frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 02 February 2026
DOI 10.3389/frai.2025.1731633

OPEN ACCESS

EDITED BY
Vinitra Swamy,
Swiss Federal Institute of Technology Lausanne, Switzerland

REVIEWED BY
DrBiswadip Basu Mallik,
Institute of Engineering and Management (IEM), India
Jinyuan Wang,
Tsinghua University, China

*CORRESPONDENCE
Sherly Alphonse
✉ sherly.a@vit.ac.in

RECEIVED 30 October 2025
REVISED 22 December 2025
ACCEPTED 25 December 2025
PUBLISHED 02 February 2026

CITATION
Deepsahith KV, Shashank B, Kumar B, Alphonse S, Subburaj B and Subramanian G (2026) Graph-enhanced multimodal fusion of vascular biomarkers and deep features for diabetic retinopathy detection.
Front. Artif. Intell. 8:1731633.
doi: 10.3389/frai.2025.1731633

Graph-enhanced multimodal fusion of vascular biomarkers and deep features for diabetic retinopathy detection

K. V. Deepsahith[1], Basineni Shashank[1], Bangipavan Kumar[1], Sherly Alphonse[1]*, Brindha Subburaj[1] and Girish Subramanian[2]

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India, [2]School of Business Administration, Penn State Harrisburg, Middletown, PA, United States

Diabetic retinopathy (DR) detection can be performed through both deep retinal representations and vascular biomarkers. This proposed work suggests a multimodal framework that combines deep features with vascular descriptors in transformer fusion architecture. Fundus images are preprocessed using CLAHE, Canny edge detection, Top-hat transformation, and U-Net vessel segmentation. Then, the images are passed through a convolutional block attention module (CBAM)-fused enhanced MobileNetV3 backbone for deep spatial feature extraction. In parallel, the segmented vasculature is skeletonized to create a vascular graph, and the descriptors are computed using fractal dimension analysis (FDA), artery-to-vein ratio (AVR), and gray level co-occurrence matrix (GLCM) texture. A graph neural network (GNN) then generates a global topology-aware embedding using all that information. The different modalities are integrated using a transformer-based cross-modal fusion, where the feature vectors from MobileNet and GNN-based vascular embeddings interact using multi-head cross-attention. The fused representation is then given to a Softmax classifier for DR prediction. The model demonstrates superior performance compared to traditional deep learning baselines, achieving an accuracy of 93.8%, a precision of 92.1%, a recall of 92.8%, and an AUC-ROC of 0.96 for the DR prediction in the Messidor-2 dataset. The proposed approach also achieves above 98% accuracy for Eyepacs and APTOS 2019 datasets for DR detection. The findings demonstrate that the proposed system provides a reliable framework compared with the existing state-of-the-art methods.

KEYWORDS

contrast limited adaptive histogram equalization (CLAHE), Convolutional Neural Networks (CNNs), deep learning, MobileNetV3, retinal images

1 Introduction

Diabetic Retinopathy (DR) is a microvascular complication of diabetes and affects the retinal vasculature. This also alters the retinal characteristics, like microaneurysms and hemorrhages, which are important biomarkers for early-stage detection. Therefore, quantifiable retinal features and vascular patterns are observable in fundus images. Artificial intelligence (AI) and deep learning approaches have significantly enhanced automated screening and diagnosis, and also led to accurate DR detection (Aljohani and Aburasain, 2024). The retinal characteristics also work as an effective biomarker for systemic pathologies like hypertension, diabetes, and cardiovascular disease(CVD) (Poplin et al., 2018; Ikram et al., 2006). Based on the retinal vascular alterations, it is possible to forecast these diseases and enable interventions on time. This resulted in incorporating AI

and deep learning algorithms for machine-based analysis that improved the efficacy and accuracy of retinal image-based diagnosis (French et al., 2022). Yet, it is time-consuming, subject to inter-observer variation, and not feasible in large-scale screening.

Despite the improvement in medical technology, traditional diagnostic methods are still mainly invasive, costly, and unavailable to some countries (Chang et al., 2020). Sophisticated medical facilities and professional expertise are needed for established methods such as coronary angiography, echocardiography, and cardiac MRI, thus restricting their applicability. Consequently, researchers have been driven to create new alternatives that exploit low-cost and noninvasive strategies for detecting early disease (Rim et al., 2021).

Deep learning allows for automatic feature extraction and classification, lowering the reliance on manual interpretation. Convolutional Neural Networks (CNNs) have proved enormously successful in detecting and classifying abnormalities on medical images, from tumor detection in radiology to retinal pathology detection in ophthalmology (Kermany et al., 2018).

In this proposed work, to improve classification accuracy, various image preprocessing methods, such as contrast limited adaptive histogram equalization (CLAHE), Canny edge detection, Top-hat transformation, and U-Net for vessel segmentation, are employed. A vascular graph is created from the segmented and then skeletonized images. Then gray-level co-occurrence matrix (GLCM) is used for regional feature extraction. GLCM offers texture-based features that assist in distinguishing normal and abnormal retinal patterns. Further, the fractal dimension analysis (FDA) is integrated to measure vascular complexity and structural abnormalities for the early detection of DR. Artery-to-vein ratio (AVR) is also an important biomarker to indicate DR severity. A graph neural network(GNN) embeds the vascular graph with other feature descriptors like GLCM, FDA, and AVR to create the graph-embedded features.

The segmented images are also given as input to a lightweight CNN model, MobileNetV3, which is optimized for high efficiency and low computational overhead, as the basis for an efficient and scalable automated DR detection. In contrast to traditional CNN models that require heavy computational resources, MobileNetV3 uses depth-wise separable convolutions, which greatly minimize the number of parameters without compromising accuracy. MobileNetV3 is highly suitable for real-time applications like mobile health systems and telemedicine platforms. Components like squeeze-and-excitation(SE), block attention mechanism, and dilated convolutions are also added to enhance it further in the proposed work. The SE attention mechanism recalibrates the feature maps dynamically, and hence the model focuses on critical vascular areas (Tseng et al., 2023). The dilated convolutions enhance the receptive field, which helps the model to detect the fine-grained vascular patterns, which are highly useful for identifying early disease. The transformer-based cross-modal fusion used in the proposed system helps in the fusion of deep features from MobileNetV3 and graph-embedded features. The contributions of the proposed work are listed as follows:

- **Preprocessing techniques:** CLAHE, Canny, and Top-hat transformation help in enhanced visibility of vessels.
- **Vessel segmentation:** U-Net helps in vessel segmentation, thereby boosting overall accuracy while extracting global and local features.
- **Local features extraction:** GLCM, FDA, and AVR calibration help in extracting local features.
- **Global features extraction:** MobileNetV3 and SE block attention mechanism enhance feature selection by dynamically recalibrating channel-wise feature responses, ensuring the model focuses on critical vascular regions such as microaneurysms, vessel narrowing, and tortuosity, which are the key indicators of DR.
- **Dilated convolutions:** Increases receptive field without elevating computational expense, allowing for the identification of fine retinal vascular abnormalities, including subtle vessel deformity and capillary dropout, that are frequently linked to DR prediction.
- **Convolutional block attention(CBAM) Module:** Helps in enhancing vessel structures, suppressing noise.
- **Graph-based embedding:** GNN helps in graph-enhanced feature embedding and also preserves the information about the vascular junctions and branches.
- **Cross-modal fusion:** The deep features from MobileNetV3 and graph-embedded features are fused using a transformer-based cross-modal fusion technique.

In the existing literature, several studies exist that focus on graph-based learning, multimodal fusion, and attention mechanisms for DR detection. But most of the existing models use only feature-level fusion across CNN streams and not physiological structures. Also, the graph-based approaches mostly rely on handcrafted descriptors, without vascular biomarkers. Most of the existing works treat the modalities as independent channels, without any standardized method for cross-modal interactions.

The proposed framework helps in addressing these gaps. (i) This work proposes a vascular biomarker graph in which nodes encode the descriptors, and edges model the anatomical relationships. This representation helps in capturing the disease-relevant dependencies that are not seen in other conventional attention-based fusion models. (ii) A graph-enhanced multimodal fusion module is proposed that uses a relation-aware fusion mechanism. Thus, the model learns complementary interactions between learned deep features and structured biomarker information, which is better than the existing hybrid pipelines. The proposed system also uses vascular biomarkers FDA, and the arteriolar-to-venular ratio (AVR) that captures the earlier microvascular changes due to DR. The transformer-based cross-modal fusion module has better interaction modeling that improves the robustness.

Section 2 discusses other existing works in the literature. Section 3 outlines the methodology, wherein preprocessing improves vascular structures before feeding them into a MobileNetV3-based model with dilated convolutions and SE attention. It also explains the proposed integrated methodology used for DR detection. Section 4 reports experimental results on the different datasets (Herrerot, 2022), providing metrics such as accuracy, precision, recall, and AUC-ROC scores. Section 5 concludes the findings.

2 Related works

(Gulshan et al., 2016) constructed a deep learning model for diagnosing DR based on retinal fundus images. The system had high specificity and sensitivity, demonstrating the viability of CNNs in automating diagnosis. The work of Gulshan et al. emphasizes the benefits of non-invasive imaging methods for high-volume screening. Limitations lie in the need for large annotated datasets and computation for training and deployment. Solutions to these might make it more viable in low-resource settings.

Das and Pumrin (2024) investigated the application of MobileNet in the classification of retinal images to diagnose DR. MobileNet's thin model supports low-cost computation, which is useful for low-resource environments. Data preprocessing methods, such as resizing and augmentation, were demonstrated in the study to significantly enhance model performance. The study, however, did not conduct an exhaustive examination of the effects of varying preprocessing approaches on prediction accuracy, leaving it for future studies to further improve these techniques.

He et al. (2015) presented the ResNet architecture that overcomes the problem of vanishing gradients in deep networks using residual connections. Litjens et al. (2017) used CNN-based architectures as a building block for processing challenging medical images such as retinal scans. Zhang et al. (2019) suggested the use of attention mechanisms with deep learning frameworks. The application of attention mechanisms improves the diagnostic performance and interpretability. Liu et al. (2024) proposed an adversarial learning-based framework for the segmentation that leads to better feature representation and edge detection. The model was good for noisy and complicated datasets. Huang et al. (2023) explored the contrastive learning methods to classify retinal images. It lowers the dependency on expert-annotated examples. Aljohani and Aburasain (2024) suggested a hybrid glaucoma detection system with Random Forest and CNNs (ResNet50, VGG-16) for glaucoma detection. Ting et al. (2017) considered the effects of automated deep learning models on early disease identification, workflow performance, and diagnostic accuracy.

Shipra et al. (2024) used explainable AI (XAI) in medical imaging. The work used Grad-CAM and SHAP values to visualize outputs that also helped the clinicians to understand and believe AI-derived predictions. The incorporation of XAI into CNN models enhanced confidence in automated diagnostic systems. Nonetheless, there were issues raised in terms of balancing explainability and predictive performance, as a few interpretable models had a slightly lower accuracy compared to their black-box variants. Future research should investigate how to improve interpretability without losing classification accuracy, perhaps through hybrid AI-human decision-making systems.

Ronneberger et al. (2015) proposed the U-Net architecture, which has been well used in medical image segmentation, including retinal vessel extraction. The experiment proved that skip connections and upsampling policies of U-Net were better at maintaining spatial details than standard CNNs, leading to better segmentation accuracy. The model's capability of performing well on small datasets was especially useful in medical applications. Nevertheless, the experiment showed a reliance on the quality of the datasets and domain-specific fine-tuning. Table 1 gives a detailed review of some of the existing works.

The transformer architectures have recently enhanced multimodal learning approaches. Shamshad et al. (2023) in their survey have highlighted the ability to model the cross-modal interactions better than CNNs. Zhou et al. (2023) introduced a transformer-based model that processes radiographs, text, and laboratory data using intra- and inter-modal attention, which performed better than image-only pipelines. Warner et al. (2024) examined multimodal machine learning in clinical biomedicine, indicating the fusion and alignment problems that actually motivate for graph-aware and transformer-based models. Dong et al. (2025) proposed a multimodal transformer system that combines fundus images with clinical data for DR diagnosis,

TABLE 1 Summary of the DR Detection methods in literature.

No.	References	Focus	Dataset	Source	Methodology	Findings	Keywords
1	Pratt et al., 2016	DR Classification	EyePACS	Variable quality	CNN + data augmentation	Robust	Retinal images, classification, CNN
2	Gulshan et al., 2016	DR Detection	EyePACS, Messidor-2	High-resolution fundus images	Inception-V3 CNN	High sensitivity	DR, deep learning, screening
3	Voets et al., 2019	Cross-domain DR performance	EyePACS, Messidor	Mixed clinical	Comparative CNN analysis	Performance drop in domain shift	Domain adaptation
4	Lam et al., 2018	DR Lesion Detection	Messidor	Good quality images	Transfer learning (ResNet)	Enhanced microaneurysm detection	Transfer learning
5	Li et al., 2019	DR Grading	DDR Dataset	Clinical dataset	Attention-based CNN	Attention maps	DR grading
6	Fang and Qiao, 2022	Early DR Detection	DIARETDB1	Medium-quality	Hybrid DL + handcrafted features	Improved early lesion detection	Hybrid ML
7	Dixit and Jha, 2025	DR Staging	APTOS, Messidor	High-quality image	EfficientNet classifier	Lightweight model	EfficientNet
8	Keel et al., 2019	DR Screening	Primary dataset	Low and variable real-world images	DL-based clinical screening system	Real-world clinical workflows	Screening system

showing the importance of cross-attention compared to retinal and systemic features to improve the performance.

Haq et al. (2024) reviewed the DR detection models, indicating the vision transformers' good performance. Bhoopalan et al. (2025) proposed a task-optimized vision transformer (TOViT) for DR detection. Mutawa et al. (2024) designed a CNN-based DR staging model with CLAHE and discrete wavelet transform to pre-process the images. Senapati et al. (2024) reviewed CNNs, hybrid, and transformer-based methods, which support the use of transformer-based multimodal fusion for DR detection. The deep learning-based models have displayed potential in classifying retinal disease (Gulshan et al., 2016; Kermany et al., 2018), they tend to be based on large annotated images and are less interpretable (Wang et al., 2020; Litjens et al., 2017). General models such as EfficientNet and ResNet need domain-level fine-tuning (He et al., 2015), while attention-based algorithms are computationally expensive. To overcome such limitations, the proposed method promotes contrast with CLAHE (Zuiderveld, 1994) and obtains clinically meaningful biomarkers, such as AVR (Seidelmann et al., 2016), FDA, and GLCM-based texture features (Haralick et al., 1973), while maintaining physiological relevance.

3 Proposed methodology

The goal of the proposed system is to enable an integrated system for the efficient and interpretable diagnosis of DR detection by analyzing retinal fundus images. The system, illustrated in Figure 1, is designed around deep learning and traditional image processing techniques to capture both macro-level and micro-level features in retinal vasculature. Initially, the system acquires high-resolution retinal images, which are subjected to a series of preprocessing steps aimed at enhancing visual quality and suppressing noise. The use of CLAHE and Canny edge detection helps in improving the vessel contrast and delineation.

After preprocessing, the U-Net creates a segmented image that's passed into an enhanced MobileNetV3 network. To complement the learned representations, handcrafted features are extracted from the same preprocessed and segmented images. These include GLCM descriptors that capture vascular texture properties and FDA, which quantifies the complexity of vessel branching. The AVR is also computed, and a vascular embedding is created using GNN. Then this is fused with deep features from MobileNetV3, using transformer-based cross-modal fusion that enhances the model's interpretability and robustness.

3.1 Dataset

Each of the DR tests in the Messidor-2 dataset consists of two macula-centered eye fundus images, one for each eye. The dataset only contained photos that were macula-centered. There are 874 examinations (1748 pictures) in Messidor-2. The excess black background has been removed from this preprocessed version of the Messidor-2 dataset, which is accessible at Messidor-2. The MESSIDOR-2 DR grades are the source of the DR grades (https://www.kaggle.com/datasets/mariaherrerot/messidor2preprocess).

Blindness detection was separated into groups for training, validation, and testing in the APTOS 2019 dataset. The Asia Pacific Tele-Ophthalmology Society 2019 Blindness Detection (APTOS 2019 BD) collection contains 3662 samples collected from numerous individuals in rural India. The Aravind Eye Hospital in India organized the dataset. The fundus images were collected from a number of locations and conditions over a long period of time. The samples were then analyzed and categorized by a group of trained medical experts using the International Clinical DR Disease Severity Scale (ICDRSS) as a reference. According to the scale system, the APTOS 2019 BD samples are divided into five groups: proliferative DR, mild DR, moderate DR, severe DR, and no DR (https://www.kaggle.com/datasets/mariaherrerot/aptos2019).

The International Clinical Diabetic Retinopathy (ICDR) grading scale, which divides retinal fundus pictures into five DR severity categories, is used in the EyePACS dataset. A healthy retina with no discernible microaneurysms or lesions is represented by class 0 (No DR). Only microaneurysms, which manifest as tiny red spots on the retina, are seen in Class 1 (Mild DR). Microaneurysms are included in Class 2 (moderate DR), which also includes moderate vascular anomalies or other hemorrhages. Intra-retinal microvascular abnormalities (IRMA) and multiple hemorrhages are characteristics of class 3 (severe DR); however, proliferative DR is not present. The most advanced stage, known as Class 4 (Proliferative DR), is characterized by neovascularization and vitreous or preretinal hemorrhages, increasing the risk of visual loss (https://www.kaggle.com/competitions/diabetic-retinopathy-detection). In the experiments, a five-fold cross-validation is used.

To further support vessel segmentation and feature validation, an additional publicly available dataset, the retina blood vessel dataset (Wagih, 2023), is incorporated. These datasets provide a broader spectrum of retinal characteristics and enhance model generalization (https://www.kaggle.com/datasets/abdallahwagih/retina-blood-vessel).

3.2 Preprocessing techniques

Preprocessing improves image quality and emphasizes diagnostically relevant structures. This study employs a series of transformations to highlight blood vessels, reduce image noise, and extract spatial texture information.

3.2.1 Vessel visibility enhancement using CLAHE

CLAHE improves local contrast by equalizing intensity values in small image tiles, avoiding over-enhancement and preserving fine details as in Figure 2. The transformation is computed using:

$$T(x) = \frac{CDF(x) - CDF_{\min}}{M - CDF_{\min}} \times (L - 1) \quad (1)$$

where:

- $CDF(x)$ is the cumulative histogram value at intensity x,
- $CDF_{\min}$ is the minimum histogram value in the tile,
- M is the number of pixels per tile,
- L is the maximum pixel intensity.

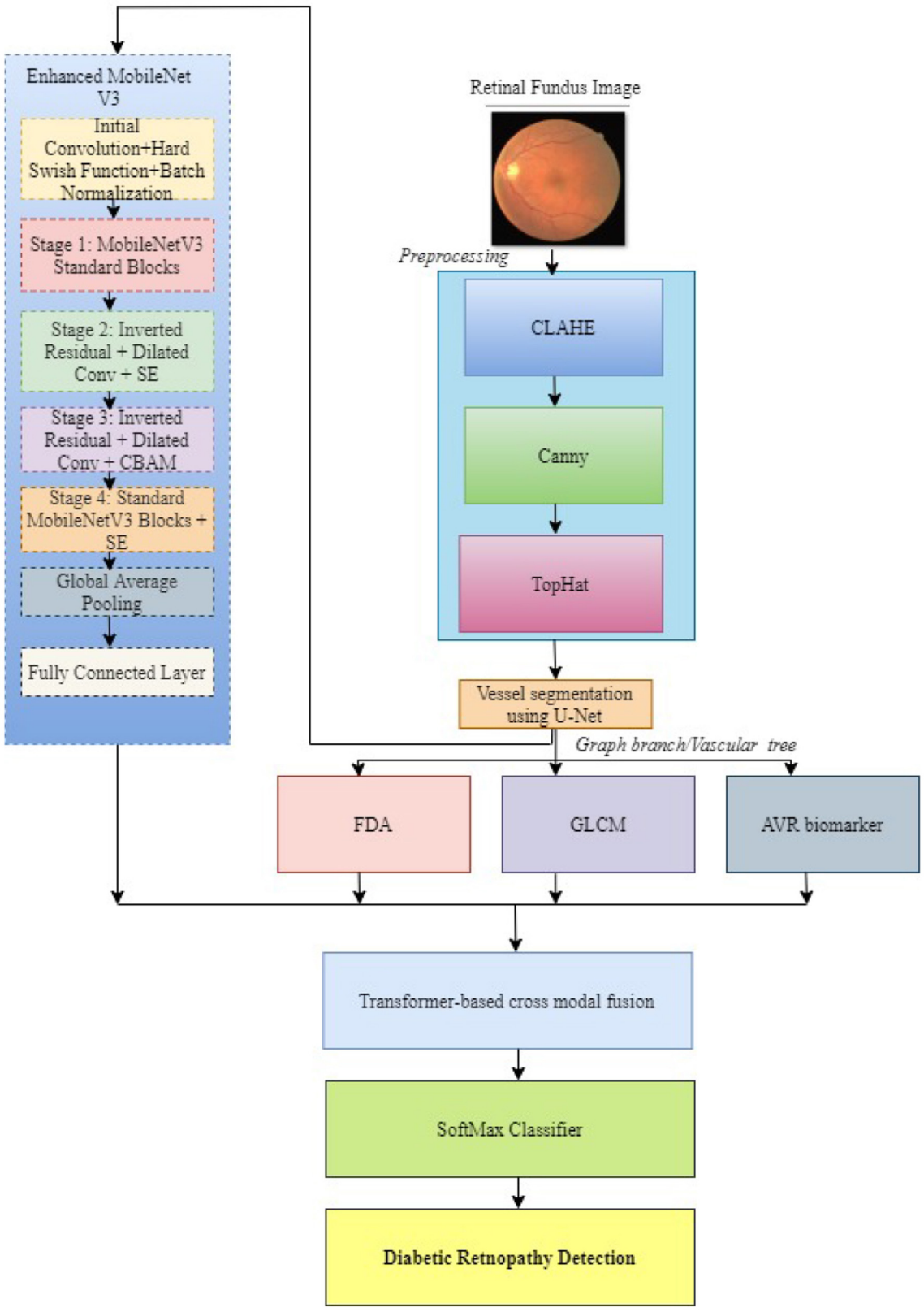

FIGURE 1
Overview of the proposed multi-modal framework with preprocessing pipeline, deep-features, graph-embedded features, and transformer-based cross-modal fusion.

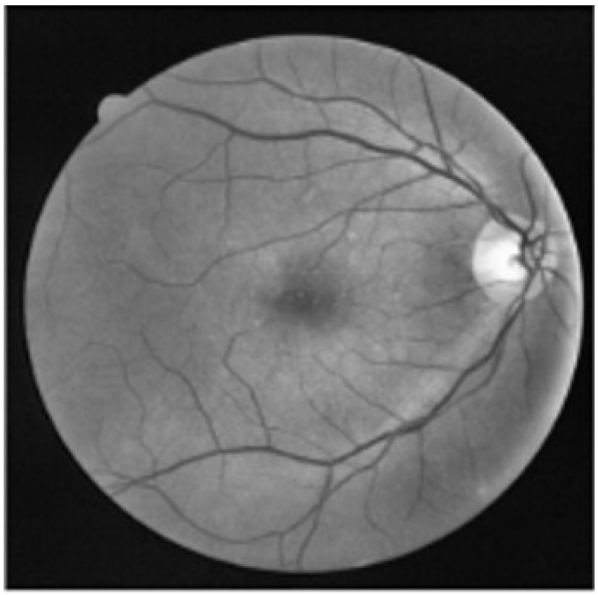

FIGURE 2
Illustration of vessel visibility enhancement in retinal fundus images using CLAHE.

FIGURE 3
Result of Canny edge detection and highlighting vessel boundaries in retinal fundus images.

3.2.2 Highlighting vessel boundaries using Canny algorithm

The Canny algorithm identifies edges by detecting gradients and applying non-maximum suppression. The steps include Gaussian smoothing and gradient estimation as in Figure 3:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2}$$

$$G = \sqrt{G_x^2 + G_y^2} \tag{3}$$

where G_x and G_y are the gradients in the $x-$ and $y-$ directions.

3.2.3 Morphological vessel enhancement (Top-hat transform)

The Top-hat transform isolates small, bright objects such as vessels. The mathematical formula is given as:

$$T_{\text{top-hat}}(I) = I - (I \circ B) \tag{4}$$

where $\circ$ denotes morphological opening.

3.2.4 U-Net for segmentation

The combination of preprocessing techniques leads to:

- Enhanced visibility of fine vascular patterns.
- Suppression of imaging artifacts and irrelevant background.
- Improved feature extraction by the deep learning backbone.

These effects collectively improve the system's diagnostic accuracy, specificity, and generalizability for real-world screening applications in DR detection. Here, the U-Net is used for segmentation purposes before extracting the features using MobileNet and handcrafted features. It also removes the noisy background and improves accuracy and interpretability (Ronneberger et al., 2015). The segmented vasculature is then converted into a skeleton representing the branching topology. Here, the nodes represent the anatomical points, the edges represent the vessel segments, and the attributes are the features. The graph representation helps to preserve the local and global properties.

3.3 Feature extraction blocks

The feature extraction blocks used help in the extraction of both the semantic features and fine-grained statistical cues from retinal images. After preprocessing, the segmented image from the U-Net is given to the MobileNetV3, which helps in capturing vessel tortuosity, branching, and lesion features. An SE attention block with the dilated convolution layer improves the focus on relevant areas within the image. Simultaneously, the segmented image is skeletonized into a vascular graph, and the features are extracted using GLCM and FDA. GLCM extracts the second-order texture information, such as contrast, correlation, and homogeneity, and FDA computes the complexity and self-similarity of the vascular structures. AVR, which is a vital biomarker used in the proposed approach, is also computed. The features are embedded in a graph-based representation using GNN, along with the topology information about the junctions and branches. The deep feature and the graph-embedded features are then fused using

a transform-based cross-modal fusion, which is then passed to a classification head that performs the final prediction.

This proposed approach has both the strength of deep features and handcrafted features that improve the sensitivity even to subtle vascular variations.

3.4 Enhanced MobileNetV3

The enhanced MobileNetV3 extracts the deep features regarding microaneurysms, hemorrhages, exudates, and vessel abnormalities, which are the primary indicators of DR.

3.4.1 Dilated convolution

Pooling is typically performed after a primary convolution operation to reduce dimensionality and strengthen the local features. Pooling also enhances the receptive field, and more global features can be extracted. However, the fine-grained information is lost in the feature maps, which can reduce image recognition accuracy. Without pooling, the receptive field may still be too limited, as it would prevent the extraction of larger spatial relations. With pooling being included, the receptive field of the convolutional kernel is larger, allowing for broader feature extraction. To overcome the disadvantage of pooling, dilated convolution was introduced as shown in Figure 4. This technique modifies the convolution process by introducing gaps (or dilation) among kernel elements, increasing the receptive field without losing the resolution of the feature maps. Unlike pooling, dilated convolution doesn't alter the sizes of input and output feature maps; therefore, no spatial information is lost.

Dilated convolution has several advantages. One, through the addition of a dilation rate, the receptive field is widened without sacrificing resolution, with the relative spatial relation between pixels remaining intact. Two, through the addition of more dilated convolutions with varied rates, multiscale contextual features are obtained. Three, computational cost is relieved because the receptive field is widened without new parameters added (Yu and Koltun, 2015). Algebraically, dilated convolution is written as:

$$z(p,q) = \sum f\left(p + d \cdot h, q + d \cdot j\right) \cdot g(h,j) \tag{5}$$

where:

- p, q are the horizontal and vertical coordinates in the feature map.
- h, j are the coordinates in the convolution kernel.
- f represents the feature map values.
- g represents the convolution kernel values.
- d is the dilation rate, determining the spacing between kernel elements.

3.4.2 Squeeze-and-Excitation block (SE) attention mechanism

The SE block improves the performance of MobileNetV3 by adaptively recalibrating channel-wise feature responses. SE blocks, as in Figure 5, enhance the informative ones while suppressing less relevant channels. This helps in detecting the vascular abnormalities in retinal images, like vessel narrowing, tortuosity, and microaneurysms, better.

The squeeze and excitation step compresses the spatial features (of size $H \times W$) in each channel using Global Average Pooling (GAP) (Jin et al., 2022), resulting in single descriptor per channel:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i,j) \tag{6}$$

Here, $X_c(i,j)$ is the activation at pixel (i,j) in channel c.

The channel-wise descriptors are given to a bottleneck consisting of two fully connected (FC) layers with non-linear

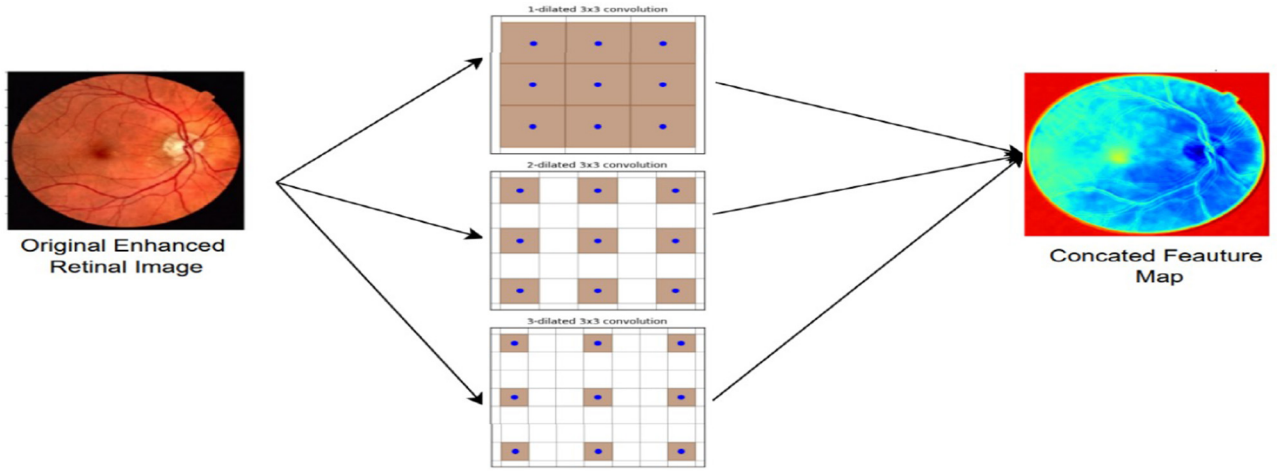

FIGURE 4
Dilated convolution used in MobileNetV3.

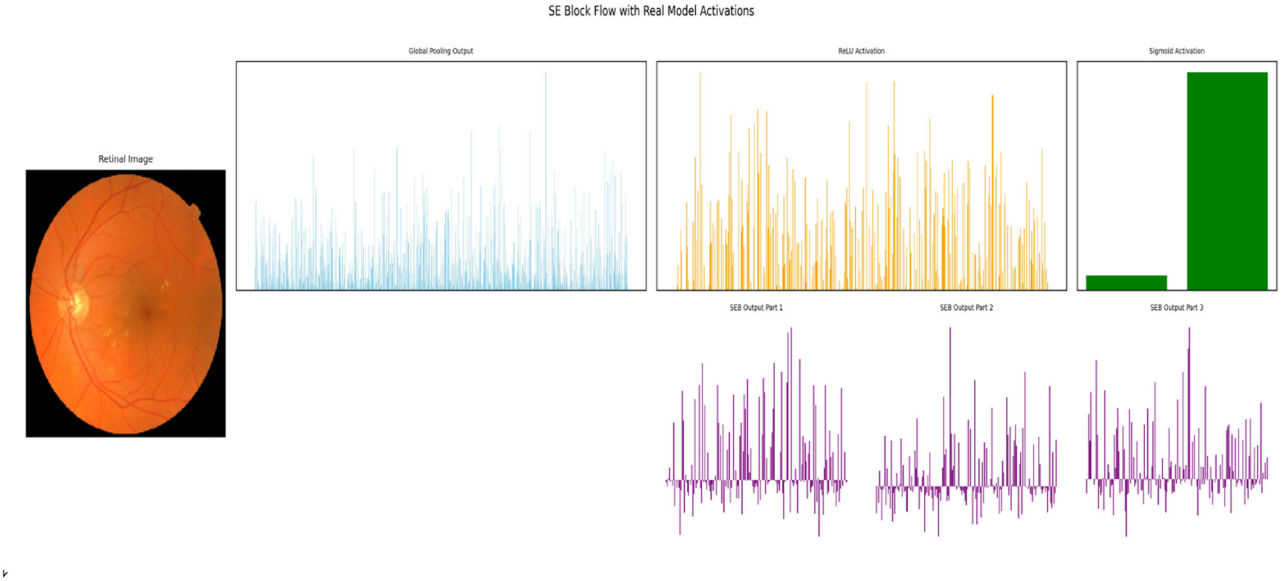

FIGURE 5
Squeeze-and-Excitation Block used in MobileNetV3.

activations (ReLU and sigmoid) that result in a learned attention weight for each channel:

$$s = \sigma(W_2 \cdot \delta(W_1 \cdot z)) \tag{7}$$

Where:

- W_1 and W_2 are the weight matrices,
- δ is the ReLU activation function,
- σ is the sigmoid activation function where output is in the range $[0, 1]$.

The original feature maps are then scaled by the learned weights.

$$\hat{X}_c = s_c \cdot X_c \tag{8}$$

where s_c is the learned attention weight, and X_c is the original feature map.

The dilated convolutions capture the multiscale spatial context without affecting the resolution. The global average pooling used in SE blocks aggregates global channel-wise statistics. This ensures that high-level contextual cues are significantly enhanced without compromising spatial details. This highlights the diagnostic features and suppresses the noisy channels.

The MobileNetV3 used in this framework has dilated convolutions and SE-block attention mechanisms. The use of SE blocks enhances feature selection. The non-linear(NL) functions, such as Hard-Swish (HS) and ReLU activation functions, enhance the efficiency. CBAM helps in focusing better on relevant features.

3.5 Vascular graph construction

The U-Net model segments the vessel structures, and the binary vessel map was obtained using morphological thinning. Bifurcation points and crossovers are identified using connectivity analysis. Each location, based on proper retinal vasculature, is a node, and the edges represent vessel continuity. Artery-vein (A/V) classification is identified using discriminative descriptors, local intensity statistics, and vessel width. GLCM-based texture descriptors are computed along each segment, and a lightweight classifier assists this identification. Each node is encoded with FDA, AVR, and GLCM-derived texture measures. Vessel segments belonging to the same branch are assigned a consistent A/V label, which is later used to augment the node attributes and other features. The graph was then processed using a GNN to create a global embedding summarizing morphology, topology, and descriptors.

The preprocessed retinal fundus image is defined as

$$I : \Omega \subset \mathbb{R}^2 \to \mathbb{R}, \tag{9}$$

where Ω is the retinal image domain. After vessel segmentation, the vessel set is obtained as:

$$S = \{x \in \Omega : \text{vessel}(x) = 1\}. \tag{10}$$

The operator $\mathsf{Skel}(\cdot)$ produces a reduced skeleton structure using:

$$\mathcal{K} = \mathsf{Skel}(S) \subset \Omega, \tag{11}$$

which preserves the vascular topology.

For each skeleton pixel $p \in \mathcal{K}$, 8-connected neighborhood is defined as

$$\mathcal{N}(p) = \{q \in \mathcal{K} : \|p - q\|_\infty = 1\}. \tag{12}$$

Nodes V are indicated as

$$V = \{p \in \mathcal{K} : |\mathcal{N}(p)| = 1 \text{ (endpoints) or } |\mathcal{N}(p)| \geq 3 \text{ (junctions)}\}. \quad (13)$$

The edges E are the maximal simple paths in $\mathcal{K}$ between two nodes, and all intermediate pixels have a degree $|\mathcal{N}(p)| = 2$.

The retinal vasculature is represented as a graph as:

$$\mathcal{G} = (V, E), \quad (14)$$

where the nodes are the endpoints/junctions, and edges are the vessel segments.

Let $A \in \{0,1\}^{|V|\times|V|}$ represent the adjacency matrix with

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E, \\ 0 & \text{otherwise}, \end{cases} \quad (15)$$

and $D = \text{diag}(d_i)$ be the degree matrix with $d_i = \sum_j A_{ij}$. The normalized adjacency is defined as:

$$\tilde{A} = A + I, \qquad \tilde{D} = \text{diag}\left(\sum_j \tilde{A}_{ij}\right), \quad (16)$$

This is used in graph neural network (GNN) processing (Kipf and Welling, 2017).

Also, deep retinal features are extracted using a MobileNet backbone enhanced with CBAM. To integrate the information from deep features and vascular-graph embeddings, a transformer-based cross-modal fusion was used. The MobileNet-CBAM feature vector and the GNN-derived vascular embedding are different modalities, and multi-head cross-attention helps in modeling the interactions. The final representation has both structural vascular biomarkers and appearance-based cues. This final fused feature vector is given to a Softmax classification layer to predict DR severity.

3.6 GLCM and FDA

Gray-Level Co-occurrence Matrix (GLCM) texture descriptors are calculated as:

$$\text{Contrast} = \sum_{i,j}(i-j)^2 G(i,j), \quad (17)$$

$$\text{Correlation} = \frac{\sum_{i,j}(i-\mu_i)(j-\mu_j)G(i,j)}{\sigma_i \sigma_j}, \quad (18)$$

$$\text{Energy} = \sum_{i,j} G(i,j)^2, \quad (19)$$

$$\text{Homogeneity} = \sum_{i,j} \frac{G(i,j)}{1+|i-j|}, \quad (20)$$

where $G(i,j)$ is the normalized GLCM, and μ_i, σ_i are the mean and standard deviation of row i. The RGB histogram is calculated as:

$$H_c(i) = \sum_{x,y} \delta(I_c(x,y) - i), \quad c \in \{R,G,B\} \quad (21)$$

where δ is the Kronecker delta, and $I_c(x,y)$ is the pixel intensity at (x,y) for channel c. GLCM captures the texture patterns regarding microaneurysms and hemorrhages effectively. Here, FDA is a non-invasive biomarker capturing retinal abnormalities. The box-counting technique is used to compute the fractal dimension in FDA as:

$$FD = \lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log(1/\epsilon)}, \quad (22)$$

where $N(\epsilon)$ is the number of boxes of size ϵ used to cover the vessel's structure, regaining the vascular branching's complexity.

3.7 AVR calibration

The vessel caliber is computed using the arteries and veins identified in the zone of 0.5–1.0 optical disc diameters from the disc margin. The central retinal arteriolar equivalent (CRAE) and central retinal venular equivalent (CRVE) are estimated using the Parr-Hubbard formulas:

$$CRAE = \sqrt{(0.87 \cdot D_1^2 + 1.01 \cdot D_2^2)}, \quad (23)$$

$$CRVE = \sqrt{(0.72 \cdot d_1^2 + 0.91 \cdot d_2^2)}, \quad (24)$$

where D_1, D_2 are the highest arteriolar diameters, and d_1, d_2 are the highest venular diameters. A lower AVR reflects narrower arterioles associated with significantly increased risk (Ikram et al., 2006; French et al., 2022). The formula for computing the AVR is given as:

$$AVR = \frac{CRAE}{CRVE}. \quad (25)$$

Canny edge detection enhances the accuracy of delineating vessel boundaries, especially in low-contrast or noisy fundus images, by finding the edges accurately. Hence, a more reliable segmentation of arterioles and venules is possible, which results in accurate CRAE/CRVE calculation (Seidelmann et al., 2016; McGeechan et al., 2008).

The AVR is a crucial retinal biomarker for DR detection. Dilated convolution modules in the architecture expand the receptive field without losing spatial resolution, thereby capturing multiscale vessel structures like fine capillaries and larger branches needed for robust vessel segmentation and AVR estimation. The inclusion of Frangi filters helps identify broken or small arterioles. DR results in lower AVR values, due to venular widening (Islam et al., 2009; Ashraf et al., 2021). The arteriolar narrowing is seen in regions of retinal non-perfusion and increased DR severity. Wider retinal venules predict the progression of DR over time (Liu et al., 2022). AVR is a helpful quantitative indicator of microvascular alterations in DR. Fused with other features, it is more effective (Quellec et al., 2017).

3.8 Graph neural network (GNN) encoder

The vascular graph $\mathcal{G} = (V, E)$ is embedded with regional features, such as GLCM descriptors, and global vascular biomarkers such as AVR, and FDA features (Kipf and Welling, 2017).

Let $X \in \mathbb{R}^{|V| \times d_x}$ denote the node feature matrix, where each node feature vector x_i includes:

$$x_i = \left[c_i,\ g_i^{\text{con}},\ g_i^{\text{ent}},\ t_i\right] \tag{26}$$

with c_i the vessel caliber, $g_i^{\text{con}}, g_i^{\text{ent}}$ GLCM contrast/entropy, and t_i the artery/vein label.

Each GNN layer embeds the node information across vessel connections:

$$H^{(0)} = X, \qquad H^{(\ell+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(\ell)}W^{(\ell)}\right), \tag{27}$$

where $\tilde{A}$ is the adjacency with self-loops, $\tilde{D}$ the degree matrix, $W^{(\ell)}$ trainable parameters, and $\sigma(\cdot)$ a non-linearity.

Along with the local encoding, this model also incorporates the global vascular biomarkers:

$$s = \left[\text{AVR},\ D_{\text{FD}},\ \bar{g}^{\text{con}},\ \bar{g}^{\text{ent}}\right], \tag{28}$$

where AVR is the arteriovenous ratio in the optic disc annulus, D_{FD} the fractal dimension of the vascular tree, and $\bar{g}^{\text{con}}, \bar{g}^{\text{ent}}$ are mean GLCM descriptors computed over the vasculature.

After L GNN layers, the node embeddings $\{h_i^{(L)}\}_{i \in V}$ are pooled to form a graph-level representation:

$$z_{\mathcal{G}} = \rho\left(\{h_i^{(L)} : i \in V\}\right) \ \| \ s,$$

where $\rho(\cdot)$ is an attention pooling, and $\|$ denotes aggregation with the handcrafted global biomarker vector s.

Thus, the final embedding $z_{\mathcal{G}}$ has both the vascular structural information acquired using GNN and clinically interpretable global biomarkers (AVR, FDA, and GLCM).

3.9 Transformer-based cross-modal fusion

To aggregate all the descriptors, a cross-modal fusion is done using a transformer encoder (Lu et al., 2019).

- $z_{\text{img}} \in \mathbb{R}^{d_{\text{img}}}$: fundus image embedding from MobileNet.
- $z_{\mathcal{G}} \in \mathbb{R}^{d_{\mathcal{G}}}$: vascular graph embedding from the GNN encoder.
- $s \in \mathbb{R}^{d_s}$: handcrafted vascular descriptors (AVR, fractal dimension, GLCM features).

All features are projected into a manifold of dimension d:

$$t_{\text{img}} = P_{\text{img}} z_{\text{img}}, \quad t_{\mathcal{G}} = P_{\mathcal{G}} z_{\mathcal{G}}, \quad t_{\text{stat}} = P_{\text{stat}} s, \tag{29}$$

where $P_{\text{img}}, P_{\mathcal{G}}, and P_{\text{stat}}$ are the trainable projection matrices.

The input token sequence is constructed as

$$T_0 = \left[t_{\text{cls}}; t_{\text{img}}; t_{\mathcal{G}}; t_{\text{stat}}\right] \in \mathbb{R}^{4 \times d}, \tag{30}$$

where t_{cls} is a learnable classification token.

Each transformer block uses multi-head self-attention (MHSA) succeeded by feed-forward layers:

$$\text{MHSA}(T) = \text{Concat}_{h=1}^{H}\left(\text{softmax}\left(\frac{Q_h K_h^{\top}}{\sqrt{d_h}}\right)V_h\right)W^O, \tag{31}$$

where $Q_h = TW_h^Q, K_h = TW_h^K, V_h = TW_h^V$.

After B transformer layers, the fused representation is obtained from the classification token:

$$t_{\text{fused}} = t_{\text{cls}}^{(B)}. \tag{32}$$

3.10 Prediction

DR is predicted using the classifier as in :

$$\hat{y} = \text{softmax}\left(W_c\, t_{\text{fused}} + b_c\right). \tag{33}$$

This architecture enables joint reasoning across image-level features, vascular topology, and handcrafted descriptors, improving robustness and interpretability.

4 Experimental results and analysis

The datasets used in the experiments, such as APTOS 2019, EyePACS, and Messidor-2, have different grading protocols, image quality, and image acquisition techniques. Class-balanced augmentation is used to manage the imbalancing problem. Messidor-2, APTOS2019, and EyePACS use different DR grading schemes, and therefore, the labels were standardized to a unified 5-class International Clinical Diabetic Retinopathy (ICDR) scale to ensure consistency. To handle dataset heterogeneity, preprocessing and normalization techniques are applied to three datasets. The preprocessing pipeline, which includes CLAHE, Canny edge detection, and Top-Hat filtering, is applied to all datasets. The parameter values can be adjusted to handle the variations in illumination, resolution, and image quality across datasets. Here, all experiments employ 5-fold cross-validation for all experiments, with patient-level splitting applied for Messidor-2 and EyePACS. All images from a single patient stay in the same folder, which prevents cross-patient data leakage. For APTOS2019, stratified 5-fold image-level splitting is used while maintaining class balance.

4.1 Dataset preprocessing and augmentation

To guarantee high-quality input data, CLAHE was utilized for contrast improvement to highlight fine retinal blood vessel pathology. Canny edge detection was used for accurate vessel segmentation, and morphological Top-hat filtering was employed to improve the measured vessel morphology. GLCM texture features were also used to determine spatial relationships among retinal microstructures. The FDA was utilized to estimate vascular complexity to provide a more quantitative measure of structural pathology.

Table 2 shows a comparison of three retinal image datasets. Messidor-2, being the main dataset used in this work, had the accuracy (93.8%), followed by EyePACS (98.2%) and APTOS 2019 (99.2%).

4.2 Model training and optimization

Training was carried out with categorical cross-entropy loss and Adam optimizer with the initial learning rate of 0.0001, which was reduced step-by-step using the ReduceLROnPlateau scheduler to avoid overfitting. Early stopping criterion tracked validation loss and stopped the training process if performance was satisfactory, avoiding repeated computation for the best convergence.

To test every component's contribution, several test runs were undertaken with and without notable enhancements like SE attention, dilated convolutions, and upgraded preprocessing. How each such component added performance is illuminated by the ablation studies (described in Section 4.5).

4.3 Performance metrics and evaluation

To empirically assess the performance of the aforementioned model, a set of evaluation performance measures was used, such as accuracy, precision, recall, specificity, F1-score, and AUC-ROC. Accuracy is one of the main measures of whether the model classifies DR classes correctly, but classification evaluation cannot be described at all by it, and a set of these measures is more crucial to obtaining the balance of false positives as well as false negatives. The hyperparameters used are given in Table 3.

TABLE 2 Accuracy obtained by the proposed model on different datasets.

Dataset	Classes	Images	Accuracy
Messidor-2	5	1,748	93.8%
EyePACS	5	88,700	98.2%
APTOS 2019	5	3,662	99.2%

TABLE 3 Hyperparameters used in the proposed model.

Parameter	Setting
Batch size	32
Epochs	100
Optimizer	Adam
Initial learning rate	1×10^{-4}
Weight decay	1×10^{-5}
Hidden dimension of transformer	256
Attention heads	4
Adam β_1, β_2	0.9, 0.999
Hardware	NVIDIA RTX 3090 GPU (24 GB), 64 GB RAM
Framework	PyTorch 2.1.0 with CUDA 12.2

Accuracy computes the number of true positives divided by all cases labeled DR, which decreases the number of false-positives as illustrated in Figure 6.

Table 4 presents the performance metrics of the proposed model, achieving 93.8% accuracy, ensuring reliable classification. The recall of 92.8 indicates strong detection of DR cases, while the specificity of 94.2% minimizes false positives. The AUC-ROC of 0.96 highlights its excellent discriminatory power, confirming the model's effectiveness in automated DR detection.

4.3.1 Output visualizations and explainability using grad-CAM heatmap

As depicted in Figure 7, the figure depicts the visualization of outputs after preprocessing, after applying FDA and GLCM.

Gradient-weighted class activation mapping (Grad-CAM) was used to visualize the key retinal regions impacting the predictions to improve the interpretability of the suggested deep learning model. Grad-CAM generates class-discriminative heat maps that highlight the geographical regions that have the most effects on model confidence, providing insights into the convolutional layers' decision-making process.

The model mainly targets areas of high vascular complexity and optic disk boundaries. These highlighted regions are clinically relevant, as microvascular irregularities in these regions are

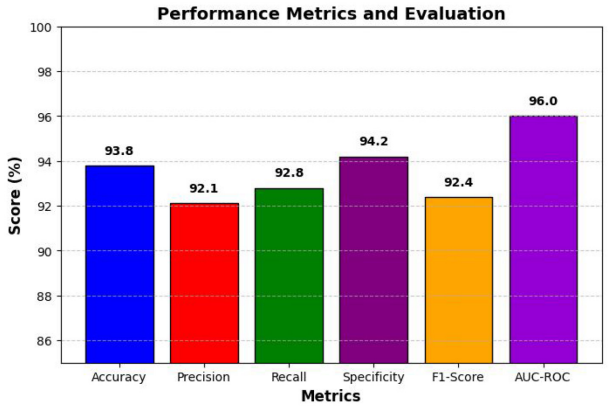

FIGURE 6
Performance analysis of the proposed model using different metrics on Messidor-2 dataset.

TABLE 4 Performance analysis of the proposed model for Messidor-2 using different metrics.

Metric	Value (%)
Accuracy	93.8
Precision	92.1
Recall	92.8
Specificity	94.2
F1-Score	92.4
AUC-ROC	0.96

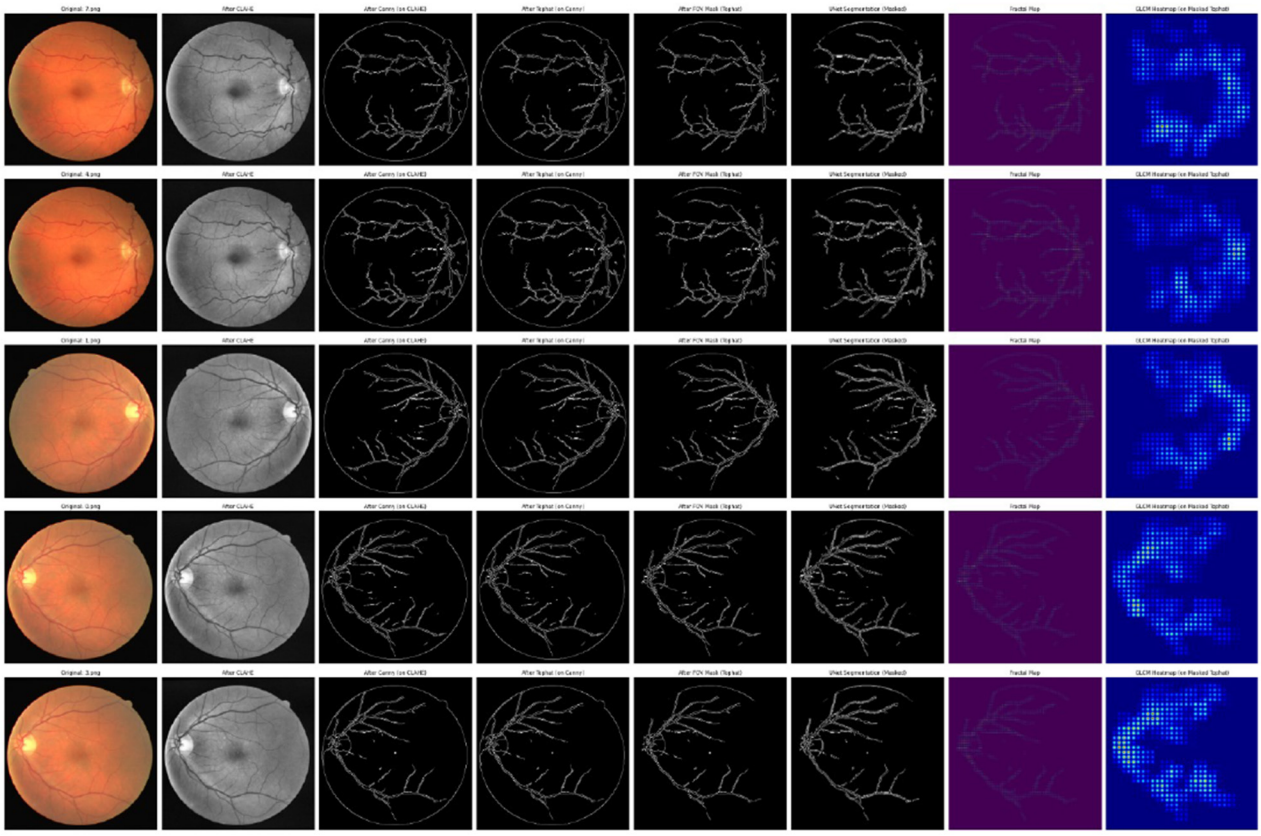

FIGURE 7
Visualization of outputs obtained during preprocessing, FDA, and GLCM.

strongly correlated with DR. The use of Grad-CAM ensures that the model's predictions align with clinically interpretable biomarkers, thereby enhancing trustworthiness for potential integration into real-world diagnostic systems.

DR is an eye disease, and therefore, models are typically interpreted through heat maps. The Grad-CAM heat map provides a pixel-level visualization of the regions that influence the decision of the model. The green, orange, and yellow regions indicate the areas of close attention. The yellow color indicates the areas that strongly contribute to the DR class. These regions coincide with hemorrhages, microaneurysms, exudates, and areas of vascular leakage. In addition, purple/blue represents areas (outer retinal periphery in which fewer lesions are visible) with less influence on the prediction. The color distribution in Figure 8 indicates that the proposed model always attends to the regions rich in lesions, which makes predictions driven by clinically important retinal features. The green zones indicate the boundaries of the vessel, the perivascular regions, and the first lesions.

4.4 Comparative analysis with baseline models

For comparison purposes, the performance of the proposed model was also compared to the existing models in the literature, including MobileNetV3 without SE augmentation, ResNet50, EfficientNet-B0, DenseNet-121, and Vision Transformer. The suggested model surpassed all the rest, with 93.8% accuracy and 0.96 AUC-ROC as illustrated in Figure 9, proving that SE is effective in recalibrating features and dilated convolutions help increase the detection of vessel pathology. Table 5 depicts that the extraction of the vessel segmentation masks by the proposed approach is good when applying it on the retina blood vessel dataset and analyzing the performance before applying it on Messidor-2. The AVR annulus map with arteries and veins identified is given in Figure 10.

4.5 Ablation study

Ablation experiments play an important role in estimating the contribution of different elements in deep learning models. In this work, the impact of the SE-Block attention mechanism, dilated convolutions, and preprocessing techniques on the accuracy of the proposed MobileNetV3-based retinal image classifier for DR detection is thoroughly analyzed. This analysis is achieved by the stepwise addition or elimination of significant components through an ablation study.

- **Baseline Model (MobileNetV3 Only)** There is no fine-grained vessel detection, and as a result has moderate accuracy.

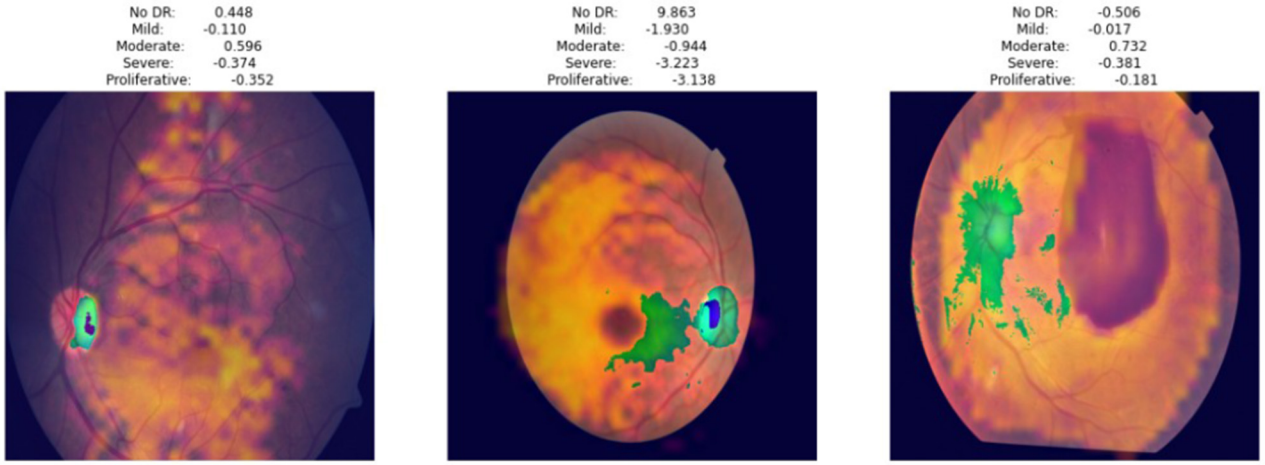

FIGURE 8
Grad-cam heatmaps obtained using the proposed model.

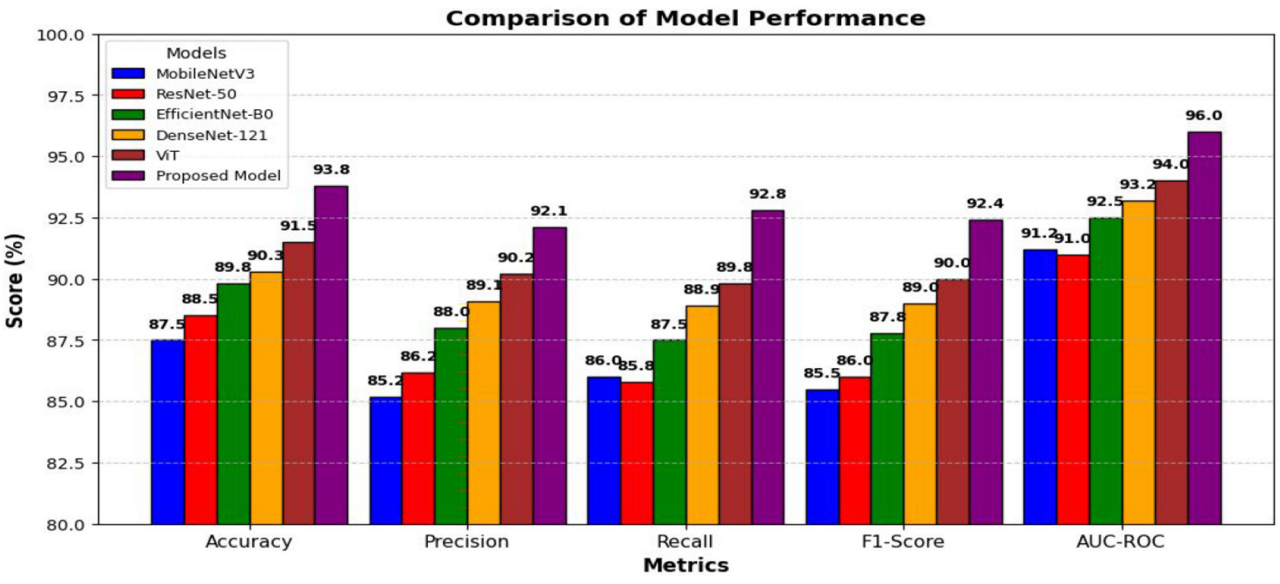

FIGURE 9
Comparative analysis with different existing models for Messidor-2.

- **+ SE-Block Attention** It has improved feature representation and thereby enhances sensitivity.
- **+ SE-Block + Dilated Convolutions** Expands the receptive field, and there is better detection of fine detailed patterns.
- **+ SE-Block + Dilated Convolutions + CBAM** Improves the vessel visibility for mild abnormalities.
- **GLCM** Helps in capturing statistical and structural details.
- **FDA** Helps in measuring the irregularity in retinal structures.
- **AVR** The decreased AVR value helps in flagging the high-risk patients by identifying the DR severity.
- **Proposed full model** Combines all enhancements, achieving the highest accuracy, proving that each added feature significantly contributes to overall performance.

Table 6 shows the results of the ablation experiment. Baseline MobileNetV3-alone model achieves 86.5% accuracy, but has low sensitivity to fine vascular pathology. The SE-Block attention mechanism improves the accuracy to 88.3%. Dilated convolutions boost the accuracy to 88.8% by capturing the fine retinal details. Preprocessing techniques such as CLAHE and edge detection significantly enhance vessel visibility, particularly in mild DR cases, boosting contrast and structural definition. Stepwise performance improvement is evidence of the necessity for combining spatial attention, multiscale feature extraction, and advanced preprocessing techniques to reach peak classification accuracy. AVR boosts the accuracy to 93%. The entire model achieves 93.8% accuracy, indicating the overall effect of feature

extraction and classification improvement. These results verify the necessity of a hybrid domain-specific and deep learning technique for medical image analysis. Also, cross-domain experiments, as in Table 7, are conducted to analyze the effect of domain shift due to variations in illumination, resolutions, and grading. Therefore, the results show the competitive performance across datasets and its ability to perform well in real-world environments. The present work focuses on publicly available datasets, but the methodology can be extended to suit naturally to hospital-based environments, which is a crucial direction for future validation.

The results in Table 7 indicate the generalization capability across datasets that have different imaging characteristics. The results of Messidor-2 show some variation because of domain shift, and the model has high accuracy and AUC even after the transfer to APTOS 2019 and EyePACS, featuring robustness. These results indicate the suitability for deployment even in heterogeneous environments.

TABLE 5 Performance analysis of vessel segmentation masks extraction by CLAHE + Canny + Top-hat + U-Net on retina blood vessel dataset.

Metric	Value
Accuracy (%)	95.85
Sensitivity (%)	92.12
Specificity (%)	90.92
Precision (%)	91.45
F1-Score (%)	93.27
Dice Coefficient	0.873
IoU	0.872
AUC	0.937

A five-fold cross-validation is performed on all three datasets. Table 8 shows the results and their 95% confidence intervals, indicating stability across folds. A paired t-test conducted on the 5 folds resulted in a statistically significant performance improvement when compared to the best baseline vision transformer, as $p < 0.05$. The proposed model also achieves

TABLE 6 Ablation study of the proposed model using Messidor-2 dataset.

Configuration	Acc.	Sens.	Spec.	AUC	F1
MobileNetV3 baseline model	86.5	81.0	91.5	0.910	81.3
+ SE	88.3	83.8	91.1	0.926	82.1
+ CBAM	88.7	84.1	91.4	0.919	82.5
+ Dilated Convolution	88.8	86.3	91.5	0.920	82.6
+ SE + CBAM	91.2	90.6	93.0	0.935	87.1
+ SE + Dilated Convolution	91.0	90.8	91.8	0.933	87.9
+ CBAM + Dilated Convolution	91.4	91.0	92.1	0.937	88.4
+ SE + CBAM + Dilated Convolution	91.8	91.1	92.5	0.941	88.9
(above) + CLAHE	91.2	91.1	93.9	0.935	89.6
+ Canny	91.5	91.4	92.3	0.920	90.2
+ U-Net	91.7	91.5	91.6	0.923	90.6
+ GLCM	92.0	92.1	92.1	0.929	90.6
+ FDA	92.1	92.3	92.6	0.924	91.1
+ AVR	93.0	92.9	92.8	0.937	92.5
Full proposed model	**93.8**	**94.2**	**94.2**	**0.960**	**92.4**

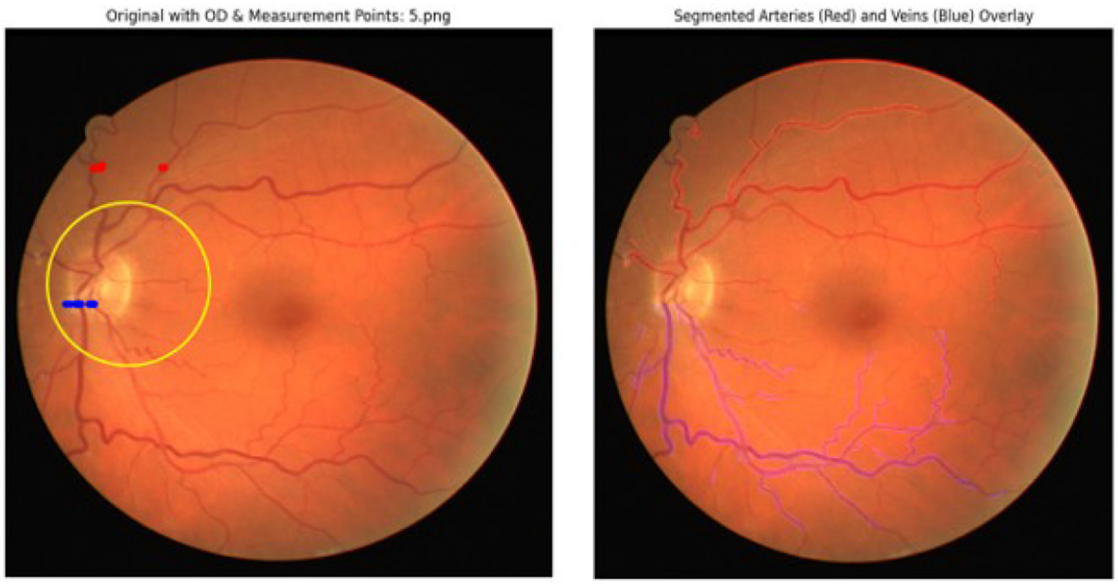

FIGURE 10
Arteries and veins for AVR biomarker.

greater than 98% accuracy on APTOS and EyePACS, with the best performance of 93.8% on Messidor-2, indicating robustness. Also, the baseline models were trained with the same preprocessing pipeline, data splits, and configuration, and the experiments are done and reported in Table 9. It indicates that the proposed work performs better than the standard CNN and transformer-based architectures for all datasets. These results also indicate the advantages of combining deep representations with vascular morphology and other descriptors.

TABLE 7 Cross-domain evaluation of the proposed model.

Training dataset	Testing Dataset	Accuracy (%)	Precision (%)	AUC
Messidor-2	EyePACS	98.2	97.8	0.98
APTOS 2019	Messidor-2	93.5	92.0	0.95
EyePACS	Messidor-2	93.5	92.1	0.95
Messidor-2	APTOS 2019	98.0	97.5	0.98

TABLE 8 Five-fold cross-validation performance of the proposed model.

Dataset	Accuracy (%)	Precision (%)	Recall (%)	AUC
Messidor-2	93.8 ± 0.7	92.1 ± 0.5	92.8 ± 0.6	0.960 ± 0.008
APTOS 2019	99.2 ± 0.5	98.8 ± 0.4	99.0 ± 0.4	0.990 ± 0.004
EyePACS	98.2 ± 0.4	97.3 ± 0.3	97.5 ± 0.4	0.982 ± 0.005

TABLE 9 Comparison with baseline models under identical conditions.

Model	Messidor-2 Acc (%)	APTOS Acc (%)	EyePACS Acc (%)
EfficientNet-B0	92.3	97.2	97.1
ResNet50	92.3	96.54	96.0
ViT-Base	92.7	97.3	97.1
Proposed model	93.8	99.2	98.2

4.6 State-of-the-art

In medical image analysis, deep learning has been a significant advancement, especially in retinal imaging, where it allows for automated evaluation of ocular disease like DR. Because of the Messidor dataset's high-resolution and detailed retinal images, this work has investigated the utilization of retinal fundus photos.

As summarized in Table 10, the proposed approach is compared to the state-of-the-art models using Messidor-2, EyePACS, and APTOS-2019 datasets for DR detection. Performance metrics, such as accuracy, sensitivity, specificity, and AUC, are used for comparison. The proposed method achieves good performance for all the datasets. The existing methods, such as ConvNeXt, EfficientNet, and vision transformer variants, are used for the comparison. The proposed approach achieves the best performance when compared to other existing works. There will be challenges due to poor illumination, demographic bias, and the presence of artifacts. In the proposed work, CLAHE eliminates poor illumination by improving the local contrast. Canny + Top-hat suppresses artifacts and highlights the vessel and lesions. GLCM and FDA quantify vascular complexity and are robust to noise. MobileNetV3 also learns discriminative features, eliminating demographic/device bias while enhancing generalization. AVR helps in normalizing vessel caliber, eliminating the demographic bias due to age, sex, and ethnicity. SE and CBAM adaptively re-weight spatial regions, eliminating the artifacts and only focusing on lesions. Dilated convolutions magnify the receptive field, maintaining good resolution, thus helping MobileNetV3 to capture information under poor illumination and varying image quality.

4.7 Limitations

There is a relatively high computational cost involved both during training and inference. This will slightly affect the deployment on low-resource systems or edge devices without hardware acceleration. Also, scalability requires more optimization strategies such as model compression. The proposed work also needs further evaluation on multi-center and handheld screening devices to verify its deployment to real-world scenarios. The

TABLE 10 Comparison with the state-of-the-art models for DR detection.

Dataset	Model used	Acc. (%)	Sens. (%)	Spec. (%)	AUC	References
Messidor-2	**Proposed**	93.8	94.2	94.2	0.96	This work
	DR-ConvNeXt	83.6	74.0	94.6	—	Song and Wu, 2025
	DRStageNet	—	—	—	0.96	Men et al., 2023
	Swin Transformer var.	—	—	—	0.95	Yao et al., 2022; Saadna et al., 2025
EyePACS	**Proposed**	98.2	98.1	98.2	0.98	This work
	EfficientNet	—	—	—	0.90	Chetoui and Akhloufi, 2020; Yi et al., 2021
	ViT / Swin	—	—	—	0.98	Huang et al., 2024; Yang et al., 2024
APTOS 2019	**Proposed**	99.2	99.1	99.3	0.99	This work
	GPMKLE-Net	—	—	—	0.98	Zhou et al., 2023
	ConvNeXt	—	—	—	0.90	Song and Wu, 2025; Nadeem et al., 2022

experiments on hospital-based data will also be done as future work, as it requires some more steps regarding domain adaptation.

5 Conclusion

This study suggests a novel and effective deep learning framework for DR prediction from retinal fundus images. The proposed architecture gathers local features from retinal fundus images using MobileNetV3, incorporating SE attention blocks and dilated convolutions to better capture fine-grained vascular features indicative of ocular disease such as DR prediction. Through comprehensive experiments and ablation studies, it is demonstrated that the inclusion of preprocessing techniques such as CLAHE-based contrast enhancement, Canny edge detection, and Top-hat transformation and segmentation using U-Net improves the performance. Also, the regional features captured using GLCM, the global biomarker features captured using AVR, and the FDA contribute significantly to improving model sensitivity, specificity, and overall robustness. The features are embedded in a graph-based representation using GNN that preserves vascular topology. The transformer-based cross-modal fusion integrates the multi-modal features so effectively. The model achieved an AUC-ROC of 0.96 on the Messidor dataset—outperforming conventional risk scoring systems and previously published deep learning benchmarks. Moreover, the model ensures feasibility for real-time screening in both hospital and remote settings. The AVR biomarker individually helps in DR detection after being fused with MobileNet, GLCM, and FDA features.

In future, it is aimed to expand the model's utility through multi-modal learning by integrating retinal image data with electronic health records, demographic information, and lifestyle factors to improve DR detection. Additionally, prospective validation in real-world clinical environments will be explored in collaboration with healthcare institutions to assess its diagnostic impact, usability, and integration into clinical workflows.

Data availability statement

Publicly available datasets were analyzed in this study. The datasets are available at the following links: https://www.kaggle.com/datasets/mariaherrerot/messidor2preprocess, https://www.kaggle.com/datasets/mariaherrerot/aptos2019, and https://www.kaggle.com/competitions/diabetic-retinopathy-detection.

Author contributions

KD: Supervision, Writing – review & editing. BS: Conceptualization, Supervision, Writing – review & editing. BK: Supervision, Writing – review & editing. SA: Formal analysis, Visualization, Writing – original draft. BS: Conceptualization, Supervision, Writing – review & editing. GS: Supervision, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. The APC charge was supported by Vellore Institute of Technology, Chennai.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Aljohani, A., and Aburasain, R. Y. (2024). A hybrid framework for glaucoma detection through federated machine learning and deep learning models. *BMC Med. Inform. Decis. Mak.* 24:115. doi: 10.1186/s12911-024-02518-y

Ashraf, M., Shokrollahi, S., Pisig, A. U., Sampani, K., Abdelal, O., Cavallerano, J. D., et al. (2021). Retinal vascular caliber association with nonperfusion and diabetic retinopathy severity depends on vascular caliber measurement location. *Ophthalmol. Retina* 5, 571–579. doi: 10.1016/j.oret.2020.09.003

Bhoopalan, R., Sekar, P., Nagaprasad, N., Mamo, T. R., and Krishnaraj, R. (2025). Task-optimized vision transformer for diabetic retinopathy detection and classification in resource-constrained early diagnosis settings. *Sci. Rep.* 15:39047. doi: 10.1038/s41598-025-25399-1

Chang, J., Ko, A., Park, S. M., Choi, S., Kim, K., Kim, S. M., et al. (2020). Association of cardiovascular mortality and deep learning-funduscopic atherosclerosis score derived from retinal fundus images. *Am. J. Ophthalmol.* 217, 121–130. doi: 10.1016/j.ajo.2020.03.027

Chetoui, M., and Akhloufi, M. A. (2020). "Explainable diabetic retinopathy using EfficientNET," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (IEEE), 1966–1969. doi: 10.1109/EMBC44109.2020.9175664

Das, P. K., and Pumrin, S. (2024). Diabetic retinopathy classification: performance evaluation of pre-trained lightweight cnn using imbalance dataset. *Eng. J.* 28, 13–25. doi: 10.4186/ej.2024.28.7.13

Dixit, R. B., and Jha, C. K. (2025). Fundus image based diabetic retinopathy detection using EfficientNetB3 with squeeze and excitation block. *Med. Eng. Phys.* 104350. doi: 10.1016/j.medengphy.2025.104350

Dong, Z., Wang, X., Pan, S., Weng, T., Chen, X., Jiang, S., et al. (2025). A multimodal transformer system for non-invasive diabetic nephropathy diagnosis via retinal imaging. *NPJ Digit. Med*. 8:50. doi: 10.1038/s41746-024-01393-1

Fang, L. and Qiao, H., 2022. Diabetic retinopathy classification using a novel DAG network based on multi-feature of fundus images. *Biomed. Signal Process. Control.* 77:103810. doi: 10.1016/j.bspc.2022.103810

French, C., Cubbidge, R. P., and Heitmar, R. (2022). The application of arterio-venous ratio (AVR) cut-off values in clinic to stratify cardiovascular risk in patients. *Ophthalmic Physiol. Optics* 42, 666–674. doi: 10.1111/opo.12967

Gulshan, V., Peng, L., Coram, M., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316, 2402–2410. doi: 10.1001/jama.2016.17216

Haq, N., Waheed, M., and Others (2024). Computationally efficient deep learning models for diabetic retinopathy detection: a systematic review. *Artif. Intellig. Rev.* 57, 1–34. doi: 10.1007/s10462-024-10942-9

Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* 3, 610–621. doi: 10.1109/TSMC.1973.4309314

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv* preprint arXiv:1512.03385. doi: 10.1109/CVPR.2016.90

Herrerot, M. (2022). *Messidor-2 Preprocessed Dataset*.

Huang, S.-C., Pareek, A., Jensen, M., Lungren, M. P., Yeung, S., and Chaudhari, A. S. (2023). Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digit. Med.* 6:74. doi: 10.1038/s41746-023-00811-0

Huang, Y., Lyu, J., Cheng, P., Tam, R., and Tang, X. (2024). Ssit: Saliency-guided self-supervised image transformer for diabetic retinopathy grading. *IEEE J. Biomed. Health Inform.* 28, 2806–2817.

Ikram, M. K., de Jong, F. J., Bos, M. J., Vingerling, J. R., Hofman, A., Koudstaal, P. J., et al. (2006). Retinal vessel diameters and risk of stroke: the rotterdam study. *Neurology* 66, 1339–1343. doi: 10.1212/01.wnl.0000210533.24338.ea

Islam, F. M. A., Nguyen, T. T., Wang, J. J., Tai, E. S., Shankar, A., Saw, S. M., et al. (2009). Quantitative retinal vascular calibre changes in diabetes and retinopathy: the Singapore Malay eye study. *Eye* 23, 1719–1724. doi: 10.1038/eye.2008.362

Jin, X., Xie, Y., Wei, X.-S., Zhao, B.-R., Chen, Z.-M., and Tan, X. (2022). Delving deep into spatial pooling for squeeze-and-excitation networks. *Pattern Recognit.* 121:108159. doi: 10.1016/j.patcog.2021.108159

Keel, S., Lee, P., Scheetz, J., and He, M. (2019). Visualizing deep learning models for the detection of referable diabetic retinopathy and glaucoma. *JAMA Ophthalmol.* 137, 288–292.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., et al. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 172, 1122–1131. doi: 10.1016/j.cell.2018.02.010

Kipf, T. N., and Welling, M. (2017). "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)* (Toulon: ICLR Conference/OpenReview).

Lam, C., Yi, D., Guo, M., and Lindsey, T. (2018). "Automated detection of diabetic retinopathy using deep learning," in *AMIA Summits on Translational Science Proceedings*, 147.

Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., and Kang, H. (2019). Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Inf. Sci.* 501, 511–522. doi: 10.1016/j.ins.2019.06.011

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88. doi: 10.1016/j.media.2017.07.005

Liu, J., Zhao, J., Xiao, J., Zhao, G., Xu, P., Yang, Y., et al. (2024). Unsupervised domain adaptation multi-level adversarial learning-based crossing-domain retinal vessel segmentation. *Comput. Biol. Med.* 178:108759. doi: 10.1016/j.compbiomed.2024.108759

Liu, T., Lin, W., Shi, G., Wang, W., Feng, M., Xie, X., et al. (2022). Retinal and choroidal vascular perfusion and thickness measurement in diabetic retinopathy patients by the swept-source optical coherence tomography angiography. *Front. Med.* 9:786708. doi: 10.3389/fmed.2022.786708

Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems (NeurIPS), Volume 32* (Red Hook: Curran Associates, Inc.).

McGeechan, K., Liew, G., Macaskill, P., Irwig, L., Klein, R., Klein, B., et al. (2008). Meta-analysis: retinal vessel caliber and risk for coronary heart disease. *Ann. Intern. Med.* 149, 404–413. doi: 10.7326/0003-4819-151-6-200909150-00005

Men, Y., Fhima, J., Celi, L. A., Ribeiro, L. Z., Nakayama, L. F., and Behar, J. A. (2023). DRStageNet: deep learning for diabetic retinopathy staging from fundus images. *arXiv* [preprint] arXiv:2312.14891. doi: 10.48550/arXiv.2312.14891

Mutawa, A. M., Al-Sabti, K., Raizada, S., and Sruthi, S. (2024). A deep learning model for detecting diabetic retinopathy stages with discrete wavelet transform. *Appl. Sci.* 14:4428. doi: 10.3390/app14114428

Nadeem, M. W., Goh, H. G., Hussain, M., Liew, S. Y., Andonovic, I., and Khan, M. A. (2022). Deep learning for diabetic retinopathy analysis: a review, research challenges, and future directions. *Sensors* 22:6780. doi: 10.3390/s22186780

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., et al. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* 2, 158–164. doi: 10.1038/s41551-018-0195-0

Pratt, H., Coenen, F., Broadbent, D. M., Harding, S. P., and Zheng, Y. (2016). "Convolutional neural networks for diabetic retinopathy," in *Procedia Computer Science, 20th Conference on Medical Image Understanding and Analysis (MIUA 2016), Vol. 90*, 200–205. doi: 10.1016/j.procs.2016.07.014

Quellec, G., Charriere, K., Boudi, Y., Cochener, B., and Lamard, M. (2017). Deep image mining for diabetic retinopathy screening. *Med. Image Anal.* 39, 178–193. doi: 10.1016/j.media.2017.04.012

Rim, T. H., Lee, C. J., Tham, Y. C., Cheung, N., Yu, M., Lee, G., et al. (2021). Deep-learning-based cardiovascular risk stratification using coronary artery calcium scores predicted from retinal photographs. *Lancet Digit. Health* 3, e306–e316. doi: 10.1016/S2589-7500(21)00043-1

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Cham: Springer International Publishing), 234–241.

Saadna, Y., Mezzoudj, S. and Khelifa, M., 2025. Efficient transformer architectures for diabetic retinopathy classification from fundus images: DR-MobileViT, DR-EfficientFormer, and DR-SwinTiny. *Informatica* 49.

Seidelmann, S. B., Claggett, B., Bravo, P. E., Gupta, A., Farhad, H., Klein, B. E., et al. (2016). Retinal vessel calibers in predicting long-term cardiovascular outcomes: the atherosclerosis risk in communities study. *Circulation* 134, 1328–1338. doi: 10.1161/CIRCULATIONAHA.116.023425

Senapati, S., Tripathy, H. K., Sharma, V., and Gandomi, A. H. (2024). Artificial intelligence for diabetic retinopathy detection: a systematic review. *Inform. Med. Unlocked*. 45:101445. doi: 10.1016/j.imu.2024.101445

Shamshad, F., Khan, S. H., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., et al. (2023). Transformers in medical imaging: a survey. *Med. Image Anal.* 88:102802. doi: 10.1016/j.media.2023.102802

Shipra, H. E., and Rahman, M. S. (2024). "An explainable artificial intelligence strategy for transparent deep learning in the classification of eye diseases," in *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)* (Cox's Bazar: IEEE), 1–6.

Song, P., and Wu, Y. (2025). DR-ConvNeXt: DR classification method for reconstructing ConvNeXt model structure. *J. X-ray Sci. Technol.* 33, 448–460.

Ting, D. S. W., Cheung, C. Y.-L., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., et al. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 318, 2211–2223. doi: 10.1001/jama.2017.18152

Tseng, R. M. W. W., Rim, T. H., Shantsila, E., Yi, J. K., Park, S., Kim, S. S., et al. (2023). Validation of a deep-learning-based retinal biomarker (reti-CVD) in the prediction of cardiovascular disease: data from uk biobank. *BMC Med.* 21:28. doi: 10.1186/s12916-022-02684-8

Voets, M., Møllersen, K., and Bongo, L. A. (2019). Replication study: development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PLoS ONE* 14:e0217541. doi: 10.1371/journal.pone.0217541

Wagih, A. (2023). *Retina Blood Vessel Dataset*.

Wang, J., Chen, Y., Li, W., Kong, W., He, Y., Jiang, C., et al. (2020). "Domain adaptation model for retinopathy detection from cross-domain oct images," in *Proceedings of the Third Conference on Medical Imaging with Deep Learning (MIDL)*, eds. T. Arbel, I. B. Ayed, M. de Bruijne, M. Descoteaux, H. Lombaert, and C. Pal (New York: PMLR), 795–810.

Warner, A., Lee, J., Hsu, W., Syeda-Mahmood, T., Kahn, C. E., Gevaert, O., et al. (2024). Multimodal machine learning in image-based and clinical biomedicine: survey and prospects. *Int. J. Comp. Vision*. 132, 3753–3769. doi: 10.1007/s11263-024-02032-8

Yang, Y., Cai, Z., Qiu, S., and Xu, P. (2024). Vision transformer with masked autoencoders for referable diabetic retinopathy classification based on large-size retina image. *PLoS One* 19: e0299265. doi: 10.1371/journal.pone.0299265

Yao, Z., Yuan, Y., Shi, Z., Mao, W., Zhu, G., Zhang, G., et al. (2022). FunSwin: A deep learning method to analysis diabetic retinopathy grade and macular edema risk based on fundus images. *Front. Physiol.* 13:961386. doi: 10.3389/fphys.2022.961386

Yi, S. L., Yang, X. L., Wang, T. W., She, F. R., Xiong, X., and He, J. F. (2021). Diabetic retinopathy diagnosis based on RA-EfficientNet. *Appl. Sci.* 11:11035. doi: 10.3390/app112211035

Yu, F., and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv* preprint arXiv:1511.07122. doi: 10.48550/arXiv.1511.07122

Zhang, S., Fu, H., Yan, Y., Zhang, Y., Wu, Q., Yang, M., et al. (2019). "Attention guided network for retinal image segmentation," in *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Cham: Springer Science and Business Media Deutschland GmbH).

Zhou, H.-Y., Yu, Y., Wang, C., Zhang, S., Gao, Y., Pan, J., et al. (2023). A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nat. Biomed. Eng.* 7, 743–755. doi: 10.1038/s41551-023-01045-x

Zhou, Q., Guo, Y., Liu, W., Liu, Y., and Lin, Y. (2025). Enhancing pathological feature discrimination in diabetic retinopathy multi-classification with self-paced progressive multi-scale training. *Sci. Rep.* 15:25705. doi: 10.1038/s41598-025-07050-1

Zuiderveld, K. (1994). *Contrast Limited Adaptive Histogram Equalization*. London: Academic Press.