

Article

MedViTNet: A Hybrid Deep Learning Framework for Accurate Classification of Ocular Diseases Using Medical Eye Images

Sufyan Danish ¹, Muhammad Fayaz ¹, Muhammad Ayyaz Akbar Khan ¹, Asmat Noreen ¹,
Amirmohammad Montazeri ¹, Chun-Ming Lai ² and Hyeonjoon Moon ^{1,*}

¹ Department of Computer Science and Engineering, Sejong University, Seoul 05006, Republic of Korea; sufyandanih@sju.ac.kr (S.D.); muhammadfayaz@sju.ac.kr (M.F.); ayazkhan@sju.ac.kr (M.A.A.K.); asmatnoreen@sju.ac.kr (A.N.); a.montazeri@sju.ac.kr (A.M.)

² Department of Computer Science, Tunghai University, Taichung City 407, Taiwan; cmlai@thu.edu.tw

* Correspondence: hmoon@sejong.ac.kr

Abstract

Timely and accurate classification of medical eye images is important for supporting early screening and computer-aided decision support. However, automated ocular image classification remains challenging. Traditional convolutional neural networks (CNNs) are effective in capturing local visual features but may have limited ability to model broader spatial relationships. On the other hand, Vision Transformer-based models can capture global dependencies but may require complementary local feature extraction to identify subtle image-level patterns. To overcome these challenges, we propose MedViTNet, a hybrid deep learning framework that integrates two specialized modules: EyeConvNet for localized feature extraction and VisionNet for global contextual representation. MedViTNet was evaluated on three publicly available datasets, Myopia, Retina, and Jaundice, achieving classification accuracies of 99.99%, 99.93%, and 95.65%, respectively. Moreover, interpretability is enhanced using Grad-CAM visualizations, which highlight image regions that contribute to MedViTNet classification decisions, thereby supporting transparent and explainable computer-aided medical eye-image analysis. The results demonstrate that MedViTNet provides a unified local–global feature learning framework for heterogeneous medical eye-image classification across multiple binary classification tasks. Future work will enhance the dataset to represent a wider variety of medical cases and adapt the framework for multiclass classification tasks.

Keywords: eye disease classification; MedViTNet; fundus imaging; vision transformers; image classification



Academic Editor: Xinshun Xu

Received: 22 April 2026

Revised: 3 July 2026

Accepted: 7 July 2026

Published: 9 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Timely diagnosis and classification of ocular conditions are essential for supporting early screening, guiding clinical decision-making, and reducing the risk of vision-related complications. Global health statistics indicate that over one billion people have their vision impaired [1,2], with almost 90% being avertable or curable in the case of early detection. A non-invasive technique that allows viewing the retina, optic disc, and macula is called fundus imaging, which has become a crucial tool in diagnosing different eye diseases because it allows visualizing key parts of the retina in detail [3,4]. Although it has diagnostic power, screening of eye diseases is still a problem because it requires special equipment, trained ophthalmologists, and a lot of resources. The latest developments

in the field of deep learning have created some opportunities in the field of automated analysis and diagnostic support, offering tools that could assist ophthalmologists in the early detection of diseases. Such developments can provide a great chance to improve and increase the routine of eye screenings on a large scale.

The quality and diversity of data can be significant in the development of accurate classification models in the world of machine learning (ML) [5]. On the other hand, deep learning (DL) models rely heavily on a large number of diverse datasets to perceive or recognize complex patterns in these datasets and achieve robust and strong accuracy. However, publicly accessible medical datasets are generally constrained due to issues and concerns about privacy and ethical and associated complexities [6,7]. These limitations are especially significant in the case of high-resolution retinal fundus images, which show that the details of anatomical characteristics and vascular networks are critical for the diagnosis of a disease. Due to the restricted availability of these types of images, developing a reliable automated diagnostic system capable of generalizing effectively across diverse demographic groups poses a serious challenge.

In order to address these types of challenges, researchers have developed some special datasets that are adapted to the requirements of specific diagnostics, i.e., retinal fundus images, myopia-related images, and external eye-region images related to jaundice. These datasets help in producing considerable advances in the development of automated systems for the classification of medical eye images, enabling enhanced and precise diagnostics that facilitate prompt interventions to prevent vision loss. When an automated model is trained on these datasets, its knowledge of eye-related pathologies is enhanced, which contributes to the growth and development of clinical science and practices, particularly in decision-support systems for clinical and research applications.

1.1. Challenges

- **CNN's Limited Contextual Awareness:** CNNs are highly efficient in capturing fine-grained textures and local feature patterns, but they fail to capture long-range dependencies and spatial relationships. This limitation makes it challenging to capture image-level patterns that require broader contextual awareness.
- **ViT's Lack of Local Details:** ViT demonstrated remarkable proficiency in modeling global contextual information. However, they exhibit limitations in effectively retaining complex local features. This drawback is particularly significant in the classification of subtle pathological features, which play an important role in the accurate classification of retinal diseases at their initial stages.
- **Balancing Between Local and Global Features:** Models, relying solely on either local or global feature extraction processes, cannot achieve a global perspective of retinal disease trends. CNN-based approaches tend to ignore larger contextual relationships, and ViT-based techniques tend to miss local-level abnormalities, making them less precise in classification and less reliable in diagnostic accuracy in clinical settings.
- **Low Data Availability:** There is limited access to large, high-quality fundus image datasets due to privacy considerations, ethical considerations, and the high cost of collecting data, which limits the amount of data available for developing models.
- **Model Transparency:** It is essential to make medical AI interpretable to enable healthcare professionals to trust these models in their decision-making, since traditional models are not always clear, and this aspect prevents their application in clinical practice.

To address these issues, we introduce MedViTNet, a state-of-the-art hybrid deep learning framework for the automated classification of ocular conditions using medical eye images. MedViTNet uses a hybrid deep learning network, which incorporates EyeConvNet and VisionNet to apply localized feature extraction with a global context representation.

The EyeConvNet module is built on the InceptionV3 architecture, which is specialized in fine-scale medical eye-image representations, whereas the VisionNet module takes image inputs in the form of a sequence of patches to add more context to spatial representations. The use of this combined methodology not only enhances the functionality and flexibility of the MedViTNet model but also supports its potential use in future screening-oriented computer-aided analyses after further validation.

1.2. Contributions

- We introduce MedViTNet, a unified hybrid deep learning framework for heterogeneous medical eye-image classification. Unlike disease-specific or modality-specific architectures, MedViTNet integrates EyeConvNet and VisionNet to combine localized multi-scale feature extraction with global contextual representations, enabling consistent classification across diverse eye-image datasets.
- We implement a ViT-style patch creation and embedding component within the VisionNet branch to convert medical eye images into non-overlapping patch tokens with positional information. This component enables VisionNet to model global contextual relationships among image regions and is integrated with the EyeConvNet branch to form the proposed hybrid MedViTNet framework.
- MedViTNet demonstrates strong classification performance on the Myopia, Retina, and Jaundice datasets, achieving accuracies of 99.99%, 99.93%, and 95.65%, respectively.
- Grad-CAM visualizations are used to provide visual explanations of MedViTNet's classification decisions by highlighting influential image regions, thereby improving interpretability and supporting transparent computer-aided analysis.

Overall, the main contribution of MedViTNet lies in its integrated local–global design, where the InceptionV3-based EyeConvNet branch extracts localized multi-scale visual patterns, while the ViT-style VisionNet branch captures global contextual dependencies. These complementary representations are combined through feature fusion to support accurate classification of medical eye images. The proposed framework contributes to automated medical eye-image analysis by providing an efficient classification approach that may support future screening-oriented decision-support systems after further validation. Figure 1 presents the graphical abstract of the proposed MedViTNet framework.

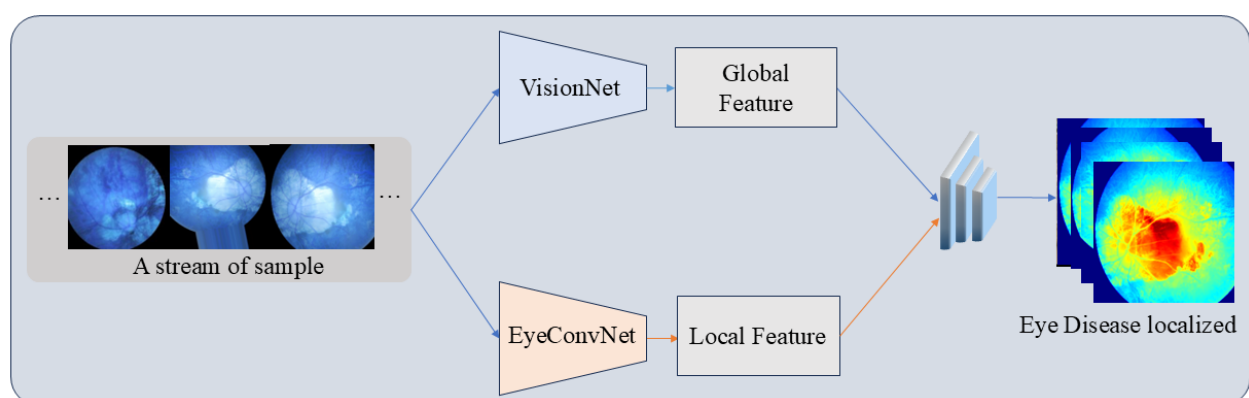


Figure 1. Conceptual overview of the proposed MedViTNet framework. The figure illustrates the high-level flow of the framework, including EyeConvNet, VisionNet, feature fusion, and Grad-CAM-based visual explanations.

This paper will be structured as follows: Section 2 will present the related work and literature review. Section 3 outlines the dataset and methodology, which involves preprocessing, extraction of features, dimensionality reduction in the data, and training of the model. In Section 4, a detailed measure of the performance of MedViTNet is provided,

comparing it to existing state-of-the-art models and outlining the key elements that make it such a precise model. Lastly, Section 6 explains the general implications of our findings, potential future research directions, and highlights the promise of large language model (LLM)-based solutions in mitigating the global burden of eye diseases.

2. Literature Review

Several studies have investigated deep learning approaches for medical eye-image classification, including myopia, retinal conditions, and jaundice-related eye manifestations, with the objective of improving automated screening and diagnostic support [8–12]. Early studies primarily focused on convolutional neural networks (CNNs) because of their strong capability to learn discriminative visual features from medical images. For example, Babaqi et al. [13] employed EfficientNet with transfer learning to classify retinal diseases and reported an accuracy of 94%, outperforming a custom CNN model. Similarly, Triwijoyo et al. [14] applied deep learning to retinal image classification using the STARE dataset and achieved an accuracy of 80.93% after dataset balancing. Jatmoko et al. [15] further demonstrated the effectiveness of transfer learning using DenseNet and EfficientNetB3, reporting accuracies of 78.59% and 95.66%, respectively. Although these studies demonstrated the effectiveness of CNN-based learning, their performance was often influenced by limited dataset diversity, relatively small sample sizes, and evaluations of specific disease-oriented datasets.

CNN-based models are still very effective in eye-image classification in medical applications, and they are able to detect local visual features like variations in texture, vessel structure, color and lesion-like features [16]. Machine learning and deep learning techniques have been widely applied in the medical domain to support diagnosis, classification, prediction, and clinical decision-making [5,16]. InceptionV3 [17] networks are especially useful within CNN architectures since these networks utilize convolutional filters of varying sizes in parallel to capture multi-scale visual information. This property is significant for analyses of visual patterns that convey information in an image, which can be present at various spatial scales. Therefore, MedViTNet adopts an InceptionV3-based EyeConvNet branch to strengthen localized and multi-scale feature extraction from medical eye images.

Although CNNs are powerful for localized feature extraction, they may have limited capacity to model long-range spatial relationships across the entire image. Vision Transformer-based models address this limitation by dividing images into patch tokens and using self-attention to learn global dependencies among different image regions. Such global contextual modeling is valuable in medical eye-image classification because condition-specific visual patterns may appear across multiple regions and may require broader spatial context for accurate classification. In MedViTNet, the VisionNet branch is used to capture these broader dependencies and complement the localized representations learned by EyeConvNet.

Recent research has therefore increasingly explored hybrid CNN–Transformer architectures, which combine convolutional feature extraction with attention-based contextual modeling [18]. These hybrid models have demonstrated promising performance by integrating local texture information with global contextual representations. In parallel, several studies have proposed CNN-based or hybrid approaches for specific ophthalmic applications. Bernabé et al. [19] achieved 99.89% accuracy for glaucoma and diabetic retinopathy classification using CNNs, although the dataset was relatively small. Rashid et al. [20] developed a CNN model for external eye disease classification, while Du et al. [21] introduced an evidence-theory-based decision fusion strategy to improve eye disease recognition. Paradisa et al. [22] combined Inception-ResNetV2 and DenseNet121 for diabetic retinopathy classification, whereas Ahmed et al. [23] employed deep CNNs for glaucoma detection.

These studies demonstrate that both CNN-based and hybrid architectures can achieve high classification accuracy; however, they are generally developed and validated for individual ophthalmic diseases or a single imaging modality.

Although significant advances have been achieved in previous studies, there are several drawbacks in medical eye-image classification approaches. The studies by Babaqi T. et al. [13] and Triwijoyo et al. [14] have shown that deep learning is a useful tool for analyzing ocular images, but the performance and generalization of such analysis was influenced by the size of the datasets involved, the variety of the datasets, and the design of the models themselves. Likewise, Paradisa R. H. et al. [22] and Ahmed M. R. [23] reported promising results on specific ophthalmic classification tasks, but most of the previous methods used only one feature learning method, such as CNN-based local feature extraction. These techniques tend to be successful in modeling small-scale visual patterns, but they may not be as effective at modeling larger-scale spatial relationships in the entire image. These observations point to the necessity for a framework that can learn from local visual information and global contextual representations from medical eye images in a joint manner.

This observation reveals an important research gap. Previous studies generally follow one of three directions, including CNN-based models focusing on local feature extraction, ViT-based models emphasizing global contextual representation, or hybrid CNN–Transformer models combining both paradigms. Despite their promising performance, most existing hybrid frameworks are designed and evaluated for a specific ophthalmic disease or a single imaging modality. Consequently, it remains unclear whether a single hybrid architecture can effectively generalize across heterogeneous medical eye-image classification tasks involving different image characteristics. Unlike previous studies, the objective of MedViTNet is not simply to combine a CNN with a Transformer. Instead, MedViTNet is designed as a unified hybrid framework that can be applied without architectural modification to heterogeneous binary medical eye-image classification tasks. Specifically, the proposed framework integrates an InceptionV3-based EyeConvNet branch for localized multi-scale feature extraction with a ViT-style VisionNet branch for global contextual representation, and it combines both feature representations through feature-level fusion before classification. The framework is evaluated on three independent datasets representing different medical eye-image classification problems, such as Myopia, Retina, and Jaundice, including both fundus images and external eye-region images. This design enables consistent local–global representation learning across heterogeneous datasets while maintaining model interpretability through Grad-CAM visualizations [24].

3. Methodology

This research study proposed a hybrid framework that combines EyeConvNet with VisionNet to perform binary classification in fundus imaging, called MedViTNet. The model adeptly captures localized and global features that are critical for distinguishing between positive and negative medical eye images. In this section, the datasets, data preprocessing, and architecture of the proposed MedViTNet framework are described. The overall framework of the MedViTNet model is shown in Figure 2.

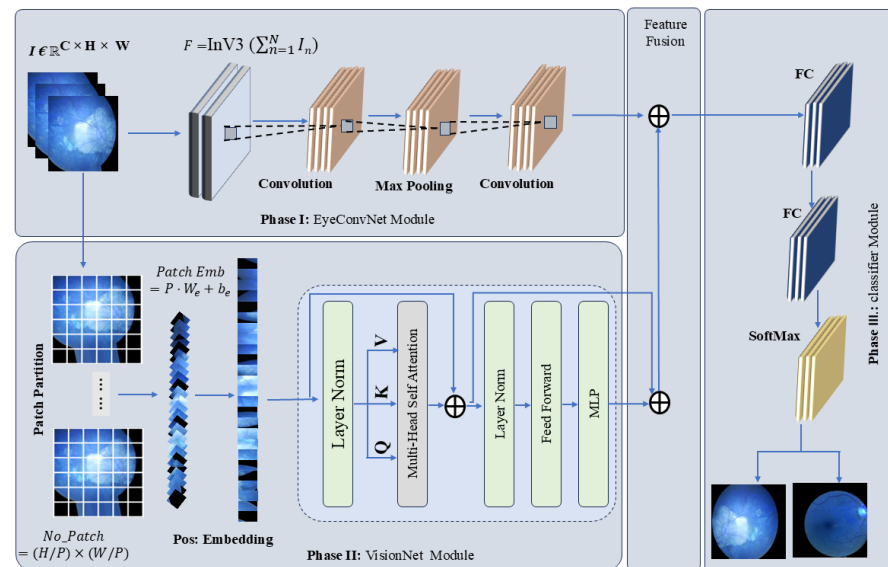


Figure 2. Detailed schematic of the implemented MedViTNet architecture. The model consists of two complementary branches: an InceptionV3-based EyeConvNet branch for local and multi-scale feature extraction and a ViT-style VisionNet contextual feature branch for global representation learning. The outputs of both branches are combined through the feature fusion module and passed to the classifier for final binary classification.

3.1. Dataset and Preprocessing

In this study, three publicly available medical eye-image datasets were used: Myopia [25], Retina [26], and Jaundice [27]. Each dataset was used for a binary classification task. The Myopia dataset contains 124,794 images, including 63,294 positive samples and 61,500 negative samples. The Retina dataset contains 7585 images, including 3662 positive samples and 3923 negative samples. The Jaundice dataset contains 464 images, including 174 positive samples and 290 negative samples. To ensure consistent evaluation, each dataset was divided using a class-stratified split, with 70% of images used for training, 10% for validation, and 20% for testing. The validation set was used for monitoring validation loss and applying early stopping, while the test set was used only for final performance evaluation. Table 1 presents the corrected class-wise dataset distribution and the corresponding training, validation, and testing split for each dataset. Sample images from each dataset are shown in Figure 3.

Table 1. Class-wise dataset distribution and class-stratified training, validation, and testing split.

Dataset Name	Class	Class Samples	Training	Validation	Testing	Total Samples
Myopia dataset [25]	Positive	63,294	44,306	6330	12,658	124,794
	Negative	61,500	43,050	6150	12,300	
Retina dataset [26]	Positive	3662	2564	366	732	7585
	Negative	3923	2746	393	784	
Jaundice dataset [27]	Positive	174	122	18	34	464
	Negative	290	203	29	58	

Each image present in the dataset undergoes a series of image preprocessing steps, i.e., resizing to the input dimension of EyeConvNet to ensure uniformity and intensity normalization to enhance image clarity while minimizing the noise present in the image. Normalization processes are particularly important in medical images where slight varia-

tions in the brightness and contrast of an image may affect model performance. The dataset was augmented with some random transformations, like rotations, flips, and shifts, to enhance the model's generalization and resilience to variations.

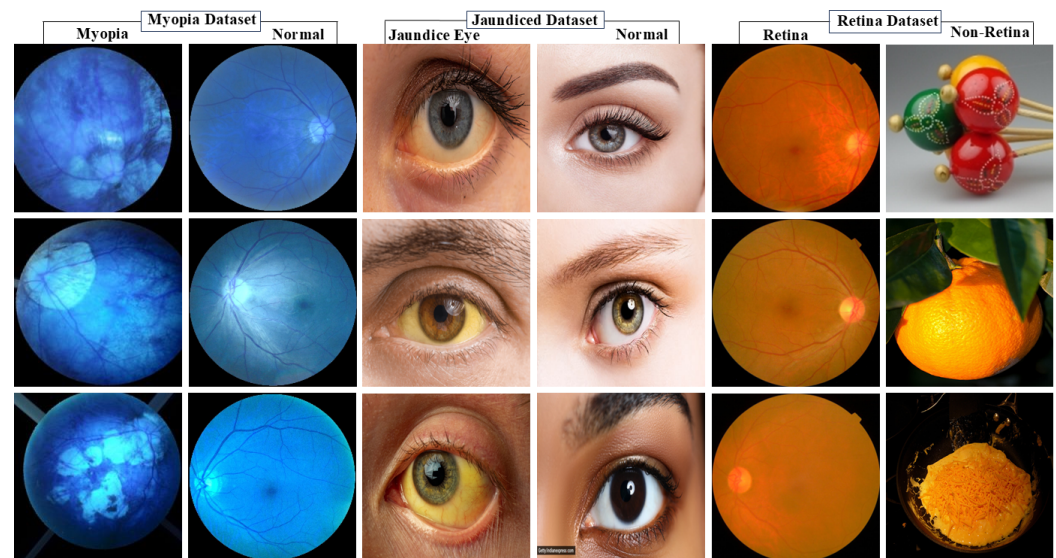


Figure 3. Sample images from the Myopia, Retina, and Jaundice datasets, showing the binary classification categories.

3.2. MedViTNet Architecture

The proposed MedViTNet architecture integrates two complementary branches, EyeConvNet and VisionNet, for binary medical eye-image classification. EyeConvNet is an InceptionV3-based convolutional branch designed to extract localized and multi-scale visual features. VisionNet is implemented as a ViT-style contextual feature branch that uses patch creation, linear patch embedding, positional embedding, multi-head self-attention, residual connections, layer normalization, and feed-forward transformation to model broader spatial relationships among image regions. VisionNet is not used as a standalone Vision Transformer classifier; instead, its contextual feature representation is fused with the EyeConvNet representation before final classification. Therefore, the distinctness of VisionNet lies in its functional role within the MedViTNet framework rather than in proposing a fundamentally new Transformer module. The following sections provide a detailed description of the EyeConvNet and VisionNet modules, as well as the fusion strategy used to combine their outputs for final classification.

3.2.1. EyeConvNet Module

EyeConvNet employs InceptionV3 as its convolutional backbone because of its strong multi-scale feature extraction capability. InceptionV3 uses parallel convolutional filters of different sizes, such as 1×1 , 3×3 , and 5×5 , enabling the model to capture localized visual patterns at multiple spatial scales. In the context of medical eye-image classification, these patterns may include vessel structures and retinal texture variations in fundus images, as well as visible color or appearance changes in external eye-region images. To further enhance local feature extraction, a dilated convolution layer is added to expand the receptive field without reducing spatial resolutions. This allows EyeConvNet to preserve fine-grained visual information while capturing structures of varying sizes. Subsequently, global average pooling is used to reduce the spatial dimensions of the feature maps while preserving discriminative feature information. Finally, a fully connected dense projection layer compresses the extracted features into a fixed-size representation that is later fused with the output of the VisionNet branch.

Mathematically, the EyeConvNet feature extraction process can be expressed as follows:

$$F_{ECN} = \text{Dense}(\text{GAP}(\text{DConv}(\text{InceptionV3}(X_{\text{image}})))) \quad (1)$$

where F_{ECN} represents the feature vector generated by the EyeConvNet module, X_{image} denotes the input medical eye image, and DConv denotes the dilated convolution operation.

3.2.2. VisionNet Module

The VisionNet module is implemented as a ViT-style contextual feature branch within MedViTNet. It processes the same input medical eye image as EyeConvNet but focuses on global contextual representation learning. The input image is divided into non-overlapping patches, which are flattened, projected into an embedding space, combined with positional information, and processed through Transformer-style encoder layers. Unlike a standard standalone ViT classifier, VisionNet does not independently generate the final prediction. Instead, the class-token representation from the final encoder layer is used as a global contextual feature vector and fused with the EyeConvNet feature vector before classification.

Patch Creation and Embedding

The first operation in the VisionNet module is the patch creation and embedding component (CPEC), which converts the input medical eye image into a sequence of patch-level tokens for global contextual representation learning. CPEC follows the patch-tokenization principle used in Transformer-style vision models. However, in the proposed MedViTNet framework, it is implemented as a configurable component of the VisionNet branch and is integrated with EyeConvNet through feature fusion. Therefore, CPEC supports global contextual feature extraction, while EyeConvNet captures localized and multi-scale visual features. The patch creation process is described as follows:

- **Patch Size Definition:** The input image is first resized to a fixed size of 224×224 pixels. CPEC then divides the image into non-overlapping patches of size $P \times P$. In this study, the patch size is set to 16×16 , and the stride is set equal to the patch size, i.e., $S = P$, to ensure non-overlapping patch extraction. The total number of patches N is determined as follows:

$$N = \frac{H}{P} \times \frac{W}{P} = \frac{224}{16} \times \frac{224}{16} = 196 \text{ patches} \quad (2)$$

where H and W represent the height and width of the input image, respectively, and P denotes the patch size. The patch size controls the granularity of the visual representation; smaller patches preserve finer spatial details, whereas larger patches provide more compact global representations.

- **Patch Flattening and Embedding:** Each extracted patch X_p with size $P \times P \times C$ is flattened into a one-dimensional vector, where C denotes the number of image channels. For an RGB image with $P = 16$ and $C = 3$, each flattened patch has a dimension of $16 \times 16 \times 3 = 768$. The flattening operation for patch X_p is expressed as follows:

$$z_p = \text{Flatten}(X_p) \quad (3)$$

where $z_p \in \mathbb{R}^{P^2C}$ represents the flattened patch vector. The flattened vector is then projected into a D -dimensional embedding space using a learnable linear projection:

$$E_p = z_p W_E + b_E \quad (4)$$

where W_E and b_E are learnable parameters, and $E_p \in \mathbb{R}^D$ represents the embedded patch token. This embedding step allows VisionNet to process each image patch as a token representation.

- **Positional Embedding:** Since self-attention does not inherently preserve the spatial order of image patches, positional information is added to the embedded patch tokens. In this study, a learnable positional embedding is used to retain the spatial arrangement of patches within the image. A learnable class token E_{class} is also prepended to the patch-token sequence to aggregate global image-level information for classification. The final input sequence to the VisionNet encoder is defined as follows:

$$E_{\text{input}} = [E_{\text{class}}, E_1, E_2, \dots, E_N] + E_{\text{pos}} \quad (5)$$

where E_{class} is the learnable class token, $E_1, E_2, \dots, E_N$ represent the embedded patch tokens, and E_{pos} denotes the learnable positional embedding.

After patch construction and embedding, the resulting token sequence is passed to the VisionNet encoder for global contextual representation learning. In the proposed implementation, VisionNet consists of four ViT-style encoder layers and eight attention heads. Each encoder layer includes multi-head self-attention, residual connections, layer normalization, and a position-wise feed-forward network. Dropout with a rate of 0.5 is applied to reduce overfitting. The class-token representation generated by the final encoder layer is used as the global image-level feature representation and is subsequently passed to the feature-fusion stage, where it is combined with the local and multi-scale representation extracted by EyeConvNet. These implementation settings are included in the Methodology Section to improve reproducibility and provide a clear description of the VisionNet architecture used in the proposed MedViTNet framework.

Multi-Head Self-Attention

VisionNet employs Multi-Head Self-Attention (MHSA) to model global dependencies across patches. This technique calculates attention scores for each patch, allowing the model to focus on the most relevant areas of the image. The attention score for each patch is calculated using the scaled dot-product formula:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

where Q , K , and V represent the query, key, and value matrices, and d_k denotes the dimension of the key vectors. Multiple attention heads are utilized to capture diverse aspects of the relationships between patches.

Residual Connections and Layer Normalization

To stabilize training, the output of each self-attention layer is combined with the input through a residual connection followed by layer normalization. The layer normalization operation is defined as follows:

$$\text{Norm}(x) = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \quad (7)$$

where μ and σ^2 represent the mean and variance of the input x , respectively, and ϵ is a small constant added for numerical stability. This normalization step improves convergence and reduces the likelihood of vanishing or exploding gradients. Each encoder block also includes a position-wise feed-forward network (FFN), which applies a non-linear transformation to each patch. The FFN is defined as follows:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \quad (8)$$

where W_1 and W_2 are weight matrices, and b_1 and b_2 are bias terms. The FFN enhances the model's ability to learn complex feature relationships by introducing non-linearity and additional capacity.

3.2.3. Feature Fusion

The EyeConvNet and VisionNet modules generate complementary feature representations from the same input medical eye image. EyeConvNet focuses on localized and multi-scale visual patterns, whereas VisionNet captures global contextual relationships among image regions. Before fusion, the feature vectors from both branches are projected into a shared latent feature space to ensure dimensional compatibility.

Let F_{ECN} denote the feature vector generated by the EyeConvNet module after global average pooling and dense projection, and let F_{VN} denote the feature vector generated by the VisionNet module. Here, $F_{\text{ECN}} \in \mathbb{R}^{d_c}$ and $F_{\text{VN}} \in \mathbb{R}^{d_v}$, where d_c and d_v denote the original feature dimensions of the EyeConvNet and VisionNet outputs, respectively. Since these two feature vectors may have different original dimensions, they are mapped into a common latent dimension using learnable projection layers:

$$\tilde{F}_{\text{ECN}} = W_c F_{\text{ECN}} + b_c \quad (9)$$

$$\tilde{F}_{\text{VN}} = W_v F_{\text{VN}} + b_v \quad (10)$$

where $\tilde{F}_{\text{ECN}}$ and $\tilde{F}_{\text{VN}}$ represent the dimension-aligned feature vectors in the shared latent space such that $\tilde{F}_{\text{ECN}}, \tilde{F}_{\text{VN}} \in \mathbb{R}^{d_f}$. Here, W_c , W_v , b_c , and b_v are learnable projection parameters, and d_f denotes the shared latent feature dimension. After dimensional alignment, the fused representation is obtained using element-wise addition:

$$F_{\text{combined}} = \tilde{F}_{\text{ECN}} + \tilde{F}_{\text{VN}} \quad (11)$$

Element-wise addition was selected because EyeConvNet and VisionNet extract complementary representations from the same image. This strategy enables the direct integration of local multi-scale features and global contextual features while maintaining a compact feature representation. Unlike concatenation, which increases the dimensionality of the classifier input, element-wise addition avoids unnecessary parameter expansion and supports a lightweight classification head. The fused representation F_{combined} is then passed through dropout and fully connected layers, followed by a sigmoid activation function for final binary classification.

3.3. Training and Optimization

MedViTNet is trained using binary cross-entropy loss:

$$L_i = -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \quad (12)$$

where y_i is the true label, and $\hat{y}_i$ is the predicted probability. The model's total loss L is averaged across all samples in each batch. The Adam optimizer, with an initial learning rate of 1×10^{-3} , iteratively updates model weights. Training incorporates a learning rate scheduler (ReduceLROnPlateau) to adjust the learning rate upon plateaus in validation performance. Additionally, early stopping with a patience of 50 epochs terminates training if performance plateaus. Through the integration of EyeConvNet-based local feature extraction and VisionNet-based global contextual representation, MedViTNet is designed to learn complementary visual features for medical eye-image classification. This training strategy supports stable optimization of the proposed hybrid architecture across the evaluated binary classification datasets.

4. Experiments and Results

This section outlines the experimental setup, performance evaluation, and analysis of the MedViTNet model with the integrated EyeConvNet and VisionNet modules. The experiments are designed to assess the model's capability in accurately classifying retinal images into normal and abnormal categories across three datasets, i.e., Myopia, Retina, and Jaundice. The evaluations include system configuration, model implementation details, ablation studies and experiments, and both qualitative quantitative and qualitative assessments of outcomes.

4.1. System Configuration

All experiments are performed on a high-performance system equipped with an NVIDIA GPU to ensure efficient DL computations. The model was implemented using PyTorch version 2.7.1, providing the flexibility to customize and optimize each component of MedViTNet. The use of GPUs enabled speedy model training, which is crucial for processing large volumes of high-dimensional and multi-channelled image data across the three datasets, i.e., Myopia, Retina, and Jaundice. The key libraries used included PyTorch for model architecture and training, NumPy version 2.2.6 for numerical operations, and OpenCV version 4.8.1 for image preprocessing, i.e., rotations, flips, and shifts, to enhance model generalization and resilience to variations. Other details about the system's configuration are given in Table 2.

Table 2. Implementation settings and hyperparameters used for MedViTNet.

Hyperparameter	Value
Learning Rate	1×10^{-3}
Batch Size	32
Epochs	200
Adam Optimizer	Yes
Weight Decay	1×10^{-4}
Dropout	0.5
Encoder Layers	4
Attention Heads	8

4.2. Implementation Details and Model Evaluation

MedViTNet was trained individually on each dataset to ensure specific optimization for each binary classification task (Myopic vs. Non-Myopic, Retinal vs. Non-Retinal, and Jaundice vs. Non-Jaundice). Training was conducted for a maximum of 200 epochs, with early stopping activated after 50 epochs with no improvements in validation loss. The learning rate was initially set to 1×10^{-3} and adjusted dynamically using the ReduceLROnPlateau scheduler to avoid overshooting optimal values as the model converged. Binary cross-entropy loss was used as the objective function, and it is defined as follows:

$$L = -\frac{1}{N} \sum_{i=1}^N (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \quad (13)$$

where y_i is the true label, $\hat{y}_i$ is the predicted probability, and N is the number of samples.

For each dataset, the model's performance was assessed using metrics such as accuracy, F1 score, recall, precision, sensitivity, specificity, false-positive rate (FPR), and false-negative rate (FNR). These metrics provide a comprehensive assessment of classification performance. They are calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (15)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (16)$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (17)$$

where TP , TN , FP , and FN represent true positives, true negatives, false positives, and false negatives, respectively.

4.3. Ablation Study

The ablation experiment described in Table 3 was used to determine the effectiveness of different backbone models on the Retina dataset, which underscores the effects of feature extraction methods on classification accuracy and strength. The proposed MedViTNet model has superior performance in all important evaluation measures, showing that it can successfully learn both localized and global features in order to improve the classification of the Retina dataset:

- **Backbone Models:** To identify the best architecture to use in localized feature extraction, we tested a number of CNN backbones, such as MobileNet, EfficientNetV2B1, EfficientNetV2S, InceptionResNetV2, and EfficientNetV2B2, replacing InceptionV3. It was found that InceptionV3 delivered the best accuracy on average due to its extraction of multi-scale features, which is essential in capturing the fine-grained features of retinal images.
- **Transformer Architecture:** The number of Transformer encoder layers and attention heads was modified to evaluate how it influences global feature learning. The most optimal trade-off between model complexity and classification accuracy was obtained with four encoder layers and eight attention heads.
- **Technique Feature Fusion:** We compared concatenation and element-wise addition to fuse the results of EyeConvNet and VisionNet. The results of element-wise addition were more successful, probably because the former can directly align localized and global feature representations without adding superfluous dimensions.

The Table 3 shows the result of the ablation study based on the Retina dataset.

Table 3. Backbone ablation study on the Retina dataset.

Backbone Model	Acc	F1 Score	Precision	Recall	Specificity
MobileNet	87.99	87.72	83.33	92.59	84.03
EfficientNetV2B1	88.98	88.77	84.62	93.35	85.17
EfficientNetV2S	89.97	89.81	85.90	94.10	86.32
InceptionResNetV2	91.95	91.88	88.46	95.57	88.66
EfficientNetV2B2	92.88	92.84	89.74	96.15	89.85
VisionTransformer	97.50	97.46	97.33	97.60	97.40
MedViTNet MODEL	99.93	99.94	99.87	1.00	99.86

All values are expressed as percentages; Bold values indicate the highest performance.

4.4. Quantitative Results

The quantitative analysis was conducted on each dataset, evaluating performance across all metrics. Table 4 presents the results, showing high classification accuracy across all datasets. MedViTNet achieved an accuracy of 99.99% and 99.93% on both the Myopia and Retina datasets and 95.65% on the Jaundice dataset respectively, demonstrating robust generalization and effectiveness in different eye disease classification tasks.

Table 4. Performance Results on Proposed MedViTNet Model.

Dataset Name	ACC (%)	F1 Score (%)	Precision (%)	Recall (%)	Specificity (%)	FPR (%)	FNR (%)
Myopia dataset	99.99	99.99	1.00	99.98	1.00	0.00	0.02
Retina dataset	99.93	99.94	99.87	1.00	99.86	0.14	0.00
Jaundice dataset	95.65	94.12	94.12	94.12	96.55	3.45	5.88

4.5. Five-Fold Cross-Validation and Fusion Robustness Analysis

To further strengthen the reliability and generalization capability of the proposed MedViTNet framework, an additional five-fold cross-validation analysis was conducted. This evaluation complements the primary train–validation–test experiments performed on all datasets (Myopia, Retina, and Jaundice), as reported in Table 4, where the proposed model achieved 99.99% accuracy on the Myopia dataset, 99.93% accuracy on the Retina dataset, and 95.65% accuracy on the Jaundice dataset. Unlike the primary evaluation, five-fold cross-validation assesses the stability of the proposed framework under multiple train–test partitions, providing additional evidence that the reported performance is consistent rather than dependent on a single data split.

For five-fold cross-validation, each dataset was independently partitioned into five mutually exclusive folds. In each iteration, one fold was used for testing while the remaining folds were used for training. Data augmentation was applied only to the training subsets after data partitioning, ensuring that augmented versions of the same image were not shared between the training and testing sets. Furthermore, dataset integrity was verified through perceptual hashing (pHash)-based duplicate analysis. Approximately 100 near-duplicate candidates were identified in the Myopia dataset, while 38 exact duplicate images were detected and removed from the Jaundice dataset before data partitioning. This preprocessing ensured that visually similar samples did not appear across different folds, thereby minimizing potential information leakage. All reported cross-validation and pooled out-of-fold results were obtained after this cleaning process.

The five-fold cross-validation results demonstrate that the performance gains observed in the primary experiments are consistently maintained across different data partitions. For the Myopia dataset (Table 5), the proposed element-wise fusion achieved a mean accuracy of 98.33%, outperforming the concatenation strategy (96.13%), while the corresponding pooled out-of-fold evaluation (Table 6) further confirmed stable predictive performance with an accuracy of 99.09% and narrow confidence intervals. These findings indicate that the proposed fusion strategy provides reliable performance beyond a single train–test split.

Similarly, for the Retina dataset, element-wise addition achieved a mean cross-validation accuracy of 99.37% compared with 98.41% for concatenation (Table 7). The pooled out-of-fold analysis (Table 8) produced an accuracy of 99.40% with a 95% confidence interval of 99.22–99.57%, demonstrating highly consistent performance across folds. Since this dataset represents a retina versus non-retina image classification task, these results further confirm the effectiveness of MedViTNet in learning discriminative anatomical image representations while maintaining excellent robustness under repeated validations.

The Jaundice dataset presents greater image variability than the Myopia and Retina datasets due to differences in external eye appearance, illumination, and image acquisition conditions. Nevertheless, the proposed framework maintained stable performance across folds. As shown in Table 9, the element-wise fusion achieved a mean accuracy of 93.65% compared with 94.37% for concatenation. The pooled out-of-fold analysis (Table 10) further achieved an accuracy of 93.72% with a narrow 95% confidence interval of 93.41–94.03%, indicating consistent generalization despite the increased variability of this dataset.

Table 5. Five-fold cross-validation performance on the Myopia dataset.

Fusion Strategy	Metric	Mean Performance	Standard Deviation
Element-wise addition	Accuracy	98.33%	1.43%
	Precision	98.12%	1.86%
	Recall/Sensitivity	98.73%	0.99%
	Specificity	77.83%	43.55%
	F1 Score	98.42%	1.31%
	AUC	99.78%	0.21%
Concatenation	Accuracy	96.13%	2.98%
	Precision	98.00%	2.44%
	Recall/Sensitivity	94.40%	5.22%
	Specificity	77.78%	43.55%
	F1 Score	96.10%	3.03%
	AUC	99.28%	0.91%

Table 6. Pooled out-of-fold performance with 95% confidence intervals.

Fusion Strategy	Metric	Performance	95% Confidence Interval
Element-wise addition	Accuracy	99.09%	99.03–99.14%
	Precision	99.74%	99.70–99.79%
	Recall/Sensitivity	98.30%	98.19–98.42%
	Specificity	99.78%	99.74–99.81%
	F1 Score	99.02%	98.96–99.08%
	AUC	99.95%	99.94–99.97%
Concatenation	Accuracy	94.44%	94.31–94.58%
	Precision	99.71%	99.66–99.76%
	Recall/Sensitivity	88.43%	88.17–88.71%
	Specificity	99.77%	99.74–99.81%
	F1 Score	93.73%	93.58–93.89%
	AUC	99.86%	99.84–99.88%

Table 7. Five-fold cross-validation performance on the Retina dataset.

Fusion Strategy	Metric	Mean Performance (%)	Standard Deviation (%)
Element-wise addition	Accuracy	99.37	0.40
	Precision	98.72	0.98
	Recall/Sensitivity	98.95	1.35
	Specificity	99.45	0.31
	F1 Score	98.83	0.95
	AUC	99.96	0.03
Concatenation	Accuracy	98.41	0.76
	Precision	98.55	0.69
	Recall/Sensitivity	98.12	0.88
	Specificity	98.70	0.63
	F1 Score	98.33	0.74
	AUC	99.21	0.31

Table 8. Pooled out-of-fold performance with 95% confidence intervals on the Retina dataset.

Fusion Strategy	Metric	Performance (%)	95% Confidence Interval (%)
Element-wise addition	Accuracy	99.40	99.22–99.57
	Precision	99.41	99.13–99.66
	Recall/Sensitivity	99.32	99.06–99.58
	Specificity	99.46	99.22–99.69
	F1 Score	99.36	99.10–99.60
	AUC	99.96	99.94–99.98
Concatenation	Accuracy	98.41	98.28–98.54
	Precision	98.55	98.42–98.69
	Recall/Sensitivity	98.12	97.98–98.27
	Specificity	98.70	98.58–98.83
	F1 Score	98.33	98.20–98.47
	AUC	99.21	99.15–99.28

Table 9. Five-fold cross-validation performance on the Jaundice dataset.

Fusion Strategy	Metric	Mean Performance (%)	Standard Deviation (%)
Element-wise addition	Accuracy	93.65	2.41
	Precision	92.12	2.87
	Recall/Sensitivity	92.12	3.05
	Specificity	94.55	1.92
	F1 Score	92.12	2.69
	AUC	96.84	0.62
Concatenation	Accuracy	94.33	3.37
	Precision	89.00	9.53
	Recall/Sensitivity	95.25	4.63
	Specificity	94.16	5.70
	F1 Score	91.66	4.39
	AUC	98.38	1.39

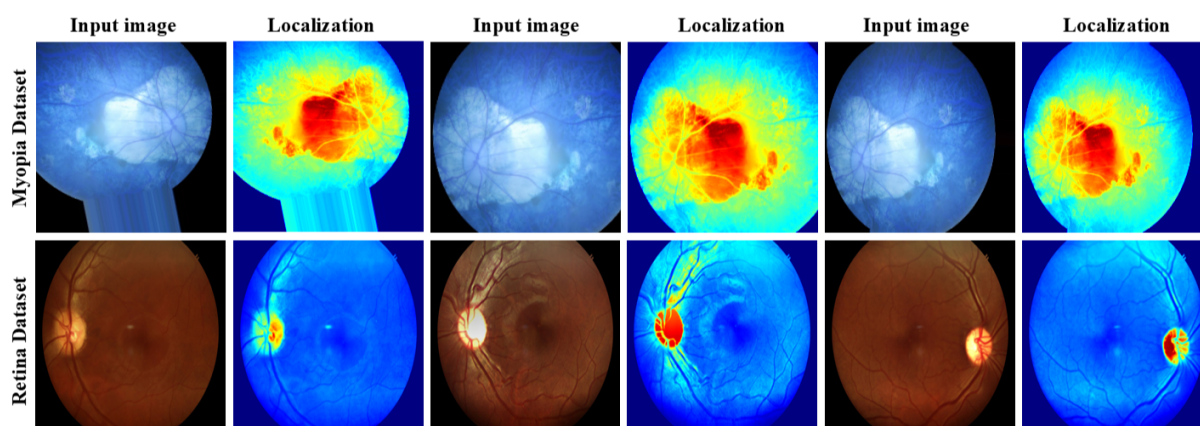
Overall, the additional cross-validation and pooled out-of-fold analyses reinforce the findings of the primary experimental results presented in Table 4. Across all three datasets, the proposed MedViTNet framework consistently maintained high predictive performance with low fold-to-fold variability, demonstrating that the reported improvements are stable and reproducible. The results further confirm that the proposed element-wise fusion strategy effectively integrates the complementary local representations learned by EyeConvNet and the global contextual representations learned by VisionNet, contributing to the superior performance of MedViTNet over existing approaches while maintaining robustness across heterogeneous medical eye-image classification tasks.

Table 10. Pooled out-of-fold performance with 95% confidence intervals on the Jaundice dataset.

Fusion Strategy	Metric	Performance (%)	95% Confidence Interval (%)
Element-wise addition	Accuracy	93.72	93.41–94.03
	Precision	92.18	91.82–92.54
	Recall/Sensitivity	92.21	91.86–92.56
	Specificity	94.61	94.32–94.90
	F1 Score	92.15	91.79–92.51
	AUC	96.86	96.72–97.00
Concatenation	Accuracy	94.37	91.75–96.19
	Precision	88.36	84.86–90.98
	Recall/Sensitivity	94.85	92.30–96.57
	Specificity	94.14	91.48–95.99
	F1 Score	91.49	88.52–93.83
	AUC	98.10	–

4.6. Qualitative Results

Qualitative evaluation was conducted using Grad-CAM visualizations, which highlight regions of each image that contributed to MedViTNet’s classification decision. For fundus-based images, Grad-CAM heatmaps emphasized relevant ocular structures such as the optic disc, macular region, and vascular patterns. For external eye-region images, the heatmaps highlighted visible regions associated with the classification outcome. These visualizations support transparent interpretation of MedViTNet’s decision-making process across the evaluated medical eye-image datasets. The visualized results of Grad-CAM given in Figure 4.

**Figure 4.** Grad-CAM heatmaps for the Myopia dataset and Retina dataset, highlighting the affected areas in the image used by the model for classification.

The confusion matrices of each dataset further demonstrate MedViTNet’s classification performance and highlighting minimal misclassifications. The MedViTNet model achieved high true-positive rates on all datasets with only a few false positives or false negatives, as given in Figures 5 and 6.

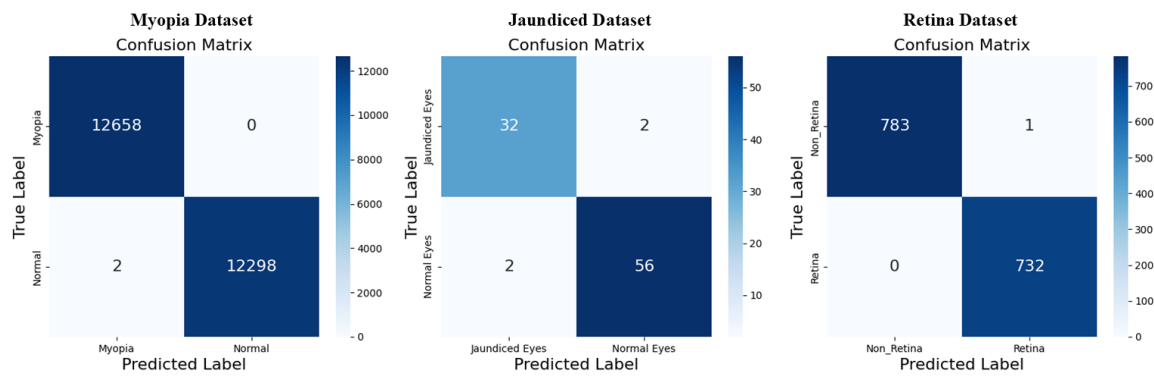


Figure 5. Confusion matrices for Myopia, Retina, and Jaundice datasets using the proposed method.

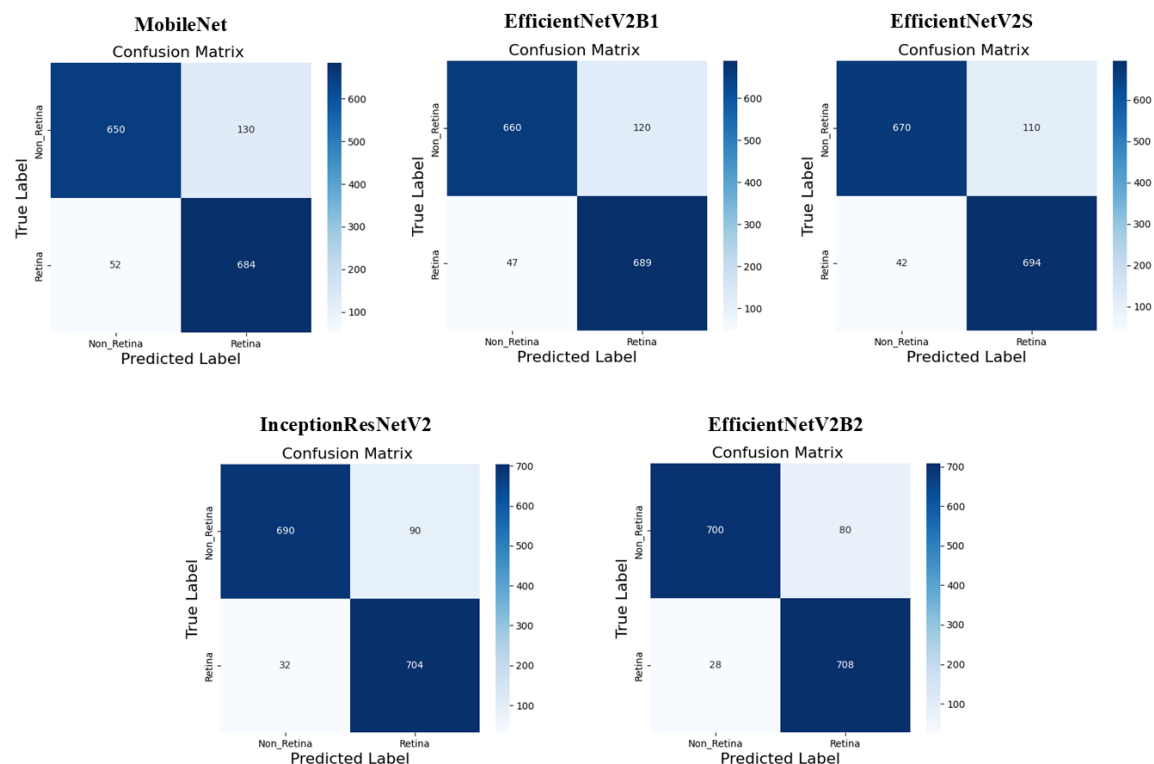


Figure 6. Confusion matrix of the model's backbone, showing the classification performance with true positives, true negatives, false positives, and false negatives.

4.7. Receiver Operating Characteristic (ROC) and Precision–Recall Analysis

Receiver operating characteristic (ROC) curves and precision–recall curves are generated to assess the model's effectiveness across range of threshold values. The area under the ROC curve (AUC) for each dataset is calculated to evaluate the discriminatory ability:

$$AUC = \int_0^1 TPR \cdot d(FPR) \quad (18)$$

where TPR and FPR are the true-positive rate and false-positive rate, respectively. The AUC scores were close to 1 for the Myopia and Retina datasets and above 0.90 for the Jaundice dataset, which means that they have strong discriminatory capabilities. Figure 7 displays the ROC performance of the Myopia, Jaundice, and Retina datasets.

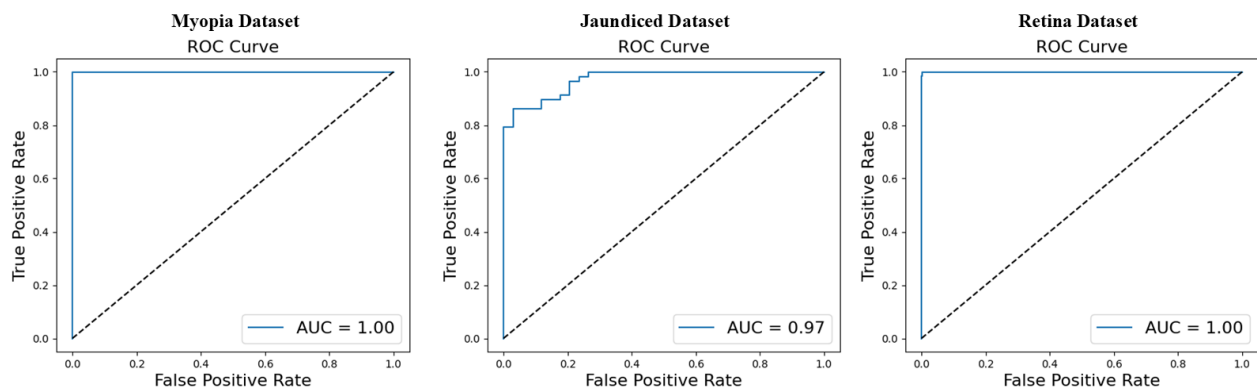


Figure 7. ROC curves for the Myopia, Retina, and Jaundice datasets, demonstrating the model's performance across different threshold settings.

The high accuracy of the Myopia, Jaundice, and Retina datasets in precision–recall analysis, as illustrated in Figure 8, also validated the high performance and stability of the model across datasets.

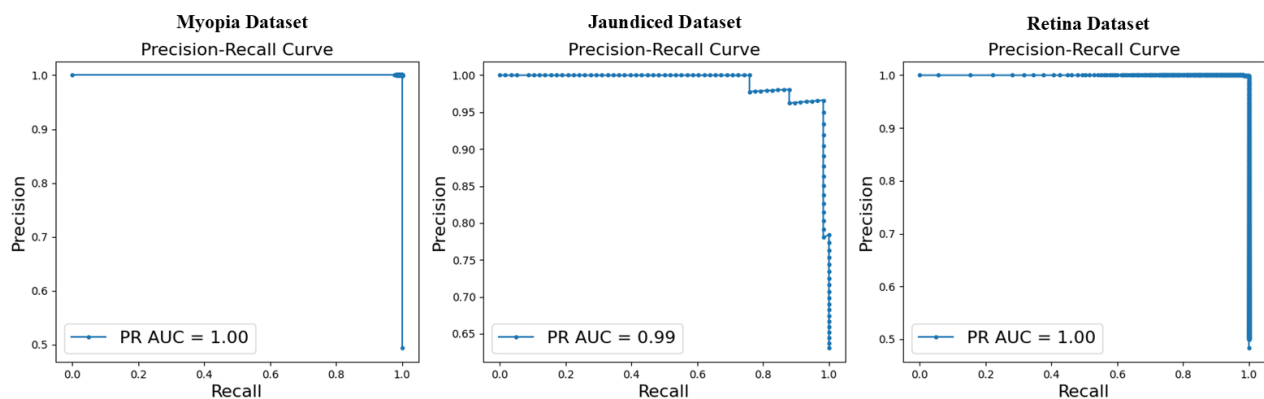


Figure 8. Precision and recall metrics for the Myopia, Retina, and Jaundice datasets, highlighting the model's accuracy in identifying true positives and minimizing false positives.

Overall, the experiments reveal that the hybrid architecture of MedViTNet, which consists of EyeConvNet and VisionNet, has higher classification accuracy, especially in the Myopia and Retina datasets, and can successfully generalize to a different dataset, namely Jaundice. The ablation experiment validates the significance of every element in the architecture, showing that the selection of EyeConvNet, the VisionNet setup, and the feature fusion scheme directly affects the performance of the model. Also, the model's interpretability based on Grad-CAM visualizations and the high scores in terms of AUC make MedViTNet a strong tool for use in automated eye disease classification, which can be applied to clinical diagnostic practice. Figure 9 shows sample images from three medical eye condition datasets, Myopia, Jaundice, and Retina, used for binary classification. In the Myopia dataset, images are classified as either myopic or normal, showcasing variations in retinal structure. The Jaundice dataset compares eyes with jaundiced (yellowed) sclera to normal eyes. The Retina dataset consists of images that represent pathological retinas, healthy retinas, and unrelated entities. Each row represents instances within each class that emphasize distinct visual features that distinguish affected eyes from healthy ones under various conditions, thereby improving the classification process.

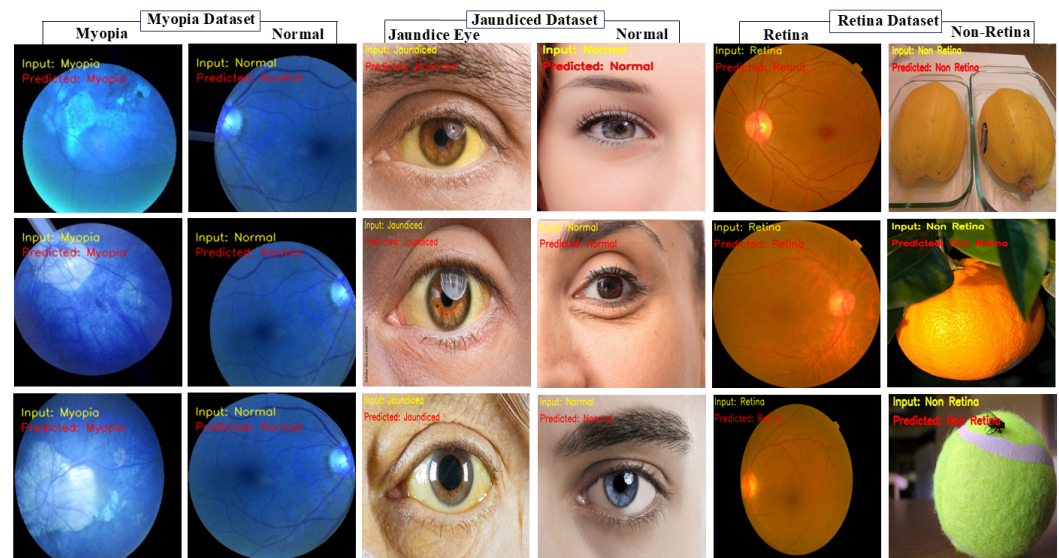


Figure 9. Sample images from Myopia, Jaundice, and Retina datasets, illustrating diseased versus normal classifications.

5. Discussion

The MedViTNet framework demonstrated strong performance in the automated classification of medical eye images across the Myopia, Retina, and Jaundice datasets. This effectiveness is supported by its hybrid local–global architecture, which integrates EyeConvNet and VisionNet. EyeConvNet, based on InceptionV3, extracts localized and multi-scale visual patterns, while VisionNet uses patch-based attention to capture broader spatial dependencies. The fusion of these complementary representations enables MedViTNet to learn discriminative features for accurate image-level classification.

In the primary evaluation, MedViTNet achieved 99.99% accuracy on the Myopia dataset, 99.93% accuracy on the Retina dataset, and 95.65% accuracy on the Jaundice dataset. The relatively lower performance on the Jaundice dataset is expected because this dataset contains only 464 images and consists of external eye-region images, which differ visually from the fundus-based Myopia and Retina datasets. Importantly, the three datasets were trained and evaluated independently as separate binary classification tasks and were not merged into a single training set; therefore, the difference in dataset size did not create a cross-dataset imbalance during model training.

To further strengthen validation reliability, a five-fold stratified cross-validation experiment was conducted on the Myopia dataset. Data augmentation was applied only to the training subset after splitting, and near-duplicate detection with perceptual-hash-based similarity checking identified only 100 near-duplicate image candidates, which were handled conservatively during validation. With element-wise addition fusion, MedViTNet achieved 98.33% mean accuracy with 1.43% standard deviation, 98.42% mean F1 score with 1.31% standard deviation, and 99.78% mean AUC with 0.21% standard deviation across the five folds. The pooled out-of-fold analysis further achieved 99.09% accuracy with a 95% confidence interval of 99.03% to 99.14%, supporting the robustness of the proposed framework.

For the Jaundice dataset, a class-stratified split was used to preserve the distribution of Jaundice and Non-Jaundice samples across the training, validation, and testing subsets. In addition, data augmentation, dropout, learning-rate scheduling, and early stopping were used to support generalization. MedViTNet achieved 95.65% accuracy, 94.12% precision, 94.12% recall/sensitivity, 96.55% specificity, and a 94.12% F1 score on this smaller dataset,

demonstrating that the proposed framework maintains strong performance under a more challenging dataset condition.

The interpretability of MedViTNet was supported using Grad-CAM visualizations, which highlight image regions contributing to the model's classification decision. Grad-CAM was used as a post hoc visual explanation method rather than supervised lesion localization. These visual explanations improve transparency and provide additional insight into the learned representations of the proposed model.

Although MedViTNet achieved strong results, some limitations remain. The Jaundice dataset is relatively small, and all datasets used in this study were publicly available datasets. Therefore, future work should validate the framework using larger, more diverse, and independent clinical datasets. Future studies will also extend MedViTNet to multiclass classification tasks. Overall, MedViTNet provides an effective hybrid framework for medical eye-image classification by combining localized multi-scale features with global contextual representations, supporting its potential for future screening-oriented computer-aided decision-support systems after further validation.

6. Conclusions

This study proposed MedViTNet, a hybrid deep learning framework that integrates EyeConvNet and VisionNet for automated classification of medical eye and eye-region images. By combining localized feature extraction from EyeConvNet with global contextual representation from VisionNet, MedViTNet achieved strong classification performance across three distinct binary classification tasks: Myopia, Retina, and Jaundice. Specifically, the Myopia dataset represents condition-based screening, the Retina dataset represents retina versus non-retina image classification, and the Jaundice dataset represents external ocular manifestation classification. Therefore, the reported results demonstrate the effectiveness of MedViTNet across heterogeneous medical eye-image classification tasks with different clinical meanings and difficulty levels. The results further show that combining CNN-based local feature learning with Transformer-style global contextual modeling improves image-level classification performance. The interpretability of MedViTNet was also demonstrated using Grad-CAM visualizations, which highlighted image regions that contributed to the model's classification decisions. These visual explanations enhance model transparency and may support future screening-oriented workflows for expert review. Although MedViTNet demonstrated strong performance, the current study was conducted using publicly available datasets, and external validation using independent clinical datasets was not included. In addition, the evaluated datasets represent different binary classification tasks and should not be interpreted as a direct clinical validation of generalized ocular disease diagnoses. Therefore, future work will extend MedViTNet to larger and more diverse datasets, including independent external clinical datasets from real-world medical settings. Future studies will also evaluate the proposed framework on multiclass eye-disease classification tasks to further assess its generalizability across broader ocular conditions. Moreover, future work may investigate the integration of large language model-based decision-support components to assist in clinical report generation, explanations of model predictions, and expert-guided interpretation. These extensions may further enhance the practical utility of MedViTNet for computer-aided medical eye-image analysis after appropriate clinical validation.

Author Contributions: Conceptualization, methodology, statistical analysis, and data analysis: S.D.; literature review, discussion, writing—original draft preparation, and data downloading: S.D. and M.F.; writing—review and editing: C.-M.L., A.N., A.M., H.M. and M.A.A.K.; visualization: S.D., C.-M.L. and A.N.; supervision: H.M. All authors have read and agreed to the published version of this manuscript.

Funding: This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government, Ministry of Science and ICT (MSIT) (RS-2026-25473871) and by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2025 (Project Name: Training Global Talent for Copyright Protection and Management of On-Device AI Models, Project Number: RS-2025-02221620, Contribution Rate: 100%) and by the "Regional Innovation System & Education (RISE)" through the Seoul RISE Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government. (2026-RISE-01-019-04) and by the National Science and Technology Council (NSTC), Taiwan, under Grants 113-2221-E-029-009-MY3 and 114-2634-F-027-001-MBK

Data Availability Statement: The data that support the findings of this study are openly available in the following public repositories: Myopia image dataset: <https://www.kaggle.com/datasets/kellysanderson/myopia-image-dataset> (accessed on 12 November 2024); Retina image dataset: <https://www.kaggle.com/datasets/sufyandanishuop/retina-image-dataset> (accessed on 30 June 2026); Jaundice and normal eye dataset: <https://www.kaggle.com/datasets/sufyandanishuop/jaundiced-and-normal-eyes> (accessed on 30 June 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. International Agency for the Prevention of Blindness (IAPB). Rethinking Global Health Means Including Eye Health. 2024. Available online: <https://www.iapb.org/news/rethinking-global-health-means-including-eye-health/> (accessed on 12 November 2024).
2. Thomsen, C.H.N.; Kronborg, T.; Hangaard, S.; Vestergaard, P.; Jensen, M.H. Developing an AI-Based clinical decision support system for basal insulin titration in type 2 diabetes in primary Care: A Mixed-Methods evaluation using heuristic Analysis, user Feedback, and eye tracking. *Int. J. Med. Inform.* **2025**, *195*, 105783. [CrossRef] [PubMed]
3. Xia, X.; Li, Y.; Xiao, G.; Zhan, K.; Yan, J.; Cai, C.; Fang, Y.; Huang, G. Benchmarking deep models on retinal fundus disease diagnosis and a large-scale dataset. *Signal Process. Image Commun.* **2024**, *127*, 117151. [CrossRef]
4. Cen, L.P.; Ji, J.; Lin, J.W.; Ju, S.T.; Lin, H.J.; Li, T.P.; Wang, Y.; Yang, J.F.; Liu, Y.F.; Tan, S.; et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nat. Commun.* **2021**, *12*, 4828. [CrossRef] [PubMed]
5. Danish, S.; Khan, A.; Dang, L.M.; Alonazi, M.; Alanazi, S.; Song, H.K.; Moon, H. Metaverse applications in bioinformatics: A machine learning framework for the discrimination of anti-cancer peptides. *Information* **2024**, *15*, 48. [CrossRef]
6. Zhang, W.; Dai, Y.; Liu, M.; Chen, Y.; Zhong, J.; Yi, Z. DeepUWF-plus: Automatic fundus identification and diagnosis system based on ultrawide-field fundus imaging. *Appl. Intell.* **2021**, *51*, 7533–7551. [CrossRef]
7. Panyarak, W.; Suttapak, W.; Mahasantipiya, P.; Charuakkra, A.; Boonsong, N.; Wantanajittikul, K.; Iamaroon, A. CrossViT with ECAP: Enhanced deep learning for jaw lesion classification. *Int. J. Med. Inform.* **2025**, *193*, 105666. [CrossRef] [PubMed]
8. Ghosh, T.; Sheikhi, F.; Guo, J.; Singh, Y.; Younis, K.; Kuanar, S.; Faghani, S.; Farina, E.M.J.d.M.; Huo, Y.; Maleki, F. Foundation Models for Volumetric Medical Imaging: Opportunities, Challenges, and Future Directions. *Electronics* **2026**, *15*, 1245. [CrossRef]
9. Man, Q.; Cho, Y.I. Transformer Based on Multi-Domain Feature Fusion for AI-Generated Image Detection. *Electronics* **2026**, *15*, 716. [CrossRef]
10. Thrishul, K.; Abugabah, A.; Salhi, A.; Ayadi, M.; Sithik, M.; Jayaprakash, D.; Ahilan, A. FIG-GAN: Fundus Image Generation via Deep Learning Based Generative Adversarial Network for AMD Disease Diagnosis. *Int. J. Comput. Intell. Syst.* **2026**, *19*, 210. [CrossRef]
11. Zong, C.; Gao, W.; Chen, S.; Xiao, F. Artificial Intelligence Models for Eye Disease Diagnosis: A Systematic Review. *Int. J. Imaging Syst. Technol.* **2026**, *36*, e70301. [CrossRef]
12. Alharthi, R.; Mehmood, R.; Albeshri, A. A Systematic, Scalable, and Interpretable Mapping of Artificial Intelligence Research in Leukemia Using a Hybrid Machine Learning and Qualitative Framework. *Electronics* **2026**, *15*, 1078. [CrossRef]
13. Babaqi, T.; Jaradat, M.; Yildirim, A.E.; Al-Nimer, S.H.; Won, D. Eye disease classification using deep learning techniques. *arXiv* **2023**, arXiv:2307.10501.
14. Triwijoyo, B.K.; Sabarguna, B.S.; Budiharto, W.; Abdurachman, E. Deep learning approach for classification of eye diseases based on color fundus images. In *Diabetes and Fundus OCT*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 25–57.
15. Jatmoko, C.; Lestiawan, H.; Agustina, F.; Erawan, L. Comparative Study of Classification of Eye Disease Types Using DenseNet and EfficientNetB3. *Kinet. Game Technol. Inf. Syst. Comput. Netw. Comput. Electron. Control* **2024**, *9*, 239–246.
16. Urooj, B.; Fayaz, M.; Ali, S.; Dang, L.M.; Kim, K.W. Large language models in medical image analysis: A systematic survey and future directions. *Bioengineering* **2025**, *12*, 818. [CrossRef] [PubMed]

17. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Rethinking the Inception Architecture for Computer Vision. *arXiv* **2016**, arXiv:1512.00567.
18. Pantelaïos, D.; Theofilou, P.A.; Tzouveli, P.; Kollias, S. Hybrid CNN-ViT models for medical image classification. In *Proceedings of the 2024 IEEE International Symposium on Biomedical Imaging (ISBI)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–4.
19. Bernabé, O.; Acevedo, E.; Acevedo, A.; Carreño, R.; Gómez, S. Classification of eye diseases in fundus images. *IEEE Access* **2021**, 9, 101267–101276. [[CrossRef](#)]
20. Rashid, F.; Abate, J.; Abdi, A. Multiclass Classification and Identification of the External Eye Diseases using Deep CNN. *Indian J. Sci. Technol.* **2024**, 17, 1107–1116. [[CrossRef](#)]
21. Du, F.; Zhao, L.; Luo, H.; Xing, Q.; Wu, J.; Zhu, Y.; Xu, W.; He, W.; Wu, J. Recognition of eye diseases based on deep neural networks for transfer learning and improved DS evidence theory. *BMC Med. Imaging* **2024**, 24, 19. [[CrossRef](#)] [[PubMed](#)]
22. Paradisa, R.H.; Bustamam, A.; Mangunwardoyo, W.; Victor, A.A.; Yudantha, A.R.; Anki, P. Deep feature vectors concatenation for eye disease detection using fundus image. *Electronics* **2021**, 11, 23. [[CrossRef](#)]
23. Ahmed, M.R.; Ahmed, S.R.; Duru, A.D.; Uçan, O.N.; Bayat, O. An expert system to predict eye disorder using deep convolutional neural network. *Acad. Platf.-J. Eng. Sci.* **2021**, 9, 47–52. [[CrossRef](#)]
24. Cho, Y.S.; Song, H.J.; Han, J.H.; Kim, Y.S. Attention Mechanism-Based Glaucoma Classification Model Using Retinal Fundus Images. *Sensors* **2024**, 24, 4684. [[CrossRef](#)] [[PubMed](#)]
25. Sanderson, K. Myopia Image Dataset. 2024. Available online: <https://www.kaggle.com/datasets/kellysanderson/myopia-image-dataset> (accessed on 12 November 2024).
26. Chhipa, H. Retina Image Dataset. 2024. Available online: <https://www.kaggle.com/datasets/hardikchhipa28/retina-image-for-binary-classification/data> (accessed on 9 December 2024).
27. Kumar, P. Jaundiced and Normal Eyes Dataset. 2023. Available online: <https://www.kaggle.com/datasets/puspendrakumar77/jaundiced-and-normal-eyes> (accessed on 9 December 2024).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.