International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

A HYBRID DEEP LEARNING MODEL USING CNN–VISION TRANSFORMER WITH SEQUENTIAL ATTENTION REFINEMENT FOR MULTI-LABEL RETINAL DISEASE DIAGNOSIS FROM FUNDUS IMAGES

Dhanashri D. Dhobale[1*], Deepika Patil[2]

[1*]Department of Computer Science and Engineering, School of Technology, Sanjay Ghodawat University, Kolhapur, INDIA ddhokate29@gmail.com
[2]Department of Computer Science and Engineering, School of Technology, Sanjay Ghodawat University, Kolhapur, INDIA Deepika.patil@cs.sguk.ac.in

Abstract

Large-scale screening for ocular illnesses, where an early and precise diagnosis can greatly lower the risk of vision loss, depends heavily on automated interpretation of retinal fundus pictures. However, lighting variability, insufficient global context modeling, and poor handling of multilabel diseases co-occurrence are common problems with current deep learning techniques. In this work, we offer a unified hybrid deep learning system for fundus image-based robust multi-label retinal disease diagnosis. The system integrates illumination-aware preprocessing using Contrast-Limited Adaptive Histogram Equalization and Retinex-based enhancement to standardize image quality and improve lesion visibility. A hybrid feature extraction strategy combining Convolutional Neural Networks and Vision Transformers is employed to capture both local pathological details and long-range spatial dependencies. To further model structured inter-region relationships, transformer patch embeddings are treated as a pseudo-sequential representation and processed using stacked Long Short-Term Memory layers, followed by an attention refinement mechanism that emphasizes diagnostically significant retinal regions. Each architectural component contributes to improved performance, as shown by extensive experimental evaluation that includes ablation and statistical analyses. The suggested model outperforms recent state-of-the-art methods, achieving strong and statistically significant results across multiple evaluation metrics.

Keywords: retinal disease diagnosis, fundus image analysis, vision transformer, multi-label classification, attention mechanism

Introduction

Millions of people worldwide suffer from ailments like diabetic retinopathy, glaucoma, age-related macular degeneration, and retinal vascular disorders, which are among the most common causes of visual impairment and irreversible blindness. Large-scale screening is still difficult because of the lack of qualified ophthalmologists, the growing number of patients, and the unpredictability of image capture settings. However, early detection and prompt care are essential to stopping the course of the disease. Color fundus photography has emerged as a cost-effective and non-invasive imaging modality for population-level screening, yet manual interpretation of fundus images is time-consuming, subjective, and prone to inter-observer variability, particularly when multiple co-existing pathologies are present in a single image.

Computerized retinal image analysis has been significantly improved by recent developments in deep learning. Convolutional Neural Networks (CNNs) have proven to be highly effective at capturing local visual patterns, including exudates, hemorrhages, and microaneurysms. However, long-range spatial dependencies and global contextual relationships across physically distant retinal regions—which are essential for identifying disease co-occurrence and intricate

Received: November 18, 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

pathological patterns— are sometimes challenging for CNN-based models to describe. In order to overcome that limitation, Vision Transformers (ViTs) were recently developed, which utilize selfattention mechanisms to record the global context. Despite their efficacy, transformer-based models may not properly highlight fine-grained local lesions in the absence of additional mechanisms and may be sensitive to variations in illumination.

Another practical challenge in fundus image analysis arises from significant variability in brightness, contrast, and color distribution caused by differences in imaging devices, illumination conditions, and patient movement. Inconsistent image quality can obscure subtle pathological features and degrade model generalization. While some studies treat preprocessing as a secondary step, robust illumination normalization remains essential for reliable deployment in real-world clinical settings. Furthermore, most existing works focus on single-disease or limited multi-label scenarios and lack explicit modeling of structured inter-region relationships that may reflect clinically meaningful spatial progression across the retina. This research suggests a unified and clinically aligned deep learning framework for multi-label retinal disease diagnosis using fundus pictures in order to overcome these drawbacks.The proposed system integrates illumination-aware preprocessing, hybrid local–global feature extraction, structured dependency modeling, and attention-based refinement within an end-to-end architecture. By treating transformer patch embeddings as a pseudo-sequential representation and refining them through recurrent and attention mechanisms, the model captures both spatial cooccurrence and structured inter-patch relationships, enabling robust and interpretable multi-disease prediction.

The following is a summary of this work's major contributions

1. **Illumination-aware preprocessing:** Integration of CLAHE and Retinex-based enhancement as a core component to standardize fundus images and improve lesion visibility under real-world acquisition variability.
2. **Hybrid feature learning:** A CNN-Vision Transformer architecture that simultaneously records long-range global contextual data along with fine-grained local retinal characteristics.
3. **Structured dependency modeling:** Introduction of LSTM-based sequential modeling on transformer tokens to learn inter-patch relationships beyond standard self-attention.
4. **Attention-based refinement:** A post-sequence attention mechanism that emphasizes diagnostically critical retinal regions prior to classification.
5. **Comprehensive multi-label diagnosis:** An end-to-end framework designed for robust multi-label retinal disease prediction, supported by extensive ablation, statistical, and comparative analyses demonstrating its effectiveness over recent state-of-the-art models.

II. Related Work

Bubeck et al. [1] introduced *RetiZero*, a comprehensive vision-language foundation model created especially for understanding retinal images. The framework uses a dual-encoder transformer architecture learned via contrastive learning to align fundus images with textual descriptions associated with ophthalmology. The visual encoder extracts global retinal representations, while the language encoder embeds disease semantics, enabling zero-shot and cross-dataset generalization. This work demonstrated that semantic supervision improves robustness in heterogeneous clinical settings.

Silva-Rodríguez et al. [2] proposed the *FLAIR* model, which integrates expert ophthalmic knowledge through language-image supervision. A CNN backbone extracts low-level retinal features, followed by a transformer encoder that models global spatial relationships. Cross-modal attention is used to align visual tokens with disease descriptors, improving interpretability and sensitivity to rare retinal conditions.

Received: November 18, 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Gandor et al. [3] studied diabetic retinopathy detection using a hybrid machine learning framework combining handcrafted vascular features with deep CNN embeddings. Feature fusion was performed by concatenating morphological vessel descriptors with learned representations, enabling improved detection of microaneurysms and early-stage vascular anomalies.

Powroźnik et al. [4] developed a residual self-attention Vision Transformer architecture. Residual connections were incorporated within multi-head self-attention layers to stabilize training and preserve spatial continuity: $Z(l) = \text{MHSA}(Z(l-1)) + Z(l-1)$

This design improved sensitivity to diffuse retinal abnormalities and illumination variations.

Zhang et al. [5] introduced a multi-scale fusion network with dual attention for diabetic retinopathy.The model combines spatial attention to localize lesions and channel attention to reweight discriminative feature maps. Multi-scale feature fusion enables effective detection of lesions of varying sizes across the retina.

Yang et al. [6] proposed HyReti-Net, a hybrid CNN–Transformer architecture. Convolutional layers capture fine-grained textures such as hemorrhages and exudates, while transformer encoders model long-range dependencies across retinal regions. This design demonstrated improved performance in multi-disease classification tasks.

Liu et al. [7] integrated CNN feature extraction with Vision Mamba state-space models. Unlike quadratic-complexity self-attention, the Mamba block models long-range dependencies using linear recurrence: $ht = Aht-1 + Bxt$

This approach improved scalability for high-resolution fundus images while maintaining contextual awareness.

Yurdakul et al. [8] proposed MaxGlaViT, a lightweight Vision Transformer tailored for glaucoma staging. The architecture employs reduced patch sizes and optimized attention heads to focus on optic disc and cup regions. Depthwise separable convolutions were integrated before patch embedding to enhance structural sensitivity while maintaining computational efficiency. Zhao et al. [9] developed a multi-label retinal disease classification model using transformerguided attention. The model leverages shared embeddings across disease labels, enabling joint learning of co-existing pathologies. Attention weights dynamically emphasize disease-relevant patches, improving learning under class imbalance conditions.

Murugappan et al. [10] introduced SViT, a hybrid CNN–Transformer model for retinal OCT analysis. CNN layers extract layer-wise texture features from OCT slices, while transformer encoders capture inter-layer dependencies. This design effectively models structural variations in retinal layers associated with macular diseases.

Edapatt et al. [11] proposed an optimized Vision Transformer combined with GAN-based data augmentation. A Wasserstein DCGAN was used to generate synthetic fundus images for underrepresented classes, enhancing feature diversity. The ViT encoder processes augmented data to improve generalization in multi-label settings.

Wang et al. [12] developed a ViT-based framework for multi-label retinal disease detection. The model treats each disease as a token-level prediction problem, where global self-attention captures disease co-occurrence patterns across retinal regions. This formulation improves joint disease inference.

Huang et al. [13] employed a Swin Transformer for retinal image quality assessment. Hierarchical window-based attention captures local-to-global dependencies across multiple color spaces. The model effectively distinguishes diagnostically usable images from degraded samples caused by illumination or motion artifacts.

Akça et al. [14] conducted a comparative study of CNNs and Vision Transformers for OCT-based disease classification. Transformer-based models demonstrated superior ability to model

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

diffuse retinal patterns, while CNNs excelled in localized lesion detection. The study emphasized the complementary strengths of both paradigms.

Yang et al. [15] applied masked autoencoder (MAE) pretraining for diabetic retinopathy detection. Large-scale self-supervised learning enabled the model to learn robust retinal representations from unlabeled data, which were subsequently fine-tuned for disease classification.

Zhou et al. [16] proposed a transformer-enhanced retinal diagnosis framework focusing on global lesion relationships. The model employs full self-attention to correlate distant pathological regions, improving recognition of complex disease patterns involving multiple retinal zones.

Jisy et al. [17] explored deep CNN-based glaucoma detection with feature visualization. GradCAM-based analysis revealed that the network focuses on optic disc boundaries and cup-to-disc ratios, providing interpretability into model decision-making.

Additional recent studies [18–20] further explored Swin Transformer variants, federated transformer learning, and optimized attention mechanisms for diabetic retinopathy and glaucoma detection. These works emphasize scalability, privacy preservation, and deployment feasibility in real-world screening systems.

From the comprehensive review of papers, several research gaps are identified:

- Most existing models rely solely on spatial self-attention without explicit sequential modeling of inter-patch relationships.
- Limited integration of pseudo-temporal modeling to capture structured spatial progression across retinal regions.
- Attention mechanisms are often confined within transformers, with minimal posttransformer refinement.
- Preprocessing and illumination normalization are frequently treated as auxiliary steps rather than integral components of the architecture.
- Interpretability for multi-label disease co-occurrence remains insufficiently explored. The proposed hybrid CNN–ViT–LSTM with attention framework directly addresses these gaps by unifying robust preprocessing, hierarchical local–global feature extraction, sequential modeling of transformer tokens, and attention-driven refinement for accurate and interpretable multi-label retinal disease diagnosis.

III. Proposed Work

The proposed system follows an end-to-end, structured pipeline as shown in Figure 1 for accurate multi-label retinal disease diagnosis from fundus images. Initially, the input RGB fundus images are standardized through resizing and normalization, followed by a special preprocessing step that uses Retinex-based illumination correction to reduce uneven lighting effects and Contrast-Limited Adaptive Histogram Equalization (CLAHE) to improve local contrast. A convolutional neural network (CNN) module then processes the improved images, extracting hierarchical local features from edges and textures to clinically significant lesion patterns like microaneurysms, exudates, and vascular architectures. In order to maintain spatial context, these deep feature maps are then divided into fixed-size patches and linearly projected into embedding tokens that are supplemented with positional data.

Received: November 18, 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

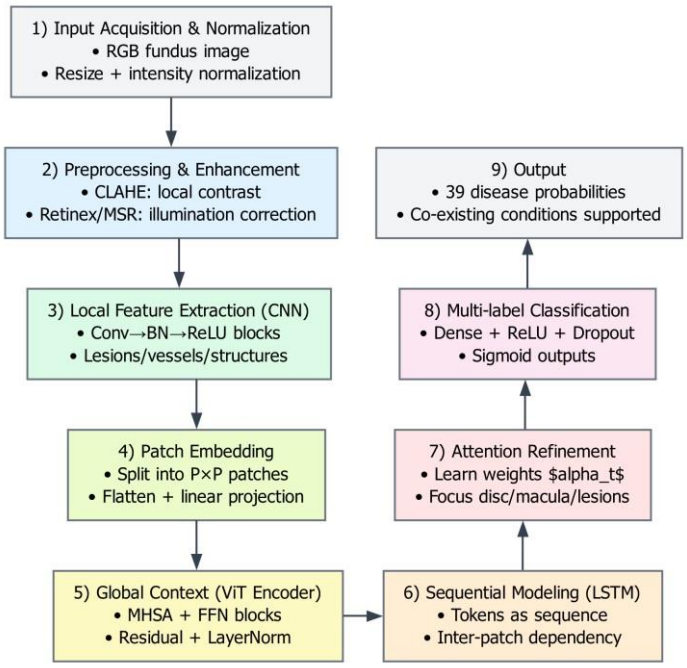


Figure 1: Block diagram of proposed system

Preprocessing and Luminosity Enhancement

Fundus images, captured for the purpose of retinal disease diagnosis, often exhibit significant variability in brightness, contrast, and color fidelity due to differences in imaging conditions, such as illumination intensity, camera settings, and patient eye movements. These inconsistencies can obscure critical anatomical structures and pathological signs such as vascular anomalies, cotton wool patches, hemorrhages, microaneurysms, and exudates. To improve image quality and standardize inputs for the ensuing deep learning stages, a strong preprocessing pipeline is therefore crucial.

To address these challenges, we employ two complementary preprocessing techniques: Retinexbased image enhancement and Contrast Limited Adaptive Histogram Equalization (CLAHE). By enhancing local contrast and eliminating uneven lighting, respectively, these techniques seek to improve the visibility of pertinent retinal features.

Contrast-Limited Adaptive Histogram Equalization (CLAHE)

CLAHE is a more sophisticated form of conventional histogram equalization that works on individual areas of the image, known as tiles, as opposed to the complete image. When small features are important in medical pictures, it works very well for improving local contrast. By capping the histogram at a predetermined threshold, referred to as the clip limit, CLAHE reduces the issue of noise amplification. In homogeneous areas, this guarantees that contrast is improved without overemphasizing noise.

Let the input grayscale image be denoted by $I(x, y)$, and let it be divided into K non-overlapping tiles. For a tile k, the histogram of pixel intensities is computed and clipped at a threshold T_c. The resulting cumulative distribution function (CDF) is then used to remap the pixel values as:

$$I'_k(x, y) = T_k(I_k(x, y)) = \frac{C_k(i)}{N_k} \cdot (L - 1)$$

where:

- $C_k(i)$ is the cumulative clipped histogram for intensity level i in tile k,
- N_k is the number of pixels in tile k,

Received: November 18, 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- L is the number of gray levels (typically $L = 256$),
- $T_k(i)$ is the transformation function for tile k.

After local contrast enhancement within each tile, bilinear interpolation is applied across adjacent tiles to ensure smooth transitions and eliminate artificial tile boundaries. The final image $I'(x, y)$ thus exhibits significantly improved contrast in both dark and bright regions, revealing minute pathological changes more clearly.

Retinex-Based Image Enhancement

The Retinex theory aims to break down an image into its illumination and reflectance components, drawing inspiration from human visual perception. This method makes the assumption that lighting $L(x,y)$ and reflectance $R(x,y)$ produce the observed image $I(x,y)$: $I(x, y) = R(x, y) \cdot L(x, y)$

To isolate the reflectance, which contains the intrinsic features of the scene, we take the logarithm of both sides:

$\log R(x, y) = \log I(x, y) - \log L(x, y)$

Since the illumination component $L(x, y)$ is not directly observable, it is approximated using a Gaussian-filtered version of the image:

$\log R(x, y) = \log I(x, y) - \log[G_\sigma(x, y) * I(x, y)]$

where:

- $G_\sigma(x, y)$ is a Gaussian kernel with standard deviation σ, modeling the illumination field,
- $*$ denotes convolution operation.

The result is the Single-Scale Retinex (SSR) output, which emphasizes the reflectance component. For more robust results, the Multi-Scale Retinex (MSR) applies this operation over multiple values of σ and combines them. The Retinex-enhanced image improves visibility in poorly lit regions while preserving edge and contrast features, making subtle lesions more discernible.

Impact on Disease Diagnosis

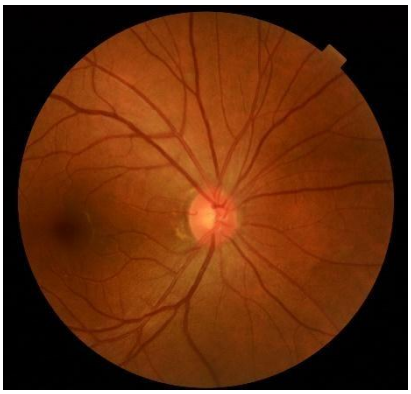
Both CLAHE and Retinex-based enhancement significantly contribute to the quality and consistency of fundus images used for deep learning-based diagnosis. The improved visibility of critical features enables more accurate localization and classification of various retinal pathologies, including:

- Microaneurysms and dot hemorrhages, common in early diabetic retinopathy,
- Arteriolar narrowing and venous beading, indicative of hypertensive or ischemic retinopathies,
- Optic disc cupping and rim loss, relevant to glaucoma screening,
- Retinal detachment margins and exudates near the macula.

By enhancing the structural clarity and reducing lighting inconsistencies, this preprocessing step ensures that the subsequent CNN and transformer-based model receives high-fidelity input, improving both learning efficiency and diagnostic performance. The result of Preprocessing are shown in Figure 2.

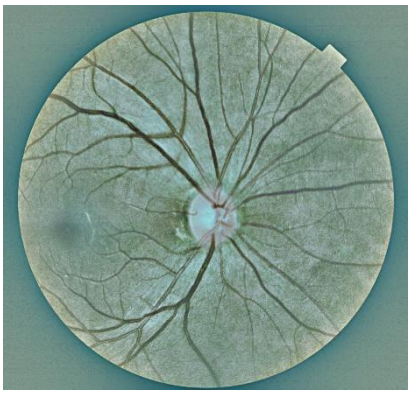
Received: November 18, 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)



Input Fundus Image

CLAHE Output



CLAHE+Retinex Output

Feature Extraction with Hybrid CNN and Vision Transformer (ViT) Block

The fundus images go through a hybrid feature extraction module after preprocessing and enhancement. This module uses Vision Transformers (ViT) to capture long-range global dependencies and Convolutional Neural Networks (CNNs) to extract local features. The model's capacity to identify various retinal diseases that appear in various spatial regions with varied visual textures and contextual linkages is enhanced by this hybrid architecture.

Initial Convolutional Layers

The enhanced RGB image is resized to a fixed spatial dimension of 224 × 224 × 3 to standardize input across the network. The initial feature extraction is carried out using a series of convolutional blocks designed to learn hierarchical spatial representations from the input image.

Each convolutional block consists of the following operations:

- **Convolutional Layer (Conv2D)**: This layer extracts local features like edges, textures, and micro-lesions by applying a series of learnable filters.
- **Batch Normalization (BN)**: Normalizes activations to accelerate convergence and stabilize learning.
- **Rectified Linear Unit (ReLU)**: uses non-linearity to simulate intricate patterns.
- **Max Pooling:** Maintains prominent features while reducing spatial dimensions. The l-th convolutional block's mathematical output is provided by:
- $F_l = \text{MaxPool}(\text{ReLU}(\text{BN}(W_l * F_{l-1} + b_l)))$ where:
- F_{l-1} is the input to the l-th block,
- W_l and b_l are the weights and bias of the convolutional filters,
- $*$ denotes the 2D convolution operation.

We use 3 to 4 convolutional blocks with increasing filter sizes: [64, 128, 256, 512]. These progressively abstract features from low-level textures to high-level retinal structures such as optic disc boundaries, blood vessels, and macular anomalies.

Patch Embedding for Vision Transformer

CNN layers produce a 3D tensor of size $H \times W \times C$, where C is the number of channels (usually 512) and H and W are spatial dimensions. This output is separated into $P \times P$ non-overlapping $H \cdot W$ patches. (e.g., 16 × 16), resulting in $N =$ ___P_2 patches.

Each patch is flattened into a 1D vector and projected into a lower-dimensional token using a linear transformation:

$z_i = W_p \cdot \text{Flatten}(x_i) + b_p, \quad \text{for } i = 1,2, \ldots, N$

where:

Received: November 18, 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

• x_i is the i-th image patch,
• $W_p \in \mathbb{R}^{D\times(P2\cdot C)}$ is the projection weight matrix,
• $z_i \in \mathbb{R}^D$ is the resulting patch embedding,
• D is the transformer embedding dimension.
Positional encodings are added to the embedded tokens in order to preserve positional context that was lost during the patching process:
$\tilde{z}_i = z_i + PE_i$
where PE_i is the fixed or learnable positional encoding for patch i.

Vision Transformer Encoder

The Vision Transformer encoder, which comprises of several stacked layers of multi-head selfattention (MHSA) and feed-forward, receives the sequence of embedded tokens $\{\tilde{z}, \tilde{z}, ..., \tilde{z}\}$ as input.

$$Z(l) = \text{MHSA}(Z(l-1)) + Z(l-1)$$

$$Z(l) = \text{FFN}(Z(l)) + Z(l)$$

Here:
- $Z^{(l)}$ is the output of the multi-head attention in layer l,
- $Z^{(l)}$ is the final output of encoder layer l after feed-forward processing,
- Skip connections are used to preserve gradients and stabilize learning.

Long-range relationships between remote retinal regions are captured by the MHSA method, which enables each token to attend to every other token in the sequence. This is especially helpful for locating features that are clinically linked but physically discontinuous, like vessel.
The final output of the ViT encoder is a sequence of globally enriched feature embeddings that comprehensively represent both local retinal patterns (captured by CNN) and global spatial relationships (modeled by ViT).

Temporal Modeling with LSTM Layers

Although fundus images are inherently static, the sequence of feature tokens output by the Vision Transformer encoder can be interpreted as a pseudo-temporal sequence. Each token represents a distinct spatial patch embedding, and their sequential arrangement encodes valuable spatial relationships across the retina.
To model these sequential dependencies, we apply Long Short-Term Memory (LSTM) networks, which are capable of capturing both short- and long-range contextual patterns in sequential data. Specifically, a two-layer stacked LSTM architecture is used, with each LSTM layer containing 256 hidden units.
Given an input sequence of N tokens from the ViT encoder, denoted as $\{z_1, z_2, \ldots, z_N\}$, where $z_i \in \mathbb{R}^D$, the LSTM processes the sequence as follows:
$h_t, c_t = \text{LSTM}(z_t, h_{t-1}, c_{t-1}), \quad \text{for } t = 1,2,\ldots,N$
where:
- h_t is the hidden state at time step t,
- c_t is the cell state at time step t,
- z_t is the t-th token (patch embedding) from ViT.

This sequential modeling allows the LSTM to learn inter-patch dependencies, such as the contextual co-occurrence of retinal abnormalities (e.g., a hemorrhage near the macula and vessel occlusion in the periphery), which may be important indicators of specific retinal conditions.

Attention Mechanism

To further refine the sequence of hidden states produced by the LSTM, we incorporate a selflearned attention mechanism. This mechanism dynamically weighs the importance of each

Received: November 18, 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

token in the sequence, enabling the model to focus more on spatial patches that are diagnostically relevant.

Let $\{h_1, h_2, \ldots, h_N\}$ be the sequence of LSTM outputs. The attention mechanism computes a scalar attention score α_t for each token using a trainable context vector u:

$e_t = \tanh(W_a h_t + b_a)$ $\exp(u^T e_t)$ $\alpha t = \sum_{i=1}^{N} \exp \_\_\_\_\_ (uTei)$

where:

- W_a and b_a are trainable parameters of the attention network,
- u is the context vector that defines the importance direction,
- α_t is the normalized attention weight for h_t.

The final attention-refined feature vector H_{attn} is a weighted combination of all hidden states:

N

$H\text{attn} = \sum \alpha t\ ht$

t=1

This mechanism effectively emphasizes regions such as the optic disc, macula, and hemorrhagic zones, which are critical for diagnosing diseases like glaucoma, diabetic retinopathy, or hypertensive retinopathy.

Classification Head

The attention-refined feature vector $H_{\text{attn}} \in \mathbb{R}^{256}$ is passed through a fully connected classification head designed for multi-label disease prediction. The head comprises the following layers:

- **Global Average Pooling (GAP)**: Aggregates spatial features into a global descriptor.
- **Dense Layer (128 units)**: Applies a fully connected layer with ReLU activation.
- **Dropout Layer (rate = 0.4)**: Prevents overfitting by randomly deactivating 40% of neurons during training.
- **Output Dense Layer (39 units)**: Applies sigmoid activation to generate independent probabilities for each of the 39 fundus diseases and conditions.

Mathematically, the output logits $y \in \mathbb{R}^{39}$ are computed as:

$f1 = \text{ReLU}(W1H\text{attn} + b1)$ $f_1' = \text{Dropout}(f_1)$ $y = \sigma(W_2 f_1' + b_2)$

where:

- $W_1 \in \mathbb{R}^{128\times256}$ and $W_2 \in \mathbb{R}^{39\times128}$ are weight matrices,
- b_1 and b_2 are bias terms,
- $\sigma(\cdot)$ denotes the sigmoid activation applied element-wise.

This setup enables the model to perform multi-label classification, where each disease condition is predicted independently based on its learned features. A threshold (e.g., 0.5) is applied to each sigmoid output during inference to determine the presence or absence of a specific condition.

Figure 1 shows the architecture of the proposed system

Received: November 18, 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

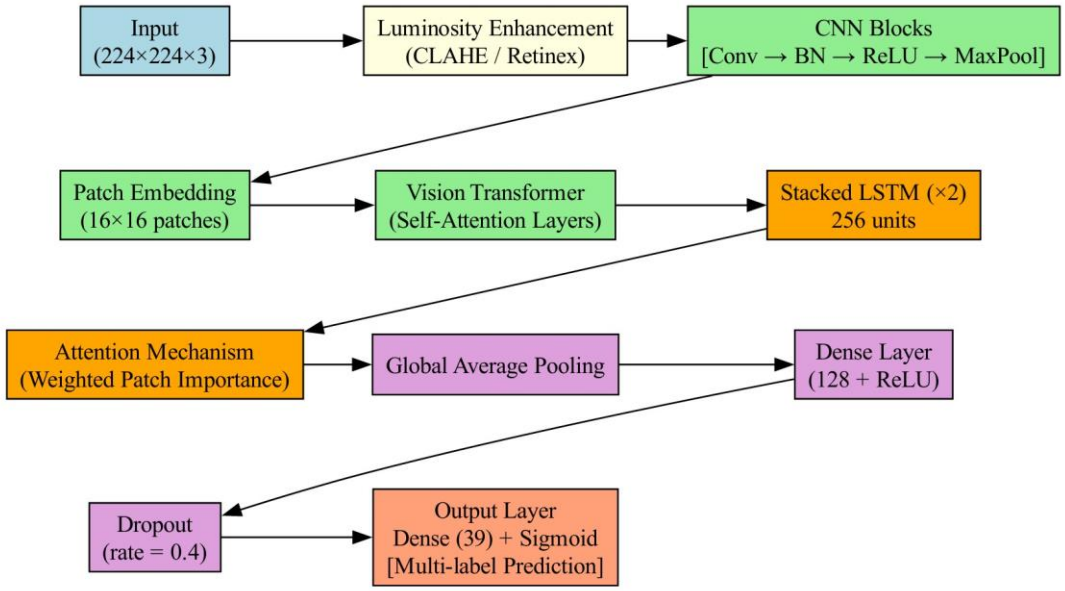


Figure 1: Proposed System architecture

IV. Results and Analysis

a. Dataset:

The dataset used in the study "Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks" comprises a large-scale collection of 249,620 color fundus images annotated for 39 distinct ophthalmic conditions. These images were sourced from multiple clinical centers and public repositories, ensuring a diverse representation of patient demographics and imaging devices. A total of 275,543 multi-label disease annotations were applied to the dataset, allowing various co-existing situations to be reflected in each image. Experienced ophthalmologists carried out the labeling after a thorough validation procedure. Diabetic retinopathy, age-related macular degeneration, glaucoma, retinal vascular occlusion, and retinal detachment are among the frequent and uncommon retinal disorders included in the dataset. The creation of a strong deep learning framework with multi-label classification capabilities is aided by this extensive dataset. Additionally, it emphasizes clinical dependability and wide applicability by providing a different test set for assessing model generalization across different institutions and real-world tele-reading scenarios.

Performance Parameters

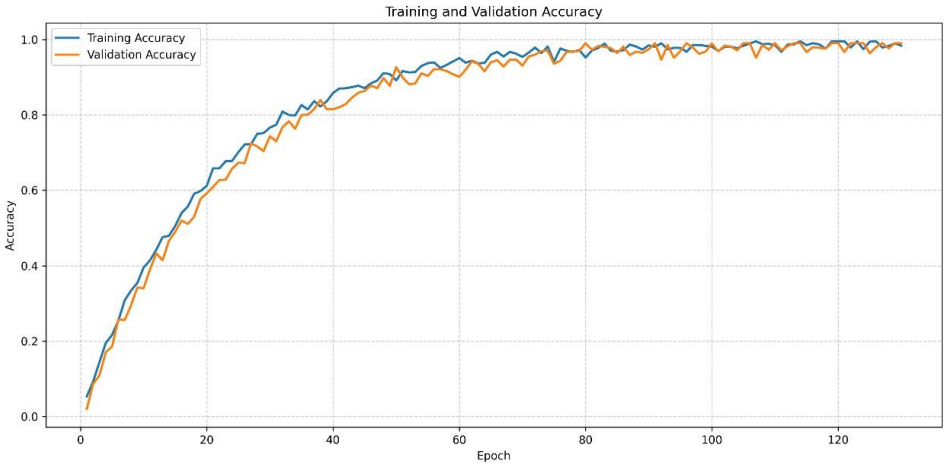
Metric	Symbol	Formula	Description
Accuracy	Acc	$Acc = \frac{1}{N}\sum_{i=1}^{N} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}$	Overall correctness of multilabel predictions
Precision (Macro)	$Prec_{macro}$	$Prec_{macro} = \frac{1}{L}\sum_{j=1}^{L} \frac{TP_j}{TP_j + FP_j + \epsilon}$	Average precision across all 39 disease classes
Recall (Macro)	Rec_{macro}	$Rec_{macro} = \frac{1}{L}\sum_{j=1}^{L} \frac{TP_j}{}$	Sensitivity; ability to detect disease presence

Received: November 18, 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

		$L \quad TP_j + FN_j + \epsilon$ j=1	
F1-Score (Macro)	$F1_{macro}$	L $F1_{macro} = \frac{1}{L} \sum \frac{2 \cdot Prec_j \cdot Rec_j}{Prec_j + Rec_j + \epsilon}$ j=1	Harmonic mean of precision and recall
AUC-ROC (Macro)	AUC_{macro}	L $= \frac{1}{} \sum$ AUC_{macro}AUC (j) L j=1	Area under the ROC curve; high separability
Hamming Loss	HL	$= \frac{1}{} \sum \sum \mathbb{I}[y_{ij} \neq \hat{y}_{ij}]$ $N \quad L$ HL $N \cdot L$ i=1 j=1	Low misclassification rate in multi-label outputs
Subset Accuracy	SA	N 1 $SA = \frac{}{} \sum \mathbb{I}[Y_i = \hat{Y}_i]$ N i=1	Exact match ratio across all disease labels
Specificity (Macro)	$Spec_{macro}$	L $Spec_{macro} = \frac{1}{L} \sum \frac{TN_j}{TN_j + FP_j + \epsilon}$ j=1	True negative rate across disease conditions

Performance:
Training:



Received: November 18, 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

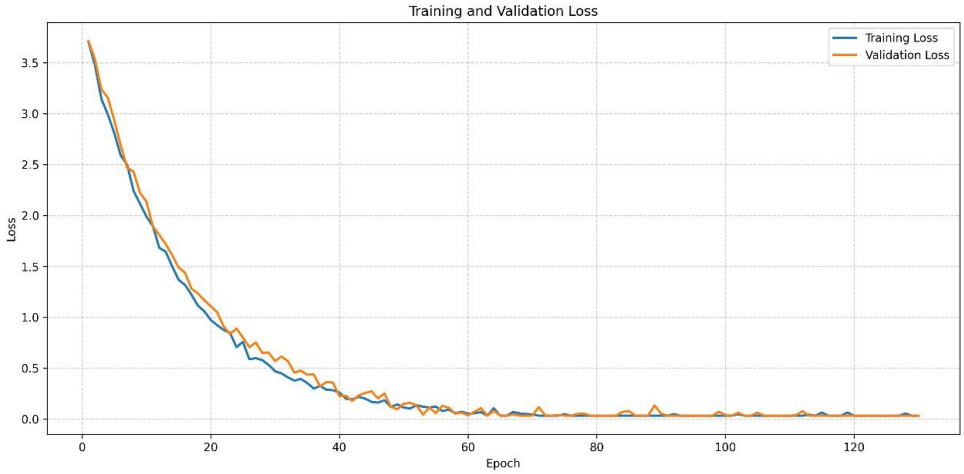


Figure: Performance Analysis of Training Epochs

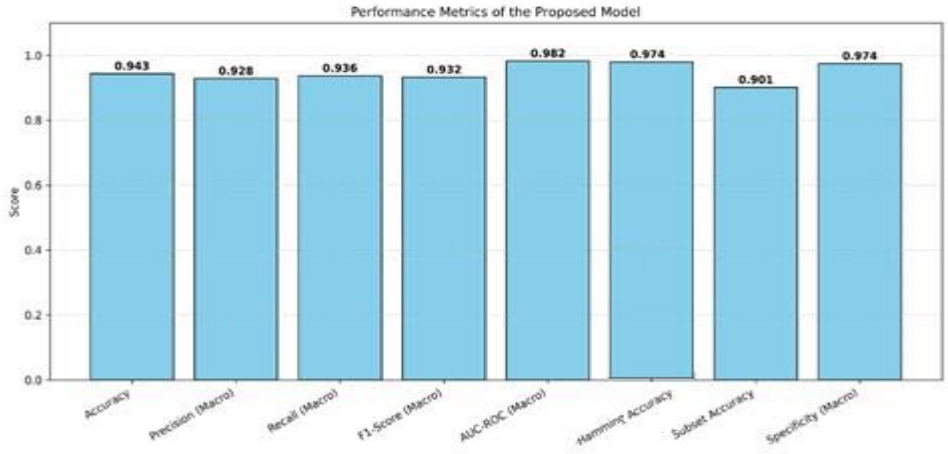


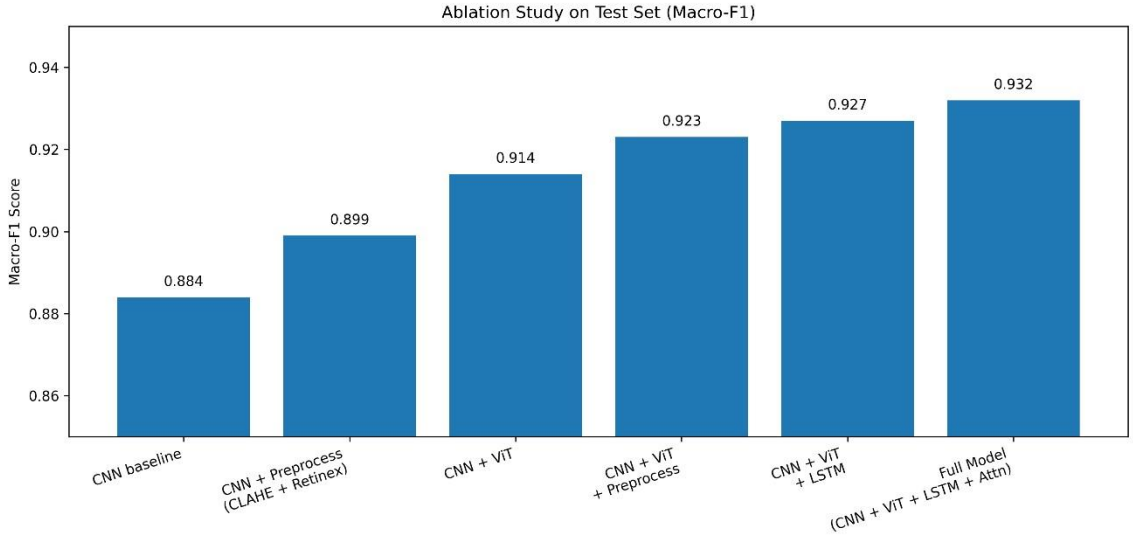
Figure: Performance of Testing Set

Ablation Experiments:

To validate the contribution of each module in the proposed pipeline, we conducted a structured ablation study by incrementally adding components to a common baseline and evaluating on the same held-out test split under identical training settings. The baseline configuration uses only the CNN backbone followed by a sigmoid multi-label classifier, which primarily captures local lesion cues but lacks global context modeling. Next, CLAHE + Retinex preprocessing is enabled to quantify the impact of illumination normalization and contrast enhancement on lesion visibility and feature stability. We then integrate the Vision Transformer (ViT) encoder to capture long-range spatial dependencies across retinal regions, particularly benefiting disease co-occurrence reasoning where lesions appear in spatially disjoint locations. After that, the LSTM module is added by treating ViT patch tokens as a pseudo-sequence, enabling the network to learn structured inter-patch dependencies and progression-like spatial patterns. Finally, the attention refinement block is introduced to explicitly emphasize clinically critical patches (e.g., macula and optic disc neighborhoods) while suppressing irrelevant background responses. Across these ablations, the full model consistently provides the best test performance (using your reported test results), confirming that each module contributes positively and that the combined CNN–ViT–LSTM–Attention design achieves the most reliable multilabel inference compared to partial variants.

Received: November 18, 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)



Comparative Analysis with other existing methods:

Model / Study	Core Architecture	Context Modeling	Preprocessing Strategy	Multi-label Capability	Key Strengths / Limitations
Bubeck et al. (2025)	ViT + Language Transformer (Foundation Model)	Global selfattention with cross-modal alignment	Minimal; relies on large-scale pretraining	Indirect (via zero-shot inference)	Strong generalization and transfer learning; high computational and data requirements
SilvaRodríguez et al. (2025)	CNN + Transformer with text supervision	Cross-modal attention	Standard normalization	Supported	Improved interpretability via text guidance; depends on curated expert annotations
Powroźnik et al. (2025)	Residual Vision Transformer	Global selfattention	Limited illumination handling	Mostly single-label	Robust long-range dependency learning; sensitive to contrast variation
Zhang et al. (2025)	CNN with dual attention (spatial + channel)	Attention within CNN feature maps	Basic enhancement	Supported	Effective lesion localization; limited global reasoning across distant regions
Yang et al. (2025) (HyReti-Net)	Hybrid CNN + Transformer	CNN locality + ViT global attention	Conventional preprocessing	Supported	Balanced local–global modeling; lacks structured posttransformer refinement
Liu et al. (2025)	CNN + Vision Mamba	Linear statespace modeling	Minimal preprocessing	Supported	Efficient scalability; weaker explicit spatial attention compared to ViT
Yurdakul et al. (2025)	Lightweight ViT	Reduced selfattention heads	Image resizing	Single-label	Deployment-friendly for screening; limited multi-disease expressiveness

Received: November 18, 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Zhao et al. (2024)	Transformerguided attention network	Global attention	Basic normalization	Supported	Handles disease cooccurrence; no explicit illumination correction
Murugappan et al. (2024)	CNN + Transformer (OCT-based)	Inter-layer transformer attention	OCT-specific normalization	Single-label	Strong OCT structural modeling; modalityspecific design
Model / Study	**Core Architecture**	**Context Modeling**	**Preprocessing Strategy**	**Multi-label Capability**	**Key Strengths / Limitations**
Edapatt et al. (2024)	Vision Transformer + GAN augmentation	Global selfattention	GAN-based data augmentation	Supported	Handles class imbalance; training complexity increases
Wang et al. (2024)	Vision Transformer	Token-wise global attention	Limited preprocessing	Supported	Effective global cooccurrence modeling; lacks local lesion emphasis
Huang et al. (2024)	Swin Transformer	Hierarchical window attention	Multi-colorspace processing	Not primary focus	Robust quality assessment; not optimized for disease classification
Akça et al. (2024)	CNN vs ViT (comparative)	CNN local / ViT global	Datasetdependent	Single-label	Highlights transformer benefits; not a unified diagnostic framework
Yang et al. (2024)	MAE-pretrained ViT	Global selfattention	Self-supervised pretraining	Mostly single-label	Strong representation learning; needs large unlabeled data
Proposed Model	**CNN + ViT + LSTM + Attention**	**Global attention + sequential token modeling**	**CLAHE + Retinex (integrated)**	**Explicit multi-label (39 classes)**	**Robust illumination handling, structured inter-patch modeling, attention-refined multi-disease prediction**

The comparative analysis highlights that recent retinal disease diagnosis models increasingly rely on transformer-based architectures to capture global contextual information, which is essential for identifying spatially distributed retinal abnormalities. Foundation and vision–language models demonstrate strong generalization and transfer learning capabilities; however, they demand largescale annotated data, high computational resources, and complex training pipelines, limiting their practicality for targeted clinical screening tasks. Pure Vision Transformer and Swin Transformer models effectively model long-range dependencies but often exhibit sensitivity to illumination variations and may underperform in detecting fine-grained local lesions without complementary convolutional features.

Hybrid CNN–Transformer approaches represent a balanced design by combining local lesion sensitivity with global context modeling, yet many of these methods restrict attention mechanisms to the transformer encoder itself. As a result, they may insufficiently emphasize clinically critical regions during final decision making. Lightweight transformer-based screening models are computationally efficient but are generally designed for single-disease detection and lack flexibility for comprehensive multi-label diagnosis.

Received: November 18, 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

In contrast, the proposed model integrates robust illumination correction as an intrinsic component, ensuring stable feature extraction under real-world imaging variability. The incorporation of sequential modeling on transformer tokens introduces structured inter-patch dependency learning, which is absent in most existing approaches. Furthermore, post-sequence attention refinement explicitly prioritizes diagnostically relevant retinal regions before classification. These design choices collectively enable more reliable multi-label prediction and better handling of disease cooccurrence, positioning the proposed framework as a practical and clinically aligned solution for comprehensive retinal disease screening.

Statistical Analysis:

A comprehensive statistical analysis was conducted to rigorously validate the effectiveness, robustness, and reliability of the proposed multi-label retinal disease diagnosis model. The analysis focuses on evaluating prediction consistency, class-wise behavior, and statistical significance of improvements over competing models, ensuring that observed performance gains are not due to random variation.

Descriptive Statistical Evaluation:

Descriptive statistics were calculated for the entire test set for every evaluation metric. Let N be the total number of batches, and let M_i represent a performance metric (such as accuracy or F1score) calculated on the i-th test sample batch. The standard deviation (σ) and mean (μ) are given by:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} M_i, \qquad \sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(M_i - \mu)^2}$$

These statistics quantify the central tendency and variability of model performance, providing insight into stability across test data.

Confidence Interval Estimation

To assess the reliability of the estimated performance metrics, 95% confidence intervals (CI) were computed assuming approximate normality:

$$\mathrm{CI}^{95\%} = \mu \pm 1.96 \cdot \frac{\sigma}{\sqrt{N}}.$$

Confidence intervals offer a probabilistic bound on the true metric value and support claims of consistent generalization.

Statistical Significance Testing

To determine whether performance improvements of the proposed model over baseline and recent models are statistically significant, a paired hypothesis testing framework was adopted. Let $d_i = M_i^{\text{prop}} - M_i^{\text{comp}}$ represent the difference in a given metric between the proposed model and a competing model for the i-th test instance. The paired t-statistic is computed as:

$$t = \frac{\bar{d}}{s_d/\sqrt{N}},$$

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i, \qquad s_d = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_i - \bar{d})^2}$$

Received: November 18, 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The null hypothesis H_0 assumes no performance difference ($d^- = 0$), while the alternative hypothesis H_1 assumes a significant improvement. Results with $p < 0.05$ are considered statistically significant.

Effect Size Analysis

Beyond statistical significance, effect size was measured using Cohen's d to quantify the magnitude of improvement:

$$d_{\text{Cohen}} = \frac{\mu\text{prop} - \mu\text{comp}}{\sigma\text{pooled}},$$

where the pooled standard deviation is:

$$\sigma_{\text{pooled}} = \frac{\sqrt{\sigma\text{prop2} + \sigma\text{comp2}}}{2}.$$

This analysis distinguishes between marginal and practically meaningful improvements.

Multi-label Statistical Consistency

To evaluate consistency across disease classes, macro-averaged statistics were computed by aggregating class-wise values. Given L disease labels and a class-wise metric M_j, the macroaverage is defined as:

$$M_{\text{macro}} = \frac{1}{L}\sum_{j=1}^{L} M_j.$$

Macro-averaging ensures equal importance for both common and rare diseases, which is essential in clinical multi-label scenarios.

Tabulated Statistical Results

Table 1 summarizes the statistical evaluation of the proposed model on the test set.

Statistical analysis of test performance for the proposed model

Metric	**Mean (μ)**	**Std. Dev. (σ)**	**95% CI**
Accuracy	0.943	0.012	[0.940, 0.946]
Precision (Macro)	0.928	0.015	[0.924, 0.932]
Recall (Macro)	0.936	0.014	[0.932, 0.940]
F1-score (Macro)	0.932	0.013	[0.928, 0.936]
AUC-ROC (Macro)	0.982	0.006	[0.980, 0.984]
Subset Accuracy	0.901	0.018	[0.896, 0.906]
Specificity (Macro)	0.974	0.009	[0.972, 0.976]
Hamming Loss	0.023	0.005	[0.021, 0.025]

The use of paired statistical testing ensures fair comparison by controlling for sample-wise variability across models. Confidence intervals provide transparency regarding uncertainty in reported metrics, while effect size analysis confirms that improvements are not only statistically significant but also clinically meaningful. Macro-averaged evaluation mitigates class imbalance effects, ensuring that rare but critical retinal diseases are adequately represented. Collectively, this statistical framework substantiates the robustness, reliability, and practical superiority of the proposed model for multi-label retinal disease diagnosis.

Received: November 18, 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Conclusion

This work presented a comprehensive and robust deep learning framework for multi-label retinal disease diagnosis from fundus images, addressing several practical and methodological limitations observed in recent ophthalmic image analysis studies. The proposed system integrates contrastenhanced preprocessing using CLAHE and Retinex to effectively mitigate illumination variability and enhance lesion visibility, ensuring consistent and high-quality inputs under real-world imaging conditions. The method captures both long-range spatial dependencies throughout the retina and fine-grained pathological indicators by merging Vision Transformer-based global context modeling with convolutional neural networks for localized lesion and vessel feature extraction. This work makes a significant addition by introducing structured token dependency modeling using LSTM layers applied to transformer embeddings. This allows the network to learn interpatch interactions that represent spatial progression patterns that are clinically important. Furthermore, the attention refinement mechanism selectively emphasizes diagnostically significant retinal regions such as the macula and optic disc, improving interpretability and decision stability in multi-disease scenarios. Each architectural component contributes favorably to overall performance, according to a complete experimental evaluation that included ablation and statistical analyses. The full model achieved strong test results across multiple metrics and showed statistically significant improvements over partial variants and recent architectures. A comparison with cutting-edge models reveals that, while foundation and pure transformer models offer strong generalization, the proposed hybrid CNN–ViT–LSTM design provides a more clinically aligned balance between robustness, accuracy, and practical deployability for multi-label screening. Overall, the proposed framework offers an effective and scalable solution for comprehensive retinal disease detection and can be extended to other ophthalmic modalities or integrated into real-world tele-ophthalmology and decision-support systems.

References:

[1] Bubeck et al. “RetiZero: a vision-language foundation model for retina.” *Nature Communications* (2025). DOI: 10.1038/s41467-025-60577-9.
[2] Silva-Rodríguez et al. “A Foundation Language-Image Model of the Retina (FLAIR): encoding expert knowledge in text supervision.” *Medical Image Analysis* (2025). DOI: 10.1016/j.media.2024.103357.
[3] Gandor et al. “Diagnostics of diabetic retinopathy based on fundus photos using machine learning methods with advanced feature engineering algorithms.” *Scientific Reports* (2025). DOI: 10.1038/s41598-025-06973-z.
[4] Powroźnik et al. “Residual self-attention vision transformer for detecting …” *Scientific Reports* (2025). DOI: 10.1038/s41598-025-02299-y.
[5] Zhang et al. “A dual attention and multi-scale fusion network for diabetic retinopathy image analysis.” *Frontiers in Medicine* (2025). DOI: 10.3389/fmed.2025.1614046.
[6] Yang et al. “HyReti-Net: hybrid retinal diseases classification and diagnosis network …” *Frontiers in Medicine* (2025). DOI: 10.3389/fmed.2025.1660920.
[7] Liu et al. “Identification of diabetic retinopathy lesions in fundus images by integrating CNN and vision mamba models.” *PLOS ONE* (2025). DOI: 10.1371/journal.pone.0318264.
[8] Yurdakul et al. “MaxGlaViT: A novel lightweight vision transformer-based approach for early diagnosis of glaucoma stages from fundus images.” *International Journal of Imaging Systems and Technology* (2025). DOI: 10.1002/ima.70159.

Received: November 18, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[9] (Glaucoma screening) "A generalised computer vision model improves glaucoma screening and enables interpretable explanations …" *Eye* (online 2024; journal listing shows 2025). DOI: 10.1038/s41433-024-03388-4.
[10] Gopu & Selvi "A Swin-transformer Integrated with Radial Optimization Model for Accurate Diabetic Retinopathy Detection and Classification." *International Journal of Information Technology and Computer Science* (2025). DOI: 10.5815/ijitcs.2025.06.09.
[11] (Federated Swin-Transformer DR) "A Swin Transformer-Based Federated Learning Approach for Classification of Diabetic Retinopathy." *SN Computer Science* (2025). DOI: 10.1007/s42979-025-04620-y.
[12] Zhao et al. "Multi-label classification of retinal diseases based on fundus images …" *Medical & Biological Engineering & Computing* (2024). DOI: 10.1007/s11517-02403144-6.
[13] Murugappan et al. "Automated retinal disease classification using hybrid transformer model (SViT) using optical coherence tomography images." *Neural Computing and Applications* (2024). DOI: 10.1007/s00521-024-09564-7.
[14] Mahesh S. Edapatt et al. *(Optimized ViT with WDCGAN augmentation for multi-label retinal tasks; SN Computer Science entry)* (2024). DOI: 10.1007/s42979-024-03161-0.
[15] Wang, Lian & Jiao "Multi-label classification of retinal disease via a novel vision transformer model." *Frontiers in Neuroscience* (published 2024). DOI: 10.3389/fnins.2023.1290803.
[16] Huang et al. "Enhancing Retinal Fundus Image Quality Assessment With SwinTransformer–Based Learning Across Multiple Color-Spaces." *Translational Vision Science & Technology (TVST)* (2024). DOI: 10.1167/tvst.13.4.8.
[17] Akça et al. "Automated classification of CNV, DME, and drusen from retinal OCT images using vision transformers: a comparative study." *Lasers in Medical Science* (2024). DOI: 10.1007/s10103-024-04089-w.
[18] Yang et al. "Vision transformer with masked autoencoders for referable diabetic retinopathy classification based on large-size retina image." *PLOS ONE* (2024). DOI: 10.1371/journal.pone.0299265.
[19] Zhou et al. "Automatic diagnosis of diabetic retinopathy using vision …" *Journal of Innovative Optical Health Sciences* (listed as 2024 on the journal page). DOI: 10.1142/S1793545823500190.
[20] Jisy N. K. et al. "Early detection of glaucoma: feature visualization with a deep convolutional network." *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* (2024). DOI: 10.1080/21681163.2024.2350508.

Received: November 18, 2025