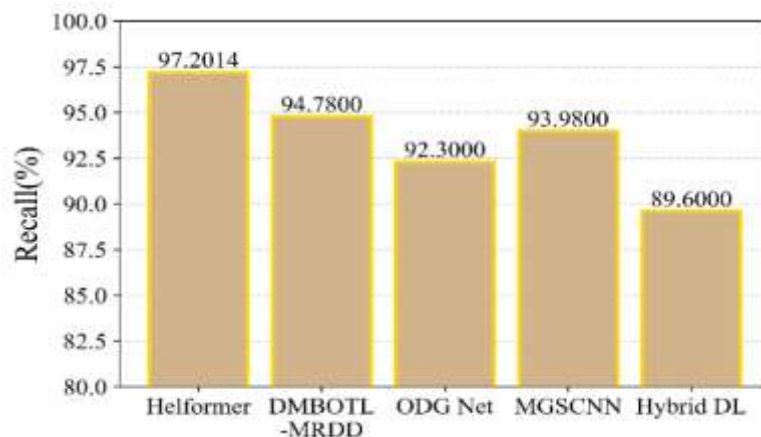


Figure 8. Recall for Helformer compared to other existing algorithms.



Figures 6, 7 and 8 provide a comparison of accuracy, precision and recall for various methods. The proposed Helformer-based classifier achieves accurate classification accuracy of 98.57%, precision of 97.05%, and recall of 97.20% which surpasses the existing methods such as DMBOTL-MRDD (97.12%, 95.60%, 94.78%), ODGNet (95.75%, 94.21%, 92.30%), MGSCNN (95.09%, 93.40%, 93.98%), and Hybrid DL (90.6%, 88.54%, 89.6%). These findings clearly demonstrate that the proposed strategy performs better than the other approaches across all three parameters. The superior performance is due to the fact that the framework can capture both global and local dependencies and, consequently, provides more reliable and accurate classification of retinal diseases.

Figure 9. Value of specificity for Helformer compared to other existing methods.

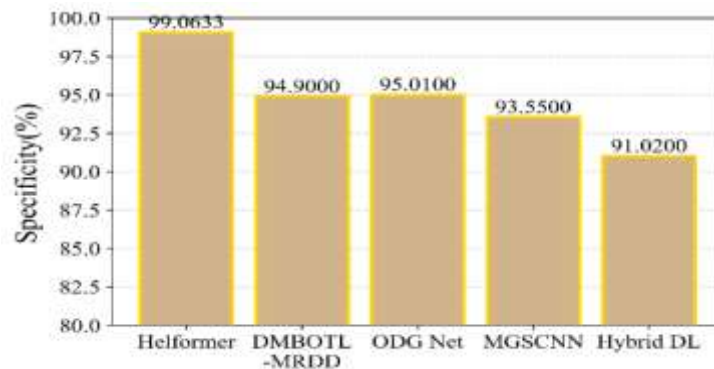


Figure 10. NPV value for Helformer compared to other existing methods.

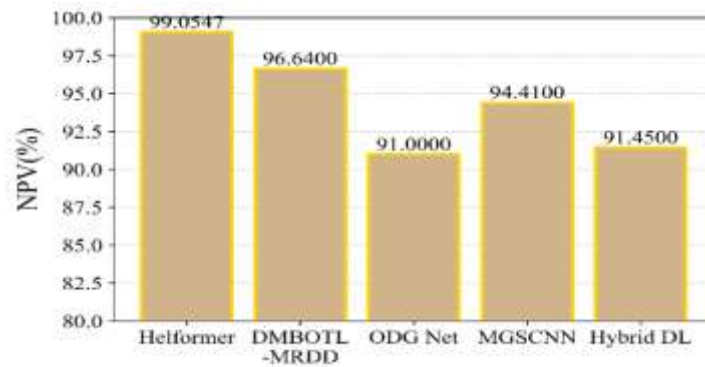
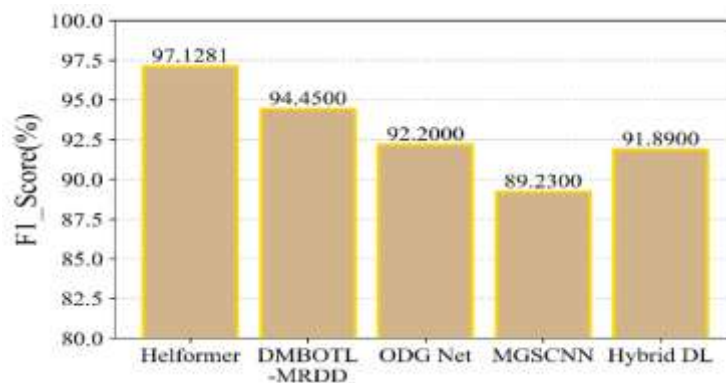


Figure 11. Helformer of F1_Score value with other current algorithms.



Figures 9, 10 and 11 show the comparison of specificity, negative predictive value (NPV), and F1 score by method. The proposed Helformer model obtains a specificity of 99.06%, NPV of 99.05%, and F1-score of 97.12% - all of which are superior to the performance of existing methods. In comparison, the existing methods such as DMBOTL-MRDD, ODGNet, MGSCNN and Hybrid DL achieved specificity values of 94.90%, 95.01%, 93.55%, and 91.02% respectively, NPV values of 96.64%, 91.00%, 94.41%, and 91.45% respectively, and F1-scores of 94.45%, 92.20%, 89.23%, and 91. These experimental results indicate the excellent discriminative power of the proposed framework, which can extract and combine the meaningful features to improve the classification robustness of retinal disease detection.

Figure 12. Helformer's error value compared to other existing techniques.

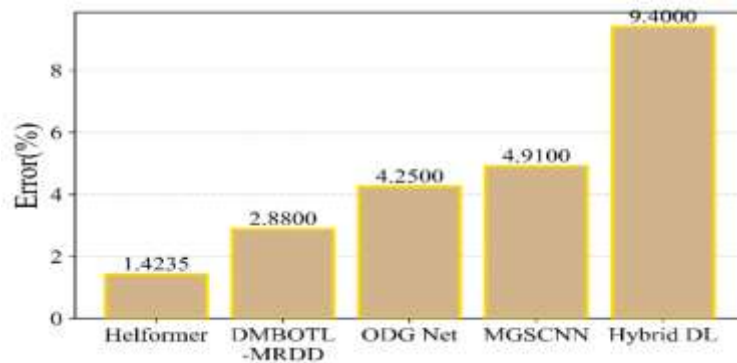


Figure 13. Value of FDR for Helformer compared to other existing methods.

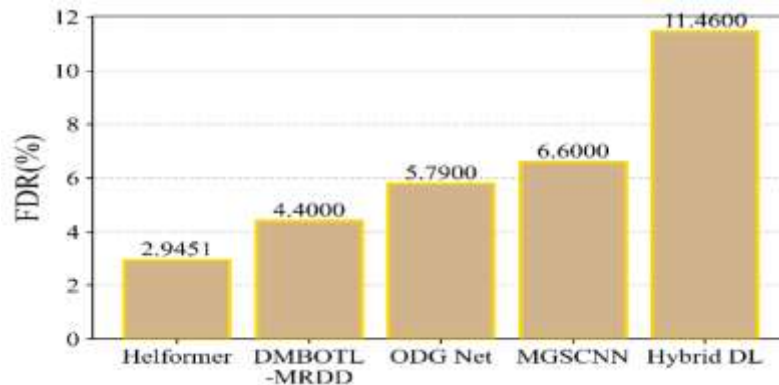


Figure 14. FNR evaluation for Helformer with other current techniques.

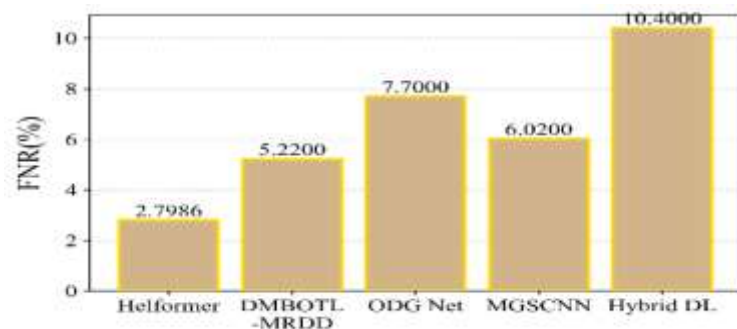


Figure 12 shows the comparison of error rate, in which the Helformer model attains an error value of 1.42% which is much lower than DMBOTL-MRDD (2.88%), ODGNet (4.25%), MGSCNN (4.91%), and Hybrid DL (9.40%). Similarly, in Figure 13 the false discovery rate (FDR) comparison shows Helformer with a value of 2.94%, where Helformer is better than DMBOTL-MRDD (4.40%), ODGNet (5.79%), MGSCNN (6.60%), and Hybrid DL (11.46%). Figure 14 shows the false negative rate (FNR) where the proposed model has proved its superiority in terms of the FNR with a value of 2.79% compared to 5.22% (DMBOTL-MRDD), 7.70% (ODGNet), 6.02% (MGSCNN), and 10.40% (Hybrid DL). These results show that the proposed method is robust, stable and performs well, even with redundant information or noisy data, hence providing a more reliable classification of retinal diseases.

Figure 15. Value of FPR for Helformer in comparison to other existing methods.

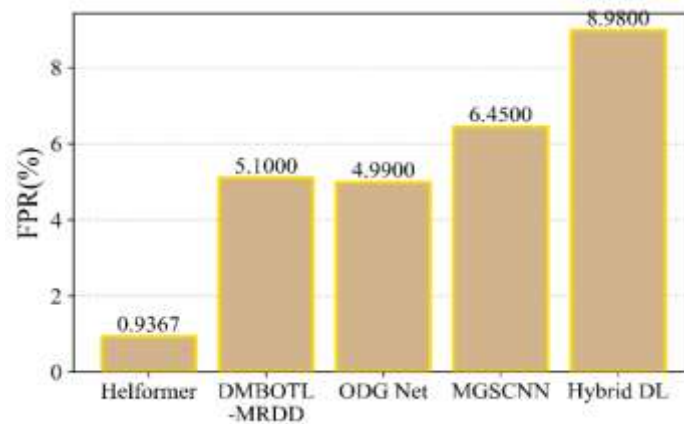


Figure 16. Value of False Omission Rate for Helformer compared to other existing methods.

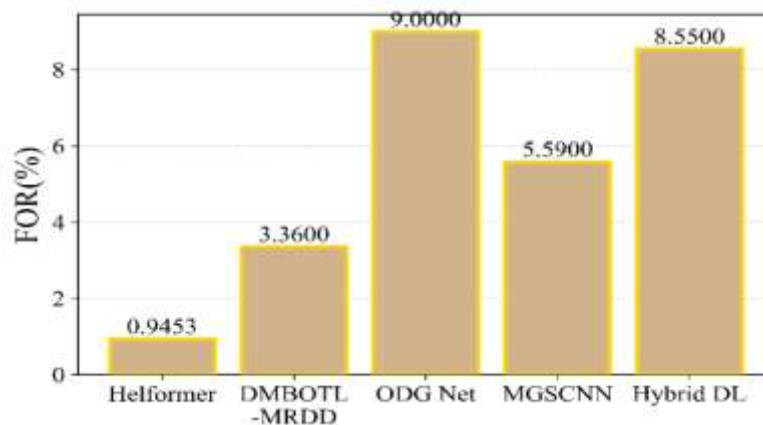


Figure 15 shows the comparison of false positive rate (FPR), in which the proposed Helformer classifier has a value of 0.93%, which is better than DMBOTL-MRDD (5.10%), ODGNet (4.99%), MGSCNN (6.45%), and Hybrid DL (8.98%). Similarly, Figure 16 shows the comparison of false omission rate (FOR) with Helformer having a significantly lower value of 0.94% than 3.36% (DMBOTL-MRDD), 9.00% (ODGNet), 5.59% (MGSCNN), and 8.55% (Hybrid DL). These results show the effectiveness of the proposed framework,

where transformer-based attention combined with hierarchical learning avoids overfitting and improve generalisation to unseen data.

Table 2. Ablation study for the proposed fusion model.

Model Variant	Accuracy (%)	Precision (%)	NPV (%)
Fundus-only (WWPF + CFNet + MFI-Net + GBRF)	94.85	92.40	95.62
OCT-only (AdResNet + SGANet + FDFNN)	95.33	93.12	96.10
Multimodal Fusion (without Helformer)	96.84	94.72	97.25
Full Model (Fundus + OCT + Helformer)	98.57	97.05	99.05

Table 2 shows the ablation study on the proposed fusion model to analyse the significance of each component. The Fundus-only pipeline (WWPF + CFNet + MFI-Net + GBRF) has an accuracy of 94.85%, precision of 92.40%, and NPV of 95.62% while the OCT-only pipeline (AdResNet + SGANet + FDFNN) slightly improves the results to 95.33% accuracy, 93.12% precision, and 96.10% NPV. A combination of both modalities via multimodal fusion without the Helformer classifier further improves the performance to 96.84% accuracy, 94.72% precision, and 97.25% NPV. Finally, the full model combining Fundus, OCT and Helformer classifier, gets the best performance with 98.57% accuracy, 97.05% precision and 99.05% NPV. These results validate that multimodal integration and the Helformer transformer-based architecture provide a considerable improvement in classification compared to the single-modality methods.

5. CONCLUSION

Retinal diseases are diseases that affect the retina (the back of the eye-the light-sensitive tissue), resulting in partial or complete vision loss. Early detection is crucial, but conventional manual diagnostic techniques are often insensitive in identifying retinal disorders at their earliest stages and hence leading to a delay in treatment. To overcome these limitations, presented a Helformer transformer-based classifier for correct classification of eye diseases based on multimodal inputs from fundus and OCT images. Fundus images are denoised with CFNet and enhanced using WWPF while OCT images are denoised with AdResNet. Furthermore, both preprocessed modalities are fine-tuned with Grad-CAM to reveal the disease affected areas. Segmentation is then performed using MFI-Net for fundus vessel extraction and SGANet for retinal layer segmentation in OCT images. Feature extraction is next where GBRF is used for fundus images and FDFNN for OCT images to retain only the most relevant features. These extracted features are concatenated into a single vector and fed into the Helformer classifier for final prediction across four classes of Normal, Diabetic Retinopathy, Diabetic Macular Edoema, and Drusen. The proposed model has an accuracy of 98.57%, precision of 97.05%, and recall of 97.20%, which obviously outperforms existing methods. Based on these results, the proposed Helformer-based framework demonstrates robustness and efficacy for early and reliable diagnosis of retinal disease.

Limitation and Future Scope

Although the proposed Helformer-based framework achieves an overall superior performance in the retinal disease classification, there are still limitations. The resulting model is less generalizable due to dependency on the dataset as population differences and differences across imaging devices are not well represented. Moreover, the multi-stage preprocessing, segmentation and fusion process is computationally expensive, which prevents its clinical deployment in real-time. While Grad-CAM does offer

interpenetrability to some degree, the decision making process of transformers is still complicated and less transparent, posing difficulties for clinical validation. The application domain for this model is also restricted to four retinal diseases and the use of several dedicated networks raises the risk of overfitting in cases where the datasets are small or have mismatched sizes. Future work will include expanding multi-institutional and multi-ethnic datasets, creating lightweight architectures to enable deployment in portable devices, and incorporating novel methods of explainable AI to increase clinical trust. Important directions are extending the model to include additional classes of retinal and systemic disease, applying cross-modality and self-supervised learning approaches for robust feature learning and making the model incrementally updateable for online adaptation. Finally, it will be possible to further improve the practical throughput of the framework, by testing it in a real-world ophthalmology workflow, and by integrating it into tele-ophthalmology interfaces.

ACKNOWLEDGEMENT

Funding. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Conflict of Interest. The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Availability of data and material. Not applicable

Code availability. Not applicable

Author contributions. The corresponding author claims the major contribution of the paper including formulation, analysis and editing. The co-authors provides guidance to verify the analysis result and manuscript editing.

Compliance with ethical standards. This article is a completely original work of its authors; it has not been published before and will not be sent to other publications until the journal's editorial board decides not to accept it for publication.

Clinical Trial Number. Not applicable

REFERENCES

- [1] Rodríguez, M. A., AlMarzouqi, H., & Liatsis, P. (2022). Multi-label retinal disease classification using transformers. *IEEE Journal of Biomedical and Health Informatics*, 27(6), 2739–2750.
- [2] Lam, B. L., Leroy, B. P., Black, G., Ong, T., Yoon, D., & Trzupek, K. (2021). Genetic testing and diagnosis of inherited retinal diseases. *Orphanet Journal of Rare Diseases*, 16(1), 514.
- [3] Heath Jeffery, R. C., Mukhtar, S. A., McAllister, I. L., Morgan, W. H., Mackey, D. A., & Chen, F. K. (2021). Inherited retinal diseases are the most common cause of blindness in the working-age population in Australia. *Ophthalmic Genetics*, 42(4), 431–439.
- [4] Holtes, L. K., de Bruijn, S. E., Cremers, F. P. M., & Roosing, S. (2025). Dual inheritance patterns: A spectrum of non-syndromic inherited retinal disease phenotypes with varying molecular mechanisms. *Progress in Retinal and Eye Research*, 104, 101308. <https://doi.org/10.1016/j.preteyeres.2024.101308>
- [5] Roman, A. J., Cideciyan, A. V., Wu, V., Garafalo, A. V., & Jacobson, S. G. (2022). Full-field stimulus testing: Role in the clinic and as an outcome measure in clinical trials of severe childhood retinal disease. *Progress in Retinal and Eye Research*, 87, 101000.
- [6] Gong, J., Cheung, S., Fasso-Opie, A., Galvin, O., Moniz, L. S., Earle, D., Durham, T., et al. (2021). The impact of inherited retinal diseases in the United States of America (US)

- and Canada from a cost-of-illness perspective. *Clinical Ophthalmology*, 15, 2855–2866.
- [7] Tayal, A., Gupta, J., Solanki, A., Bisht, K., Nayyar, A., & Masud, M. (2022). DL-CNN-based approach with image processing techniques for diagnosis of retinal diseases. *Multimedia Systems*, 28(4), 1417–1438.
- [8] Hemalakshmi, G. R., Murugappan, M., Sikkandar, M. Y., Begum, S. S., & Prakash, N. B. (2024). Automated retinal disease classification using hybrid transformer model (SViT) using optical coherence tomography images. *Neural Computing and Applications*, 36(16), 9171–9188.
- [9] Sheet, S. S. M., Tan, T.-S., As'ari, M. A., Hitam, W. H. W., & Sia, J. S. Y. (2022). Retinal disease identification using upgraded CLAHE filter and transfer convolution neural network. *ICT Express*, 8(1), 142–150.
- [10] Yan, A. L., Du, S. W., & Palczewski, K. (2023). Genome editing, a superior therapy for inherited retinal diseases. *Vision Research*, 206, 108192.
- [11] Upadhyay, P. K., Rastogi, S., & Kumar, K. V. (2022). Coherent convolution neural network based retinal disease detection using optical coherence tomographic images. *Journal of King Saud University - Computer and Information Sciences*, 34(10), 9688–9695.
- [12] Schofield, D., Kraindler, J., Tan, O., Shrestha, R., Jelovic, D., West, S., Ma, A., Grigg, J., & Jamieson, R. V. (2022). Patient-reported health-related quality of life in individuals with inherited retinal diseases. *Ophthalmology Science*, 2(1), 100106.
- [13] Kayadibi, İ., & Gürakın, G. E. (2023). An explainable fully dense fusion neural network with deep support vector machine for retinal disease determination. *International Journal of Computational Intelligence Systems*, 16(1), 28.
- [14] Chowh, S. S., Bhuiyan, M. R. I., Payel, I. J., Karim, A., Khan, I. U., Montaha, S., Hasan, M. Z., Jonkman, M., & Azam, S. (2025). A low complexity efficient deep learning model for automated retinal disease diagnosis. *Journal of Healthcare Informatics Research*, 9(1), 1–40.
- [15] Woof, W. A., de Guimarães, T. A. C., Al-Khuzaei, S., Daich Varela, M., Sen, S., Bagga, P., Mendes, B., et al. (2025). Quantification of fundus autofluorescence features in a molecularly characterized cohort of >3500 patients with inherited retinal disease from the United Kingdom. *Ophthalmology Science*, 5(2), 100652.
- [16] Menaouer, B., Dermane, Z., Kebir, N. E. H., & Matta, N. (2022). Diabetic retinopathy classification using hybrid deep learning approach. *SN Computer Science*, 3(5), 357.
- [17] Chavan, R., & Pete, D. (2024). Automatic multi-disease classification on retinal images using multilevel glowworm swarm convolutional neural network. *Journal of Engineering and Applied Science*, 71(1), 26.
- [18] Latif, J., Tu, S., Xiao, C., Rehman, S. U., Imran, A., & Latif, Y. (2022). ODGNet: A deep learning model for automated optic disc localization and glaucoma classification using fundus images. *SN Applied Sciences*, 4(4), 98.
- [19] Al-Kahtani, N., Varela-Aldás, J., Aljarbough, A., Ishak, M. K., & Mostafa, S. M. (2025). Discrete migratory bird optimizer with deep transfer learning aided multi-retinal disease detection on fundus imaging. *Results in Engineering*, 26, 104574.
- [20] Al-Antary, M. T., & Arafa, Y. (2021). Multi-scale attention network for diabetic retinopathy classification. *IEEE Access*, 9, 54190–54200.
- [21] Zhang, W., Zhou, L., Zhuang, P., Li, G., Pan, X., Zhao, W., & Li, C. (2023). Underwater image enhancement via weighted wavelet visual perception fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4), 2469–2483.
- [22] Zuo, Y., Yao, W., Zeng, Y., Xie, J., Fang, Y., Huang, Y., & Jiang, W. (2024). CFNet: Conditional filter learning with dynamic noise estimation for real image denoising. *Knowledge-Based Systems*, 284, 111320.

- [23] Gopalan, K., Srinivasan, S., Singh, M., Mathivanan, S. K., & Moorthy, U. (2025). Corn leaf disease diagnosis: Enhancing accuracy with ResNet152 and Grad-CAM for explainable AI. *BMC Plant Biology*, 25(1), 440.
- [24] Ye, Y., Pan, C., Wu, Y., Wang, S., & Xia, Y. (2022). MFI-Net: Multiscale feature interaction network for retinal vessel segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(9), 4551–4562.
- [25] Jorreia, O., Gonçalves, N., & Cortesão, R. (2025). Graph-based radiomics feature extraction from 2D retina images. *IEEE Access*, 13, 125359–125373.
- [26] Sheeba, M. C., & Christopher, C. S. (2025). Adaptive deep residual network for image denoising across multiple noise levels in medical, nature, and satellite images. *Ain Shams Engineering Journal*, 16(1), 103188.
- [27] Li, P., Liu, Y., Cui, Z., Yang, F., Zhao, Y., Lian, C., & Gao, C. (2022). Semantic graph attention with explicit anatomical association modeling for tooth segmentation from CBCT images. *IEEE Transactions on Medical Imaging*, 41(11), 3116–3127.
- [28] Kayadibi, İ., & Güraksın, G. E. (2023). An explainable fully dense fusion neural network with deep support vector machine for retinal disease determination. *International Journal of Computational Intelligence Systems*, 16(1), 28.
- [29] Kehinde, T. O., Adedokun, O. J., Joseph, A., Kabirat, K. M., Akano, H. A., & Olanrewaju, O. A. (2025). Helformer: An attention-based deep learning model for cryptocurrency price forecasting. *Journal of Big Data*, 12(1), 81.
- [30] Linchundan. (2019). Fundus image 1000 [Data set]. Kaggle. <https://www.kaggle.com/datasets/linchundan/fundusimage1000>
- [31] Mooney, P. (2018). Kermany 2018 [Data set]. Kaggle. <https://www.kaggle.com/datasets/paultimothymooney/kermany2018>