International Journal of Computer Applications (0975 - 8887)
Volume 187 - No.71, January 2026

Local–Global Feature Fusion using CNN and Vision Transformer with Ensemble Post-Classification for Diabetic Retinopathy Diagnosis

Padmashree G.
Department of Computer Science and Engineering,
A J Institute of Engineering and Technology,
NH-66, Kottara Chowki,
Mangalore, Karnataka, India

Murali G. Rao
Department of Forensic Medicine,
PGIMER (Post Graduate Institute of
Medical Education and Research),
Chandigarh, India

ABSTRACT

Diabetic retinopathy is a leading cause of vision impairment globally, necessitating timely and accurate diagnosis to prevent irreversible damage. This paper proposes a novel hybrid deep learning framework that combines local and global feature representations for robust DR classification from retinal fundus images. Local features are extracted using a convolutional neural network branch that captures fine-grained pathological patterns such as microaneurysms and hemorrhages. Simultaneously, global contextual features are learned through a Vision Transformer, which models long-range dependencies across the retinal image. The extracted features from both branches are fused and passed through a series of dense layers for initial classification. To further enhance generalization and interpretability, features from the Global Average Pooling layer are used to train a Random Forest classifier. The proposed methodology is evaluated on a benchmark DR dataset with five severity classes. Extensive experiments and ablation studies demonstrate the effectiveness of our architecture in capturing both fine-grained and holistic features, leading to improved classification performance. Our results suggest that the fusion of local and global features, combined with ensemble post-classification, can provide a robust and scalable solution for automated DR diagnosis.

General Terms

Deep learning, Diabetic Retinopathy

Keywords

Diabetic Retinopathy; Convolutional Neural Network; Vision Transformer; Feature Fusion; Random Forest; Medical Image Classification

1. INTRODUCTION

Diabetic retinopathy (DR) is a common and serious microvascular complication of diabetes mellitus, recognized as a leading cause of vision loss and blindness among working-age adults worldwide. Approximately 30–36% of individuals with diabetes are affected by DR, with prevalence rates varying by region and population characteristics [17][18]. In China, for example, the prevalence of diabetic retinopathy among adults with diabetes is estimated at 16.3%, with vision-threatening forms affecting 3.2% of this population [5]. Among children with type 2 diabetes, the prevalence is lower but increases significantly with disease duration, highlighting the importance of early and regular screening [2]. DR develops as a result of chronic hyperglycemia, which leads to damage of the retinal blood vessels, neurodegeneration, and chronic inflammation. If left untreated, DR can progress to proliferative diabetic retinopathy, which is associated with a high risk of irreversible blindness [18][8]. The risk of developing DR is influenced by factors such as poor glycemic control, hypertension, duration of diabetes, and socioeconomic status. [5][17][4] The global burden of DR is expected to rise in the coming decades, particularly in low- and middle-income countries, due to increasing diabetes prevalence and limited access to effective screening and treatment [16]. Early detection and timely intervention are critical to prevent vision loss, and recent advances in imaging, artificial intelligence, and personalized medicine are poised to transform DR management in the near future[16][3].

Convolutional neural networks (CNNs) have been extensively applied to DR classification due to their ability to learn hierarchical spatial features. However, CNNs primarily focus on local receptive fields and may fail to capture global contextual information that is essential for distinguishing between closely related DR severity levels. On the other hand, Vision Transformers (ViTs) have recently emerged as powerful architectures capable of modeling long-range dependencies using self-attention mechanisms, thus providing a global view of the image. Nonetheless, ViTs often require large datasets and may underperform in the absence of sufficient training data. To address these limitations, a hybrid deep learning framework that combines the strengths of CNNs and ViTs for comprehensive feature extraction is proposed. The model consists of two parallel branches: a CNN-based local feature extractor and a ViT-based global feature extractor. The outputs of both branches are concatenated and passed through multiple fully connected layers with dropout and regularization to ensure robust learning. Furthermore, to improve interpretability and enhance classification performance, features are extracted from the Global Average Pooling

(GAP) layer and used to train a traditional Random Forest (RF) classifier. This hybrid approach benefits from the expressive power of deep feature extraction while leveraging the generalization capabilities of ensemble learning. The key contributions are as follows:

(1) To develop a deep learning model that combines CNN and Vision Transformer to capture both detailed local features and overall image patterns in retinal images.
(2) The model merges local (CNN) and global (ViT) features, helping to improve the accuracy of classifying different stages of diabetic retinopathy.
(3) To use features from the deep model to train a random forest classifier, which adds reliability and makes the results easier to understand.
(4) The model was tested on a diabetic retinopathy dataset with five severity levels and showed good results across all standard performance measures.
(5) The approach is designed to be easy to extend to other medical problems and suitable for use in real clinical settings.

2. LITERATURE SURVEY

Recent years have witnessed significant progress in the automated detection and classification of diabetic retinopathy, driven by the rapid development of deep learning and machine learning techniques. Early and accurate identification of DR is critical, as manual grading of retinal images is time-consuming, subjective, and often limited by resource constraints. Deep learning models, particularly convolutional neural networks (CNNs) and their variants, have emerged as powerful tools for analyzing retinal images, offering improved accuracy, efficiency, and scalability compared to traditional methods.

[7] demonstrated that clinicians can use automated ML and public datasets (Messidor-2, EyePACS) to develop high-performing DR models, with AUROC up to 0.951 and accuracy up to 96.7%, supporting democratization of AI in healthcare. [1] introduced a hybrid CNN-SVD model with improved SVM-RBF, DT, and KNN classifiers, achieving 99.18% accuracy, 98.15% sensitivity, and 100% specificity on the IDRiD dataset for vision-threatening DR, surpassing existing methods. [6] developed and validated code-free AutoML models for DR classification using 17,829 handheld retinal images, achieving 97% accuracy and high sensitivity/specificity in both internal and external validation, demonstrating feasibility for community-based screening.

[13] proposed a parallel CNN for feature extraction and ELM for classification, achieving 91.78% accuracy on Kaggle DR 2015 and 97.27% on APTOS 2019. The model is efficient, robust to dataset size and balance, and outperforms state-of-the-art models in speed and accuracy. [11] evaluated transfer learning models (VGG16, InceptionV3, DenseNet121, MobileNetV2) on combined datasets (APTOS, Messidor2, IDRiD), with DenseNet121 achieving 98.97% accuracy, showing that combining datasets improves performance. [9] revised ResNet-50 with improved preprocessing and adaptive learning, achieving 83.95% train and 74.32% test accuracy, outperforming other common CNNs and reducing overfitting and loss fluctuation. [10] compared a hybrid VGG16-XGBoost model and DenseNet 121 for DR detection on APTOS 2019. DenseNet 121 achieved 97.3% accuracy, significantly outperforming the hybrid model (79.5%), highlighting the effectiveness of advanced deep learning.

[3] developed DeepDR Plus, a deep learning system trained on over $800,000$ fundus images, predicting time to DR progression with concordance indexes of 0.754–0.846, enabling personalized screening intervals and validated on large, multiethnic datasets. [15] proposed a dual-branch deep learning model using transfer learning, trained on a large multi-center dataset including APTOS 2019. Achieved 98.5% accuracy (binary), 89.6% (stage grading), and a QWK of 93.0, outperforming established literature. [12] introduced MAPCRCI-DMPLC, a deep multilayer perceptive learning model with novel preprocessing and feature extraction, outperforming five state-of-the-art approaches on a retinal image dataset. To provide a concise comparison of recent advancements in diabetic retinopathy detection, Table 2 summarizes the key contributions from various studies and compares state-of-the-art approaches for DR detection, including parallel CNN architectures, hybrid deep learning models, automated machine learning frameworks, and transfer learning strategies. These studies demonstrate the evolution of model architectures, the impact of dataset diversity, and the growing feasibility of deploying AI-driven DR screening in real-world clinical and community settings.

3. METHODOLOGY

The proposed framework, termed **ViT–Local Global Fusion**, integrates convolutional and transformer-based architectures to achieve a comprehensive feature representation for diabetic retinopathy (DR) diagnosis from retinal fundus images. The model captures both *fine-grained local structures* (microaneurysms, hemorrhages, and exudates) and *global contextual relationships* (vascular patterns, optic disc structure) by fusing local and global feature representations within a unified deep learning framework. The overall architecture comprises three major components: (i) a local feature extraction branch based on a Convolutional Neural Network (CNN), (ii) a global feature extraction branch based on a Vision Transformer (ViT), and (iii) a fusion and classification module that integrates and interprets these features for final decision-making as shown in Figure 1.

3.1 Input Representation

Each input image $X \in \mathbb{R}^{H\times W\times C}$ represents a color fundus photograph, where $H = W = 224$ and $C = 3$. To ensure numerical stability, each image is normalized to the range $[0, 1]$ as follows:

$$X_{\text{norm}} = \frac{X}{255} \tag{1}$$

The dataset is partitioned into training, validation, and test subsets using a 70:30 split, ensuring stratification across DR severity levels. Data augmentation, including rotation ($\pm 15^{\circ}$), flipping, zooming (0.9–1.1), and brightness variation, is applied during training to improve model generalization.

3.2 Local Feature Extraction (CNN Branch)

The **local feature extraction branch** learns fine structural and textural details indicative of DR severity. It consists of three convolutional blocks, each composed of a convolutional layer, batch normalization, and ReLU activation. The local feature mapping process can be represented as:

$$F_{\text{local}} = f_{\text{CNN}}(X_{\text{norm}}; \Theta_{\text{CNN}}) \tag{2}$$

where f_{CNN} denotes the CNN function parameterized by weights Θ_{CNN}. Each convolutional layer performs spatial filtering as:

Table 1. Summary of recent literature on diabetic retinopathy detection using deep learning and machine learning techniques. The table includes model types, best performance metrics, and datasets used.

Paper Title	Year	Main Method/Model	Best Performance	Dataset(s) Used
[7]	2023	Automated ML (self-training)	AUROC 0.951, accuracy 96.7%	Messidor-2, EyePACS, Egypt
[1]	2024	Hybrid CNN-SVD + ISVM-RBF	99.18% accuracy, 100% specificity	IDRiD
[6]	2023	AutoML (code-free deep learning)	97% accuracy, high sensitivity/specificity	Handheld retinal images (17,829), APTOS
[13]	2023	Parallel CNN + ELM	97.27% (APTOS 2019), 91.78% (Kaggle DR 2015)	Kaggle DR 2015, APTOS 2019
[11]	2023	DenseNet121, VGG16, InceptionV3, MobileNetV2	98.97% (DenseNet121, combined)	APTOS, Messidor2, IDRiD
[9]	2023	Revised ResNet-50	83.95% train, 74.32% test accuracy	Not specified
[10]	2023	DenseNet121, VGG16-XGBoost	97.3% (DenseNet121)	APTOS 2019
[3]	2024	DeepDR Plus (DL system)	Concordance index 0.754–0.846	717,308 pretrain, 118,868 multiethnic validation
[15]	2023	Dual-branch DL, transfer learning	98.5% (binary), 89.6% (grading), QWK 93.0	APTOS 2019, multi-center dataset
[12]	2024	MAPCRCI-DMPLC	Outperforms 5 state-of-the-art models	Retinal image dataset (not specified)

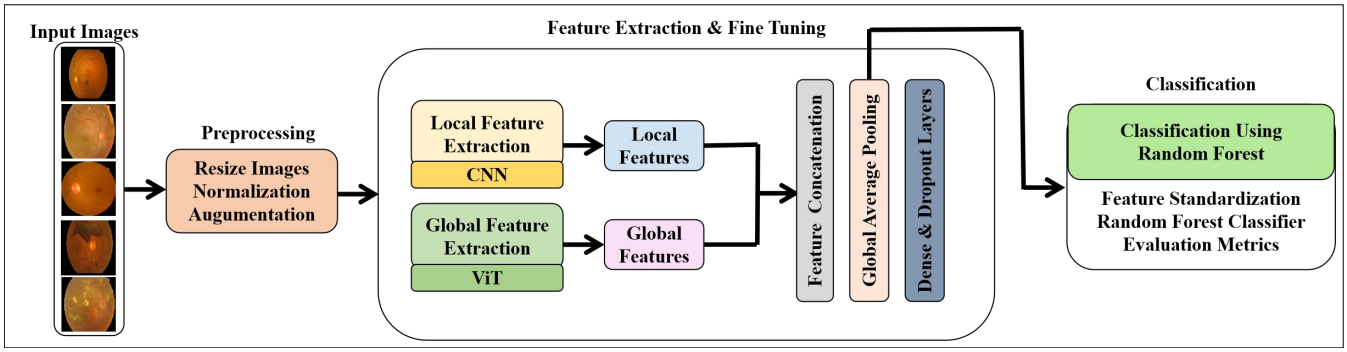

Fig. 1. Proposed model "ViT-Local Global Fusion" for diabetic foot ulcer classification

$$Y_{i,j,k} = \sigma\left(\sum_{m,n} X_{i+m,j+n} \cdot W_{m,n,k} + b_k\right) \tag{3}$$

where $W_{m,n,k}$ and b_k represent kernel weights and biases, and $\sigma(\cdot)$ denotes the ReLU activation. Filters of size 3×3 are used with increasing depths of 64, 128, and 256. Batch normalization stabilizes learning, while a *Global Average Pooling* (GAP) layer compresses spatial dimensions into a compact representation:

$$F_{\text{local}} = \frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} f_{i,j} \tag{4}$$

resulting in a 256-dimensional local feature vector summarizing spatial lesion-level features.

3.3 Global Feature Extraction (Vision Transformer Branch)

The **Vision Transformer (ViT)** branch captures global spatial dependencies within the retinal image. The input image is divided into non-overlapping patches of size $P \times P$ (with $P = 16$), resulting in:

$$N = \left(\frac{H}{P}\right) \times \left(\frac{W}{P}\right) = 196 \tag{5}$$

Each patch is flattened and projected into a 64-dimensional embedding space using a learnable matrix $E \in \mathbb{R}^{(P^2 \cdot C) \times D}$:

$$Z_0 = [x_1 E; x_2 E; \ldots; x_N E] + E_{\text{pos}} \tag{6}$$

where $E_{\text{pos}} \in \mathbb{R}^{N \times D}$ denotes the positional embedding. The embedded sequence is processed through $L = 8$ transformer encoder blocks, each composed of layer normalization, multi-head self-attention (MHSA), feed-forward network (MLP), and residual skip connections. For each encoder layer ℓ:

$$Z'_\ell = \text{MHSA}(\text{LN}(Z_{\ell-1})) + Z_{\ell-1} \tag{7}$$

$$Z_\ell = \text{MLP}(\text{LN}(Z'_\ell)) + Z'_\ell \tag{8}$$

The MHSA operation computes contextual attention between all patch pairs as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{9}$$

where $Q, K, V \in \mathbb{R}^{N \times D}$ denote the query, key, and value matrices, and d_k is the dimensionality of the key vector. After the final encoder block, a Global Average Pooling operation yields the global feature vector:

$$F_{\text{global}} = \text{GAP}(Z_L) \tag{10}$$

resulting in a 64-dimensional global embedding summarizing long-range dependencies and global contextual cues.

3.4 Feature Fusion and Classification

The local and global features are concatenated to form a joint representation:

$$F_{\text{fused}} = [F_{\text{local}}; F_{\text{global}}] \tag{11}$$

The fused feature vector is passed through a series of dense layers to refine discriminative capability:

$$h_1 = \sigma(W_1 F_{\text{fused}} + b_1) \tag{12}$$
$$h_2 = \sigma(W_2 h_1 + b_2) \tag{13}$$
$$h_3 = \sigma(W_3 h_2 + b_3) \tag{14}$$

where $\sigma(\cdot)$ denotes the ReLU activation. Each dense layer employs L2 regularization ($\lambda = 0.001$) and dropout ($p = 0.5$) to minimize overfitting. Finally, a softmax classifier predicts the probability distribution across the five DR severity classes:

$$\hat{y} = \text{softmax}(W_o h_3 + b_o) \tag{15}$$

where $\hat{y} \in \mathbb{R}^5$ corresponds to the class probability vector.

3.5 Model Optimization

The model is trained end-to-end using the Adam optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-5}. The objective function is the categorical cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{16}$$

where y_i and $\hat{y}_i$ denote the true and predicted probabilities of the i^{th} class, respectively, and $C = 5$. Model performance is evaluated using accuracy, precision, recall, F1-score, and AUC metrics.

3.6 Architectural Summary

The ViT–Local Global Fusion model effectively integrates two complementary feature spaces: (i) local lesion-level features learned through CNN convolutional operations and (ii) global contextual dependencies captured by Vision Transformer encoders. The final fused representation leverages both detailed and holistic information for improved diagnostic accuracy. The overall forward process of the model can be summarized as:

$$\hat{y} = f_{\text{fusion}}\left([f_{\text{CNN}}(X_{\text{norm}}), f_{\text{ViT}}(X_{\text{norm}})]\right) \tag{17}$$

This architecture enhances interpretability and classification performance, making it well-suited for real-world diabetic retinopathy screening and clinical decision-support systems.

3.7 Evaluation strategy

Model performance was evaluated using standard classification metrics, including overall accuracy, class-wise precision, recall, and F1-score. Additionally, a confusion matrix was generated to visualize the performance of the model in differentiating between the five classes. These metrics provide a comprehensive understanding of the model's diagnostic effectiveness, particularly in reducing false negatives for higher severity stages.
The trained RF model was evaluated using multiple metrics. Accuracy was calculated as the proportion of correctly predicted instances over the total number of predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. To assess the agreement between predicted and actual labels beyond chance, Cohen's Kappa Score was computed. Additionally, a confusion matrix and a detailed classification report (including precision, recall, and F1-score for each class) were generated to provide insights into class-wise performance.
To evaluate the model's discriminatory power, Receiver Operating Characteristic (ROC) curves were plotted using a one-vs-rest strategy, and Area Under the Curve (AUC) values were calculated for each class i as follows:

$$AUC_i = \int_0^1 TPR_i(FPR_i) dFPR_i \tag{19}$$

where TPR_i and FPR_i are the true positive rate and false positive rate for class i, respectively. These curves help visualize the trade-off between sensitivity and specificity for each class. Finally, the overall and per-class effectiveness of the classifier was validated using these ROC curves and the confusion matrix. This post-classification approach using Random Forest not only improved the prediction performance but also enhanced the model's transparency, making it more suitable for real-world clinical decision support systems.

4. RESULTS AND DISCUSSION

4.1 Dataset

The dataset [14] used in this study comprises a total of 3,662 color fundus images categorized into five diabetic retinopathy (DR) classes: No_DR (1,805), Mild (370), Moderate (999), Severe (193), and Proliferative_DR (295), and Figure 2 shows some sample images from the dataset. To ensure balanced evaluation and effective model training, the dataset was stratified into training, testing, and validation sets in a 60:20:20 ratio. The training set contains 2,197 images, the test set includes 733 images, and the validation set comprises 732 images. The class-wise distribution across each subset is detailed in Table 4.1.
To enhance the diversity of the training samples and improve model generalizability, a set of real-time data augmentation techniques was applied. These included horizontal flipping, a zoom range of 0.2, a shear range of 0.2, width and height shifts up to 20%, and random rotations up to 30°. This augmentation strategy was designed to simulate various real-world imaging conditions and reduce the risk of model overfitting.

4.2 Experimental setup

The model was compiled using the Adam optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5} to ensure optimal convergence and avoid overfitting. The categorical cross-entropy loss function was selected due to the multiclass nature of the problem, where the true label is one of five possible classes. Training was performed using a batch size of 8 over 20 epochs, with early stopping and model checkpointing enabled to preserve the best-performing model based on validation loss. A validation split of 20% from the training dataset was used for real-time monitoring of performance metrics during training.

4.3 Results

This section presents and analyzes the results obtained through two experimental phases of the proposed model. The primary aim was to classify retinal fundus images into five diabetic retinopathy (DR) classes. The first phase involved end-to-end training of the deep network with a softmax output layer. The second phase involved

Table 2. Class-wise distribution of images across training, testing, and validation sets for the diabetic retinopathy dataset.

Data split/Classes	Mild	Moderate	No_DR	Proliferate_DR	Severe	TOTAL IMAGES
Train	222	599	1083	177	115	**2197**
Test	74	201	361	59	40	**733**
Validation	74	199	361	59	38	**732**
TOTAL IMAGES	**370**	**999**	**1805**	**295**	**193**	**3662**

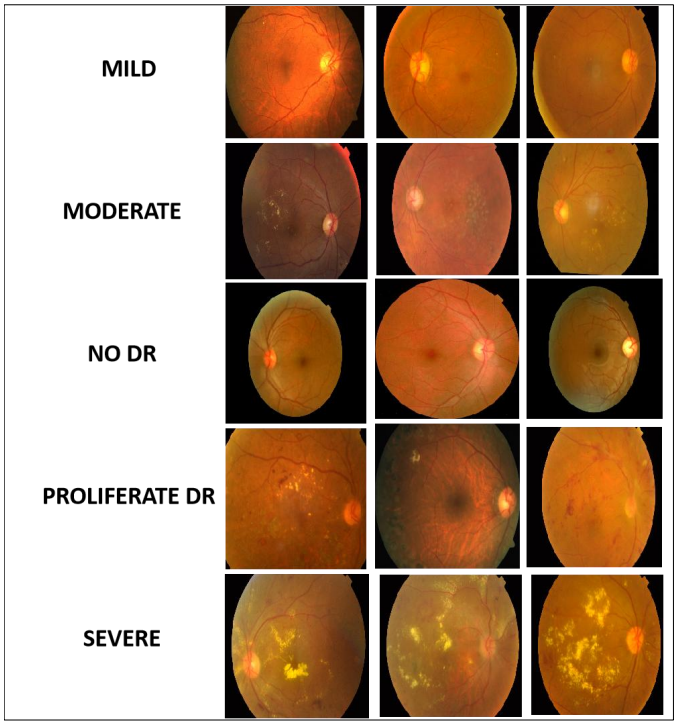

Fig. 2. Representative fundus images from the diabetic retinopathy dataset showing different severity levels. The rows correspond to the five classes: No_DR, Mild, Moderate, Severe, and Proliferative_DR (from top to bottom). Each column displays distinct examples within a class, capturing the variation in lesion presentation and image quality across the dataset.

extracting features from the Global Average Pooling (GAP) layer of the trained model and classifying them using machine learning algorithms such as Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT).

4.3.1 Performance of End-to-End Model with Softmax Output. The initial experiment evaluated the complete hybrid model (CNN + ViT) using the softmax activation for direct five-class classification. As shown in the learning curves in Figure 3, the model converged well with minimal overfitting. The final training accuracy achieved was 73.37%, with a corresponding training loss of 0.7226. On the validation dataset, the model achieved an accuracy of 72.95% and a validation loss of 0.7925, indicating balanced generalization.

The model was further tested on the unseen test set, where it achieved a test accuracy of 70.12% and a test loss of 0.7820. The class-wise performance, represented via the confusion matrix in Figure 4, demonstrates that the model correctly classified most of the samples across all five categories. However, minor misclassifications were observed, particularly in intermediate classes (e.g., Moderate vs. Severe), which share overlapping visual features. The ROC curve for each class depicted in Figure 5 indicated strong performance, with AUC values ranging from 0.77 to 0.98. Notably, the model achieved the highest AUC of 0.98 for Class 2 (Moderate DR), suggesting strong discriminative capability for mid-stage severity detection.

4.3.2 Feature-Based Classification using GAP Layer and Machine Learning Models. To further enhance classification performance and investigate the discriminative power of learned features, deep features are extracted from the GAP layer of the trained model. These features were then fed into classical machine learning classifiers—Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), and Gaussian Naive Bayes (GNB). Among these, both Random Forest and Decision Tree classifiers achieved a perfect classification performance, with 100% accuracy, precision, recall, F1-score, and an AUC of 1.0, as shown in Figure 6 and the confusion matrix in Figure 7. This indicates that the hybrid architecture was highly effective in learning discriminative latent representations, even though the original softmax-based classifier had not fully exploited this capability.

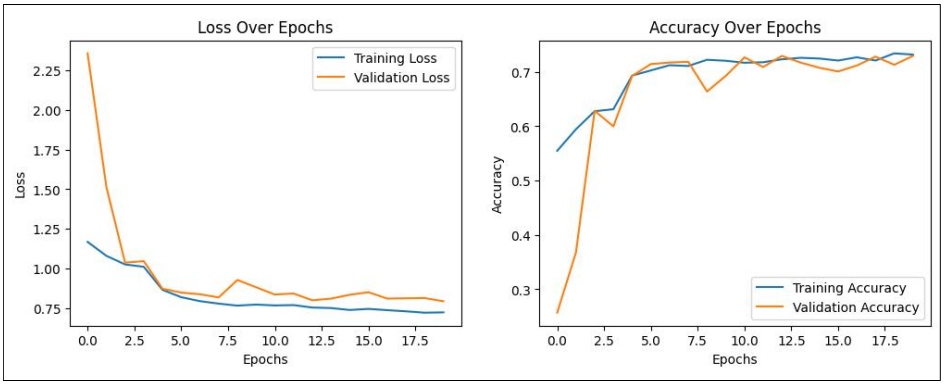

Fig. 3. Training and validation performance curves over 20 epochs of the proposed model. The left plot shows the decrease in training and validation loss, indicating model convergence. The right plot presents the training and validation accuracy trends, demonstrating consistent improvement and generalization performance across epochs.

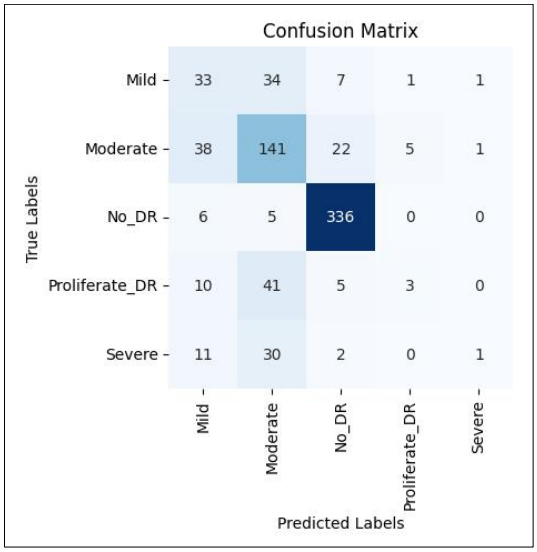

Fig. 4. Confusion matrix illustrating the classification performance of the diabetic retinopathy model before extraction of features from the GAP layer.

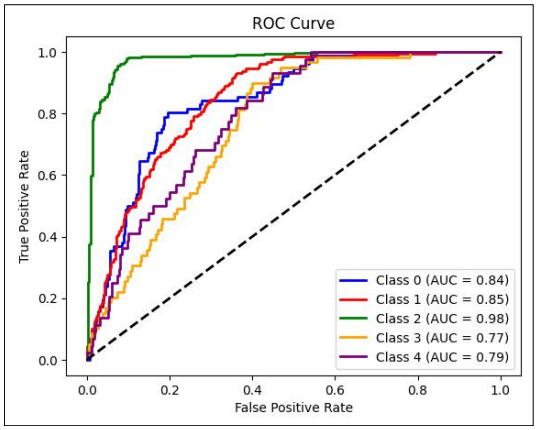

Fig. 5. Receiver Operating Characteristic (ROC) curves for the five-class diabetic retinopathy classification model before extraction of features from the GAP layer.

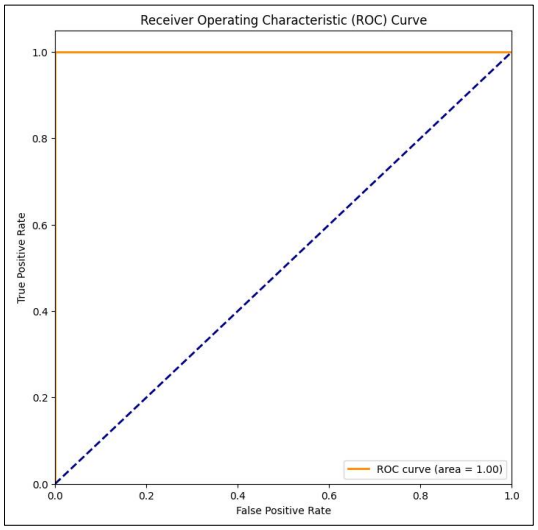

Fig. 6. Post-classification ROC analysis using the Random Forest classifier to evaluate model discrimination capability.

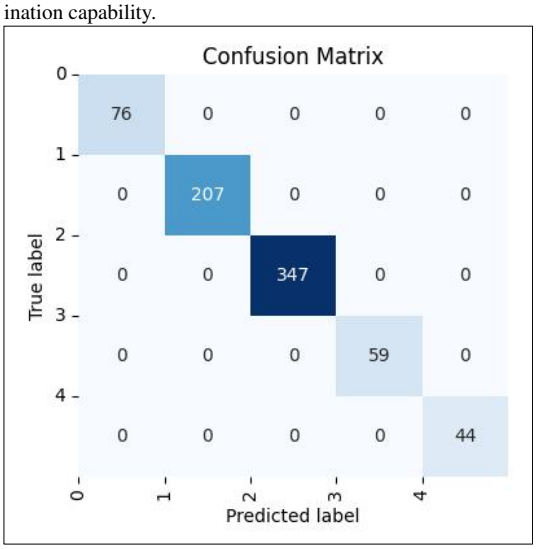

Fig. 7. Confusion matrix illustrating the classification performance of the proposed diabetic retinopathy model.

Further comparative analysis using SVM and Gaussian Naive Bayes showed AUC values of 0.94 and 0.80, with slightly lower classification metrics, thus reinforcing the superiority of ensemble tree-based methods for this particular feature space. This stage-wise decoupling of feature extraction and classification not only boosts performance but also opens up the possibility for using the extracted features in clinical decision support systems.

These findings suggest that while end-to-end deep learning models are effective, the extracted features from intermediate layers (like GAP) can be highly informative and better leveraged by traditional classifiers. The proposed ViT-LocalGlobalFusion model demonstrates strong potential in real-world clinical settings, especially when combined with interpretable classifiers like decision trees. Furthermore, the hybrid fusion of local (CNN) and global (ViT) features enables better encoding of retinal structures ranging from microaneurysms to widespread hemorrhages, thereby supporting robust multi-stage DR diagnosis.

4.3.3 Ablation study. To understand the impact of each component in the proposed architecture, a step-by-step ablation study was conducted. Initially, the performance of the individual branches—CNN for local features and Vision Transformer (ViT) for global features was evaluated. The local feature branch achieved an accuracy of 68.89% with an F1-score of 60.16%, while the global feature branch yielded a slightly better performance with 72.31% accuracy and an F1-score of 64.36%. This suggests that global contextual information captured by the ViT plays a significant role in classifying diabetic retinopathy (DR) severity levels.

Subsequently, the local and global features extracted from both branches were fused, and a classifier was trained on the resulting combined feature vector. This fusion substantially improved performance, resulting in an accuracy of 78% and an F1-score of 73%, highlighting the complementary nature of local and global representations.

Finally, the proposed method was evaluated by extracting features from the Global Average Pooling (GAP) layer after fusion and feeding them into a post-classification stage using machine learning classifiers. The Random Forest (RF) and Decision Tree (DT) classifiers both achieved 100% across all evaluation metrics—precision, recall, F1-score, and accuracy—indicating the strong discriminative power of the fused features. These results, as illustrated in Figure 8, demonstrate that the combination of deep feature extraction and classical ensemble classifiers significantly enhances performance and reliability in DR classification.

To further enhance classification reliability, features were extracted from the Global Average Pooling (GAP) layer after fusion and were used as input to classical machine learning classifiers: Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), and Naïve Bayes (NB). As shown in Figure 9, RF and DT yielded perfect scores across all metrics (100% precision, recall, F1-score, and accuracy), highlighting the strong discriminative power of the extracted features. SVM performed moderately well, achieving 76.94% accuracy and a 72.64% F1-score, while NB underperformed with 49.52% accuracy and an F1-score of 51.58%. These findings confirm that the fusion of local and global features produces highly informative representations and that ensemble-based classifiers such as RF and DT can effectively leverage these for superior classification in diabetic retinopathy diagnosis. Additionally, the confusion matrices for each of the machine learning classifiers are presented in Figure 10, providing a detailed view of the class-wise performance and further validating the classification outcomes.

The discriminative capability of the machine learning classifiers was further assessed using Receiver Operating Characteristic (ROC) curves and their corresponding Area Under the Curve (AUC) values. As illustrated in Figure 11, Random Forest and Decision Tree both achieved an AUC of 1.00, reflecting perfect classification performance. The SVM model followed closely with an AUC of 0.94, whereas Gaussian Naïve Bayes exhibited a lower AUC of 0.80. These findings reinforce the effectiveness of the ensemble-based models (RF and DT) in fully exploiting the fused feature representations for accurate diabetic retinopathy detection.

5. CONCLUSION

This study presents a comprehensive evaluation of the proposed ViT-LocalGlobalFusion framework for diabetic retinopathy (DR)

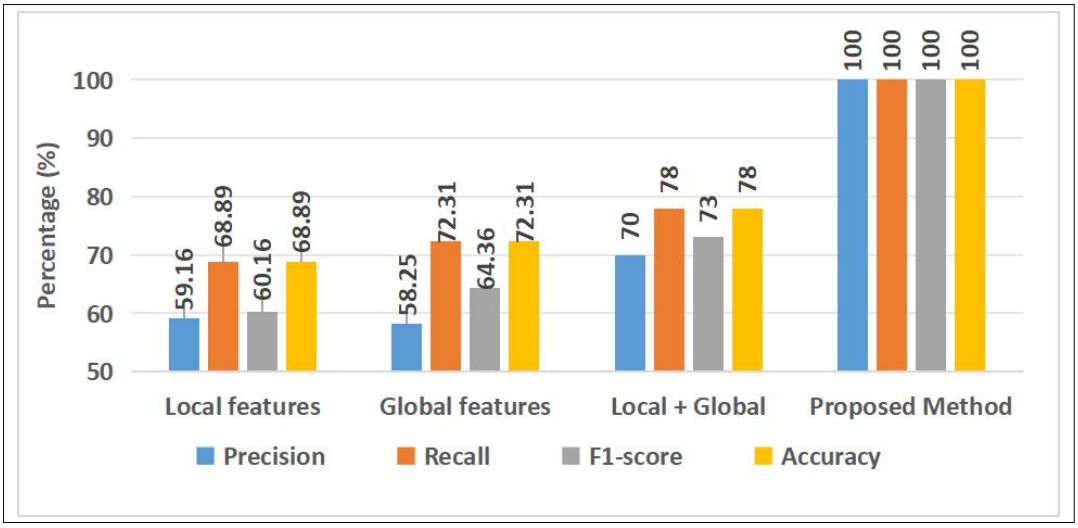

Fig. 8. Bar chart illustrating the results of the ablation study. The performance metrics are compared across different stages of the proposed architecture—CNN branch, ViT branch, fused features, and post-classification using Random Forest (RF).

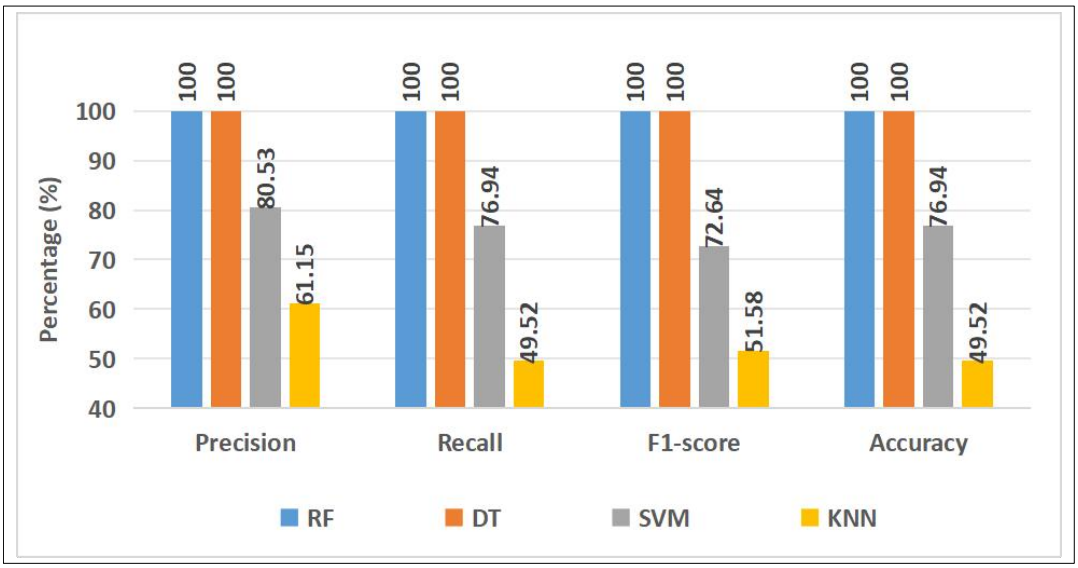

Fig. 9. Performance comparison of different machine learning classifiers (RF, DT, SVM, KNN) on fused features extracted from the proposed hybrid deep learning model.

classification. Through a series of controlled experiments, the individual and combined contributions of local features extracted by CNN and global contextual features captured by Vision Transformer (ViT) were systematically assessed. While both branches demonstrated effectiveness independently, their fusion led to a significant boost in classification performance, particularly in challenging intermediate stages such as Moderate DR.

The unified representation achieved an accuracy of 78.23% and an F1-score of 73.23%, with ROC analysis showing AUC values exceeding 0.90 across all classes—highlighting the complementary nature of local and global features. Furthermore, when the fused features were input into classical machine learning classifiers, ensemble-based models like Random Forest (RF) and Decision Tree (DT) achieved perfect scores across all evaluation metrics. In contrast, SVM and Naïve Bayes performed less effectively, with NB showing considerable performance limitations.

These results, supported by confusion matrices and ROC curves, validate the robustness and discriminative strength of the learned feature representations. Overall, the proposed framework not only enhances classification reliability but also demonstrates strong potential for integration into real-world clinical decision support systems aimed at early and accurate diagnosis of diabetic retinopathy. Future work will focus on extending the evaluation of the proposed ViT–Local Global Fusion framework to multiple retinal image datasets such as APTOS 2019, Messidor-2, and IDRiD, en-

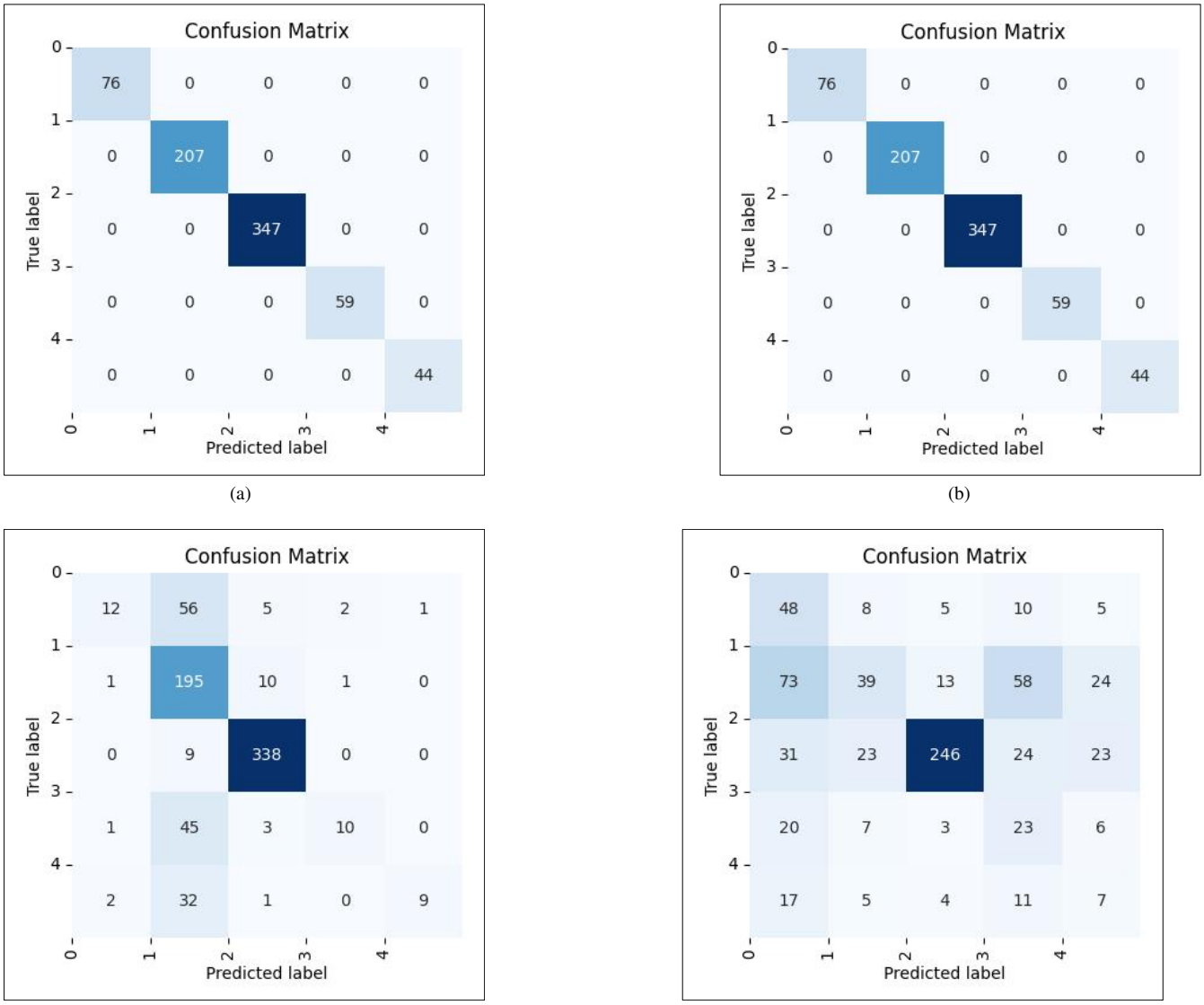

Fig. 10. Confusion matrices of machine learning classifiers (Random Forest, Decision Tree, Support Vector Machine, and Naïve Bayes) applied to the fused features. The matrices illustrate the class-wise prediction performance, with RF and DT showing perfect classification, while SVM and NB exhibit varying degrees of misclassification. (a) RF (b) DT (c) SVM (d) NB

abling a more comprehensive analysis of its robustness and generalization capability across varied imaging conditions.

6. REFERENCES

[1] Anas Bilal, Azhar Imran, Talha Imtiaz Baig, Xiaowen Liu, Haixia Long, Abdulkareem Alzahrani, and Muhammad Shafiq. Improved support vector machine based on cnn-svd for vision-threatening diabetic retinopathy detection and classification. *Plos one*, 19(1):e0295951, 2024.

[2] Milena Cioana, Jiawen Deng, Ajantha Nadarajah, Maggie Hou, Yuan Qiu, Sondra Song Jie Chen, Angelica Rivas, Parm Pal Toor, Laura Banfield, Lehana Thabane, et al. Global prevalence of diabetic retinopathy in pediatric type 2 diabetes: a systematic review and meta-analysis. *JAMA network open*, 6(3):e231887–e231887, 2023.

[3] Ling Dai, Bin Sheng, Tingli Chen, Qiang Wu, Ruhan Liu, Chun Cai, Liang Wu, Dawei Yang, Haslina Hamzah, Yuexing Liu, et al. A deep learning system for predicting time to progression of diabetic retinopathy. *Nature Medicine*, 30(2):584–594, 2024.

[4] Sebastian Dinesen, Lonny Stokholm, Yousif Subhi, Tunde Peto, Thiusius Rajeeth Savarimuthu, Nis Andersen, Jens Andresen, Toke Bek, Javad Hajari, Steffen Heegaard, et al. Five-year incidence of proliferative diabetic retinopathy and associated risk factors in a nationwide cohort of 201 945 danish

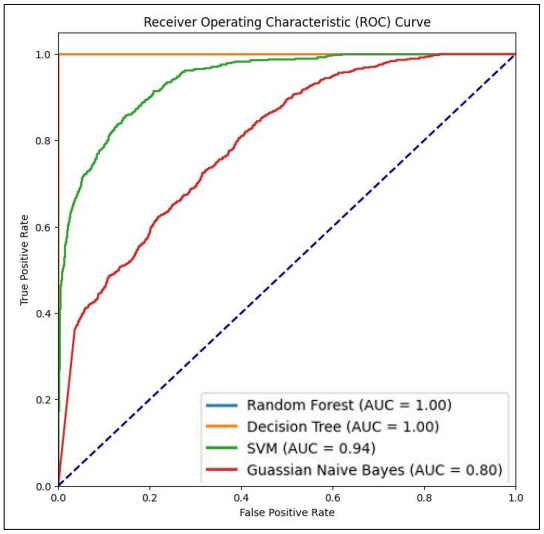

Fig. 11. Performance comparison of different machine learning classifiers (RF, DT, SVM, KNN) on fused features extracted from the proposed hybrid deep learning model.

patients with diabetes. *Ophthalmology science*, 3(3):100291, 2023.

[5] Xuhong Hou, Limin Wang, Dalong Zhu, Lixin Guo, Jianping Weng, Mei Zhang, Zhiguang Zhou, Dajin Zou, Qiuhe Ji, Xiaohui Guo, et al. Prevalence of diabetic retinopathy and vision-threatening diabetic retinopathy in adults with diabetes in china. *Nature Communications*, 14(1):4296, 2023.

[6] Cris Martin P Jacoba, Duy Doan, Recivall P Salongcay, Lizzie Anne C Aquino, Joseph Paolo Y Silva, Claude Michael G Salva, Dean Zhang, Glenn P Alog, Kexin Zhang, Kaye Lani Rea B Locaylocay, et al. Performance of automated machine learning for diabetic retinopathy image classification from multi-field handheld retinal images. *Ophthalmology Retina*, 7(8):703–712, 2023.

[7] Edward Korot, Mariana Batista Gonçalves, Josef Huemer, Sara Beqiri, Hagar Khalid, Madeline Kelly, Mark Chia, Emily Mathijs, Robbert Struyven, Magdy Moussa, et al. Clinician-driven ai: code-free self-training on public data for diabetic retinopathy referral. *JAMA ophthalmology*, 141(11):1029–1036, 2023.

[8] Martina Kropp, Olga Golubnitschaja, Alena Mazurakova, Lenka Koklesova, Nafiseh Sargheini, Trong-Tin Kevin Steve Vo, Eline de Clerck, Jiri Polivka Jr, Pavel Potuznik, Jiri Polivka, et al. Diabetic retinopathy as the leading cause of blindness and early predictor of cascading complications—risks and mitigation. *Epma Journal*, 14(1):21–42, 2023.

[9] Chun-Ling Lin and Kun-Chi Wu. Development of revised resnet-50 for diabetic retinopathy detection. *BMC bioinformatics*, 24(1):157, 2023.

[10] Cheena Mohanty, Sakuntala Mahapatra, Biswaranjan Acharya, Fotis Kokkoras, Vassilis C Gerogiannis, Ioannis Karamitsos, and Andreas Kanavos. Using deep learning architectures for detection and classification of diabetic retinopathy. *Sensors*, 23(12):5726, 2023.

[11] AM Mutawa, Shahad Alnajdi, and Sai Sruthi. Transfer learning for diabetic retinopathy detection: A study of dataset combination and model performance. *Applied Sciences*, 13(9):5685, 2023.

[12] Dharmalingam Muthusamy and Parimala Palani. Deep learning model using classification for diabetic retinopathy detection: an overview. *Artificial Intelligence Review*, 57(7):185, 2024.

[13] Md Nahiduzzaman, Md Robiul Islam, Md Omaer Faruq Goni, Md Shamim Anower, Mominul Ahsan, Julfikar Haider, and Marcin Kowalski. Diabetic retinopathy identification using parallel convolutional neural network based feature extractor and elm classifier. *Expert Systems with Applications*, 217:119557, 2023.

[14] Sovit Ranjan Rath. Diabetic retinopathy 224x224 gaussian filtered dataset. `https://www.kaggle.com/datasets/sovitrath/diabetic-retinopathy-224x224-gaussian-filtered`, 2022. Accessed: July 21, 2025.

[15] Hossein Shakibania, Sina Raoufi, Behnam Pourafkham, Hassan Khotanlou, and Muharram Mansoorizadeh. Dual branch deep learning network for detection and stage grading of diabetic retinopathy. *Biomedical Signal Processing and Control*, 93:106168, 2024.

[16] Tien-En Tan and Tien Yin Wong. Diabetic retinopathy: Looking forward to 2030. *Frontiers in Endocrinology*, 13:1077669, 2023.

[17] Alebachew Ferede Zegeye, Yemataw Zewdu Temachu, and Chilot Kassa Mekonnen. Prevalence and factors associated with diabetes retinopathy among type 2 diabetic patients at northwest amhara comprehensive specialized hospitals, northwest ethiopia 2021. *BMC ophthalmology*, 23(1):9, 2023.

[18] Jing Zhou and Bo Chen. Retinal cell damage in diabetic retinopathy. *Cells*, 12(9):1342, 2023.