

Early Detection of Retinal Microaneurysms Through Neurovascular Remodeling-Aware Transformer-LSTM Network

R. Deepa^{1,2,*} and N. K. Narayanan³

¹Department of Information Technology, Kannur University, Kannur, Kerala 670567, India

²Department of Computer Applications, College of Engineering, Vadakara, Kozhikode, Kerala 673104, India

³Indian Institute of Information Technology, Valavoor, Kottayam, Kerala 686635, India

*E-mail: deepashaaju@gmail.com

Received for publication
December 26, 2025.

Abstract

Retinal microaneurysms (MAs), crucial for early diabetic retinopathy detection via fundus imaging, are difficult to identify due to the complexity of retinal pathology. Hence, a novel “Neurovascular Fourier-Otsu Vision Support Transformer Network with Long Short-Term Memory (NFOVSTN-LSTM)” is proposed to improve retinal MA detection and analysis using retinal fundus imaging. However, macular degeneration caused by neurovascular remodeling leads to vascular alterations and intensity fluctuations that are not directly associated with visible lesions. Thus, a novel Neurovascular Remodeling Detection Network (NRDN) is presented to enhance vascular contrast and texture by capturing subtle variations and detecting irregular branching, capillary dropout, and fragmented vascular patterns. In addition, extracting MA candidates is challenging due to the dense neural architecture of the region, which generates significant background clutter and variability in image texture. Hence, a Fourier-Otsu Hough Graph Transformer Network (FOHGTN) is used to improve MA localization by reducing background clutter, identifying subtle circular lesions, and capturing spatial and anatomical relationships. Meanwhile, silent ischemia complicates early detection of MAs, due to reduced perfusion in the absence of visible retinal biomarkers. Hence, a Local Vision Support Machine Transformer (LVSMT) is introduced to accurately distinguish MAs from ischemic regions by combining local texture features with global context, enabling precise classification even in the absence of visible biomarkers. Experimental results demonstrate improved accuracy, higher sensitivity, and lower loss, confirming the proposed method’s superior performance in MA detection compared to existing approaches.

Keywords

Microaneurysm detection, diabetic retinopathy, vision loss, fovea, ischemic, etc.

Abbreviations

AI: Artificial Intelligence; ALS: Amyotrophic Lateral Sclerosis; BiLSTM: Bidirectional Long Short-Term Memory; GRU: Gated Recurrent Unit; CFP: Color Fundus Photography; CPU: Central Processing Unit; DCNN: Deep Convolution Neural Network; FCN: Fully Convolution Network; FPN: Feature Pyramid Network; ISR: Image Super

Resolution; KNN: K-Nearest neighbor; MAFP: Magnified Adaptive Feature Pyramid Network; MA-YOLO: Microaneurysms You Only Live Once; MLP: Multilayer Perceptron; NFOVSTN: Neurovascular Fourier-Otsu Vision Support Transformer Network; NPDR: Nonproliferative Diabetic Retinopathy; OCT-A: Optical Coherence Tomography Angiography; PSO: Particle Swarm Optimization; RCNN: Region-based Convolutional Neural Network; RBF: Radial Basis Function; RGB: Red Green Blue; RNN: Recurrent Neural Network; RMSE: Root Mean Square Error; SBF: Sliding Band Filters; SSD: Single Shot multibox Detector; SwinIR: Swin Transformer Image Restoration; VCG: Vector Cardiogram; VGG: Visual Geometry Group; VGG-Net: Visual Geometry Group.

I. Introduction

Microaneurysms (MAs) are the earliest clinical signs of diabetic retinopathy (DR) and occur primarily in the inner retinal layers, especially within the capillary networks of the retina. These small, round outpouchings of capillary walls result from the weakening of blood vessel walls due to prolonged hyperglycemia in diabetic patients. Most commonly found in the posterior pole of the retina near the macula, MAs can leak fluid and blood, leading to retinal edema and vision impairment. Their presence indicates localized vascular damage and increased vascular permeability. Over time, if untreated, they contribute to progressive retinal degeneration and can significantly compromise visual acuity. MAs are typically detected through fundus photography or fluorescein angiography. Their development reflects an early and critical phase in the pathogenesis of DR. Early detection and management are essential to prevent further complications. Understanding their role helps in the timely diagnosis and treatment of diabetic retinal diseases [1–4].

To support the early and accurate detection of retinal MAs, image preprocessing becomes a crucial initial step in the diagnostic pipeline. However, several challenges arise during preprocessing, including variability in image quality, low contrast between MAs and the retinal background, uneven illumination, and noise introduced during image acquisition. These factors can significantly hinder accurate detection by causing false positives or missed lesions. Preprocessing techniques, such as histogram equalization, Gaussian filtering, and green channel extraction, help to enhance image clarity by normalizing brightness, improving contrast, and isolating key structures. As a result, preprocessing enhances the performance of subsequent detection algorithms. Existing models, such as LeNet, U-Net, ResNet,

and DeepDR, have incorporated these techniques effectively, while hybrid approaches combining convolutional neural network (CNN) with support vector machines (SVMs) or attention mechanisms are also being explored. Nonetheless, these methods often struggle to generalize across diverse datasets due to differences in imaging equipment and patient demographics, making robust and real-time detection a continuing challenge [5–7].

Following preprocessing, feature extraction serves as the next vital phase by isolating and identifying significant attributes of MAs, such as size, shape, intensity, and spatial location. These features are essential for differentiating MAs from similar retinal structures, such as blood vessels or hemorrhages. Both traditional methods, such as morphological analysis and texture descriptors, and advanced deep learning techniques, such as convolutional feature embeddings, are employed to strengthen model performance. Prominent architectures, including VGG16, InceptionV3, EfficientNet, and U-Net++, have demonstrated promising results in this domain. Additionally, hybrid approaches that integrate handcrafted and deep features offer a balanced framework. Nevertheless, challenges remain, including overlapping features, inconsistencies in lesion appearance, and sensitivity to image noise. Moreover, these models often suffer from overfitting and lack interpretability, while being highly dependent on well-annotated datasets, highlighting the need for further optimization and adaptability in real-world settings [8–11].

Building upon preprocessing and feature extraction, the final and critical stage is the accurate detection of retinal MAs. This detection step plays a vital role in the early diagnosis of DR, allowing clinicians to monitor disease progression and automate screening efforts. While this step provides significant clinical advantages, it also faces persistent challenges. The

small size, subtle appearance, and similarity of MAs to other artifacts contribute to high rates of false positives or missed detections. Advanced object detection models, such as Faster R-CNN, YOLOv3, DeepDRiD, and RetinaNet, have been employed to address these issues. However, these models frequently fall short when dealing with variability in image quality, lighting conditions, and lesion diversity across patient populations. Moreover, the need for large volumes of annotated data and computational resources further limits their practical application, especially in low-resource healthcare environments. These challenges underscore the need for continued innovation in developing reliable, interpretable, and resource-efficient detection models [12–14]. Hence, a need for advanced detection models that are not only robust to image variability but also capable of accurate lesion localization in diverse clinical settings. Such models must integrate intelligent preprocessing, adaptive feature extraction, and scalable learning strategies to handle real-world challenges. Addressing these demands can significantly improve early diagnosis, enhance screening efficiency, and ultimately reduce the burden of vision loss due to DR. The main contribution of the research is

- To tackle preprocessing difficulties in early macular degeneration, the
- Neurovascular Remodeling Detection Network (NRDN) was designed to identify irregular vascular structures and capillary loss, enhancing the accuracy of vessel boundary detection even in fragmented and noisy retinal images.
- To address the challenge of feature extraction in the dense, highly pigmented foveal region, the Fourier-Otsu Hough Graph Transformer Network (FOHGTN) was developed, which precisely detects subtle circular MAs and captures their spatial relationships within complex retinal textures.
- To differentiate MAs from ischemic retinal regions without obvious biomarkers, the Local Vision Support Machine Transformer (LVSMT) combines detailed texture extraction with global context analysis, enabling accurate classification and improving early detection of MAs under silent ischemia conditions.

The rest of the paper is organized as follows: Section II describes the literature survey, Section III explains the research work and models, Section IV illustrates the performance evaluation, and comparison with existing works, and is concluded in Section V.

II. Literature Survey

Hervella et al. [15] introduced a novel deep-learning methodology leveraging unlabeled multimodal image pairs to enhance MA detection in retinography. The approach employed adversarial multimodal pretraining using Generative Adversarial Networks (GANs) to synthesize fluorescein angiography from retinography, enabling feature learning without manual annotations. This method was significant as it addressed the challenges of low contrast and tiny lesion sizes in non-invasive imaging. Furthermore, the authors modeled detection as a heatmap regression task, facilitating precise localization of multiple MAs. Despite its advantages, the model's reliance on GAN-generated angiography introduces synthetic artifacts that affect accuracy.

Soares et al. [16] introduced a multi-scale algorithm for the automatic detection of MAs in retinal fundus images. The method was structured into image preprocessing, MA candidates' detection with retinal vessel detection, and candidate labeling, all designed within a multi-scale framework. This approach was significant as it enhanced the detection of small and subtle MAs across different image resolutions. The use of four public databases demonstrated the model's generalizability and robustness. However, the method depended heavily on handcrafted features and traditional image processing, limiting adaptability to complex or unseen patterns.

Mateen et al. [17] presented a hybrid feature embedding approach for early detection of MAs in DR using pretrained CNN models VGG-19 and Inception-v3. The method extracted deep features from both networks to enhance classification performance. This approach was significant as it leveraged the strengths of two powerful architectures, improving the accuracy of MA detection on challenging datasets. The model achieved impressive classification accuracies on E-Ophtha and DIARETDB1, demonstrating its potential for clinical application. Nevertheless, the method did not explicitly address the localization of MAs, which is crucial for clinical interpretability.

Li et al. [18] offered a robust model for automatic MA detection in fluorescein fundus angiography (FFA) images, employing preprocessing techniques, such as Top-hat transformation, Gray-stretching, and Gaussian filtering, to reduce noise. They used an improved matched filter for coarse segmentation and a morphological strategy for accurate detection. This method was significant for its enhanced performance across various metrics, including accuracy, sensitivity,

and intersection-over-union, outperforming several traditional approaches. However, the models' pre-processing and segmentation steps require manual parameter tuning, which affects consistency and scalability.

Zhang et al. [19] presented the MA-YOLO model for the automatic detection of MAs in fluorescein angiography images, incorporating SwinIR for super-resolution enhancement and a customized detection layer in YOLOv8 to better capture small features. The model was significant for addressing common issues, such as missed detection and feature loss, while also improving generalization through transfer learning and optimizing the bounding box regression with Wise-IoU loss. MA-YOLO demonstrated superior performance compared to SSD, RetinaNet, YOLOv5, YOLOX, and YOLOv7. However, the model is optimized for FFA images and does not transfer effectively to standard color fundus images without significant retraining or adaptation.

Pundikal and Holı [20] introduced a six-phase model for detecting MAs in fundus retinal images, incorporating image enhancement, vessel removal using Hessian-based filtering and Otsu thresholding, segmentation validation with a grey wolf optimizer, and classification using a Modified K Nearest Neighbor. The model was significant for achieving high detection accuracy on the e-optha and DiaRetDB1 datasets, outperforming existing methods. It effectively improved MA's visibility and segmentation reliability. Nonetheless, the model's multi-stage design introduces cumulative processing errors and increases computational complexity.

Jabbar et al. [21] modeled an assistance system for DR patients to manage their health better, which employed fundus imaging to gather retinal images, and then transmit them to the processing unit, thereby effectively detecting and classifying disease severity. The model with a more straightforward and lighter approach required less CPU computing power, making the model suitable for smaller devices. The system supported real-time DR detection with the use of deep transfer learning algorithms, specifically VGGNet. This model was designed for general classification; hence, it was not suitable for multi-stage DR grading or differential diagnosis.

Qiao et al. [22] proposed an early diagnosis system for MA and NPDR using a deep CNN with semantic segmentation of fundus images. The system optimizes matching filter response (MFR), mutual information (MI), and Laplacian of Gaussian (LoG) in a 2D function space via Differential Evolution. However, limited class balancing and data augmentation result

in CNN bias toward dominant classes, reducing detection accuracy.

Bhimavarapu and Battineni [23] employed fuzzy logic techniques in digital image processing for the effective detection of DR. The MAs are identified by PSO by segmenting the digital fundus images. The particle swarm optimization clustering grouped the high-similarity data into clusters and combined the membership functions. Model testing was conducted on the publicly available dataset called DIARETDB0, and probability-based (PBPSTO) clustering algorithms enable image segmentation. Meanwhile, PSO easily gets trapped in local minima, especially in high-dimensional search spaces or with noisy medical images.

Long et al. [24] proposed a machine-learning approach for early DR detection using Directional local contrast (DLC) to identify MAs. Blood vessels were enhanced and segmented using a Hessian-based function, and then excluded from analysis. The remaining image was divided into patches, from which handcrafted features were extracted and classified as MA or non-MA using three machine learning models. While effective, the method produces false positives in areas with textures or intensities similar to MAs due to reliance on connected component analysis and handcrafted features.

AL-Dossary et al. [25] developed a deep learning framework to identify ALS using vowel phonation, and it demonstrated the capability of deep learning to detect nuanced biomedical patterns. Between ALS patients and healthy individuals, this approach achieved reliable discrimination by modeling temporal and spectral variations. The framework exemplifies effective design in biomedical AI with the strategies to handle class imbalance and prevent overfitting. The approach offered advantages such as detecting fine-grained physiological changes and providing a non-invasive diagnostic tool, while data augmentation improved its generalizability. Despite its strength, the model is dataset-dependent but it does not generalize well to different languages. Additionally, its domain-specific nature reduces clinical interpretability.

Maseleno et al. [26] introduced an explainable deep learning framework for automatic classification using skin lesion images. The study evaluated CNN, long short-term memory (LSTM), GRU, and BiLSTM models and integrated explainable AI techniques to highlight discriminative lesion regions supporting clinical decision-making. The framework achieved near-perfect classification performance, indicating strong discriminative capability of deep features

for lesion analysis. While the model offers improved interpretability compared to conventional black-box classifier but, its evaluation is restricted to a single dataset without external validation, which affects robustness and limits its generalizability to diverse imaging conditions.

Xiao et al. [27] have made an investigation about the Ensemble Transitive Bidirectional Decouples Self-Distillation (ETBiDecSD) method to address limitations in existing deep learning approaches for time-series classification (TSC), where traditional models often emphasize high-level semantic features extracted from lower layers but undervalue the contribution of lower-level information back to higher-level representations. ETDiDecSD was developed to enhance mutual learning throughout the model's representation levels by facilitating bidirectional knowledge communication between high- and low-level features. The technique integrates two components, such as average feature ensemble (AFE) that aggregates outputs from all intermediate levels, and another one is transitive bidirectional decoupled distillation (TBDD) structures that systematically transfers both target-class and non-target-class knowledge between levels. This experiment demonstrated that it achieved superior performance on diverse benchmark time-series datasets, suggesting better generalization compared to traditional self-distillation techniques. Even though it performs well, it possesses certain limitations like architectural complexity, which result in higher training overhead, and its direct applicability to domain-specific tasks requires additional adaptation, while effective for TSC.

Xiao et al. [28] examined an attentional knowledge-based state-space model to improve the classification of electrocardiogram (ECG) signals. This model integrates state-space representation learning with multi-head attention mechanisms to form a hierarchical framework that dynamically models the evolution of ECG signal states over time. The traditional models for ECG classification often rely on static feature extraction, which limits their ability to capture both local and global temporal dynamics. The major advantage of this model is better representations of both short and long-range temporal dependencies in ECG signals, and also, attention mechanisms allow selective focus on informative waveform patterns. The major limitation is model complexity, when compared to simpler CNN or RNN approaches. Additionally, this model introduces computational overhead and architectural complexity, which requires tuning and higher resources.

The analysis of recent studies shows several critical limitations that have been emerged due to some inabilities. Especially, the reliance on GAN-generated angiography in [15] introduces synthetic artifacts that affects accuracy, while [16] limits the adaptability to complex or unseen patterns. Further, the methods in [17] fails to explicitly address the localization of microaneurysms (MAs) and in [18] the required preprocessing and segmentation steps depends on manual parameter tuning, which severely affects consistency and scalability. Also, the techniques employed in [19] did not transfer effectively to standard color fundus images without significant retraining, whereas [20] introduced cumulative processing errors and also increases the computational complexity. Additionally, in [21] provided the lightweight nature that makes unsuitable for multi-stage diabetic retinopathy (DR) as well as differential diagnosis. Meanwhile, the detection accuracy in [22] had been hindered by the lack of adequate class balancing or data augmentation and further, in [23] as Particle Swarm Optimization (PSO) induces local minima problems, and the combination of connected component analysis and handcrafted feature in [24] identifies excessive false positives, especially in regions sharing similar intensity or texture characteristics with MAs. Furthermore, in [25], the model struggles in dataset dependency and does not generalize well to different languages, and its evaluation is restricted [26] to a single dataset without external validation. Domain-specific tasks require additional adaptation [27]. Xiao et al. [28] introduced additional computational overhead and requires tuning and higher resources. Hence, there is a need to develop a model for the efficient detection of retinal MAs at a very early stage.

III. Proposed Methodology

MAs, the earliest signs of DR, can be visualized through fundus imaging, but their accurate detection is challenging due to the retina's complex structure. To overcome these limitations and enhance the detection and analysis of retinal MAs, this research proposes a novel architecture, a *Neurovascular Fourier-Otsu Vision Support Transformer Network with LSTM (NFOVSTN-LSTM)*. In the proposed NFOVSTN-LSTM, the first hidden layer includes Neurovascular Re-modeling Detection Network, the second hidden layer includes Fourier-Otsu Vascular Enhanced Attention Network (FOVEAN), and the third hidden layer includes LVSMT. Initially, the macular degeneration presents significant challenges as pathological

changes like drusen, pigmentary alterations, and macular edema obscure or copycat MAs, leading to misclassification or loss of critical information. These changes stem from neurovascular remodeling processes driven by glial cell activation and low-grade inflammation, which induce subtle, non-lesion-related alterations in the capillary network. These alterations shift the vascular texture and local intensity distributions without forming overt lesions. Therefore, existing preprocessing algorithms often fail to incorporate vascular network connectivity disruption caused by neurovascular remodeling, where early-stage macular degeneration results in irregular branching and capillary dropout, creating fragmented vascular patterns that confuse traditional vessel detection methods not designed to handle such disruptions. To address this issue, a novel NRDN is incorporated in the first hidden layer of LSTM to learn and model the temporal dependencies in vascular patterns. This network integrates a Co-occurrence Gray-Level Edge Network (CGLEN), which employs the Gray-Level Co-occurrence Matrix (GLCM) to capture subtle texture variations in vascular regions impacted by neurovascular remodeling, alongside Canny Edge Detection to enhance boundary localization even in irregular, fragmented vessels. Additionally, the Frangi Contrast Enhancement Module (FCEM) is introduced to further improve vessel visibility. This module combines the Frangi Filter, which emphasizes tubular vascular patterns despite dropout, with Contrast-limited Adaptive Histogram Equalization (CLAHE) to enhance local contrast and reveal low-intensity vascular details obscured by pigmentary changes or edema. Thus, it captures temporal dynamics and accurately identifies the subtle progression patterns characteristic of early-stage macular degeneration.

Furthermore, extracting MA candidates, small, round, dark red lesions indicating localized capillary outpouchings, from the foveal region using retinal fundus imaging presents a significant challenge due to the region's dense and complex neural architecture. The fovea's specialization for high-acuity vision results in tightly packed cone photoreceptors and ganglion cells, creating a highly heterogeneous and optically intricate environment marked by intense light scattering and high pigmentation variability. These factors introduce background clutter, due to the minimal presence of retinal vasculature in the Foveal Avascular Zone (FAZ), which disrupts contextual vascular cues and complicates lesion differentiation, making it difficult to isolate and identify subtle vascular abnormalities such as MAs. As a result, early detection of MAs becomes highly unreliable,

potentially delaying timely intervention and affecting the accuracy of clinical evaluations. To address this, a novel FOHGTN is employed in the LSTM's second layer to learn fine-grained temporal and spatial dependencies across sequences of retinal patches. This integrates the FOVEAN, which combines Otsu's adaptive thresholding for segmenting the image in response to high pigmentation variability, Morphological Reconstruction (MR) for suppressing background clutter and stabilizing lesion shapes, and Fourier-based template matching for emphasizing the periodicity of MA structures despite intense texture and pigmentation variability.

In parallel, the Hough-Graph Transformer Network (HGTN) employs the probabilistic Hough transform (PHT) to strongly extract subtle circular lesion patterns and a heterogeneous graph transformer (HGT) to model spatial-anatomical relations in the avascular zone. Thus, overcoming the fovea's optical complexity and enhancing the sensitivity, precision, and early reliability of MA candidate detection. Moreover, in retinal MA detection, a significant challenge arises from the phenomenon of silent ischemia, where MAs are present, but perfusion in the affected retinal areas is already compromised due to endothelial dysfunction and pericyte loss, which is primarily caused by increased oxidative stress. This stress often manifests as hyperreflective areas in retinal fundus imaging, resulting in endothelial cell destruction and reduced blood flow. These non-perfused regions typically lack distinct biomarkers or visible signs of damage, making early detection of ischemic changes difficult. As a result, the absence of clear clinical indicators in these regions delays the identification of MAs, complicating early diagnosis and intervention and increasing the risk of irreversible retinal damage and vision loss in DR.

To overcome this limitation, the model incorporates an LVSMT within the third hidden layer of the LSTM to enhance the detection of slow-developing ischemic changes. LVSMT includes the Local Binary Transformer Network (LBT-Net), using local binary patterns (LBP) for extracting critical texture features in identifying subtle ischemic changes, and the Vision Transformer (ViT), for focusing on ischemic regions without visible biomarkers using self-attention mechanisms to learn long-range dependencies in the retinal image. The features extracted by the LBP and ViT are classified by the SVM classifier as either MAs or ischemic regions accurately at the third hidden layer of LSTM. This comprehensive approach enhances early diagnosis by concentrating on local texture patterns and global vascular context, overcoming the challenges posed by silent ischemia and providing

more accurate and timely intervention in retinal damage and vision loss.

Figure 1 illustrates the block diagram of the proposed NFOVSTN with LSTM for MA detection.

The input fundus image is passed to an LSTM network, which captures sequential dependencies in retinal data. In the first hidden layer, preprocessing is

done using GLCM for texture, Canny edge detection for boundaries, Frangi filter for vessel enhancement, and CLAHE for contrast boosting. This layer helps identify early macular degeneration and vascular remodeling. The second hidden layer employs FOHGTN, where Otsu thresholding, MR, and Fourier-based template matching extract lesion-related

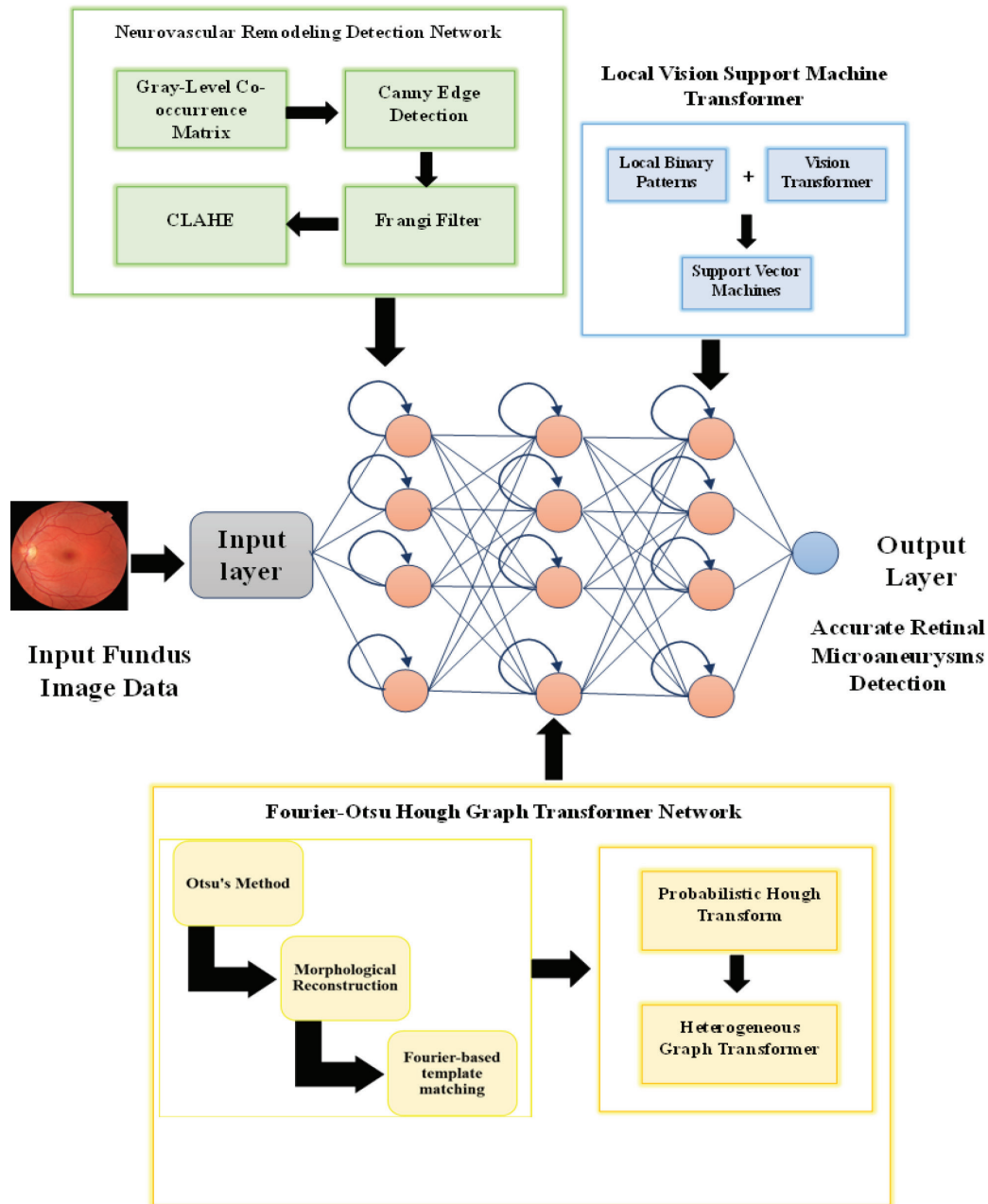


Figure 1: Block diagram of proposed NFOVSTN-LSTM. NFOVSTN-LSTM, Neurovascular Fourier-Otsu Vision Support Transformer Network with Long Short-Term Memory.

features. PHT detects circular MA patterns, and a HGT enhances contextual awareness. In the third hidden layer, LVSMT handles classification tasks between MA and ischemic zones. LBT-Net extracts ischemic texture features, while ViT leverages self-attention to focus on key retinal areas. SVM then classifies lesions accurately into MAs or ischemic regions. The final output layer produces the detected MAs with high precision. Although the proposed architecture is a multi-stage architecture, the NFOVSTN-LSTM is designed under a unified theoretical principle of progressive representation refinement guided by neurovascular physiology and scale-specific inductive bias. Each module corresponds to a distinct level of retinal information abstraction, vascular remodeling enhancement, NRDN, spatial anatomical relational modeling, FOHGTN, and fine-grained lesion discriminative LVSMT. Rather than jointly optimizing all representations in a single step, the architecture enforces a coarse-to-fine learning hierarchy that stabilizes training by isolating weak lesion cues before high-level contextual reasoning. This staged design ensures that clinically meaningful structures are preserved throughout learning. Empirical ablation results further demonstrate that each stage contributes complementary and non-redundant information, confirming that the pipeline is rather than heuristic.

a. Hierarchical architecture of NFOVSTN-LSTM

The proposed NFOVSTN-LSTM framework is designed as a hierarchical learning architecture that progressively transforms retinal image representations across multiple preprocessing stages. Rather than learning all features simultaneously, the framework organizes the pipeline into three sequential stages, such as neurovascular structural enhancement, MA candidate extraction, and lesion discrimination. In the first stage, the NRDN enhances vascular structures and captures subtle texture variations associated with neurovascular remodeling, including vessel continuity, intensity gradients, and capillary irregularities. Then, the resulting vascular feature maps are then forwarded to the second stage.

In the second stage, to isolate lesion-related features from complex retinal backgrounds, the FOHGTN is proposed, which mainly focuses on isolating potential MA candidates from complex retinal backgrounds using frequency-domain analysis, adaptive segmentation, and graph-based relational modeling.

In the final stage incorporates the LVSMT and performs high-level discriminative learning to differentiate

MAs from ischemic regions. This stage integrates transformer-based contextual learning with texture-aware descriptors and as SVM classifier to produce robust final predictions.

Although the architecture contains multiple processing modules, all the parameters are jointly optimized through end-to-end training using gradient-based optimization. Using backpropagation, gradients propagate sequentially through the LVSMT, FOHGTN, and NRDN stages. The gating mechanisms within the LSTM layer regulate information flow and stabilize gradient propagation across the hierarchical layers, thereby mitigating vanishing or exploding gradient effects during multi- θ_{t+1}

$$= \theta_t - \eta \nabla_{\theta} F(\theta). \quad (1)$$

Here, θ represents the model parameters, η is the learning rate, and $F(\theta)$ denotes the loss function. To further clarify the training procedure, the overall hierarchical optimization strategy of the NFOVSTN-LSTM framework is summarized in Algorithm 1.

b. Retinal image preprocessing and vessel structure enhancement using neurovascular re-modeling detection network

While LSTMs are traditionally utilized for temporal data, the NFOVSTN-LSTM framework employs the recurrent architecture to model spatial-sequential dependencies inherent in the retinal topography. Instead of treating the retinal image as a temporal signal, the spatial feature maps generated by the NRDN are modernized into anatomically guided spatial sequences based on retinal topology and vessel continuity. This spatial ordering is guided by structural relationships among retinal vessels and lesion candidates, ensuring that the sequence reflects meaningful anatomical connectivity rather than random spatial reshaping. The LSTM therefore models long-range structural dependencies across spatially distant vascular regions, which is critical for classifying subtle pathological variations associated with MAs. Through gated memory propagation, the LSTM accumulates contextual evidence from multiple retinal regions, allowing the model to differentiate subtle MA signatures from surrounding vascular textures by examining structural continuity patterns across the neurovascular network. Consequently, the LSTM operates as a context-aware spatial modeling component rather than a temporal processor. Avoiding

artificial temporal dependency assumptions while enabling progressive contextual aggregation of vascular features within a static fundus image. This spatial-sequence formulation aligns with representation learning strategies used in structured biomedical imagery. In anatomical connectivity and spatial relationships, direct feature dependencies rather than temporal evolution. To tackle the challenges arising from early-stage macular degeneration, a novel NRDN is implemented at the first hidden layer of the LSTM architecture, integrating CGLEN and FCEM. The NRDN is a fully learnable, parameterized deep module integrated into the proposed framework. It comprises convolutional layers whose parameters are jointly optimized with subsequent stages using task-driven supervision and backpropagation from the final loss function. Unlike conventional preprocessing techniques, NRDN does not apply fixed enhancement or filtering operations but instead learns discriminative neurovascular representations relevant to MA detection. The effectiveness of NRDN is quantitatively validated through ablation analysis, demonstrating its contribution beyond standard preprocessing. This hybrid preprocessing enhances vascular contrast and texture, enables the detection of subtle, non-lesion-related changes in the vascular structures, thereby improving the visibility of subtle, early-stage pathological features. To progressively refine vascular representations from the retinal image, the NRDN module follows a hierarchical feature abstraction strategy. In the initial stage, low-level features such as vessel edges, local intensity gradients, and micro-contrast variations are captured using edge-based feature extraction and texture-aware operators. These representations are then transformed into intermediate vascular feature maps that encode vessel continuity, branching patterns, and localized capillary irregularities associated with neurovascular remodeling. In the final stage, contextual refinement layers integrate spatial neighborhood information to isolate lesion-specific characteristics from surrounding vascular structures. This hierarchical separation of background retinal structures, vascular patterns, and lesion-related features explicitly defines the feature abstraction boundaries within NRDN and reduces contextual modeling stages. Formally, the vascular feature transformation performed by the NRDN module is expressed as given in Eq. (2) [29].

$$V_I = \Phi_{en}(R) \quad (2)$$

where I represent the input retinal fundus image and Φ_{en} denotes the composite vascular en denotes

enhancement operator consisting of edge enhancement, texture extraction, and vessel structure filtering. The resulting feature representation V_I emphasizes vascular structural information while suppressing background retinal variations, thereby ensuring that subsequent contextual modelling modules operate on well-defined vascular abstractions. Similar hierarchical vascular feature abstraction strategies have been widely adopted in retinal vessel analysis to isolate vessel-specific structural information prior to higher-level contextual learning.

a.i. CGLEN grayscale segmentation and edge detection

To capture subtle retinal texture variations in vascular regions and boundary localizations, CGLEN is proposed. CGLEN employs the GLCM to extract texture descriptors that characterize the non-lesion-based changes since the neurovascular remodeling doesn't always create visible lesions but subtly shifts the texture. GLCM detects these shifts by revealing changes in pixel pair distributions that wouldn't be evident in raw images. Initially, the RGB fundus images are converted into grayscale while preserving critical intensity features. The grayscale images are represented as 2D matrices, where each pixel value denotes brightness within a specified range. To quantify retinal texture, GLCM is computed, which captures the frequency of pixel intensity transitions across spatial relationships. The GLCM [23] feature includes homogeneity in Eq. (3) to measure local uniformity with high homogeneity suggesting smooth, healthy retinal tissue, and low homogeneity in vascular areas indicates disordered vessel growth due to neurovascular remodeling.

$$\text{Homogeneity} = \sum_x \sum_y \frac{1}{1 + (x - y)^2} P_{xy} \quad (3)$$

where P_{xy} is the normalized co-occurrence probability of gray level x occurring with the gray level y , and GLCM measures variation in gray levels where the elevated dissimilarity suggests greater grayscale transitions typical in pathological regions such as MAs or hemorrhages using Eq. (4).

$$\text{Dissimilarity} = \sum \sum |x - y| P_{xy} \quad (4)$$

In addition, the GLCM feature contrast in Eq. (5) is sensitive to edge sharpness and grayscale intensity differences.

$$\text{Contrast} = \sum_{x=0}^{\infty} \sum_{y=0}^{\infty} P(x-y)^2 \quad (5)$$

GLCM integrates broader contextual patterns over a region, making it more robust in distinguishing pathological textural shifts from imaging noise or pigment variations. Thus, GLCM captured subtle texture variations in vascular regions induced by neurovascular remodeling, which traditional intensity-based methods often overlook. In the presented framework, to complement the contextual representations learned by the deep LSTM network, GLCM-based texture descriptors are incorporated. These co-occurrence statistics quantify spatial relationships between neighboring pixel intensities, capturing subtle, localized texture variations that are characteristic of MAs and other vascular abnormalities. GLCM features provide a quantitative link between texture representation and structural Vision information, by explicitly modeling the distribution and correlation of pixel intensities around lesion boundaries. When GLCM-based features are combined with deep contextual features, these textual descriptors strengthen lesion boundary delineation and improve the discrimination between MAs and surrounding retinal tissue. Consequently, this complementary integration enhances the interpretability of the learned representations and ensures that subtle texture variations associated with neurovascular remodeling are preserved, thereby supporting more accurate detection of MAs. Following texture quantification using GLCM descriptors, boundary refinement is performed using the Canny Edge Detection algorithm to ensure precise localization of vascular structures highlighted by the extracted texture patterns.

Due to intensity variations in lesion edges, initial Gaussian Blurring is applied to suppress high-frequency noise by averaging local pixel intensities, preventing false edge responses. To attain pixel-wise intensity change, the images are convolved using the Canny operator to calculate the gradient of each pixel in the l direction, G_l and in the m direction, G_m . The overall gradient magnitude G and edge orientation $\varnothing$ are derived in Eqs (6) and (7) [30].

$$G = \sqrt{G_l^2 + G_m^2} \quad (6)$$

$$\varnothing = \arctan \frac{G_m}{G_l} \quad (7)$$

Subsequently, non-maximum suppression is used to thin edges to a single pixel by retaining only local

maxima in the gradient direction, suppressing weak, irrelevant edge points. This is crucial for localizing disrupted or fragmented vessel boundaries caused by neurovascular remodeling. Finally, dual thresholding is performed so that strong edges exceeding a high threshold are retained, weak edges below a low threshold are discarded, and intermediate edges are conditionally preserved if connected to strong edges. This hysteresis step ensures vessel continuity is preserved even when intensity is reduced due to capillary dropout or foveal signal variability. Through this multi-step enhancement, Canny edge detection enhances boundary localization even in irregular, fragmented vessels.

a.ii. FCEM structural and contrast enhancement in preprocessing

The FCEM is then incorporated to enhance visibility by fusing the Frangi Filter, which plays a crucial role in enhancing tubular vascular structures, particularly in cases of capillary dropout, irregular branching, or pigmentary occlusion, obscuring vessel boundaries. The Frangi filter analyzes the retinal fundus image using the Hessian matrix at each pixel as in Eq. (8), by capturing second-order intensity variations to describe the curvature of intensity changes in both spatial directions.

$$H = \begin{bmatrix} \frac{\partial^2 I}{\partial p^2} & \frac{\partial^2 I}{\partial p \partial q} \\ \frac{\partial^2 I}{\partial q \partial p} & \frac{\partial^2 I}{\partial q^2} \end{bmatrix} \quad (8)$$

where p, q are the spatial positions, $I(p, q)$ represents the image intensity. Then, the eigenvalues λ of the Hessian matrix are calculated, characterizing the local shape and structure of the image.

- $\lambda \gg 0$ indicates bright tubular structures, typically the central axis of vessels
- $\lambda \geq 0$ corresponds to vessel edges or weak tubular structures
- $\lambda < 0$ suggests non-vascular structures or background regions.

By evaluating bright, elongated features while suppressing noise and blob-like structures, the Frangi filter generates a Vesselness Enhancement Image (VEI) [31]. This VEI selectively enhances retinal vessels by emphasizing tubular patterns despite capillary

dropout and also preserves anatomical accuracy, making it especially effective in cases of fragmented vasculature due to neurovascular remodeling. Although the Frangi vesselness filter is highly effective for emphasizing tubular vascular structures, it exhibits sensitivity to curvature variations and local noise fluctuations in highly heterogeneous retinal regions. To address these challenges, the proposed framework incorporates Frangi filtering within the FCEM module as part of a broader enhancement strategy rather than as an isolated operation. Specifically, the preceding texture-guided preprocessing stage stabilizes intensity transitions and suppresses high-frequency noise components before vesselness computation. Additionally, the multiscale Hessian analysis enables the detection of vessels with different diameters and curvature characteristics, which is particularly important in regions affected by neurovascular remodeling. The enhanced vesselness response is subsequently refined using contrast normalization and structural context provided by the surrounding NRDN layers to further preserve vascular continuity. The integrated strategy ensures that the vesselness optimization stage emphasizes anatomically consistent vascular patterns while reducing noise amplification and minimizing fragmentation of delicate capillary structure. The preprocessing step-by-step procedure is illustrated in Figure 2.

The VEI output is fused with CLAHE to improve local contrast while preventing over-enhancement of noise. Here, the input image VEI is divided into small, non-overlapping contextual regions or tiles, and then histogram equalization is applied individually within each tile. Technically, CLAHE limits contrast amplification by setting a clip limit threshold, especially in areas with low variance (e.g., uniform regions), which prevents noise amplification. If any gray-level bin exceeds this threshold, the excess pixels are redistributed evenly across all gray levels within the local tile. This mechanism boosts local contrast to reveal low-intensity vascular details obscured by pigmentary changes or edema and prevents noise over-amplification. To address block artifacts commonly seen in adaptive histogram equalization, interpolation techniques are applied between neighboring tiles, smoothing grayscale transitions, and preserving fine vascular details. Careful tuning of these parameters ensures optimal enhancement of vascular features critical for detecting MAs while maintaining image quality. CLAHE, in addition to boosting the visibility of low-intensity vessels, boosts local contrast to reveal low-intensity vascular details obscured by pigmentary changes or edema. This combination enables the detection of irregular branching, capillary dropout, and fragmented vascular patterns, which are often overlooked by traditional vessel detection

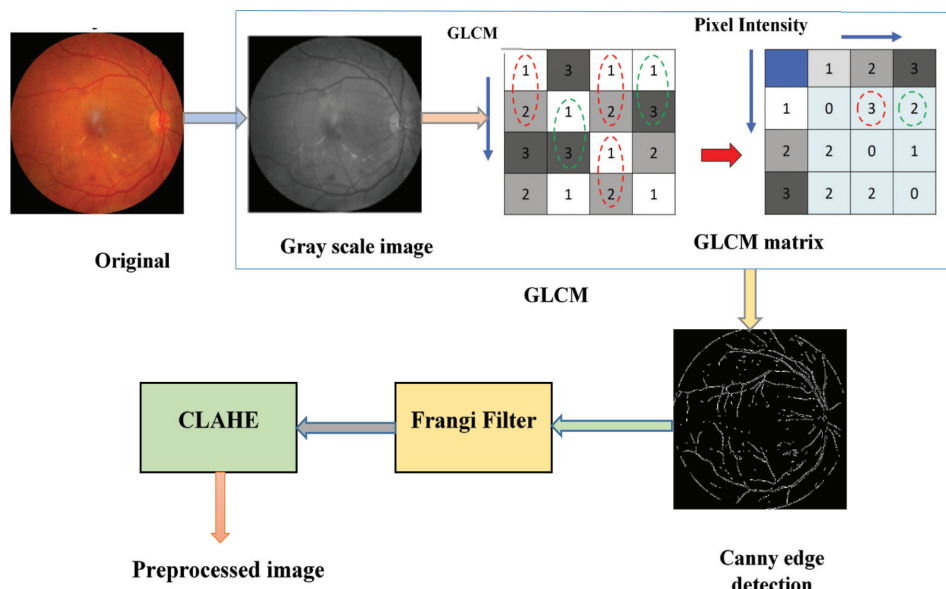


Figure 2: Preprocessing of the retinal fundus image. CLAHE, contrast-limited adaptive histogram equalization; GLCM, Gray-Level Co-occurrence Matrix.

methods. In this proposed framework, remodeling-related intensity variations are distinguished from true MAs based on their structural and spatial characteristics rather than raw intensity alone. Neurovascular remodeling typically tends to low frequency, smoothly varying intensity patterns aligned with vessel structures, while MAs exhibit high frequency and localized texture responses. By explicitly enhancing vascular structures and modeling remodeling characteristics in the initial stage, the framework suppresses non-lesion intensity fluctuations while preserving compact lesion-specific responses. This separation ensures that the final SVM classifier operates on features that are representative of true pathological MAs rather than background vascular intensity changes.

b. Extracting MA features

To address the challenges in extracting MA candidates from the highly complex and optically variable foveal region in retinal fundus imaging, a novel FOHGTN approach is proposed and integrated into the LSTM second hidden layer by combining FOVEAN and HGTN.

b.i. FOVEAN in feature extraction

To effectively mitigate the background clutter issues and to enhance the periodic lesion patterns of MAs, FOVEAN is proposed, thus solving the issues induced by structural complexity. FOVEAN incorporates Otsu's method for image segmentation in response to high pigmentation variability. Although classical Otsu thresholding is theoretically formulated for bimodal histogram distributions, its effectiveness in the presented framework is reinforced through prior contrast normalization and structural enhancement stages. Due to heterogeneous illumination conditions, pigmentation variability, and complex anatomical structures such as vessels and macular textures, retinal images frequently exhibit multimodal grayscale distributions. To address this limitation, the proposed framework incorporates contrast normalization and structural filtering before applying Otsu segmentation. Specifically, CLAHE locally redistributes grayscale intensities and reduces illumination imbalance across retinal regions. This normalization step suppresses extreme intensity variations and improves the separability of lesion-related intensity responses from the surrounding background. In addition to that, the subsequent MR stages remove spurious background structures and stabilize the geometric representation

of potential lesion candidates before segmentation. As a result of this, the intensity distribution presented to the Otsu operator becomes more discriminative between MAs candidates and non-lesion retinal structures. Consequently, the segmentation process becomes more robust in isolating MA candidates even under heterogeneous illumination and multi-modal grayscale conditions.

The Otsu algorithm is employed to perform adaptive thresholding for robust segmentation of potential MA candidates in high pigmentation variability. The grayscale intensity distributions of retinal imaging vary widely due to pigmentary changes, edema, or background clutter, especially in the macular and foveal regions. The enhanced fundus image, post-CLAHE, has dimensions $L \times M$ and is analyzed to partition pixels into two distinct classes: background with M_b pixels and target with M_t pixels. The ratio of the background pixels P_b and the ratio of target pixels P_t in the entire image is measured using Eqs (9) and (10) [32].

$$P_b = \frac{M_b}{L.M} \quad (9)$$

$$P_t = \frac{M_t}{L.M} \quad (10)$$

$$\text{Variance of the entire image, } \sigma^2 = (P_b)(P_t)(\mu_b - \mu_t)^2 \quad (11)$$

where μ_b and μ_t are the grayscale values of the background and target, respectively. Otsu's method dynamically selects a threshold to maximize this inter-class variance in Eq. (9) between background and target, whereas the intra-class variance remains minimal. By optimizing thresholding, Otsu's method ensures the effective segmentation of subtle lesion structures from retinal background variations in response to high pigmentation variability.

To suppress background clutter and stabilize lesion shapes, FOVEAN employs MR. This operation suppresses background clutter by selectively preserving the shapes and boundaries of relevant vascular lesions while eliminating spurious, non-lesion-related variations caused by texture noise, pigmentation shifts, or anatomical overlap. It operates using three key inputs: the marker image from Otsu-based segmentation, the mask image from the CLAHE-enhanced grayscale fundus image, and a structuring element that defines pixel connectivity. The process begins with opening-by-reconstruction, where the marker is eroded to eliminate small-scale

noise, and then reconstructed under the mask to restore only valid anatomical structures. This ensures the retention of lesion and vessel features while suppressing non-informative artifacts. To address both fine and coarse noise patterns, the MR filtering operates in a looped structure with a progressively increasing kernel size [33]. The filtering continues until the branching structure of the retinal vessels remains intact and further noise removal does not eliminate relevant features. Each cycle removes increasingly larger noise clusters and includes complementary closing-by-reconstruction to bridge fragmented vessels and reinforce vascular continuity. As a result, MR suppresses background clutter and stabilizes lesion shapes.

Fourier-based template matching emphasizes the periodicity of MA structures by transforming the spatial domain image into the frequency domain, where repeating or circular patterns like MAs are more distinguishable from irregular background textures. MAs typically appear as small, round, mid-frequency spectral responses in the Fourier domain, and their periodic shape leads to identifiable frequency components. In contrast, background noise due to pigmentation variability or retinal texture appears as high-frequency components or non-uniform patterns. The Fourier Transform is defined by an integral Eq. (12) that transforms a function $x(t)$ into its frequency domain representation.

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt \quad (12)$$

where $x(t)$ is the function of time, $X(\omega)$ is the Fourier transform (function of frequency), ω is the angular frequency, j is the imaginary unit [34]. By applying a Fourier Transform to the morphologically reconstructed image, frequency-domain templates (representing idealized MA shapes) are used to identify areas with similar spectral characteristics. This process suppresses the impact of non-periodic background variations and enhances the detection of consistent, circular lesion patterns. An inverse FT of this filtered spectrum in Eq. (13) produces an MA-enhanced frequency map, which highlights potential lesion regions for downstream segmentation.

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{j\omega t} d\omega \quad (13)$$

Through inverse transformation, the matched locations are mapped back to the spatial domain,

emphasizing the periodicity of MA structures despite intense texture and pigmentation variability. The entire process in the feature extraction stage is represented in Figure 3. From the perspective of signal-processing, the effectiveness of Fourier domain analysis for MA detection can be attributed to the spectral distinction between localized vascular lesions and slowly varying retinal background structures. In the Fourier spectrum, MAs typically appear as compact circular intensity depressions, and their localized spatial characteristics generate relatively stable mid-frequency components. Conversely, illumination changes, pigmentation variations, and retinal texture heterogeneity generally manifest as slowly low-frequency components. As a result of this, while suppressing background variations, performing template matching in the frequency domain helps emphasize lesion-related spectral responses. Consequently, this property improves the robustness of candidate MA localization even in fundus images exhibiting heterogeneous pigmentation and complex retinal textures [35].

b.ii. HGTM in feature extraction

To further refine MA candidate extraction from the structurally complex and optically heterogeneous foveal region, the PHT is employed to enhance the detection of subtle circular patterns that characterize MAs. PHT selectively samples edge points to localize partial circular structures, enabling robust detection of faint or incomplete MA-like patterns embedded within noisy and low-contrast backgrounds, particularly within the FAZ, where vessel-based context is minimal. PHT detects arcs and elliptical contours with variable radii that align with MA morphology, even when boundaries are disrupted by pigmentation or retinal curvature. Detected circular regions are then geometrically normalized using an affine transformation computed from arc edge control points as in Eq. (14) [36].

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & b_0 \\ a_{21} & a_{22} & b_1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (14)$$

Where x, y are the original coordinates, x', y' are the transformed coordinates and a_{11}, a_{12}, a_{21} and a_{22} are the linear transformation parameters (rotation, scaling, shearing). b_0 and b_1 are the translation offsets in x, y directions. This transformation compensates for spatial distortion due to retinal curvature or illumination artifacts, aligning MA candidates into a

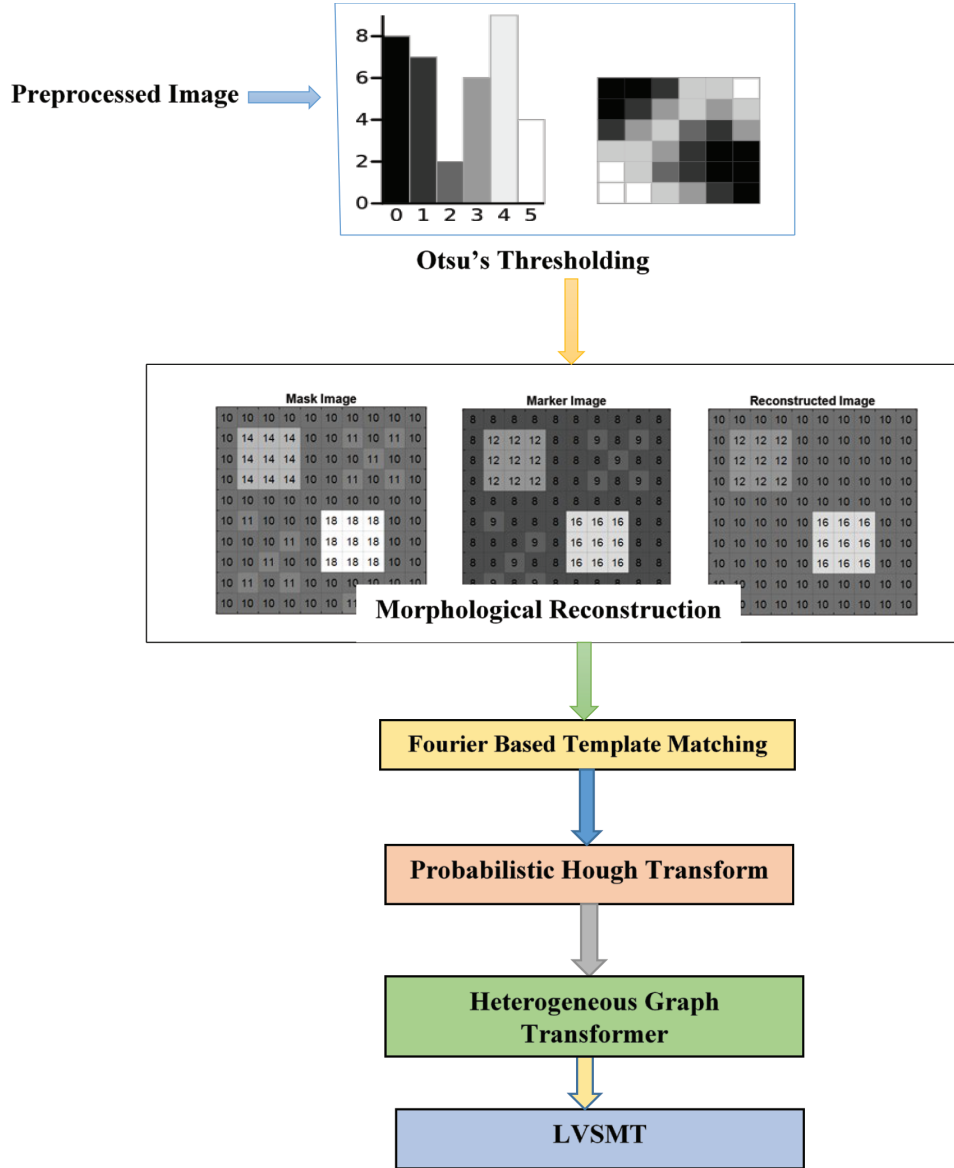


Figure 3: FOHGTN feature extraction. FOHGTN, Fourier-Otsu Hough Graph Transformer Network; LVSMT, Local Vision Support Machine Transformer.

planar and uniform coordinate space while preserving local vascular geometry. Thus, the PHT enables strong extraction of subtle circular lesion patterns even with weak vascular cues.

To effectively model interrelations among retinal anatomical structures and pathological features, including MAs, the FAZ, macular texture, and vascular patterns, a HGT) is incorporated. In this framework, a heterogeneous graph is constructed where each node encodes a distinct feature type, such as vessel segments, lesion candidates, texture descriptors,

and regional zones, while edges represent diverse spatial and functional relationships, such as anatomical adjacency, symmetry, co-occurrence, or contrast transitions. Each node-pair connection is defined through a meta-relation triplet that captures the node types and the specific nature of their relationship, enabling semantically rich and structurally grounded graph modeling. To model these relationships, the HGT performs type-aware attention-based message passing, where, unlike standard transformers, it assigns unique attention weights and transformation

matrices per node and edge type. This enables the model to contextualize each node's features not only based on its properties but also on the heterogeneous roles and spatial configurations of its neighbors, particularly useful in the FAZ, where vascular cues are sparse but contextual geometry is critical. Through multi-head relation-specific attention, message propagation, and aggregation, the HGT dynamically learns spatial and semantic dependencies that reflect the true anatomical structure of the retina [37]. Finally, the spatially aligned embeddings generated by the HGT are aggregated into a comprehensive global representation, thereby modelling the spatial and anatomical relationships to enrich contextual understanding within the avascular zone.

To further strengthen anatomical relational reasoning, the FOHGTN module encodes structural constraints by representing retinal structures as a $S = NR$ where N represents nodes corresponding to retinal regions and R represents edges encoding spatial relationships between neighboring anatomical structures. Here each node $n_i \in N$ corresponds to a local retinal region containing vascular or lesion-related features, and edges $r_{xy} \in R$ represent anatomical connectivity derived from vessel continuity and spatial proximity. This formulation enables the network to model interactions between neighboring vascular regions and lesion candidates while preserving anatomical topology. Graph feature propagation is performed using a structure aggregation mechanism expressed in Eq. (15) [38].

$$F^{L+1} = \sigma(A F^L T^L) \quad (15)$$

where A denotes the adjacency matrix encoding retinal structural relationships, F^L represents node feature embeddings at layer L , T^L is the learnable transformation matrix, and $\sigma(\cdot)$ denotes a non-linear activation function. Though this graph-based relational modelling FOHGTN captures spatial dependencies among vascular segments and surrounding retinal regions, thereby improving interpretability and structural consistency of learned feature representations during MA detection. In a similar way, graph-based reasoning strategies have been successfully applied in medical image analysis to model structural relationships among anatomical regions. To ensure a consistent and reproducible graph representation across the experimental framework, the heterogeneous graph is formally defined as $M = (P, T, B, J)$, where V represents the node set, T the edge set, B the

node types, and J the relation types. In this graph formulation, the nodes include MA candidates (P_{MA}) detected through PHT, the foveal avascular Zone (P_{FAZ}), the macular texture descriptions (P_{Tex}), and vascular segments (P_{Ves}). Using the affine transformation in Eq. (12) each node extracts a feature extracted from regions normalized. This normalization maps all detected anatomical and pathological structures into a planar coordinate space, ensuring the spatial relationships across retinal curvature and imaging conditions. Edges are constructed using fixed meta-relations J that encode anatomical adjacency and functional co-occurrence. Especially, an edge exists if the Euclidean distance between normalized coordinates (S', V') is below a predefined threshold. These node definitions and edge construction rules are uniformly applied across all images, guaranteeing consistent graph topology throughout the dataset. The HGT performs type-aware message passing using relation-specific attention with distinct transformation matrices L_T for each node type $G \in C$. This allows effective aggregating of spatial and anatomical context, particularly in regions with sparse vascular cues such as the FAZ. Likewise, these node definitions, relation categories, and edge construction rules are deterministically defined and applied uniformly across all images, ensuring a consistent and reproducible graph structure throughout the experimental framework.

c. MA detection using LVSMST

To enhance MA detection by focusing on fine-grained spatial patterns and region-specific attention, an LVSMST is introduced at the third hidden layer of the LSTM, enabling reliable differentiation of MAs from ischemic regions lacking clear biomarkers in the retinal fundus.

c.i. LBT-Net in MA detection

To capture binary texture patterns and local spatial transitions crucial for the identification of MAs and subtle retinal abnormalities, LVSMST integrates an LBT-Net. The LBT-Net includes LBP principles into a transformer architecture wherein this module encodes local texture changes by comparing each pixel to its neighborhood in a binary manner. These binary maps are then tokenized into patches and embedded as input sequences to the transformer. Then, the embedded binary patterns are passed

through a binary-aware transformer that applies multi-head self-attention (MHSA) over the encoded texture descriptors. LBT-Net highlights binary transitions, such as edge-like structures and localized dissimilarities, making it especially sensitive to micro-lesions and noise-robust features. At the output of LBT-Net, two heads are introduced as a classification head for determining the presence of retinal pathologies (e.g., MAs and hemorrhages) and a detection head that predicts lesion locations through bounding box regression or pixel-wise segmentation, depending on the dataset annotations [37]. This dual-head setup allows the model to not only recognize diseased instances but also localize clinically relevant regions, even in challenging zones such as the macula or overlapping vascular networks. This LBT-Net module, by incorporating binary spatial encoding and transformer-based pattern discrimination, extracts critical texture features to identify subtle ischemic changes, such as hyperreflective areas and vessel irregularities.

To improve the recognition and differentiation of critical lesion types, specifically MAs and ischemic regions, ViT is integrated in the feature classification stage of the retinal MA detection framework. ViT utilizes the core principles of the transformer architecture, originally designed for natural language processing, and adapts them for high-resolution image analysis. Initially, the input, which is a feature map derived from prior modules, is divided into a series of non-overlapping patches. Each patch is embedded with positional encoding to retain the spatial relationships between anatomical structures,

such as the macula, FAZ, and retinal vessels. These position-aware patches, together with a special classification token, are then passed through the MHSA layers of the transformer [39] and are shown in Figure 4. This allows ViT to detect complex spatial patterns and subtle visual cues associated with MAs, such as small red dots near capillary beds, as well as ischemic regions.

Through the attention mechanism, ViT distinguishes between these lesion types by learning to assign higher attention weights to regions that exhibit characteristics of either r , such as localized vascular bulges, or ischemia marked by texture flattening and vascular rarefaction. The resulting classification token embedding represents a high-level summary of the image context, encoding both morphological and spatial features relevant to pathology. This learned representation is refined by an MLP for dimensionality reduction and further abstraction.

c.ii. SVM in detecting MA

To effectively separate MAs from ischemic regions, an SVM classifier is employed based on the high-dimensional features output from the ViT and LBP for the final decision-making step. SVM is a powerful supervised learning algorithm known for its high performance in binary classification tasks, particularly in high-dimensional feature spaces. It works by identifying the optimal hyperplane that separates data points of different classes (in this case, MAs and ischemic features) with the maximum margin, thus ensuring better generalization and minimizing

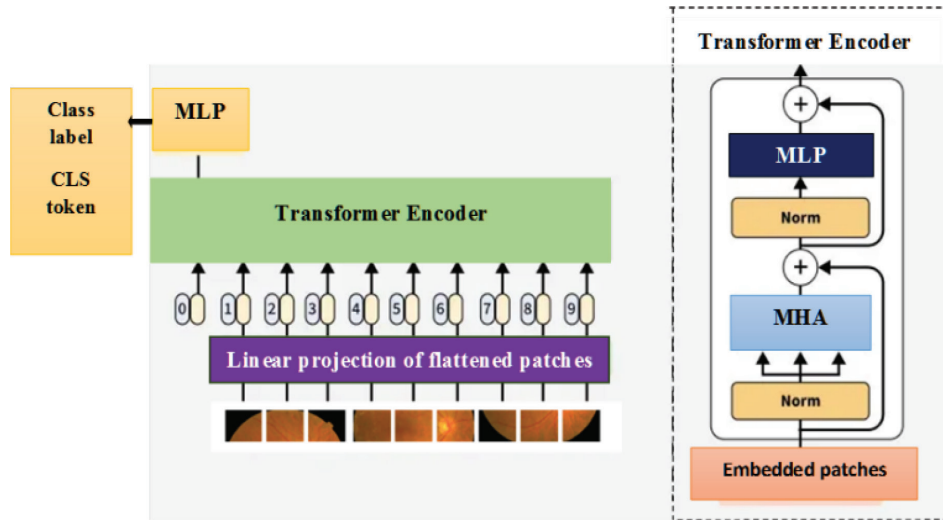


Figure 4: ViT in MA detection. MA, microaneurysm; ViT, vision transformer.

classification errors. In the proposed architecture, the SVM is utilized as a local discriminative refinement layer within the LVSMT module rather than as an independent classifier. As shown in Figure 1, the integration of traditional LBP alongside the deep features extracted by the ViT is a deliberate architectural choice designed to robustly capture both local textural information and global contextual dependencies within the input fundus images. In the presented framework, the high-level contextual embeddings are extracted by the ViT are fused with LBP descriptors that encode fine-grained micro-textural characteristics of MAs. LBP is a powerful descriptor, and it is especially designed to encode the microstructure and texture patterns of an image with invariance to grayscale variations, making it highly suitable for analyzing subtle pathological changes in retinal tissue. While ViTs excel at modeling long-range interactions across image patches through their self-attention mechanisms, ViT learns discriminative contextual representations, when the LBPs capture subtle intensity and texture variations associated with MAs. By fusing these complementary feature sets, our model mitigates the ViT's inherent bias toward global features and enhances the representation with explicit, noise-resistant texture data, leading to a more comprehensive and discriminative feature space for the final classification of accurate retinal MAs. The SVM uses the transformed feature representations from ViT to construct a decision boundary that maximizes inter-class separation. Its ability to handle non-linear data through kernel functions (e.g., radial basis function and polynomial kernels) makes it particularly effective when the distinction between lesions is subtle or overlapping in feature space. By combining the ViT's contextual embedding capabilities with the SVM's discriminative classification strength, the proposed architecture achieves high sensitivity and specificity in identifying and differentiating between early MAs and ischemic changes, even in complex or ambiguous retinal regions. Overall, the proposed Neurovascular Remodeling-Aware Transformer-LSTM Network isolates and highlights critical retinal structures, captures complex anatomical and vascular dependencies, and contextualizes spatial features to distinguish MAs from ischemic regions. This comprehensive pipeline enhances early detection capabilities, particularly in cases of silent ischemia, and provides a precise, automated foundation for clinical decision support in DR screening. Although the NFOVSTN-LSTM architecture is multi-layered, each component serves a specific purpose in addressing the challenges of retinal MA detection. The NRDN

in the first hidden layer captures subtle vascular changes and enhances vessel visibility in early-stage macular degeneration. Then the second hidden layer, FOHGTN, which combines with adaptive thresholding, MR, Fourier-based template matching, and HGT modeling to suppress background clutter and capture the spatial-anatomical relationships of MAs in the foveal region. The third layer is LVSMT, which uses LBT-Net and ViT to differentiate MAs from ischemic regions with SVM for classifying the features accurately. Before SVM classification, to ensure scale consistency across heterogeneous descriptors, the fused ViT and LBP feature vectors are added and normalized using Z-score normalization. To reduce redundancy, overfitting, computational complexity, and control feature dimensionality, a linear projection layer is applied before classification. The final SVM classifier uses an RBF kernel; this is selected for its ability to model non-linear class boundaries common in subtle lesion-ischemia discrimination. This design smoothly ensures that the SVM operates on compact, normalized, and discriminative derived from transformer-based contextual embeddings.

Algorithm 1: Training Procedure of NFOVSTN-LSTM

Input: Retinal fundus image dataset D

Output: Trained NFOVSTN-LSTM model

1. Initialize the parameters of the NRDN, FOHGTN, LVSMT, and LSTM modules.
 2. For each retinal fundus image, $I \in D$, performing the following steps:
 3. Apply the NRDN to enhance vascular structures and generate vascular feature representation V_i .
 4. Pass V_i to the FOHGTN to extract potential microaneurysm candidates.
 5. Construct spatial-anatomical relationships using the HGT to obtain refined lesion candidate features.
 6. Feed the refined features to the LVSMT to learn contextual and texture representations.
 7. Fuse the contextual embeddings from ViT with texture descriptors derived from LBP.
 8. Normalize the fused feature representation using Z-score normalization.
 9. Classify the features using an SVM classifier with RBF kernel to predict microaneurysm or ischemic regions.
 10. Compute the loss between predicted outputs and ground-truth labels.
-

11. Perform backpropagation to propagate gradients through LVSTM, FOHGTN, and NRDN modules.
12. Initialize the parameters of the NRDN, FOHGTN, LVSTM, and LSTM modules.
13. For each retinal fundus image, $I \in D$, performing the following steps:
14. Apply the NRDN to enhance vascular structures and generate vascular feature representation V_i .
15. Pass V_i to the FOHGTN to extract potential microaneurysm candidates.
16. Construct spatial-anatomical relationships using the HGT to obtain refined lesion candidate features.
17. Feed the refined features to the LVSTM to learn contextual and texture representations.
18. Fuse the contextual embeddings from ViT with texture descriptors derived from LBP.
19. Normalize the fused feature representation using Z-score normalization.
20. Classify the features using an SVM classifier with RBF kernel to predict microaneurysm or ischemic regions.
21. Compute the loss between predicted outputs and ground-truth labels.
22. Perform backpropagation to propagate gradients through LVSTM, FOHGTN, and NRDN modules.
23. Update model parameters using gradient-based optimization.
24. Repeat the process for all training samples until convergence criteria are satisfied.

IV. Results and Discussion

In this part, the outcomes of the suggested model have been shown. When compared to other current methods, the findings show that the suggested model offers superior MA detection and classification to distinguish from ischemic regions, as well as prediction accuracy.

a. Experimental setup

This section provides a comprehensive overview of the performance and implementation results of the suggested system, which was simulated in Python 3.9 using PyTorch 2.0 deep learning framework. Additionally, a proportional analysis is provided to verify that the proposed system functions as planned. Moreover, each module was introduced with a specific function objective, reducing reliance on a single deep model to learn all representations simultaneously. All the experiments

were conducted under a controlled computational environment to ensure reproducibility and transparency.

a.i. Computational environment

The experimental platform consisted of the following hardware and software specifications:

Operating system	: Windows 11 (Version 22H2, 64-bit)
Processor	: Intel® Core™ i7-10750H @ 2.60 GHz
RAM	: 16 GB DDR4
GPU	: NVIDIA GeForce RTX 3060
CUDA version	: 11.6
CuDNN version	: 8.4
Deep learning framework	: PyTorch 2.0

Here, GPU acceleration was enabled using CUDA 11.6 and cuDNN 8.4 to optimize training efficiency and memory utilization.

a.ii. Model configuration

Backbone architecture	: Vision transformer (ViT-Base-Patch16-224)
Feature extraction	: LBP combined with ViT deep embeddings
Sequence modeling	: LSTM with 3 stacked layers and 256 hidden units per layer
Optimize	: Adam
Initial learning rate	: 0.001
Batch size	: Full-batch gradient descent
Number of epochs	: 150
Loss function	: BCE loss
Final classifier	: SVM (RBF kernel)

BCE, binary cross-entropy; LBP, local binary patterns; LSTM, long short-term memory; SVM, support vector machine; ViT, vision transformer.

These settings were selected to ensure stable training convergence and effective feature learning for accurate cardiovascular disease prediction.

b. Dataset description

The proposed NFOVSTN-LSTM dataset uses the Retinal Lesions Dataset to effectively support the

design, training, and evaluation. This dataset was strategically selected for its compatibility with the multi-stage vascular feature analysis, temporal modeling, and lesion classification pipeline described in this research. The Retinal Lesion Segmentation Dataset comprises 1,906 color fundus images annotated at the pixel level for two clinically significant lesion types: MAs (1,435 images) and Ischemic Lesions (471 images). Each image is paired with a correspondence between the original images and annotations for supervised segmentation. To enhance model generalization, the dataset includes images acquired under diverse illumination settings and varying pathological conditions. All samples are organized in paired original-mask format with an 80:20 train-test split and standardized preprocessing, making the dataset well-suited for deep learning-based retinal lesion segmentation and localization tasks. <https://www.kaggle.com/datasets/ajithdari/retinal-lesion-segmentation-dataset/data>

c. Simulated results of the proposed model

This section depicts the simulation model and results of the suggested model, highlighting its performance.

Figure 5 shows the preprocessing and vascular enhancement stages applied to fundus images in the proposed system. The process begins with the original retinal image, which includes subtle signs of MAs and ischemic areas, particularly around the macula. GLCM is first applied to segment the image and capture fine texture variations. Canny Edge Detection follows, extracting vessel boundaries and microvascular structures. The Frangi Filter then enhances tubular features, improving the visibility of the vascular network and dropout zones. Finally, CLAHE increases local contrast to highlight faint lesions and perfusion irregularities near the macular region.

Figure 6 presents the feature extraction and lesion detection stages following preprocessing. The original retinal fundus image shows microvascular abnormalities. Otsu thresholding is applied to segment lesion-intense regions using histogram-based grayscale analysis, isolating hyperreflective and hypoperfused zones indicative of MAs or ischemia. The Fourier spectrum then transforms the image to the frequency domain, highlighting high-frequency texture anomalies linked to MAs. Lesion candidates are identified via template matching and spatial filtering, with green circles marking detected MAs and ischemic patterns. These outputs are processed by Heterogeneous

Graph and Vision Transformers for contextual lesion classification. Final MA vs. ischemia classification is performed by SVM, as shown in Figure 7, comparing the original and the detected lesion image.

Thus, the proposed framework highlights critical retinal structures, captures complex anatomical and vascular dependencies, and contextualizes spatial features to distinguish MA from ischemic regions.

Figure 8 depicts the confusion matrix of the proposed NFOVSTN-LSTM framework, representing the classification performance of the proposed ViT integrated with the SVM classifier for distinguishing between MA and ischemic regions. The matrix reveals that out of the ischemic cases, 1,615 were correctly identified, while 22 were misclassified as MA. Similarly, 978 MA cases were accurately detected, and 12 were incorrectly labeled as ischemic. Despite the inherent difficulty in distinguishing overlapping retinal features, such as clustered MAs and silent ischemic regions, the hybrid ViT-SVM architecture successfully differentiates subtle lesion patterns, validating its robustness in clinical retinal image analysis.

The receiver operating characteristic (ROC) curve further evaluates the discriminative capability of the proposed model. The model achieves an area under the curve (AUC) of 98.23%, indicating strong separability between MA and Ischemic lesion classes. The ROC curve lies significantly above the diagonal reference line (AUC-0.5), confirming that the model performs substantially better than random classification. This result demonstrates a high sensitivity-specificity balance across different decision thresholds. Despite achieving a strong overall performance (AUC = 98.23%), the confusion matrix indicates 34 misclassifications out of 2,627 samples ($\approx 1.29\%$). Specifically, 22 ischemic regions were predicted as MAs, and 12 MAs were classified as ischemic. These errors mainly occur in small, low-contrast MA lesions and diffuse ischemic regions with irregular boundaries. Peripheral localization and illumination variability further contribute to classification uncertainty. From the error analysis, the low error rate confirms the robustness of the proposed model while identifying clinically challenging areas for future enhancement. Figure 9 illustrates the AUC curve of the proposed model.

d. Performance evaluation

The performance metrics, such as accuracy, sensitivity, precision, mean absolute error (MAE), RMSE, etc., are calculated based on the simulation results and are explained in this section.

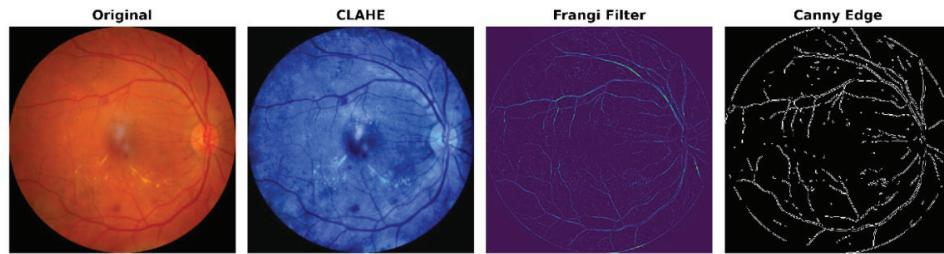


Figure 5: Retinal image preprocessing and vascular enhancement pipeline. CLAHE, contrast-limited adaptive histogram equalization.

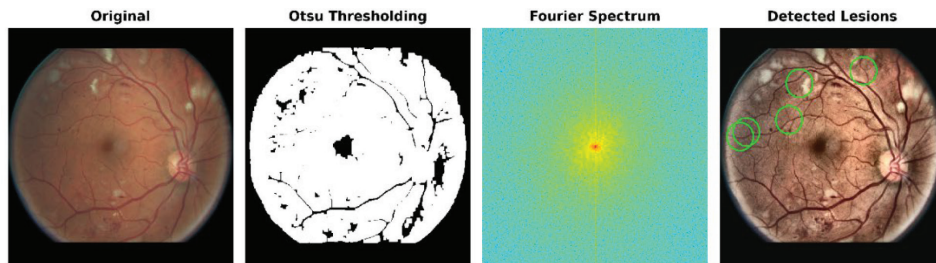


Figure 6: Feature extraction for retinal MA detection. MA, microaneurysm.

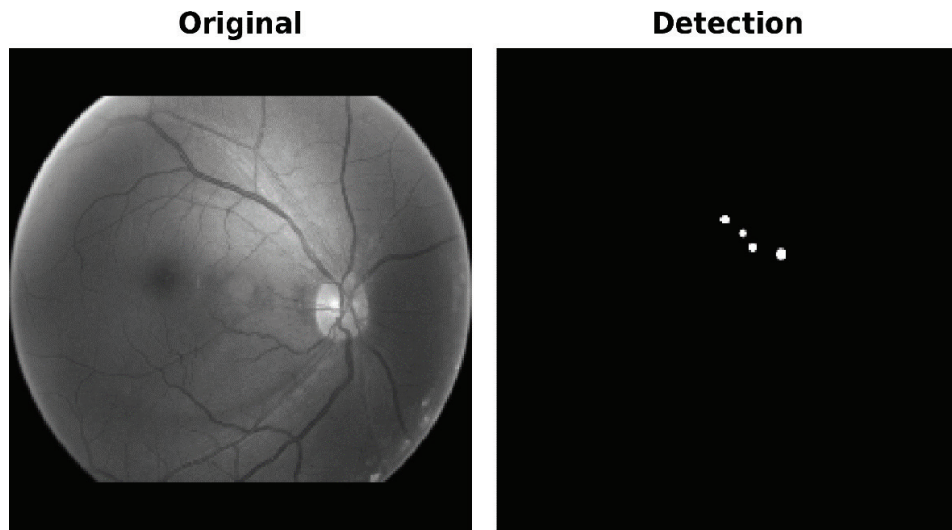


Figure 7: The original image and the MA-detected image. MA, microaneurysm.

Figure 10 illustrates the accuracy of the proposed NFOVSTN-LSTM framework, where the accuracy is calculated by increasing epochs. The accuracy increases with an increasing number of epochs, starting from the minimum value of 0.45, the accuracy increases and reaches a maximum of 0.987 at 120–140 epochs. LVSM-T using contextual attention in ViT minimizes misclassifications across diverse lesion

appearances, thereby raising the overall percentage of correct predictions. Also, it consistently yields a higher ratio of true positives and true negatives, which directly translates into improved accuracy.

Figure 11 analyzes the loss of the proposed framework with increasing epochs. Initially, the loss is 0.7, and on increasing the epochs, it reduces and reaches a minimum value of 0.01 at epoch

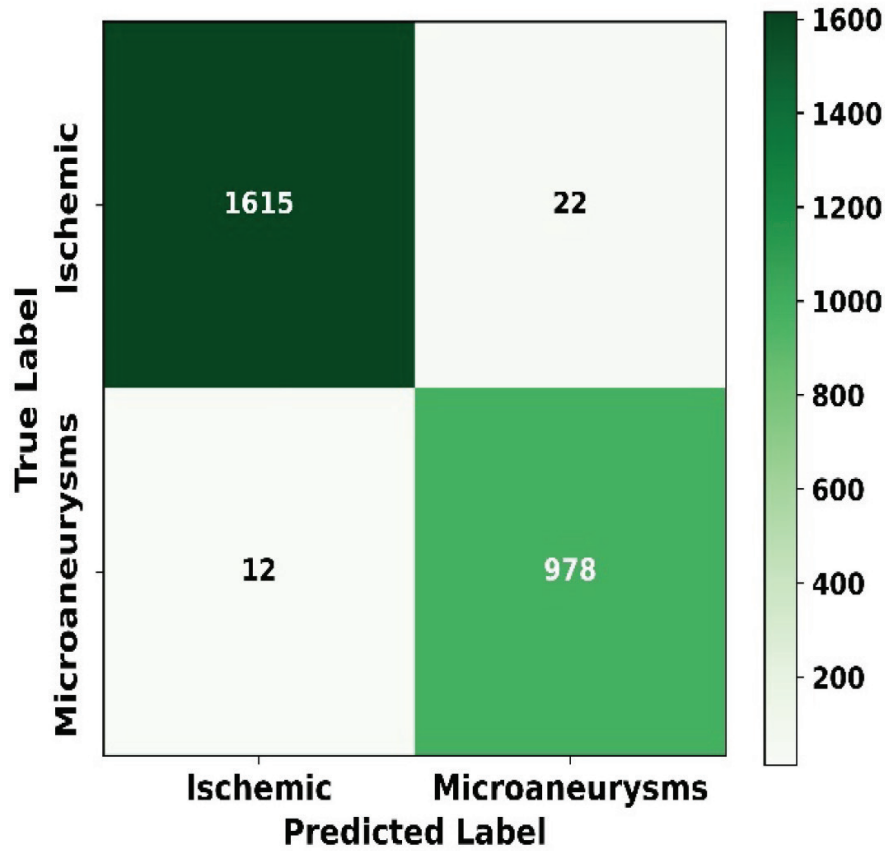


Figure 8: Confusion matrix of the proposed model.

150. The loss graph is directly tied to LVSMT incorporating ViT, which extracts robust features, and SVM ensures clear decision boundaries, lowering classification error. Thus, the model minimized classification error over epochs, making the model effective in learning to differentiate between MAs and ischemic regions, even in complex retinal environments.

Figure 12 illustrates the sensitivity of the proposed framework, where the sensitivity is initially low at 24% and gradually increases with increasing epochs and reaches a maximum value of 98.5% at epoch 150. The NRDN enables the detection of subtle texture and contrast variations by employing GLCM and CLAHE, thereby improving sensitivity to early-stage and non-lesion neurovascular remodeling, ensuring early anomalies are not missed.

Figure 13 illustrates the sensitivity of the proposed framework, where the specificity is initially low of 17% and gradually increases with increasing epochs and reaches a maximum value of 99.1% at epoch 150. This increased specificity is attributed to FOHGTN

as its MR step suppresses background clutter and non-MA textures, ensuring that only genuine circular lesions pass through, thus reducing false positives in MA detection.

Figure 14 analyzes the precision of the proposed NFOVSTN-LSTM framework, where the precision is initially low of 35%, and with increasing epoch, it gradually increases and reaches a maximum of 97.8% at epoch 150. This trend is attributed to LVSMT by the LBP-based texture encoding isolating texturally distinct ischemic regions, preventing misclassification of normal vascular variations as pathology, which raises the positive predictive value.

Figure 15 depicts the average time taken by the proposed NFOVSTN-LSTM framework to complete the entire process with increasing iterations. Initially, the time is 3 s, and on increasing epochs, it takes more time, and it takes a maximum time of 25 s, and this trend is attributed to FOHGTN. Instead of deep segmentation networks, Fourier-based template matching, and Otsu thresholding are used for fast, efficient segmentation.

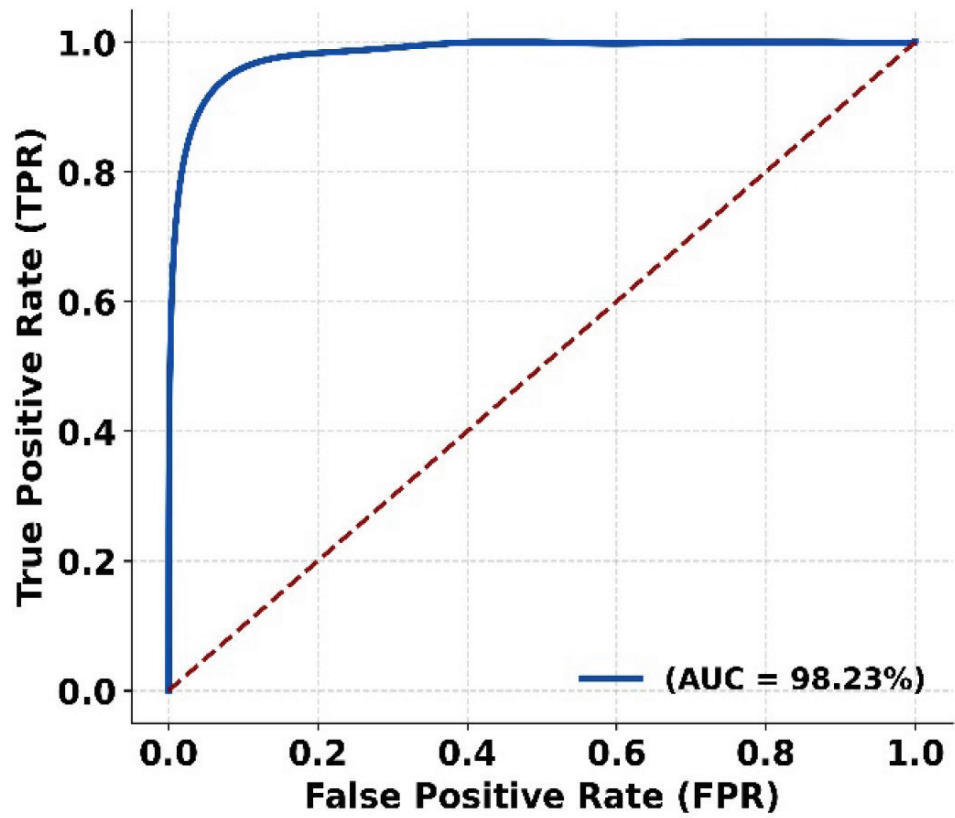


Figure 9: ROC curve of the proposed model. AUC, area under the curve; FPR, false positive rate; ROC, receiver operating characteristic; TPR, true positive rate.

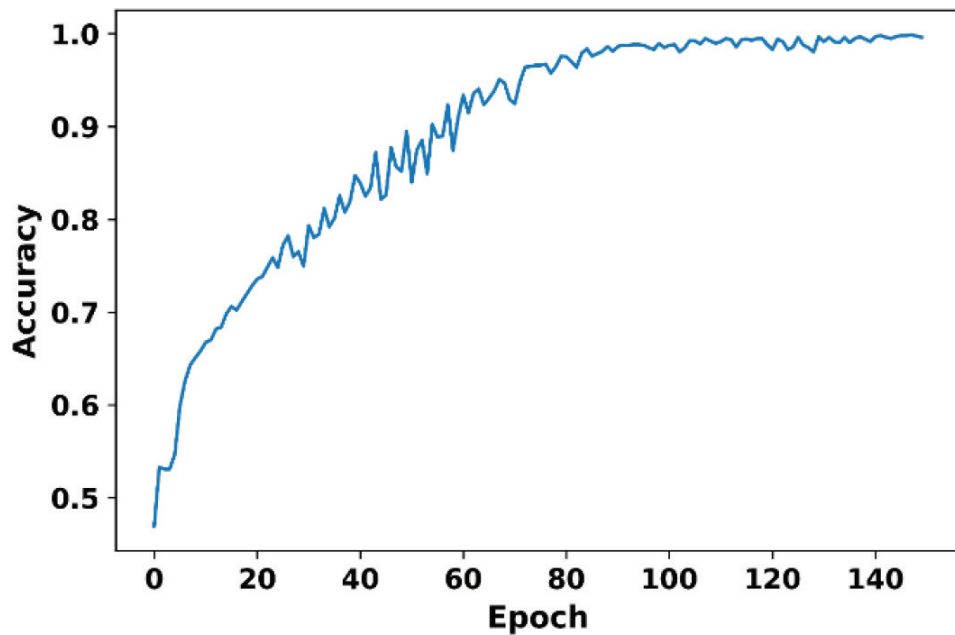


Figure 10: Accuracy of the proposed framework.

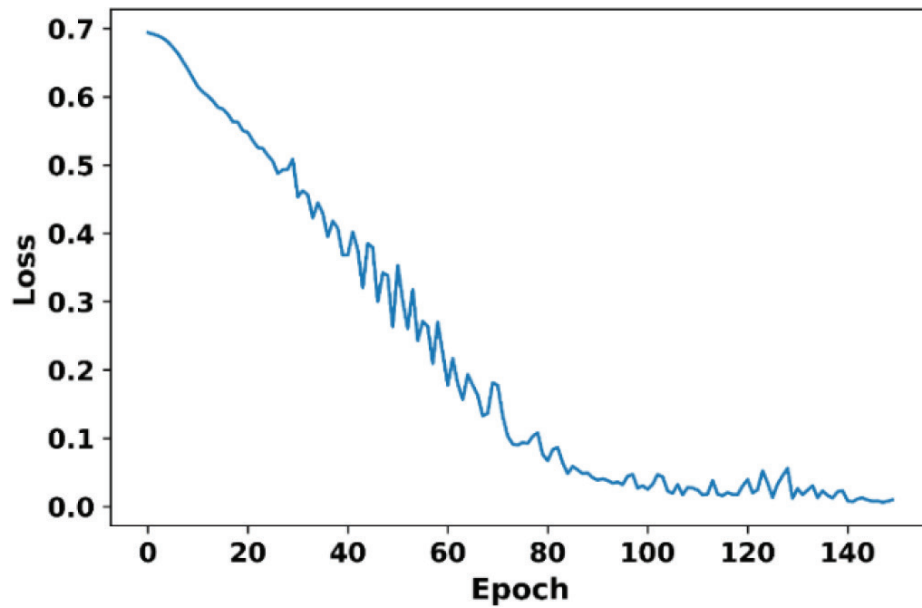


Figure 11: Loss analysis of the proposed framework.

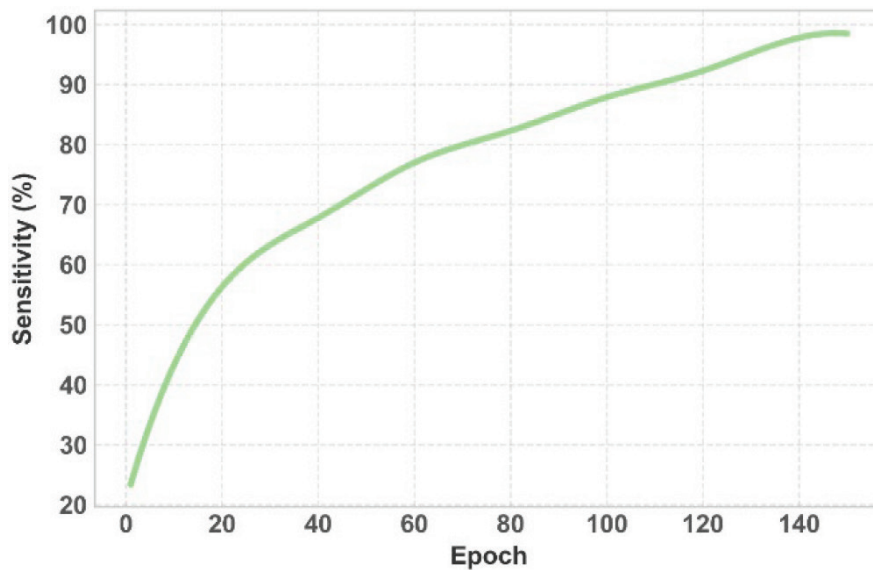


Figure 12: Sensitivity of the proposed framework.

Figure 16 illustrates the F1-score of the proposed NFOVSTN-LSTM framework, where the initial value is low at 35%, and it increases with increasing epoch and reaches a maximum of 98.4% at epoch 150. FOHGTN balances the f1-score by the fusion of spatial-frequency features with temporal information. It shows that the system can be reliably used for early diagnosis and clinical decision support, even in cases where visual cues are subtle or ambiguous.

Figure 17 illustrates the MAE of the proposed NFOVSTN-LSTM, where the model has a higher MAE of 1.57 at the beginning, and with increasing iterations, the MAE decreases to 0.012 at epoch 140. LVSM's ViT attention mechanism reduces MAE by focusing on the correct spatial regions, ensuring that predicted lesion severity aligns closely with actual pathology. Lower MAE indicates improved lesion boundary prediction and change detection.

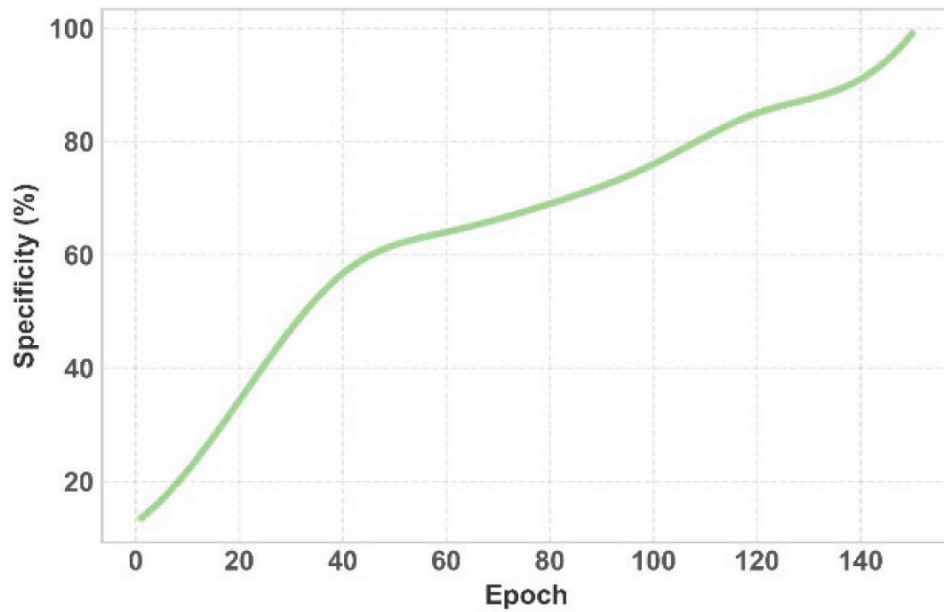


Figure 13: Specificity of the proposed framework.

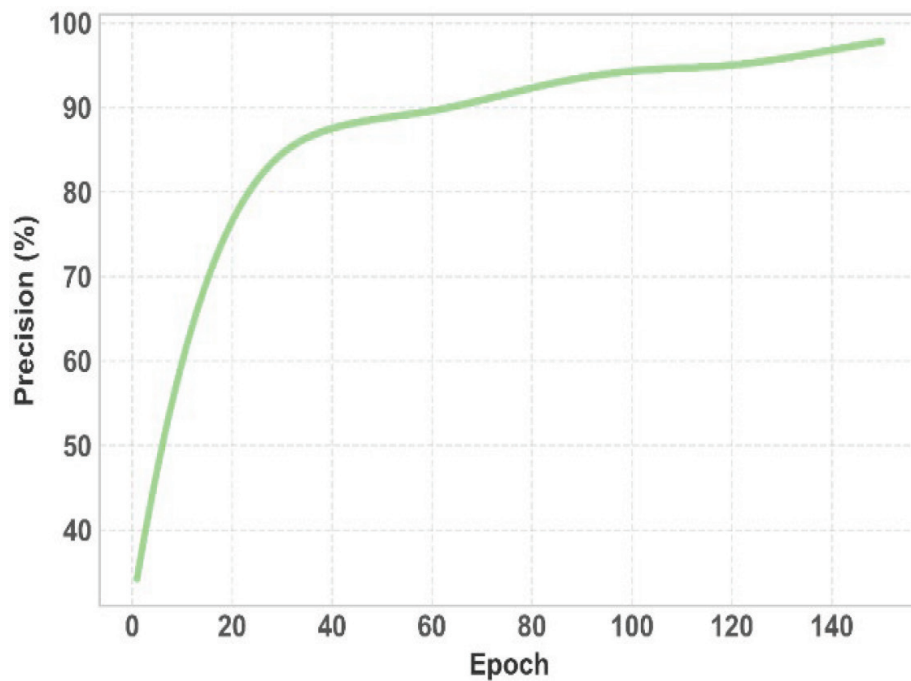


Figure 14: Precision of the proposed framework.

Figure 18 represents the RMSE of the proposed framework, which has a maximum value of 1.85 at the initial epoch, gradually decreases with increasing epoch, and attains a lower value of 0.034 at epoch 150. LVSMs with improved feature stability with ViT attention, lowering high-error outliers, help fine-tune

the predictions, ensuring that occasional outliers are less extreme, contributing to a smaller squared-error average.

Figure 19 represents the area under the precision-recall curve (AUPR) of the proposed framework, which begins with a lower value of

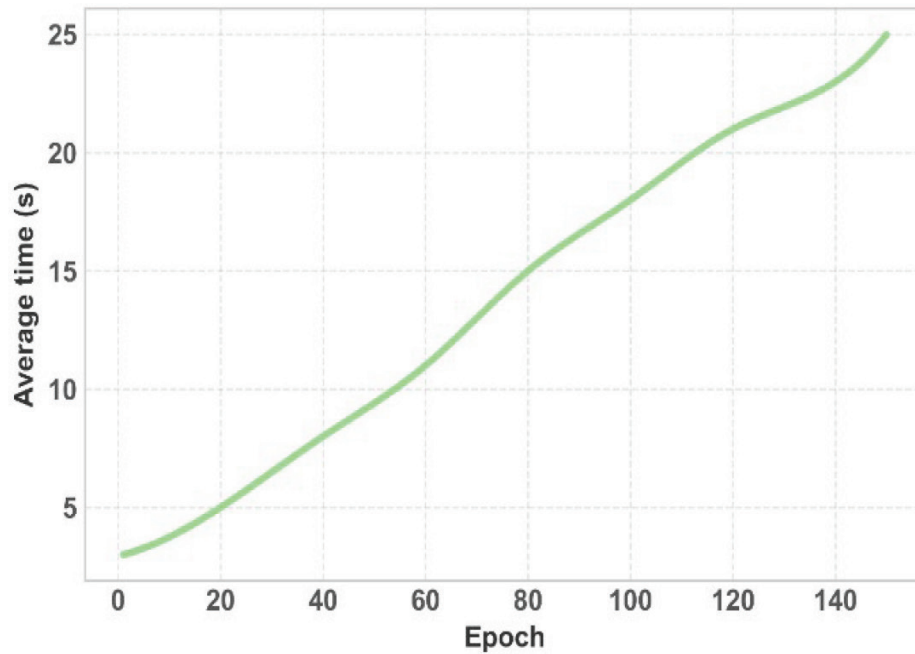


Figure 15: Average time of the proposed NFOVSTN-LSTM. NFOVSTN-LSTM, Neurovascular Fourier-Otsu Vision Support Transformer Network with Long Short-Term Memory.

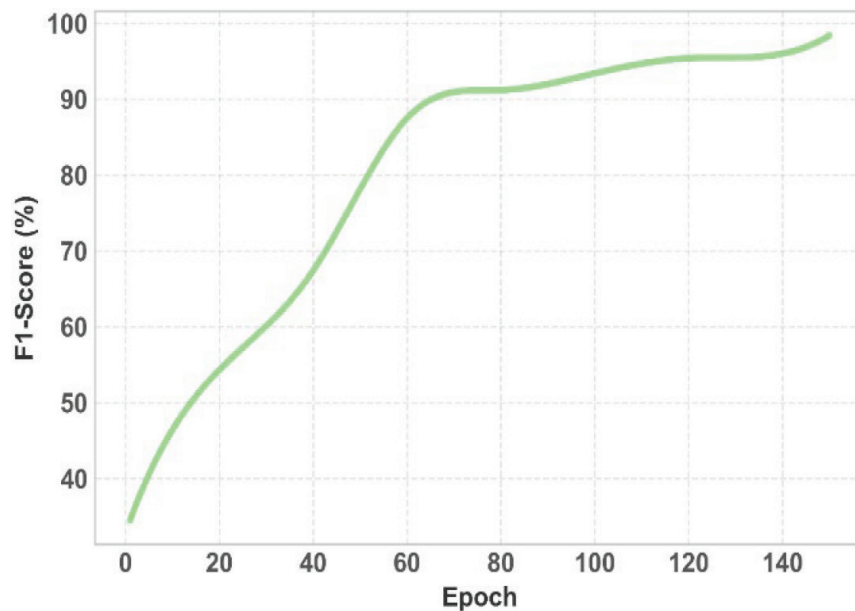


Figure 16: F1-score of the proposed NFOVSTN-LSTM framework. NFOVSTN-LSTM, Neurovascular Fourier-Otsu Vision Support Transformer Network with Long Short-Term Memory.

approximately 13% at the initial epoch and gradually increases with increasing epochs, reaching nearly 50% at epoch 100. The steady upward trend indicates consistent improvement in precision-recall performance as training progresses. The progressive

enhancement reflects improved discrimination of true MAs regions while maintaining false positives. The smooth convergence pattern further confirms stable optimization and reliable-level detection capability of the proposed NFOVSTN-LSTM framework.

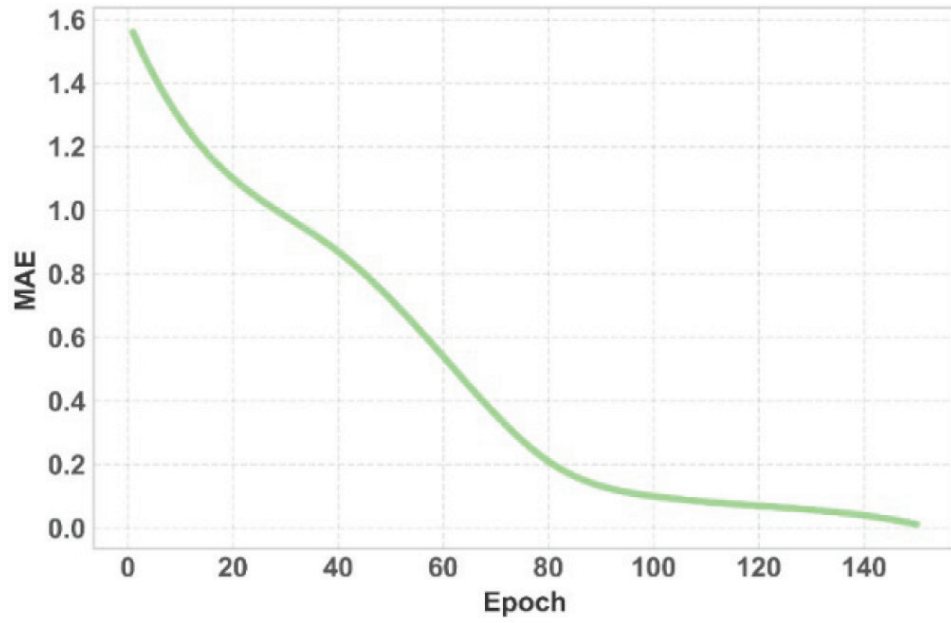


Figure 17: MAE of the proposed NFOVSTN-LSTM framework. MAE, mean absolute error; NFOVSTN-LSTM, Neurovascular Fourier-Otsu Vision Support Transformer Network with Long Short-Term Memory.

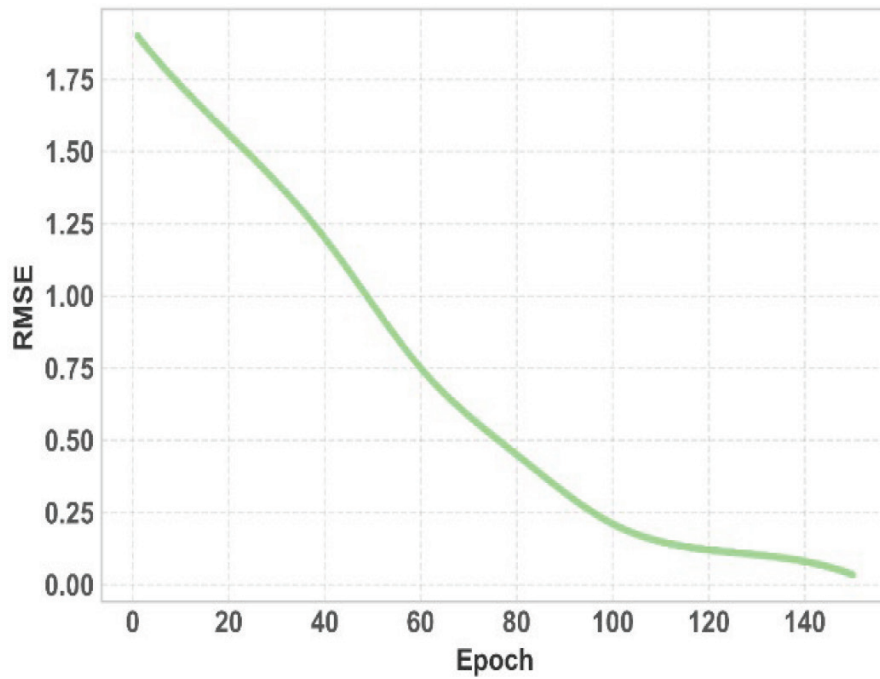


Figure 18: RMSE of the proposed NFOVSTN-LSTM. NFOVSTN-LSTM, Neurovascular Fourier-Otsu Vision Support Transformer Network with Long Short-Term Memory.

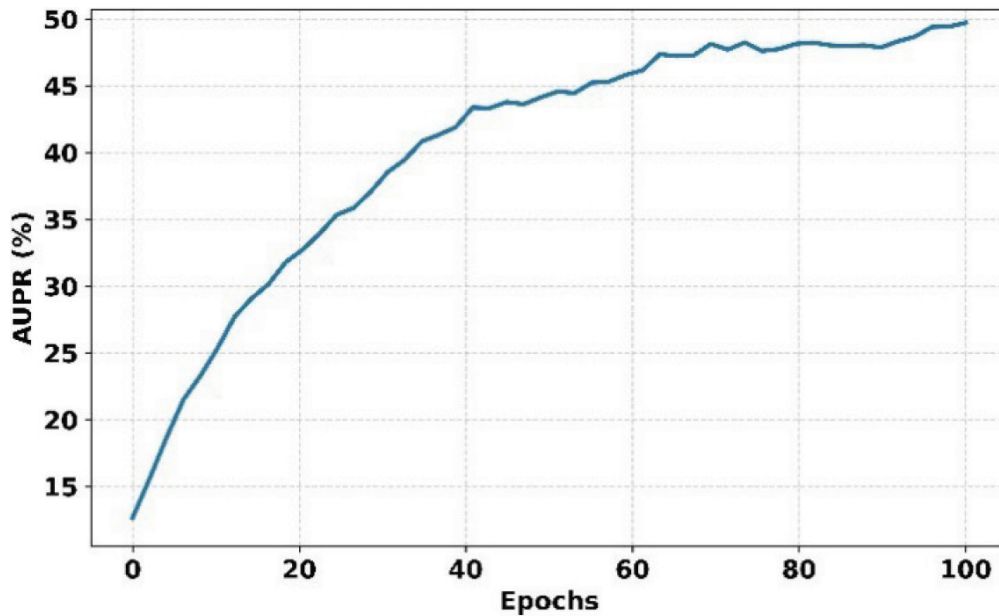


Figure 19: AUPR of the proposed NFOVSTN-LSTM framework. AUPR, area under the precision-recall curve; NFOVSTN-LSTM, Neurovascular Fourier-Otsu Vision Support Transformer Network with Long Short-Term Memory.

Figure 20 illustrates the Dice score progression across epochs, where the initial value is 9% during the early training phase and consistently improves, attaining around 36% at epoch 100. This incremental improvement demonstrates the model's enhanced capability in capturing fine-grained lesion structures and reducing segmentation mismatch. The stable upward trend confirms effective optimization and improved structural agreement between predictions and annotations.

e. Analysis of the proposed model with other approaches

This section presents a comparative performance analysis of the proposed model with several existing approaches widely used in retinal image analysis and MA detection. The selected methods represent different methodological categories, including traditional machine learning models, deep convolutional networks, segmentation architectures, and object detection frameworks. DCNN is considered a baseline deep learning model due to its ability to automatically extract hierarchical features from fundus images, which aligns with the core concept of the proposed framework. The SBF-based MA method is included to compare against handcrafted feature-based techniques that rely on structural and

vascular characteristics. VGG-19 and Inception-V3 [18], along with GoogLeNet, AlexNet, and NAD VGNet [20], are well-established deep CNN architectures known for strong feature representation and multi-scale learning, which are essential for detecting small lesions such as MAs. FCN is incorporated for its pixel-wide segmentation capability, while KNN and Tree Ensemble methods [25] are conventional classifiers dependent on manual feature extraction. Additionally, Faster R-CNN with FPN, ISR-FPN, and MAFP-Net [41] are incorporated due to their effectiveness in multi-scale and small-object detection, directly relevant to MA identification. The comparison is performed using accuracy, sensitivity, specificity, F1-score, MSE, and RMSE to validate the effectiveness of the proposed model.

Figure 21 compares the accuracy of the proposed model with existing models in MA detection. The accuracy of DCNN, SBF-based MA, and VGG-19& Inception V3 are 69%, 83%, and 94%, whereas the proposed framework achieves a higher accuracy of 98.7%. This layered approach significantly boosts diagnostic precision, particularly in early or silent ischemic stages, and leads to a marked improvement in overall detection performance compared to existing systems.

Figure 22 compares the sensitivity of the proposed NFOVSTN-LSTM framework with existing

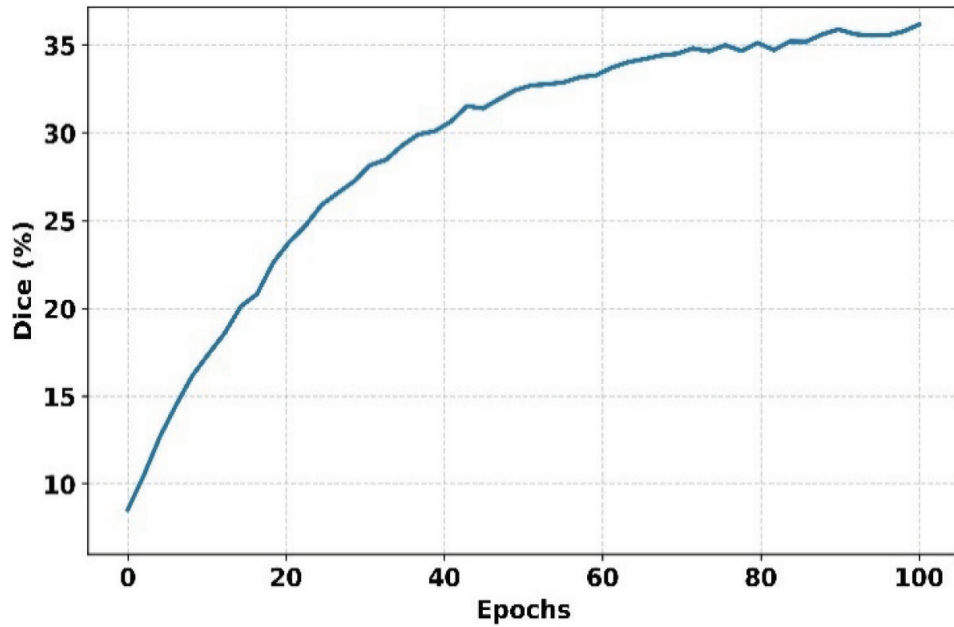


Figure 20: Dice coefficient of the proposed NFOVSTN-LSTM Framework. NFOVSTN-LSTM, Neurovascular Fourier-Otsu Vision Support Transformer Network with Long Short-Term Memory.

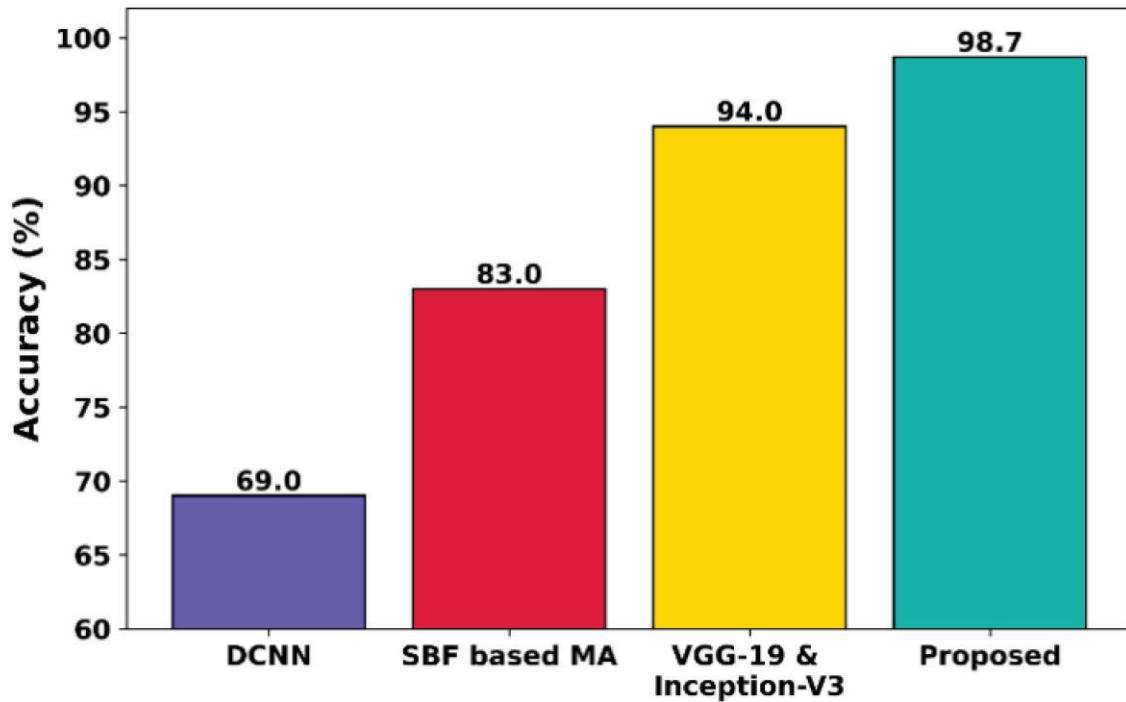


Figure 21: Accuracy comparison with existing models. MA, microaneurysm.

approaches, where the sensitivity of DCNN, SBF-based MA, and VGG-19& Inception V3 are 64%, 64%, and 92%, whereas the proposed MA detection model achieves 98.5%, which is very high compared

to other models. This fine-grained classification ensures that subtle lesions and ischemic changes are detected early and accurately, reducing false negatives significantly.

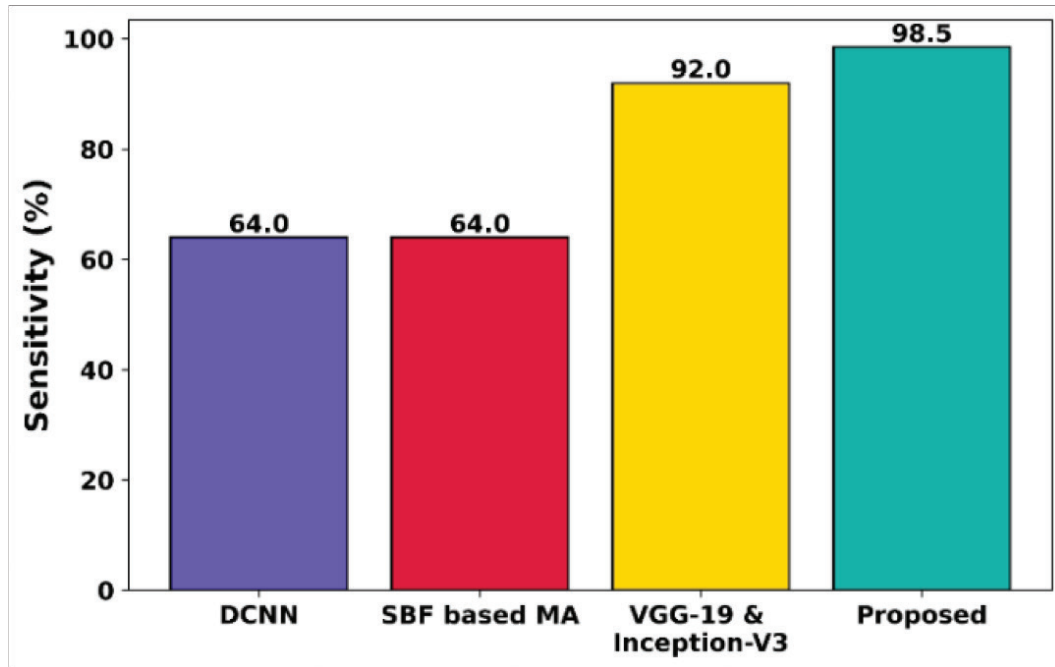


Figure 22: Sensitivity comparison of the NFOVSTN-LSTM framework. MA, microaneurysm.

Figure 23 illustrates the specificity comparison of the proposed NFOVSTN-LSTM framework with existing models, where the specificity of DCNN, SBF-based MA, and VGG-19& Inception V3 are 88%, 82%, and 94%, whereas the proposed MA detection model achieves 99.1%, which is very high compared to other models. This indicates fewer false positives, leading to more clinically reliable detection and improving screening accuracy for early DR by ensuring that only true pathological signs are flagged, thus reducing unnecessary follow-ups or overtreatment.

Figure 24 compares the precision of the proposed NFOVSTN-LSTM framework with existing works, where the precisions of faster RCNN FPN, faster RCNN ISR FPN, and MAFP-Net are 42.1%, 40.2%, and 63.3%, respectively, and the proposed model has a higher precision of 97.8%. This higher precision indicates fewer false positive detections, which is critical in clinical screening to avoid overdiagnosis or misidentification of healthy regions.

Figure 25 compares the F1-score of the proposed NFOVSTN-LSTM framework with existing works, where the F1-score of faster RCNN FPN, faster RCNN ISR FPN, and MAFP-Net are 49.8%, 53.6%, and 62.4%, respectively, and the proposed model has a higher F1-score of 98.4%. The above trend significantly improves contrast and vessel clarity

and enhances the separability of MAs and ischemic regions from the background, increasing true positive detection and reducing false negatives.

Figure 26 compares the AUC of the proposed NFOVSTN-LSTM framework with existing works, where the AUCs of faster RCNN FPN, faster RCNN ISR FPN, and MAFP-Net are 0.521, 0.785, and 0.721, respectively, and the proposed model has a higher AUC of 0.99. This higher AUC reflects the effectiveness of the transformer-based classification module and the precision of SVM in distinguishing retinal lesions.

Figure 27 illustrates the MAE comparison of the proposed framework with the existing works, where the MAEs of GoogleNet, AlexNet, and VGNet are 0.021, 0.023, and 0.018, respectively, whereas the proposed model has a lower MAE of 0.012. The observed trend is due to the precise localization of MAs and ischemic zones, and misclassification is reduced.

Figure 28 illustrates the RMSE comparison of the proposed framework with the existing works, where the RMSEs of GoogleNet, AlexNet, and VGNet are 0.109, 0.115, and 0.117, respectively, whereas the proposed model has a lower RMSE of 0.034. This performance indicates strong generalization capability, suggesting the system's potential for reliable clinical deployment in screening and diagnosing DR.

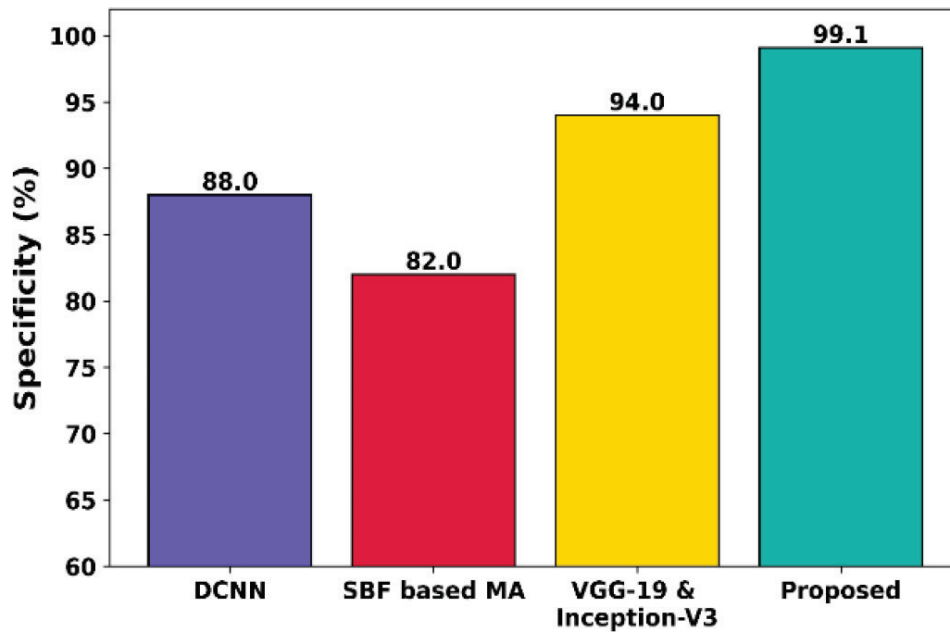


Figure 23: Specificity comparison of the proposed framework. MA, microaneurysm.

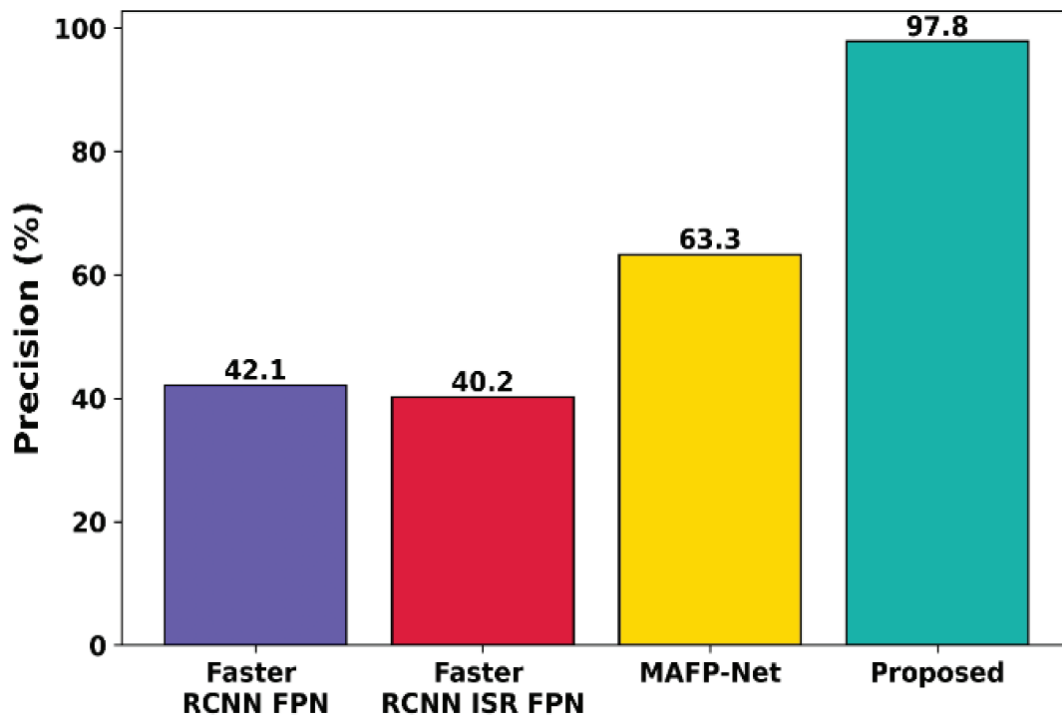


Figure 24: Precision comparison of the proposed framework.

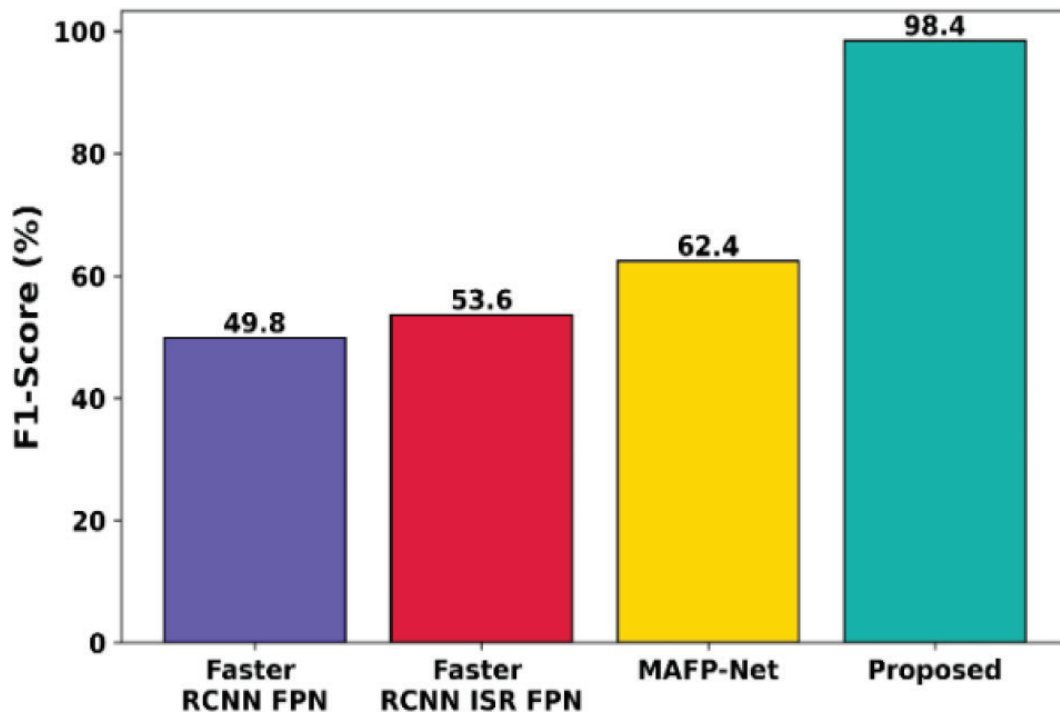


Figure 25: F1-score comparison of the proposed framework.

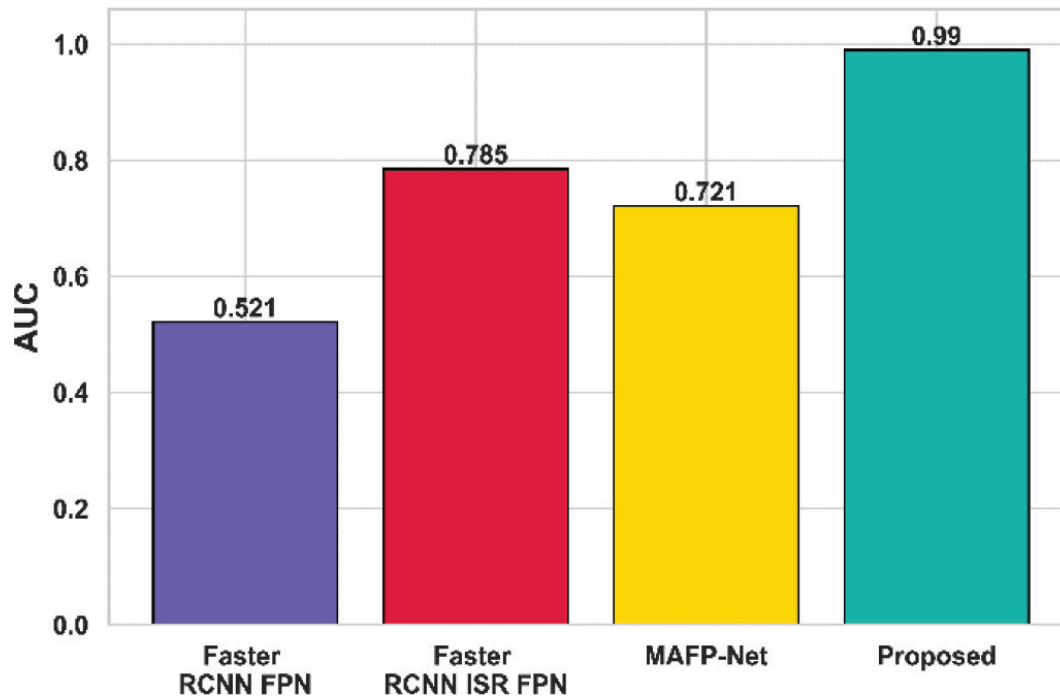


Figure 26: AUC comparison of the proposed framework. AUC, area under the curve.

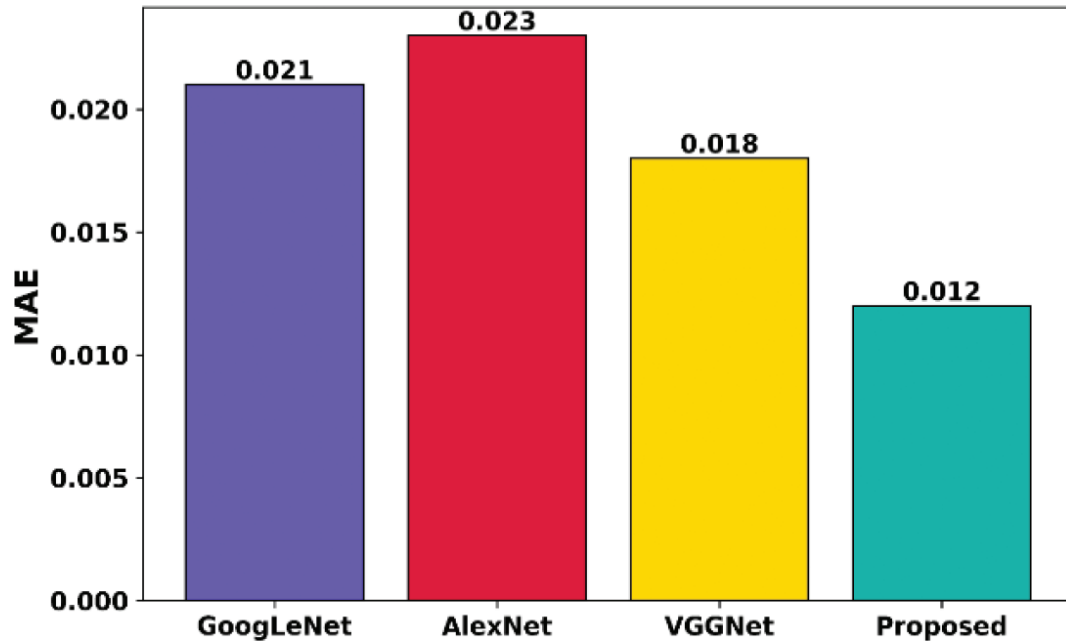


Figure 27: MAE comparison of the proposed framework. MAE, mean absolute error.

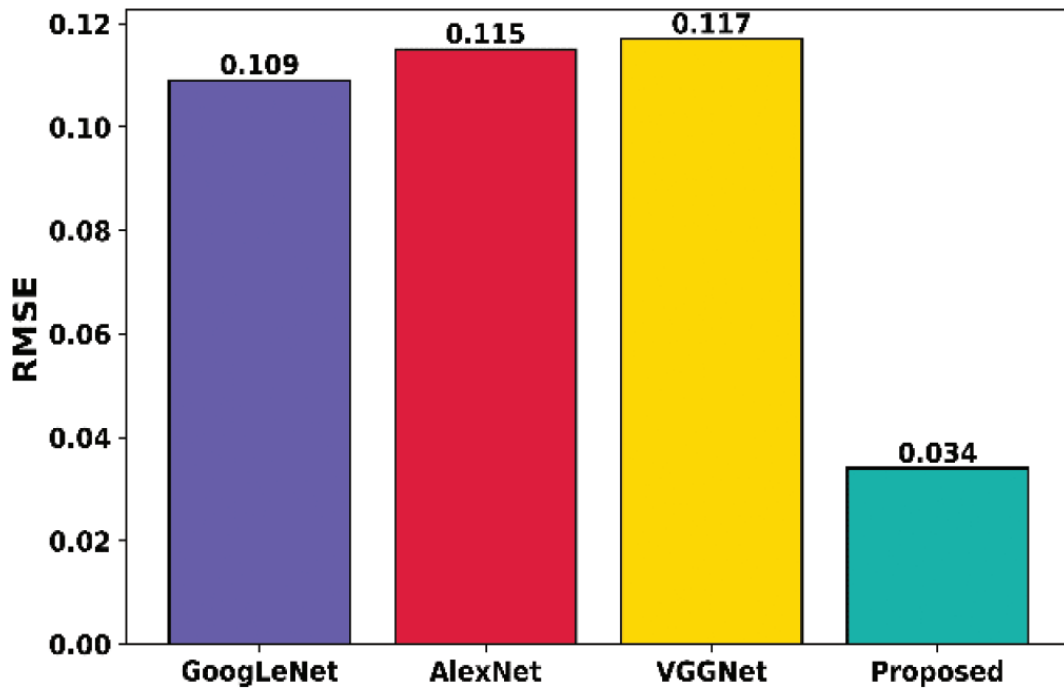


Figure 28: RMSE comparison of the proposed framework.

Figure 29 compares the average time of the proposed framework with the existing works, where the time taken by FCN, KNN, and Tree Ensemble are 220 s, 60 s, and 65 s, respectively, whereas the proposed model takes the shortest time of 25 s. The observed reduction in average execution time is primarily driven by the integration of efficient preprocessing, localized feature extraction, and transformer-based attention mechanisms, which collectively reduce the dimensionality and computational complexity of the data while preserving diagnostic accuracy.

Figure 30 compares the AUPR of the proposed NFOVSTN-LSTM framework with existing works, where the AUPR of UNet, UNet++, M2MRF, and MCA-UNet are 18.76, 29.98, 28.46, and 46.41, respectively. The proposed model has a higher AUPR of 49.72, and AUPR is more informative than AUC-ROC in class-imbalanced settings common in retinal imaging. Higher AUPR signifies higher lesion detection sensitivity while avoiding overfitting to false alarms.

Figure 31 compares the Dice coefficient of the proposed NFOVSTN-LSTM framework with existing works, where the Dice scores of UNet, UNet++, M2MRF, and MCA-UNet are 21.59%, 25.35%, 27.81%, and 33.32%, respectively, whereas the proposed model achieves a higher Dice score of 36.18%.

This improvement confirms the effectiveness of the proposed multi-stage architecture in accurately segmenting MA regions compared to conventional segmentation models.

Figure 32 illustrates the cross-dataset generalization ability of the proposed framework across multiple public and private retinal datasets like E-Ophtha-MA, DIARETDBI, FFA Images, and Private DR, which are acquired using different imaging devices and acquisition protocols. Despite variations in image quality, resolution, and population characteristics, the presented model consistently maintains high accuracy, precision, and F1- score of about 98.7, 97.8, 98.4. This performance stability indicates that the learned representations are robust to dataset-specific biases and demonstrate strong generalizability across diverse retinal imaging conditions.

f. Ablation study

Table 1 presents an ablation study evaluating the contribution of each architectural component. The full NFOVSTN-LSTM achieves the highest performance with 98.70% accuracy, 97.80% precision, and 98.40% F1-score. The most significant performance drop occurs when the NRDN module is removed that reduces its accuracy to 93.62% and F1-score

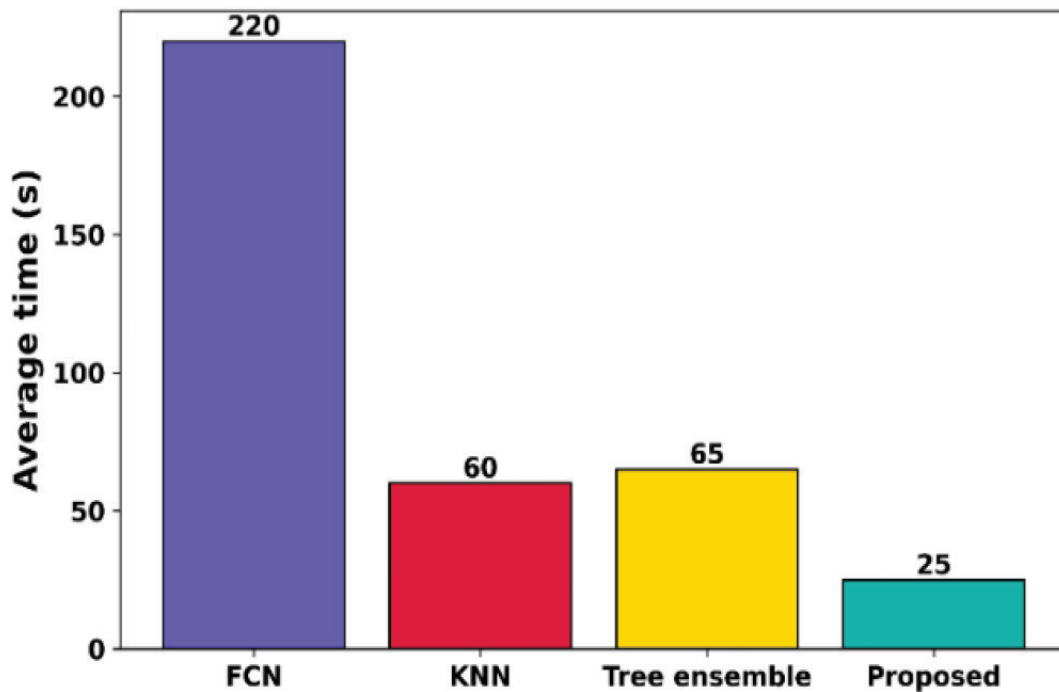


Figure 29: Average time comparison of the proposed framework.

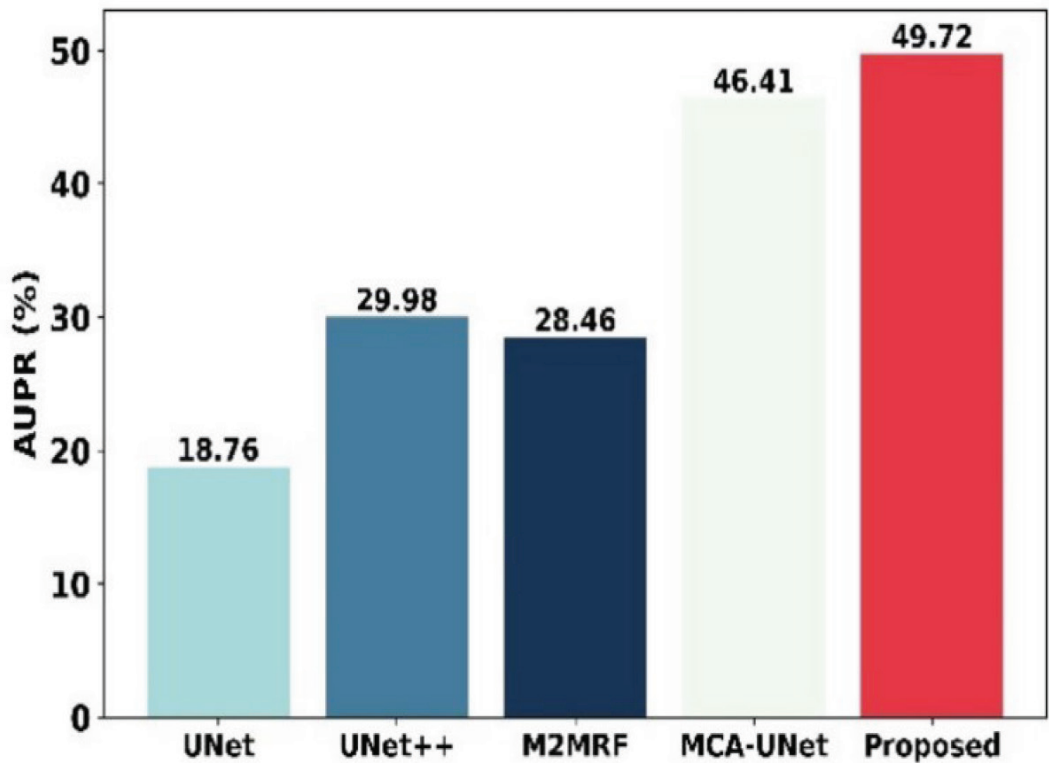


Figure 30: AUPR comparison of the proposed framework. AUPR, area under the precision-recall curve.

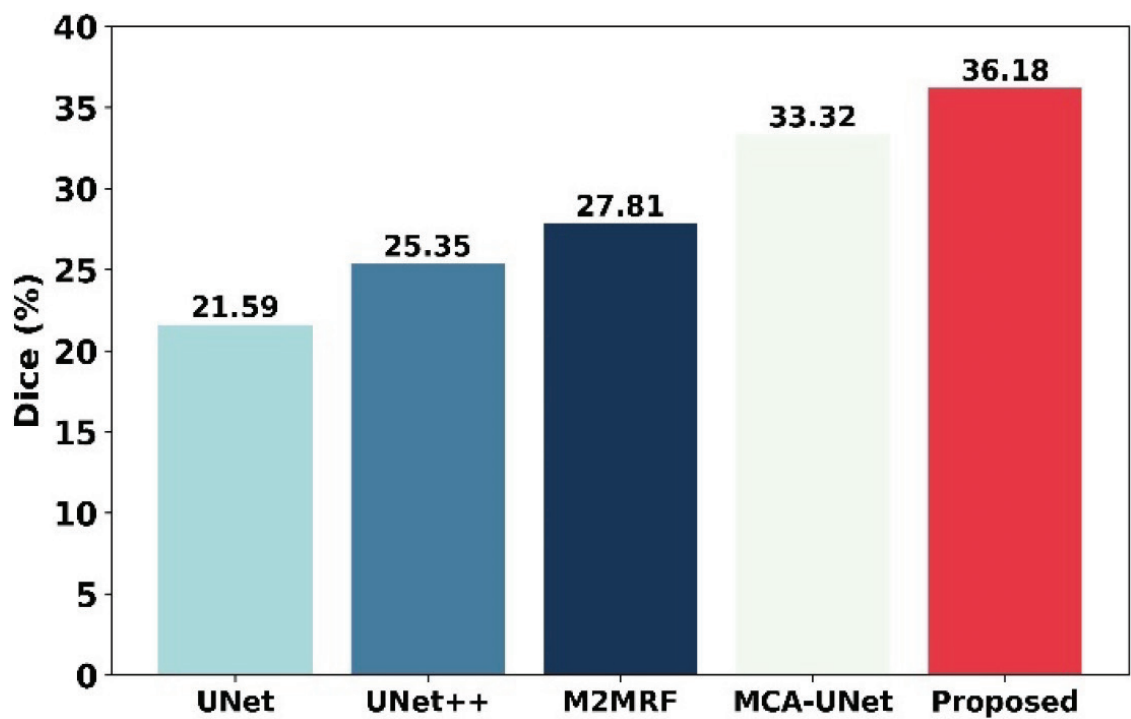


Figure 31: Dice coefficient comparison of the proposed framework.

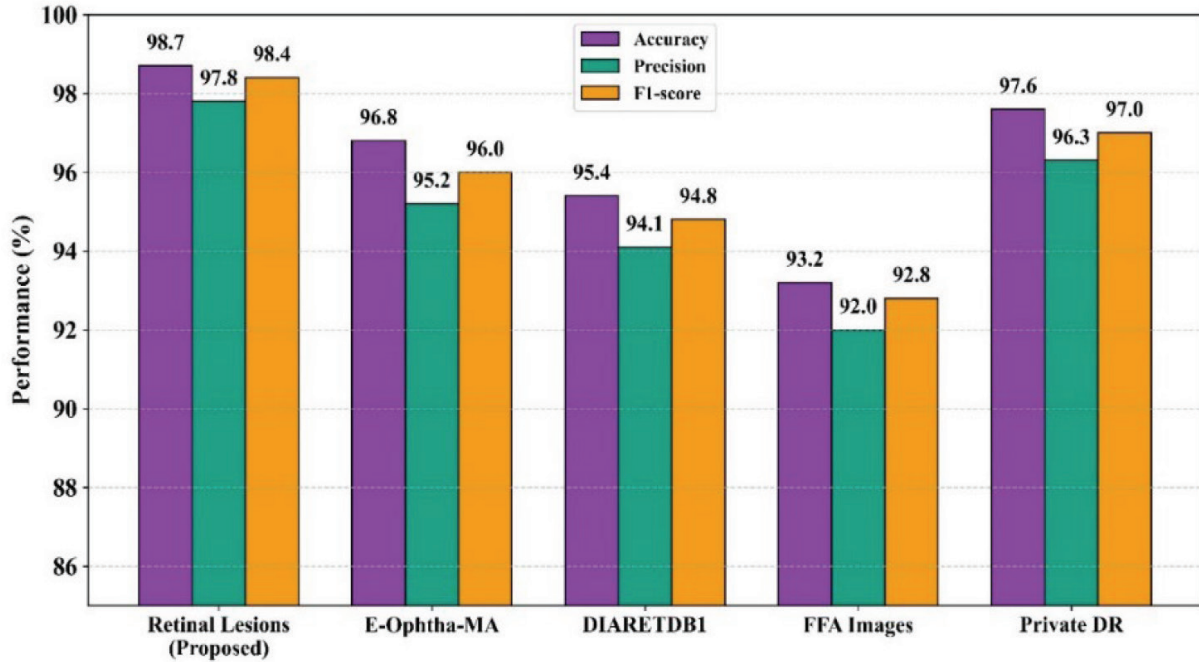


Figure 32: Cross-dataset performance analysis of the proposed retinal lesion detection framework. DR, diabetic retinopathy; FFA, fluorescein fundus angiography.

Table 1: Ablation study.

Configurations	Accuracy (%)	Precision (%)	F1-Score (%)
Without NRDN in the LSTM first layer	93.62	92.80	92.95
Without CGLEN	94.48	93.70	93.90
Without FCEM	95.31	94.60	94.80
Without FOHGTN in the second LSTM layer	95.96	95.20	95.40
Without FOVEAN	96.42	95.70	95.90
Without HGTN	96.88	96.20	96.35
Without LVSMT in the third LSTM layer	97.34	96.70	96.90
Without LBT-Net	97.62	96.90	97.10
Without ViT	97.95	97.20	97.40
Without an SVM classifier	98.12	97.40	97.60
Full NFOVSTN-LSTM	98.70	97.80	98.40

CGLEN, Co-occurrence Gray-Level Edge Network; FCEM, Frangi Contrast Enhancement Module; FOHGTN, Fourier-Otsu Hough Graph Transformer Network; FOVEAN, Fourier-Otsu Vascular Enhanced Attention Network; HGTN, Hough-Graph Transformer Network; LBT-Net, Local Binary Transformer Network; LSTM, long short-term memory; LVSMT, Local Vision Support Machine Transformer; NFOVSTN-LSTM, Neurovascular Fourier-Otsu Vision Support Transformer Network with Long Short-Term Memory; NRDN, Neurovascular Remodeling Detection Network; SVM, support vector machine; ViT, vision transformer.

to 92.95%, this highlights its critical role in capturing vascular features. These results confirm that each module contributes significantly and the architectural complexity of NFOVSTN-LSTM is both necessary and effective.

V. Discussion

In this section, to evaluate the reliability of the results and effectiveness of the proposed NFOVSTN-LSTM for MA detection beyond standard classification metrics, it is crucial to evaluate both the consistency of predictions and the spatial accuracy of lesion localization. To achieve this, an analysis is carried out with respect to quantitative performance and statistical significance, highlighting how each architectural component contributes to robust and precise detection. Although MAE and RMSE are traditionally associated with regression tasks, in the proposed framework, NFOVSTN-LSTM quantifies errors in the continuous-valued lesion likelihood maps and confidence scores generated by the network before threshold-based MA classification. The MAE decreases from 1.57 at the initial epoch to 0.012 at epoch 140. While the RMSE reduces from 1.85 to 0.034 by epoch 150, this reflects reduced high-magnitude errors and improved spatial precision of lesion predictions. These reductions demonstrate enhanced spatial consistency, reduced prediction variance, and improved alignment between predicted and ground-truth lesion responses. Consequently, MAE and RMSE serve as complementary indicators of optimization quality and boundary refinement, which supports the robustness

of the proposed framework beyond discrete classification metrics. Next, the performance of the proposed framework is evaluated quantitatively against existing methods to highlight its accuracy, precision, and overall reliability.

Table 2 presents the statistical comparison of the proposed model with baseline approaches in terms of mean performance, standard deviation (SD), and 95% confidence intervals. The proposed model consistently achieves the highest accuracy, precision, and F1-Score with the lowest variability ($\pm$ SD), indicating stable and reliable performance. Moreover, its narrow confidence intervals demonstrate strong statistical confidence and robustness compared to the other models, which exhibit lower mean values and higher dispersion. The proportional benefit of the proposed multi-stage architecture is further illustrated through an ablation study in Table 2. Basically, retinal MAs are extremely small, sparsely distributed, and visually similar to surrounding vascular structures, which limits the effectiveness of conventional end-to-end models that primarily learn global representations. In the proposed framework, NFOVSTN-LSTM is used to explicitly preserve vascular contrast and capture subtle neurovascular remodeling patterns by incorporating task-specific inductive biases through its modular stages, and it is essential for MA detection. All comparative experiments are conducted under identical training data and evaluation protocols, and the ablation results confirm that each stage contributes consistent and non-linear performance gains, particularly in sensitivity and lesion-level detection accuracy. These results cannot be replicated by simpler end-to-end

Table 2: Statistical analysis

Metrics	Model	Mean $\pm$ SD (%)	95% Confidence interval (%)
Accuracy	Proposed Model	98.7 $\pm$ 0.6	[97.52, 99.88]
	VCG-19% Inception-V3	94.0 $\pm$ 1.2	[91.65, 96.35]
	SBF-based MA	83.0 $\pm$ 1.8	[79.47, 86.53]
	DCNN	69.0 $\pm$ 2.3	[64.49, 73.51]
Precision	Proposed Model	97.8 $\pm$ 0.7	[96.43, 99.17]
	VCG-19% Inception-V3	63.3 $\pm$ 2.0	[59.38, 67.22]
	SBF-based MA	40.2 $\pm$ 2.5	[35.30, 45.10]
	DCNN	42.1 $\pm$ 2.1	[37.98, 46.22]
F1-Score	Proposed Model	98.4 $\pm$ 0.5	[97.42, 99.38]
	VCG-19% Inception-V3	62.4 $\pm$ 1.9	[58.68, 66.12]
	SBF-based MA	52.6 $\pm$ 2.2	[48.29, 56.91]
	DCNN	49.5 $\pm$ 2.4	[44.80, 54.20]

SD, standard deviation.

alternatives under the same conditions, thereby justifying the necessity of the multi-stage architecture. Additionally, to mitigate the risk of overfitting arising from the multi-stage architecture, the model was trained using data augmentation, regularization strategies, and cross-validation during evaluation.

Moreover, each module is introduced with a specific functional objective that reduces reliance on a single deep model to learn representations simultaneously. Even though the multi-stage pipeline introduces additional computational overhead compared to conventional end-to-end models, this design enables explicit modeling of vascular structures and components that were executed in a sequential but lightweight manner, ensuring that the overall inference process remains feasible for offline clinical screening scenarios. Moreover, the MA and ischemic classes in the dataset are inherently imbalanced and reflect real-world clinical prevalence. To avoid distributional bias during training and testing the issue was explicitly addressed through class-weighted loss functions that emphasize minority-class learning, along with stratified portioning of the dataset. Additionally, to ensure a reliable assessment of MA detection performance during evaluation imbalance-aware metrics such as recall, F1-Score and lesion level sensitivity were emphasized rather than accuracy. Although ischemia is conventionally confirmed using Fluorescein Angiography or OCT-A, the study does not claim direct ischemia detection from color fundus photography. Instead, ischemia-suspected regions were referenced using FA/OCT-A guided annotations that were spatially mapped onto corresponding CFP images. Consequently, the proposed framework learns fundus-visible correlates of ischemia, such as vascular rarefaction and localized pallor, enabling discrimination from MAs without replacing angiographic validation.

a. Impact of contextual embedding on subtle lesion discrimination

An accurate detection of MAs requires very careful preservation of localized structural characteristics because these lesions typically appear as very small intensity with weak boundary contrast in retinal fundus images. ViT architectures model relationships between image patches, which helps the network to capture range contextual dependencies across retinal images [40]. Although this contextual modeling enhances the understanding of global anatomical structures, the dominance of global token interactions reveals the sensitivity of representations to highly localized pathological patterns. Particularly

subtle lesions such as MA become less distinguishable when global feature aggregation suppresses fine-grained cues that are important for boundary characterization. This limitation has been noted in the transformer-based visual representations framework, where strong contextual modeling sometimes dilutes localized features' specificity when patch-level information is aggregated through a global attention mechanism. Consequently, maintaining a balance between contextual feature integration and localized structural preservation becomes important for reliable lesion discrimination. In this study, the challenges are addressed by retaining localized feature information during the feature extraction and lesion candidate refinement stages. These stages help stabilize representations and preserve spatial variations associated with MAs boundaries. This framework still benefits from the modeling capability of the transformer-based attention mechanism. As a result, the framework achieves improved discrimination of subtle lesion structures without compromising the advantages of global contextual representations.

b. Margin optimization behavior and feature scaling sensitivity in SVM classification

The classification of the proposed framework employs SVM learning to distinguish lesion candidates from non-lesion patterns. SVM classifiers determine class separation by maximizing the margin between support vectors belonging to different classes while mapping feature representations into higher-dimensional spaces using kernel transformations. However, the interpretability of the resulting decision boundaries can be influenced by the sensitivity of the model to hyperparameter selection and feature scaling conditions. Regularization parameters and kernel settings influence the shape of the non-linear decision boundary, and variations in feature scaling may alter the distribution of features within the transformed space. As a result of that, the non-linear decision boundary generated by the classifier appears complex or less interpretable when the margin configuration is highly sensitive to these factors. For this high-sensitivity problem, the proposed framework applies standardized feature normalization and systematic hyperparameter tuning during the training process to reduce the high sensitivity. Overall, the proposed strategy stabilizes the margin configuration and improves the consistency of non-linear decision boundaries used for lesion candidate discrimination [42].

VI. Conclusion

The proposed NFOVSTN-LSTM detected MA efficiently, which is vital for the treatment of DR. The NRDN preprocessed the images by capturing subtle texture variations and by enhancing the structure and contrast. FOHGTN extracted the MA features, suppressed the background clutter, and modeled the spatial and anatomical relationships. Finally, an LVSMST classified MAs from ischemic regions in retinal imaging. The experiments are simulated in Python, and performance evaluation metrics are calculated where the system achieved 98.7% accuracy, 98.5% sensitivity, 99.1% specificity, 97.8% precision, 98.4% F1-score, 0.99 AUC, 0.98 AUPR, 0.012 MAE, 0.034 RMSE, and an average time of 25 s. This improved performance showed the proposed work efficiency in the detection of MA for early screening of DR, which is crucial for timely intervention before irreversible retinal damage occurs. The advanced modeling of neurovascular relationships and perfusion patterns enables the identification of ischemic regions even in the absence of visible biomarkers. The proposed framework assists ophthalmologists in analyzing fundus images, reducing diagnostic workload and human error, and improving clinical throughput.

a. Future scope

The modular nature of the design also supports future extensions such as multimodal imaging integration or temporal progression analysis. Future work will explore multi-class lesion classification and focus on enhancing the interpretability of the proposed framework to better support clinical decision-making. Specifically, explicit visual explanation mechanisms such as lesion probability maps, attention-based heatmaps, and confidence visualization will be integrated to highlight regions contributing to the model's predictions. In addition, incorporating clinician-in-the-loop validation and evaluating the usefulness of these explanations in real-world screening settings will be explored to improve transparency and trustworthiness for practical deployment.

References

[1] S.B. Patil and B.P. Patil, "Automatic detection of microaneurysms in retinal fundus images using modified high boost filtering, line detectors and OC-SVM",

Proc. Int. Conf. Industry 4.0 Technology (I4Tech), pp. 148–153, 2020.

[2] V. Mayya, S. Kamath and U. Kulkarni, "Automated microaneurysms detection for early diagnosis of diabetic retinopathy: A comprehensive review", *Comput. Methods Programs Biomed. Update*, vol. 1, pp. 100013, 2021.

[3] S. Parrulli, F. Corvi, M. Cozzi, D. Monteduro, F. Zicarelli and G. Staurengi, "Microaneurysms visualisation using five different optical coherence tomography angiography devices compared to fluorescein angiography", *Br. J. Ophthalmol.*, vol. 105, no. 4, pp. 526–530, 2021.

[4] M. Tavakoli, A. Mehdizadeh, A. Aghayan, R.P. Shahri, T. Ellis and J. Dehmeshki, "Automated microaneurysms detection in retinal images using radon transform and supervised learning: Application to mass screening of diabetic retinopathy", *IEEE Access*, vol. 9, pp. 67302–67314, 2021.

[5] J. Li, Q. Ma, M. Yao, Q. Jiang, Z. Wang and B. Yan, "Segmentation of retinal microaneurysms in fluorescein fundus angiography images by a novel three-step model", *Front. Med.*, vol. 11, pp. 1372091, 2024.

[6] V. Mayya, S.K. S., U. Kulkarni, D.K. Surya and U.R. Acharya, "An empirical study of preprocessing techniques with convolutional neural networks for accurate detection of chronic ocular diseases using fundus images", *Appl. Intell.*, vol. 53, no. 2, pp. 1548–1566, 2023.

[7] R. Sarki, K. Ahmed, H. Wang, Y. Zhang, J. Ma and K. Wang, "Image preprocessing in classification and identification of diabetic eye diseases", *Data Sci. Eng.*, vol. 6, no. 4, pp. 455–471, 2021.

[8] A.A. Abdulsahib, M.A. Mahmoud, H. Aris, S.S. Gunasekaran and M.A. Mohammed, "An automated image segmentation and useful feature extraction algorithm for retinal blood vessels in fundus images", *Electronics*, vol. 11, no. 9, pp. 1295, 2022.

[9] R. Ramli, K. Hasikin, M.Y.I. Idris, N.K.A. Karim and A.W.A. Wahab, "Fundus image registration technique based on local feature of retinal vessels", *Appl. Sci.*, vol. 11, no. 23, pp. 11201, 2021.

[10] M.E. Hoque and K. Kipli, "Deep learning in retinal image segmentation and feature extraction: A review", *Int. J. Online Biomed. Eng.*, vol. 17, no. 14, pp. 1–18, 2021.

[11] A.E. Ilesanmi, T. Ilesanmi and G.A. Gbotoso, "A systematic review of retinal fundus image segmentation and classification methods using convolutional neural networks", *Healthc. Anal.*, vol. 4, pp. 100261, 2023.

[12] C. Munn, L. Kamintsky, Y. Serlin, A. Winter, A. Cruess, A. Friedman, J. Levy and O. Tomkins-Netzer, "An algorithm for detecting leaky retinal microaneurysms", *Invest. Ophthalmol. Vis. Sci.*, vol. 64, no. 8, pp. 253–253, 2023.

- [13] C. Santos, M. Aguiar, D. Welfer and B. Belloni, "A new approach for detecting fundus lesions using image processing and deep neural network architecture based on YOLO model", *Sensors*, vol. 22, no. 17, pp. 6441, 2022.
- [14] Z. Wang, X. Li, M. Yao, J. Li, Q. Jiang and B. Yan, "A new detection model of microaneurysms based on improved FC-DenseNet", *Sci. Rep.*, vol. 12, no. 1, pp. 950, 2022.
- [15] Á.S. Hervella, J. Rouco, J. Novo and M. Ortega, "Retinal microaneurysms detection using adversarial pre-training with unlabeled multimodal images", *Inf. Fusion*, vol. 79, pp. 146–161, 2022.
- [16] I. Soares, M. Castelo-Branco and A. Pinheiro, "Microaneurysms detection in retinal images using a multi-scale approach", *Biomed. Signal Process. Control*, vol. 79, pp. 104184, 2023.
- [17] M. Mateen, T.S. Malik, S. Hayat, M. Hameed, S. Sun and J. Wen, "Deep learning approach for automatic microaneurysms detection", *Sensors*, vol. 22, no. 2, pp. 542, 2022.
- [18] J. Li, Q. Ma, M. Yao, Q. Jiang, Z. Wang and B. Yan, "Segmentation of retinal microaneurysms in fluorescein fundus angiography images by a novel three-step model", *Front. Med.*, vol. 11, pp. 1372091, 2024.
- [19] B. Zhang, J. Li, Y. Bai, Q. Jiang, B. Yan and Z. Wang, "An improved microaneurysm detection model based on SwinIR and YOLOv8", *Bioengineering*, vol. 10, no. 12, pp. 1405, 2023.
- [20] M. Pundikal and M.S. Holi, "Microaneurysms detection using Grey Wolf Optimizer and modified K-nearest neighbor for early diagnosis of diabetic retinopathy", *Int. J. Intell. Eng. Syst.*, vol. 15, no. 1, pp. 1–10, 2022.
- [21] A. Jabbar, S. Naseem, J. Li, T. Mahmood, M.K. Jabbar, A. Rehman and T. Saba, "Deep transfer learning-based automated diabetic retinopathy detection using retinal fundus images in remote areas", *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, pp. 135, 2024.
- [22] L. Qiao, Y. Zhu and H. Zhou, "Diabetic retinopathy detection using prognosis of microaneurysm and early diagnosis system based on deep learning algorithms", *IEEE Access*, vol. 8, pp. 104292–104302, 2020.
- [23] U. Bhimavarapu and G. Battineni, "Automatic microaneurysms detection for early diagnosis of diabetic retinopathy using improved discrete particle swarm optimization", *J. Pers. Med.*, vol. 12, no. 2, pp. 317, 2022.
- [24] S. Long, J. Chen, A. Hu, H. Liu, Z. Chen and D. Zheng, "Microaneurysms detection in color fundus images using machine learning based on directional local contrast", *Biomed. Eng. Online*, vol. 19, pp. 1–23, 2020.
- [25] Al-Dossary, H., Rahamathulla, M.P. and Sha, M., 2025. Unleashing Effective Identification of ALS Based on Vowel Phonation: A Deep Learning Approach. *Emerging Science Journal*, 9(4), pp.1962-1989.
- [26] Maselena, A., Huda, M. and Ratanamahatana, C.A., 2024. An explainable deep learning approach for classifying monkeypox disease by leveraging skin lesion image data. *Emerging Science Journal*, 8(5), pp. 1875-1897.
- [27] Xiao, Z., Xing, H., Qu, R., Li, H., Feng, L., Zhao, B. and Wan, Q., 2026. Ensemble Transitive Bidirectional Decoupled Self-Distillation for Time-Series Classification. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.
- [28] Xiao, Z., Wan, Q., Tong, H. and Xing, H., 2025. Attentional Knowledge-Based State-Space Model for Electrocardiogram Signal Classification. *IEEE Transactions on Instrumentation and Measurement*, 74, pp. 1-15.
- [29] Nguyễn, Đ., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science*.
- [30] Y. Cheng, W. Liu, C. Wang, X. Gu, Y. Cheng, S. Wang and J. Peng, "A fast localization and extraction of microaneurysm for early detection of diabetic retinopathy", *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 16400–16405, 2020.
- [31] G. Shi, H. Lu, H. Hui and J. Tian, "Benefit from public unlabeled data: A Frangi filter-based pretraining network for 3D cerebrovascular segmentation", *Med. Image Anal.*, pp. 103442, 2025.
- [32] X. Liu, Z. Zhang, Y. Hao, H. Zhao and Y. Yang, "Optimized OTSU segmentation algorithm-based temperature feature extraction method for infrared images of electrical equipment", *Sensors*, vol. 24, no. 4, pp. 1126, 2024.
- [33] K. Mardani and K. Maghooli, "Enhancing retinal blood vessel segmentation in medical images using combined segmentation modes extracted by DBSCAN and morphological reconstruction", *Biomed. Signal Process. Control*, vol. 69, pp. 102837, 2021.
- [34] Y. Zhu, J. Chen, L. Liang, Z. Kuang, L. Jin and W. Zhang, "Fourier contour embedding for arbitrary-shaped text detection", *Proc. IEEE/CVF CVPR*, pp. 3123–3131, 2021.
- [35] Walter, T., Massin, P., Erginay, A., Ordonez, R., Jeulin, C. and Klein, J.C., 2007. Automatic detection of microaneurysms in color fundus images. *Medical image analysis*, 11(6), pp.555-566.[four]
- [36] H. Li, Y. Ma, H. Bao and Y. Zhang, "Probabilistic Hough transform for rectifying industrial nameplate images: A novel strategy for improved text detection

and precision in difficult environments”, *Appl. Sci.*, vol. 13, no. 7, pp. 4533, 2023.

[37] Z. Hu, Y. Dong, K. Wang and Y. Sun, “Heterogeneous graph transformer”, *Proc. Web Conf.*, pp. 2704–2710, 2020.

[38] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C. and Sun, M., 2020. Graph neural networks: A review of methods and applications. *AI open*, 1, pp.57-81.

[39] H.N. Vu, M.H. Nguyen and C. Pham, “Masked face recognition with convolutional neural networks

and local binary patterns”, *Appl. Intell.*, vol. 52, no. 5, pp. 5497–5512, 2022.

[40] M. Vafaezadeh, H. Behnam and P. Gifani, “Ultrasound image analysis with vision transformers”, *Diagnostics*, vol. 14, no. 5, pp. 542, 2024.39

[41] S. Sun, Z. Cao, D. Liao and R. Lv, “A magnified adaptive feature pyramid network for automatic microaneurysms detection”, *Comput. Biol. Med.*, vol. 139, pp. 105000, 2021.

[42] Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning*, 20(3), pp.273-297.