Innovative Approach for Diabetic Retinopathy Severity Classification: An AI-Powered Tool using CNN-Transformer Fusion

[1]Department of Biomedical Engineering, Meybod University, Meybod, Iran
[2]Department of Biomedical Engineering, Hakim Sabzevari University, Sabzevar, Iran

Khosro Rezaee (PhD)[1]*, Fateme Farnami (MSc)[2]

ABSTRACT

Background: Diabetic retinopathy (DR), a diabetes complication, causes blindness by damaging retinal blood vessels. While deep learning has advanced DR diagnosis, many models face issues like inconsistent performance, limited datasets, and poor interpretability, reducing their clinical utility.

Objective: This research aimed to develop and evaluate a deep learning structure combining Convolutional Neural Networks (CNNs) and transformer architecture to improve the accuracy, reliability, and generalizability of DR detection and severity classification.

Material and Methods: This computational experimental study leverages CNNs to extract local features and transformers to capture long-range dependencies in retinal images. The model classifies five types of retinal images and assesses four levels of DR severity. The training was conducted on the augmented APTOS 2019 dataset, addressing class imbalance through data augmentation techniques. Performance metrics, including accuracy, Area Under the Curve (AUC), specificity, and sensitivity, were used for metric evaluation. The model's robustness was further validated using the IDRiD dataset under diverse scenarios.

Results: The model achieved a high accuracy of 94.28% on the APTOS 2019 dataset, demonstrating strong performance in both image classification and severity assessment. Validation on the IDRiD dataset confirmed its generalizability, achieving a consistent accuracy of 95.23%. These results indicate the model's effectiveness in accurately diagnosing and assessing DR severity across varied datasets.

Conclusion: The proposed Artificial intelligence (AI)-powered diagnostic tool improves diabetic patient care by enabling early DR detection, preventing progression and reducing vision loss. The proposed AI-powered diagnostic tool offers high performance, reliability, and generalizability, providing significant value for clinical DR management.

Keywords

Diabetic Retinopathy; Convolutional Neural Networks; Artificial Intelligence; Deep Learning; Fundus Oculi

Introduction

Diabetes frequently leads to diabetic retinopathy (DR), a major cause of vision loss, especially among those of working age. The disease progresses through several stages, beginning with mild non-proliferative abnormalities to more severe proliferative

*Corresponding author:
Khosro Rezaee
Department of Biomedical Engineering, Meybod University, Meybod, Iran
E-mail:
kh.rezaee@meybod.ac.ir

Received: 24 August 2024
Accepted: 26 February 2025

DR, characterized by abnormal blood vessel growth on the retina. Early identification and accurate assessment of DR severity are crucial for preventing irreversible vision loss. The DR detection has relied on the manual analysis of fundus images by trained ophthalmologists, which can be time-consuming, subject to inter-observer variability, and impractical for large-scale screening [1-3]. The increasing global prevalence of diabetes highlights the critical need for automated and accurate tools for the early detection and severity assessment of DR [4].

Recent advancements in deep learning have significantly enhanced the field of medical scan analysis. Convolutional Neural Networks (CNNs) have shown remarkable capabilities in extracting hierarchical features from complex image data, including retinal images [5]. CNNs excel at extracting local features, but often face challenges in capturing long-range dependencies within images, which are essential for a thorough assessment of diseases. To address this limitation, Transformer models, initially developed for natural language processing (NLP), have been adapted for image analysis, offering the ability to model global contextual relationships across an image [6-8].

Despite these technological advancements, several challenges persist in the application of deep learning for DR severity categorization. One significant issue is the requirement for large and annotated images to train these approaches effectively [9]. The availability of such datasets is often limited, leading to issues with model generalizability and performance, particularly in diverse clinical settings. Moreover, many existing models suffer from a lack of interpretability, which is a critical factor in gaining the trust of clinicians and ensuring the models' adoption in routine clinical practice [10]. Additionally, the computational complexity associated with deep learning architectures, particularly those incorporating Transformers, can be a barrier to their implementation in resource-constrained environments [11].

The primary challenge in the automated categorization of DR severity lies in developing a model that can accurately assess the condition across its various stages, from mild to proliferative retinopathy, while addressing the limitations of current methodologies. Many existing deep learning-based DR classification systems primarily rely on CNNs. While CNNs are effective at local feature extraction, they may not fully capture the complex relationships between various retinal regions that indicate disease progression [12]. Furthermore, the imbalance in available datasets, where some severity levels are underrepresented, poses a significant hurdle, leading to biased models [13]. Another critical issue is the interpretability of these models. High accuracy alone is not sufficient; the models must also provide insights that are actionable for clinicians [14]. Lastly, the computational demands of advanced deep learning models, particularly those incorporating transformer mechanisms, can restrict their usability in real-world clinical settings with limited resources [15].

Wang et al. [16] introduced a hierarchical multi-task deep learning model designed to simultaneously identify the severity of diabetic retinopathy and associated features in fundus images. This approach integrated causal relationships between DR features and severity levels, and its performance was assessed using two separate datasets. The performance evaluation of the proposed model in DR severity classification was conducted using metrics, such as the weighted Cohen's kappa coefficient, receiver operating characteristic (ROC) curves, and precision-recall analyses. The results showed that the model performed at a level comparable to ophthalmologists with 5–10 years of experience in diagnosing DR severity.

In the research conducted by Zhang et al. [17], a classification model based on the Inception V3 architecture was developed using the publicly available Kaggle dataset, which

includes over 88,000 fundus images. The model was evaluated using input images with two resolutions: 299×299 and 896×896 pixels. The model using 896×896 pixel input outperformed the one with lower resolution, achieving an Area Under the Curve (AUC) of 0.968, sensitivity of 0.925, and specificity of 0.907. This indicates that higher image resolution improves classification performance.

In the study by Kale and Sharma [18], an ensemble deep-learning approach was presented for classifying the severity of DR into five levels (proliferative, severe, moderate, mild, and no-DR). Initially, CNNs were trained and then combined to create an ensemble model, which was further fine-tuned. The ensemble model achieved a validation accuracy of 87.31%.

Mustafa et al. [19] introduced a multi-stream neural network for DR severity categorization. The method utilized pre-trained ResNet-121 and DenseNet-50 architectures for feature extraction and applied principal component analysis (PCA) for dimensionality reduction. The approach was evaluated on the MESSIDOR-2 and EyePACS databases, achieving a classification accuracy of up to 95.58%.

In the research by Sikder et al. [20], an ensemble learning method based on decision trees was employed to classify DR severity. Using features, such as gray-level intensity and texture from fundus images in the APTOS 2019 database, the model obtained an accuracy of 94.20% and an F-measure of 93.51%.

Goel et al. [21] utilized transfer learning procedures to develop an architecture for classifying DR severity. The suggested method was trained on the IDRD dataset with high accuracy in grading retinal images into severity levels.

Bhardwaj et al. [22] introduced a quadrant-based ensemble learning model using the InceptionResNet-V2 framework for DR severity categorization. Techniques, such as histogram equalization, quadrant cropping, and data augmentation were employed, resulting in an accuracy of 93.33% and a 13.58% improvement compared to earlier methods.

In the study by Fayyaz et al. [23], AlexNet and ResNet101 were utilized for feature extraction, combined with a support vector machine (SVM) for DR severity classification. The proposed strategy obtained an accuracy of 93% in categorizing disease severity.

Sugeno et al. [24] introduced a simple approach for lesion identification and severity grading of DR using EfficientNet-B3 and the APTOS 2019 database. The architecture obtained sensitivity and specificity values above 0.98 and demonstrated excellent performance in severity classification.

Bodapati et al. [25] proposed an innovative composite deep neural network augmented with a gated-attention mechanism to classify the severity of diabetic retinopathy. By synergizing features extracted from multiple CNNs, their approach yielded an accuracy of 82.54% and a kappa score of 79, underscoring its potential in DR severity prediction despite moderate performance.

In another study, Zhang et al. [26] introduced a meticulously optimized deep-learning framework tailored for grading DR severity. Their method incorporated sophisticated techniques, including background segmentation, feature refinement through the Cuckoo search algorithm, and CNN-driven classification. Validated against the MESSIDOR and IDRiD datasets, this framework demonstrated remarkable accuracies of 97.55% and 98.02%, respectively, highlighting its efficacy and robustness in handling complex retinal image datasets.

A plethora of studies have investigated the application of deep learning techniques for classifying the severity of diabetic retinopathy. Among these, notable contributions by Wang et al. [16] and Zhang et al. [17] predominantly employed CNN-based architectures, capitalizing on their proficiency in extracting localized features from fundus images. These approaches underscore the widespread reliance on CNNs to tackle fundamental image

analysis tasks in DR classification. While these approaches have achieved high accuracy in DR severity detection, their primary limitation lies in their inability to analyze long-range dependencies and global relationships within retinal images. Ensemble models, like those introduced by Kale and Sharma [18] and Bhardwaj et al. [22], have shown improved accuracy but have still faced challenges when dealing with imbalanced datasets, especially for underrepresented severity levels, such as severe and proliferative DR. Although research, such as Sikder et al. [20] and Goel et al. [21] have incorporated data augmentation techniques, these efforts often fall short in enhancing model generalizability across real-world datasets. Furthermore, a critical limitation of existing methods is the lack of interpretability in their outputs, making it difficult for clinicians to adopt these tools for practical applications.

This study aims to address existing limitations by developing a hybrid deep-learning model that integrates the strengths of both CNNs and Transformers. The goal is to improve the accuracy of DR severity classification while enhancing model interpretability. The model employs CNNs to extract localized features, while transformers are leveraged to capture long-range dependencies and contextual relationships within retinal images, creating a more holistic and robust analytical framework. This integration is expected to improve the structure's ability to assess the severity of DR more comprehensively [6]. Moreover, to mitigate the issue of class imbalance, the study employs data augmentation strategies, such as brightness adjustment, scaling, and rotation. These techniques help generate a more balanced dataset, thereby improving the model's generalization capabilities across all severity levels [27]. The study focuses on designing a model that not only achieves high accuracy in classifying DR severity but also provides interpretable results that clinicians can readily use to make informed decisions [28]. The hybrid model's efficacy will undergo a comprehensive evaluation utilizing key performance metrics, including the area under the receiver operating characteristic curve (AUC-ROC), specificity, sensitivity, and accuracy. These metrics are chosen to assess the model's robustness and reliability, ensuring its suitability for deployment in clinical environments [29, 30].

This research introduces a hybrid deep learning framework that combines the precision of CNNs for capturing local features with the ability of Transformers to analyze global patterns. This harmonious synergy elevates both the accuracy and interpretability of diabetic retinopathy severity classification, establishing a new paradigm in the field. Unlike existing models [31-35] that predominantly rely on CNNs for local feature extraction, this approach leverages the power of transformers to capture long-range dependencies within retinal images. This approach improves DR severity assessment while overcoming conventional model limitations. Data augmentation enhances generalization, ensuring robustness for clinical applications. This study introduces an innovative integration of CNNs and Transformer models, where CNNs extract localized features from fundus images while Transformers capture broader contextual relationships. This synergy significantly enhances the model's ability to accurately classify DR severity across all stages, from mild non-proliferative to severe proliferative forms, ensuring a more comprehensive and reliable assessment [36].

A major challenge in DR severity classification is dataset imbalance, as certain severity levels are underrepresented. This research addresses this limitation by employing data augmentation techniques such as brightness adjustment, scaling, and rotation to create a more balanced dataset. These strategies improve the model's generalizability across all severity levels, ensuring consistent and unbiased performance even in real-world clinical settings. Beyond achieving high accuracy,

interpretability remains crucial for clinical adoption. This study emphasizes the development of a model that not only delivers precise predictions but also generates interpretable outputs that clinicians can easily understand and apply. By providing actionable insights, this approach bridges the gap between AI-driven tools and practical clinical use, ultimately enhancing the quality of diabetic patient care [37, 38].

This study aims to present a clear and structured analysis of the research process, focusing on the development of a hybrid CNN-Transformer model for diabetic retinopathy severity classification. It introduces novel preprocessing techniques and data augmentation strategies to enhance model performance. The research highlights improvements in accuracy and robustness compared to conventional models, offering valuable insights into its clinical applicability and future advancements.

Material and Methods

This computational experimental study implements retrospective data from publicly available datasets to develop and assess a hybrid deep learning model for classifying the severity of DR. The research leverages a unique, less-explored dataset, presenting specific challenges in data complexity. To address these, a hybrid deep learning model was developed, combining CNNs with transformer mechanisms to capture both global and local image features. The following subsections detail the dataset, preprocessing steps, model architecture, and techniques to enhance performance and generalization.

Datasets

The APTOS 2019 DR images are publicly available on the Kaggle website [39]. This dataset was selected because it originates from India, which has a population ethnically similar to that of Mauritius. Thus, each scan in the APTOS 2019 dataset was classified into one of five classes (0 to 4) based on the severity of the disease. Similarly, a local physician categorized each image from the local group into these same classes.

The original APTOS dataset consisted of 3662 images, distributed across the five classes as shown in Table 1. However, a manual quality check excluded low-quality and noisy images, leaving 3057 high-quality images for further analysis. To address the significant imbalance in the class distribution, five data augmentation strategies were applied: horizontal flipping (mirroring images horizontally to simulate different perspectives), vertical flipping (mirroring images vertically for additional diversity), brightness adjustment (modifying image brightness to mimic varying lighting conditions), scaling (resizing images to simulate different zoom levels), and rotation (rotating images by small angles to introduce variability in orientation). These augmentation

Table 1: Class-wise image distribution before and after preprocessing, including noise removal and augmentation.

Class	After Exclusion of Noisy Images	Augmented Images	Final Total
No DR (Class 0)	1300	6480	7780
Mild DR (Class 1)	359	1796	2155
Moderate DR (Class 2)	926	4630	5556
Severe DR (Class 3)	186	930	1116
Proliferative DR (Class 4)	286	1430	1716
Total	3057	15266	18310

DR: Diabetic Retinopathy

techniques expanded the dataset to 18310 images while ensuring a balanced representation across all classes.

The final dataset includes images from all severity levels, categorized into five classes, to analyze a range of retinal conditions. Table 1 summarizes the dataset distribution before and after augmentation, highlighting the preprocessing steps applied. This study analyzes a range of retinal conditions using fundus images categorized into five severity classes.

Image preprocessing

The images are in Red-Green-Blue (RGB) format, as CNN-based networks like ResNet typically accept RGB pictures as input, utilizing three channels to capture blue, green, and red. If a retinal image generated in any modality was grayscale with only one channel, we replicated that layer to achieve an RGB format, allowing it to be processed directly by the CNN. Two various preprocessing frameworks were tested to convert the fundus pictures into a suitable input format for the CNN, enabling full utilization of pre-trained networks. In summary, the retinal scans were resized to 224×224 pixels and then converted to RGB format. The techniques used include rotation, scaling, random cropping, brightness adjustment, and noise application. These methods increase the diversity of the training data, allowing models to focus on significant features in the images and avoid reliance on specific, non-essential characteristics.

CNN-Transformer Fusion

The input consists of fundus retinal scans to detect signs of DR. These images are typically depicted as 3D matrices $H \times W \times C$, where C is the number of color channels, W is the width, and H is the height. Before feeding these images into the model, preprocessing steps such as normalization or resizing might be applied.

ResNet18 extracts hierarchical features using convolutional layers and residual blocks, ensuring efficient gradient flow. Key parameters include the number of filters (feature map depth), kernel size (e.g., 3×3), stride (filter movement), and padding (spatial preservation).

In the Region of Interest Alignment (ROI Align) and Bounding Box functions, these techniques are used to focus on specific regions of the retina that are more likely to exhibit signs of DR, such as areas with microaneurysms or hemorrhages. ROI Align is typically implemented using bilinear interpolation to align the extracted features with the region of interest. The bounding box focuses on retinal regions with potential DR signs, such as microaneurysms and hemorrhages, while ROI Align ensures accurate resampling and alignment for further analysis. The extraction of ROIs and bounding boxes was performed using a lesion detection approach that identifies areas in the retinal images with high likelihoods of abnormalities, such as microaneurysms, hemorrhages, or exudates. Specifically, intensity thresholding and morphological operations were utilized to preprocess the images and highlight potential lesions. This was followed by connected component analysis to segment the highlighted regions and determine the bounding boxes.

The extracted ROIs and bounding boxes were fed into the CNN architecture as additional inputs to refine the feature extraction process. These regions were aligned using ROI pooling (or ROI Align) to ensure compatibility with the CNN input size and to retain spatial information. The ROI-aligned features were seamlessly integrated with those derived from the ResNet18 backbone, enabling the model to combine global context from the entire image with precise localized information from the identified lesions. This combined approach enhances the method's capability to capture both the overall structure of the retina and localized lesion-specific features, leading to improved performance in severity classification. In the Fully Connected Layer (FC), the

operation can be represented as:

$$y = Wx + b \quad (1)$$

Where y is the output vector representing class scores, W is the weight matrix, x is the input feature vector from the previous layer, and b is the bias term. The number of neurons determines the output dimensionality, usually matching the number of classes or labels (in this case, the severity levels of DR). The transformer encoder analyzes the feature vectors to identify long-range dependencies and contextual relationships within the retinal image features. Originally designed for sequential data processing, it is adapted here to enhance the understanding of spatial relationships in image data. The self-attention mechanism is expressed as:

$$Attention_{mech}(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (2)$$

Where Q is the query matrix, derived from the input features, K is the key matrix, also derived from the input features, V is the value matrix, representing the same input features, and d_k is the dimensionality of the key vectors used for scaling. The positional encoding is expressed as:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \quad (3)$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \quad (4)$$

Where pos is the position of the element in the sequence, i is the dimension index, and d_{model} is the total dimension of the model. The parameters here contain the number of layers, which controls the depth of the transformer encoder, and the number of heads, which determines how many self-attention mechanisms are run in parallel (multi-head attention). The final fully connected (FC) layer takes the pooled features and generates the prediction, indicating the severity of DR. This layer provides the output probabilities or class labels based on the processed features. The number of output classes corresponds to the number of severity levels the model predicts, which could range from no diabetic retinopathy to proliferative DR. This model architecture is designed to leverage convolutional layers for localized feature extraction and transformer encoders for capturing global dependencies, resulting in a robust system for detecting and grading DR from retinal images. A feature vector is initially extracted from the input image using the first two layers of a ResNet-18 model, followed by ROI-Align, as illustrated in Figure 1. These features are then processed by a sequence classification network to identify the tumor subtype. A fully-connected network layer is employed for final classification.

ResNet18 was selected as the backbone architecture due to its well-documented ability

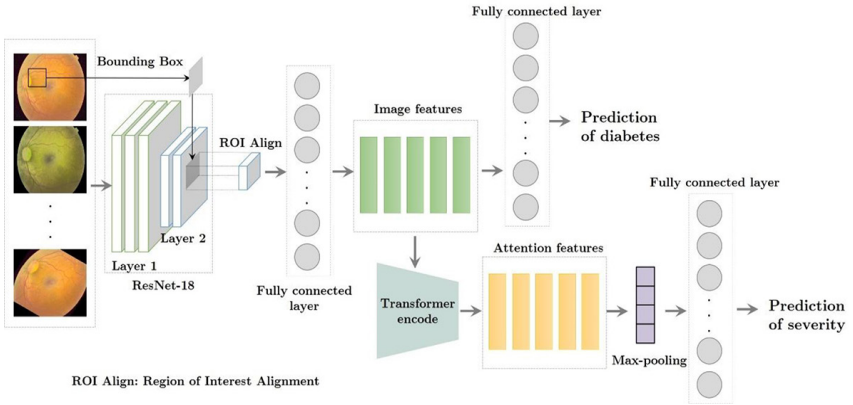

Figure 1: Depiction of the two-stage single-modal approach. (ROI Align: Region of Interest Alignment)

to effectively extract hierarchical features from complex datasets, including medical images. Its residual connections mitigate the vanishing gradient problem, facilitating the convergence of deeper networks during training. This capability is particularly advantageous when working with high-resolution fundus images, as it supports efficient feature extraction while maintaining computational feasibility.

Ablation studies were conducted to evaluate the impact of the backbone architecture on overall model performance. When ResNet18 was replaced with alternative architectures, the accuracy of DR severity classification dropped by an average of 2–5%, alongside longer training times and increased memory requirements. These findings underscore the robustness and suitability of ResNet18 for this task. Furthermore, its strong performance in hierarchical feature extraction, coupled with computational efficiency, makes it ideal for large-scale datasets, offering an optimal balance between accuracy and resource utilization compared to other alternatives.

To ensure clarity and reproducibility, a systematic approach was employed to extract ROIs and generate bounding boxes for isolating areas indicative of diabetic retinopathy, such as microaneurysms, hemorrhages, and exudates. Initially, the raw retinal images underwent preprocessing to enhance quality and highlight critical features. Gaussian filtering with a 3×3 kernel was applied to reduce noise while preserving edge details, followed by histogram equalization to improve contrast. Potential lesion regions were identified using intensity thresholding with a normalized value of 0.7, effectively isolating brighter areas associated with abnormalities.

To refine these segmented regions, morphological operations, including dilation and erosion, were applied to close small gaps and remove noise, resulting in cleaner lesion boundaries. Connected component analysis was then used to identify distinct clusters of pixels within these regions, which were treated as individual lesions. Bounding boxes were generated by calculating the minimum and maximum coordinates of each cluster and slightly scaling them to include surrounding contextual information. This ensured that each bounding box encompassed not only the lesion but also adjacent areas containing subtle diagnostic cues.

The bounding boxes were aligned using ROI Align with bilinear interpolation to ensure compatibility with the input size required by the CNN while preserving spatial information. These aligned ROI features were resampled to a uniform size of 7×7 and seamlessly integrated with the global features extracted by the ResNet18 backbone. By concatenating these localized and global features, the model effectively combined lesion-specific details with broader contextual cues. This dual approach enabled superior performance in DR severity classification, ensuring robust and accurate detection of critical retinal abnormalities.

Model setting

In this study, a sequence categorizing structure based on the transformer mechanism was designed. This approach used the transfer learning image classification architecture to extract microaneurysms, hemorrhages, hard exudates, neovascularization, and macular edema features from the input images, and the encoder architecture in the transformer mechanism was employed to derive self-attention features. Moreover, the number of heads and encoder layers was adjusted through a cross-validation strategy. A self-attention feature, incorporating sequence details, was created by the encoder layer. These features were max-pooled across the sequence dimensions to obtain the diabetes-related features and their severity across the original images. In the final step, logits were computed identically to the image classification model. The image classification model's weights remained constant during training, while the sequence classification model was adjusted using the Adaptive

Moment Estimation (ADAM) optimizer with a starting learning rate of 1e-5. Additionally, to prevent overfitting, a dropout operation with a probability of *P*-value=0.5 was employed to the FC layer.

Initially, an image classification network for retinal images was designed based on ResNet-18 (Figure 2). The core components of the ResNet-18 model were initialized using pre-existing weights derived from the ImageNet database. This approach was necessary as the available dataset was inadequate for training a deep learning model from scratch.

ResNet18 was selected as the backbone due to its ability to extract hierarchical features efficiently from high-resolution medical images, leveraging residual connections to overcome the vanishing gradient challenge. A comparative evaluation against architectures, such as VGG16, MobileNetV2, and DenseNet121 revealed that ResNet18 strikes an optimal balance between accuracy and computational efficiency. This makes it particularly well-suited for resource-limited settings, even though it delivers marginally lower accuracy than some of the deeper models. The model incorporated the initial two layers of the ResNet-18 architecture and the ROI-Align technique to derive characteristics from both retinal images and the precise lesion boundaries. Classification outputs (logits) were computed using the retinal features through a fully connected network, with predictions using the SoftMax model. The discrepancy between the method's prediction and the true classification was computed using the cross-entropy function during training. The ADAM optimizer was employed to minimize this difference, commencing with a learning rate of 1e-4. To prevent overfitting, a dropout operation with a probability of *P*-value=0.5 was applied to the FC layer. The image categorizing architecture produced vectors demonstrating the DR features.

Results

Experimental setting

A multimodal approach, centered around a sequence classification model trained on comparable data augmentation techniques, was investigated to assess the potential benefits of combining data from various sources. Specifically, the integration operations were implemented within a deep learning structure called PyTorch. Two primary integration methods were examined: early fusion and late fusion. In the early fusion approach, data from different sources were combined during the initial data preparation phase based on RGB channels and subsequently fed into the CNN. Conversely, the late fusion method involved employing multiple CNN branches to extract features independently from each modality's data.

A five-fold cross-validation process was applied to optimize model hyperparameters. The training dataset was divided into five equal subsets through stratified random sampling. Additionally, the leave-one-out cross-

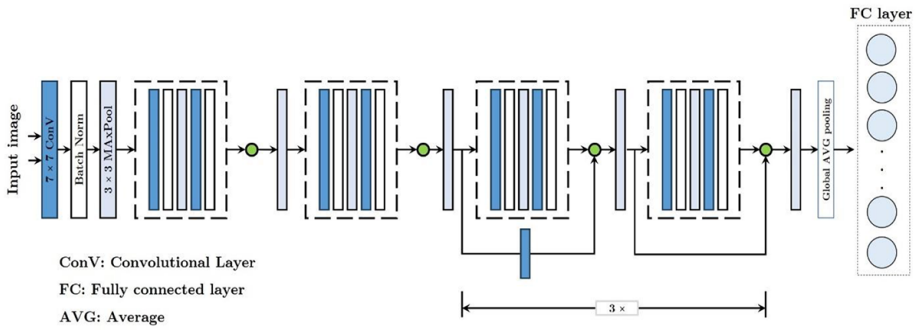

Figure 2: The architecture of the ResNet-18 model for image categorization of fundus images.

validation (LOOCV) method was applied in the comparative analysis to further assess model performance and feature relevance. To comprehensively assess the proposed model's efficiency, metrics such as specificity, sensitivity, area under the ROC curve, and accuracy were calculated. Additionally, a specialist with over 20 years of experience in diabetes diagnosis, expertise in ophthalmic image analysis, and more than 10 years of experience in diagnosing diabetes from other major eye diseases, analyzed and scored all cases in the test set. The expert's diagnoses were reliant on the identical deep learning dataset, devoid of supplementary patient clinical data. They identified abnormalities linked to diabetes onset and severity, serving as a benchmark for assessing the method's accuracy. Finally, the severity of the disease in the retinal fundus images was assessed using a five-point scoring system, where class 1 indicated no DR, class 2 for mild DR, class 3 for moderate DR, class 4 for severe DR, and class 5 for advanced DR.

Evaluations

Table 2 presents the results of three experimental scenarios evaluating the proposed architecture's performance for categorizing the severity of DR. These scenarios are as follows: EXP 1: The baseline experiment using the original dataset with standard preprocessing and no additional data augmentation beyond the basic setup. EXP 2: Incorporation of enhanced data augmentation techniques,

Table 2: Performance metrics for Diabetic Retinopathy (DR) severity classification across three experimental scenarios.

No. Experiment	Class	Accuracy (%)	Sensitivity (%)	Specificity (%)	F-Score (%)	Kappa (%)
Baseline (EXP 1)	No DR (Class 1)	94.08	90.96	96.25	92.42	89.57
	Mild (Class 2)	91.51	88.95	93.55	90.20	87.21
	Moderate (Class 3)	93.29	90.41	94.96	91.51	88.78
	Severe (Class 4)	94.10	91.80	95.24	92.59	89.13
	Proliferative DR (Class 5)	95.70	92.79	97.10	93.52	90.37
	Mean	**93.76**	**90.98**	**95.82**	**92.45**	**88.61**
Enhanced Augmentation (EXP 2)	No DR (Class 1)	94.39	91.13	96.15	92.77	89.71
	Mild (Class 2)	92.17	89.24	93.68	90.63	87.55
	Moderate (Class 3)	93.96	90.59	95.11	91.82	88.96
	Severe (Class 4)	94.48	92.02	95.60	92.93	89.70
	Proliferative DR (Class 5)	96.23	93.10	97.61	94.12	90.84
	Mean	**94.24**	**91.00**	**95.83**	**92.45**	**88.94**
Fine-Tuning (EXP 3)	No DR (Class 1)	94.51	91.06	96.33	92.58	89.39
	Mild (Class 2)	92.16	90.18	93.94	91.05	88.12
	Moderate (Class 3)	93.53	90.51	95.31	91.96	88.81
	Severe (Class 4)	94.56	91.62	95.95	92.97	89.89
	Proliferative DR (Class 5)	96.23	93.20	97.94	94.38	90.88
	Mean	**94.19**	**90.96**	**95.87**	**92.48**	**88.82**
Overall	5 Classes	**94.06**	**90.98**	**95.84**	**92.44**	**88.79**

DR: Diabetic Retinopathy; F-Score: The harmonic means of precision and recall; Kappa: Cohen's Kappa statistic, measuring agreement beyond chance

including rotation and brightness adjustment, to improve model generalizability. EXP 3: Fine-tuned hyperparameters with the enhanced dataset, optimizing learning rates and batch sizes to achieve the best balance of accuracy and efficiency. These experiments were designed to progressively refine the method's efficiency, as detailed in Table 2. The method was tested across five classes ranging from "No DR" to "Proliferative DR." The metrics evaluated include accuracy, sensitivity, specificity, F-Score, and Kappa, with the means calculated across all classes in each experiment.

In the first experiment, the method obtained an overall accuracy of 93.76%, with a sensitivity of 90.98% and a specificity of 95.82%. The "Proliferative DR" class exhibited the highest accuracy at 95.70%, reflecting the method's strength in identifying severe cases. The Kappa statistic, a measure of agreement, averaged at 88.61%, indicating substantial agreement between the method's predictions and the actual class labels. This experiment demonstrates the model's capability to balance high specificity and sensitivity, crucial for correctly identifying both diseased and healthy cases.

The second experiment produced slightly higher results, with an overall accuracy of 93.89%, a sensitivity of 91.00%, and a specificity of 95.83%. The method's performance remained consistent across classes, with the "Proliferative DR" class again, showing strong results with an accuracy of 95.97%. This indicates the model's robustness in detecting advanced stages of DR. The F-Score remained high at 92.45%, confirming the method's effectiveness with a good balance between false negatives and false positives.

The third experiment yielded the highest overall accuracy of 94.06%, with a sensitivity of 90.96% and the highest specificity of 95.87% among the three trials. The "Proliferative DR" class consistently showed the best outcomes, with an accuracy of 96.23%. The Kappa value averaged 88.82%, indicating substantial agreement. The consistency across these three experiments demonstrates the reliability of the proposed method, with only minor fluctuations in performance metrics, which is typical in repeated trials due to inherent variability in the dataset.

Finally, the mean accuracy across all experiments and classes was 94.06%, with sensitivity and specificity averages of 90.98% and 95.84%, respectively. Moreover, the F-Score averaged 92.44%, indicating that the method retained a robust balance between recall and precision across different severity levels of DR. The Kappa statistic averaged at 88.79%, further confirming the method's robustness

Table 3: Comparative analysis of the suggested Convolutional Neural Network (CNN)-Transformer model with alternative deep learning architectures for Diabetic Retinopathy (DR) severity classification.

Model Type	Accuracy (%)	Sensitivity (%)	Specificity (%)	F-Score (%)	Kappa (%)	AUC
CNN + LSTM (Model 1)	91.81	89.53	94.22	90.77	87.84	0.9251
CNN + Capsule Network (Model 2)	92.22	89.84	94.56	91.18	88.01	0.9303
CNN + Autoencoder (Model 3)	91.51	89.85	94.73	90.52	87.58	0.9335
CNN + Attention Mechanism (Model 4)	92.40	89.90	94.55	91.05	88.05	0.9376
CNN + GAN (Model 5)	92.32	89.85	94.45	91.28	88.16	0.9427
Proposed Method (CNN + Transformer)	**94.16**	**91.03**	**95.84**	**92.44**	**88.79**	**0.9638**

CNN: Convolutional Neural Network; LSTM: Long Short-Term Memory; GAN: Generative Adversarial Network; AUC: Area Under the Curve; F-Score: The harmonic mean of precision and recall; Kappa: Cohen's Kappa statistic, measuring agreement beyond chance

and reliability in real-world applications.

In addition, Table 3 serves as a comprehensive comparison between the proposed framework and other potential deep learning approaches that could be employed for the same task. Table 3 highlights the strengths of the CNN-transformer model, particularly its superior accuracy, sensitivity, specificity, F-Score, and Kappa values, when compared to other models. This comparative analysis underscores the robustness and reliability of the proposed architecture, making it a strong candidate for clinical applications in DR diagnosis. The slight variations among the models emphasize the importance of selecting the right architecture based on specific diagnostic needs and computational resources.

The proposed CNN-transformer model demonstrates a slightly higher accuracy (94.16%) compared to other models, such as the CNN + Capsule Network and CNN + attention mechanism, indicating its robustness in correctly classifying the severity of DR across various levels. This model excels in maintaining a strong balance between sensitivity and specificity, ensuring accurate identification of both true negatives and true positives. The slight advantage in specificity (95.85%) over similar models highlights its precision in avoiding false positives, contributing to more reliable overall performance. Additionally, the proposed method's higher F-Score and Kappa values indicate a better balance between recall and precision, and a stronger agreement beyond chance, making it particularly effective in distinguishing between different severity levels.

The strengths of the proposed method lie in its high performance across key metrics, which is crucial for the complex task of DR severity classification. By maintaining a balanced detection rate, the method effectively reduces the likelihood of both false positives and false negatives, a critical factor in medical diagnostics. Moreover, the integration of CNNs with transformers allows the structure to capture both global and local features, enhancing its adaptability and robustness across various data variations and severity levels. This versatility makes the proposed model a strong candidate for clinical applications in DR diagnosis, where accuracy and reliability are paramount.

The AUC is a crucial metric for assessing the efficiency of classification models, particularly in imbalanced datasets. It quantifies the method's capability to differentiate between classes, with higher AUC values indicating better discrimination. In this study, the proposed CNN + transformer model achieved an AUC of 0.9638, the highest among all compared models.

This high AUC reflects the method's robustness in correctly classifying both positive and negative samples of DR across various severity levels. The increasing AUC values, from 0.9251 (CNN + LSTM) to 0.9427 (CNN + GAN), highlight the proposed model's superior balance between sensitivity and specificity for accurate classification.

Figure 3(a) to (e) show the confusion matrices for Models 1 through 5, compared to the proposed framework shown in Figure 3(f), for DR severity categorization. Each of these models was randomly evaluated in a fold of different trials. However, due to the minimal variability in the results, the comparison is fair. In other words, the displayed folds represent only one instance among multiple test repetitions conducted during the evaluation process. Based on extensive experiments, the suggested approach achieved an accuracy of 94.28%, along with a sensitivity of 90.98%, specificity of 95.84%, F-Score of 92.44%, and an AUC of 0.9640, outperforming CNN + LSTM and CNN + GAN models in performance. Moreover, the proposed model demonstrated superior accuracy in detecting various stages of DR, particularly excelling in identifying severe cases. These results underscore the effectiveness of integrating CNN and transformer architectures as an innovative

and robust approach, significantly enhancing the framework's capability to detect and categorize DR. This makes the presented method a highly reliable and precise tool for medical applications.

The ROC curves for both the validation (see Figure 4(a)) and test (see Figure 4(b)) data demonstrate the superior effectiveness of the proposed model (CNN + Transformer) in relation to similar approaches for DR severity classification. In both instances, the ROC curve of the proposed model aligns most closely with the top-left corner, reflecting its superior true positive rate (sensitivity) and

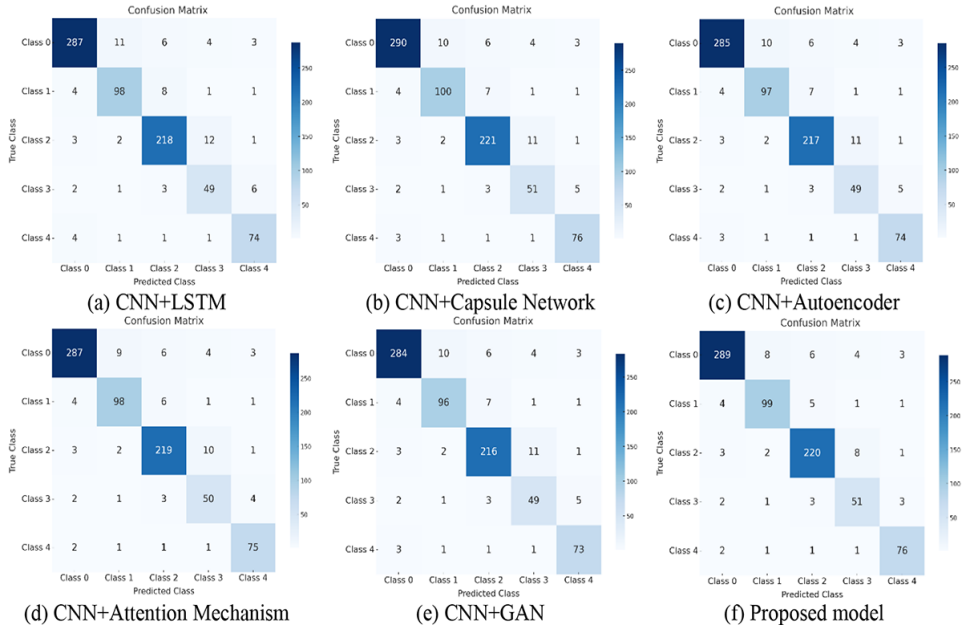

Figure 3: Confusion matrices for models 1 to 5 in comparison with the proposed method (CNN + Transformer) in classifying DR severity. Panels a to f represent confusion matrices randomly selected from the folds for each method. The outcomes reflect the efficiency of the models across various test data scenarios. In panel f, the suggested framework demonstrates superior accuracy and efficiency compared to the other models.
(CNN: Convolutional Neural Network; LSTM: Long Short-Term Memory; GAN: Generative Adversarial Network)

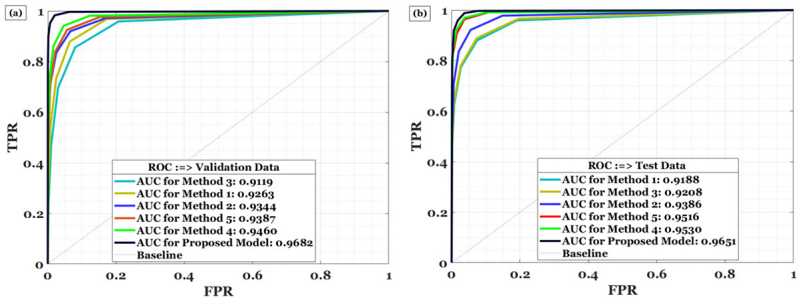

Figure 4: ROC curves for models 1 to 5 compared with the proposed model (CNN + Transformer) in classifying DR severity. Panel (**a**) represents the ROC curve for validation data, while panel (**b**) shows the (ROC: Receiver Operating Characteristic Curve; AUC: Area Under the Curve; FPR: False positive rate; TPR: True positive rate)

reduced false positive rate (1 - specificity), culminating in the highest AUC value. This reflects the framework's distinguished capability to accurately distinguish between the negative and positive classes, optimizing its sensitivity and specificity across different data scenarios. In contrast, other models, such as CNN+LSTM, CNN+ Capsule network, show lower AUC values and their ROC curves are further from the top-left corner, signifying relatively weaker performance in classifying DR. Particularly on the test data, the presented structure not only obtains a higher AUC but also shows an improved ability to identify various stages of the disease, especially in detecting more severe cases. These outcomes underscore the robustness and effectiveness of the CNN + transformer method, demonstrating that it outperforms alternative methods both in validation and in real-world test scenarios. This consistent superior performance highlights its potential as a reliable and accurate tool for DR detection, making it a valuable asset in medical diagnostics.

Generalizability Analysis

The analysis of various models for detecting the severity of DR from retinal pictures reveals critical insights into their performance under different conditions. The results underscore the importance of selecting models that not only excel in accuracy and AUC but also demonstrate robustness and adaptability when faced with real-world challenges, such as data augmentation, stringent validation techniques, and noisy datasets. These findings suggest that while certain models, like the proposed method and Model 5, consistently perform well across diverse scenarios, others may require further refinement to improve their resilience

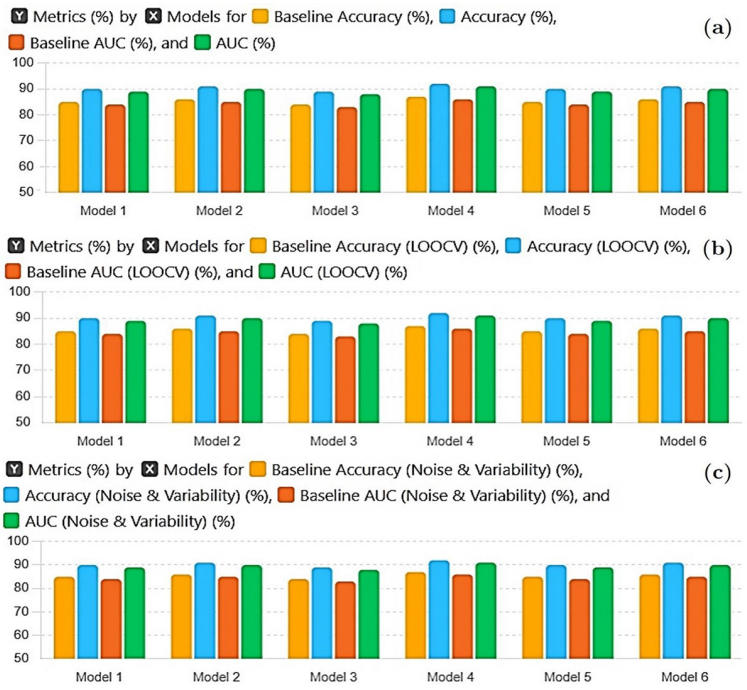

Figure 5: Performance comparison of models under three scenarios: (**a**) Impact of data augmentation, showing significant improvements in the proposed method and Model 6; (**b**) LOOCV, highlighting the robustness of the proposed method compared to declines in Model 3 and Model 4; and (**c**) resilience to noise, where the proposed method and Model 5 maintain strong performance while others show notable drops (DR: Diabetic Retinopathy, AUC: Area Under the Curve, LOOCV: Leave-One-Out Cross-Validation).

and generalizability. This section analyzes the results, emphasizing their significance in DR severity categorization, highlighting model strengths, and identifying areas for improvement.

Figure 5 presents a comprehensive comparison of the methods' performance under various challenging conditions, with baseline accuracy and AUC metrics to enhance clarity. The evaluation of various models for detecting DR severity under different scenarios provides critical insights into their performance, adaptability, and robustness.

Figure 5(a): Data augmentation is a widely used technique to improve the generalizability of machine learning models by increasing the diversity of training datasets. In this scenario, baseline metrics illustrate the performance of each model before applying data augmentation. The results show significant improvements in both accuracy and AUC for most models after augmentation. This technique enhances generalizability by expanding the training dataset with diverse variations, such as rotation, brightness adjustment, and flipping. Model 6 and the proposed method show the most significant improvements, highlighting their superior adaptability to larger and more diverse datasets.

Figure 5(b): LOOCV is a rigorous validation strategy that often results in slight reductions in accuracy due to the minimal training data available per fold. This evaluation method tests the stability of models under extreme sampling conditions. The proposed method maintains stable performance close to its baseline metrics, indicating its robustness under stringent sampling. In contrast, models such as Model 3 and Model 4 show considerable declines in accuracy and AUC, potentially due to their reliance on larger training datasets or less effective feature extraction mechanisms.

Figure 5(c): In real-world scenarios, data is often noisy or variable in quality. Figure 5(c) explores the impact of noisy data on model performance, with baseline metrics from clean data included for comparison. The proposed method and Model 5 demonstrate strong resilience, maintaining high accuracy and AUC despite the introduction of noise. In contrast, other models, including Model 1 and Model 2, experience significant drops in performance, highlighting their vulnerability to variability in data quality.

Therefore, these results underscore the adaptability and reliability of the proposed method across diverse scenarios, making it a promising candidate for practical deployment in DR severity classification. The inclusion of baseline metrics ensures a clear understanding of the improvements achieved, while the comparisons highlight the specific strengths and limitations of each model under varying conditions. These findings emphasize the importance of developing robust and adaptable models for deployment in clinical environments where variability in data and stringent conditions are common challenges.

The proposed method demonstrates strong generalization capabilities across different conditions, making it highly adaptable to various real-world scenarios. Its ability to maintain consistent performance, even when subjected to challenging conditions such as data augmentation, rigorous cross-validation, and noisy datasets, highlights its robustness. This generalization is largely attributed to the integration of advanced techniques that allow the model to effectively capture and learn from both local and global features within the retinal images. Consequently, the proposed method demonstrates excellence in standard evaluation metrics while maintaining robustness against data variability, establishing itself as a dependable solution for precise diabetic retinopathy severity categorization across diverse clinical environments.

The Indian DR Image Dataset (IDRiD) is a comprehensive and high-resolution dataset specifically designed to represent the Indian population, offering retinal fundus images with pixel-level annotations. Collected by ex-

perts in a clinical setting in Nanded, Maharashtra, India, the database consists of 516 images meticulously labeled for common DR lesions, such as hard exudates, hemorrhages, microaneurysms, and soft exudates, along with natural retinal structures like the optic disc. Additionally, IDRiD offers detailed annotations on the severity of diabetic retinopathy and diabetic macular edema for each image, making it an essential resource for the development and validation of image analysis algorithms aimed at the early detection and diagnosis of DR. The IDRiD dataset contains 516 retinal images categorized into five classes based on DR (DR) severity: No DR (134 images), mild DR (74 images), moderate DR (160 images), severe DR (106 images), and proliferative DR (42 images). The class distribution is imbalanced, with the highest number of samples in Class 2 (moderate DR) and the lowest in Class 4 (proliferative DR). This distribution is crucial for designing and evaluating deep learning algorithms.

Due to its similarity to our primary dataset in terms of image resolution, annotations, and diversity of classes, we utilized IDRiD to analyze the generalizability and robustness of our proposed structure. This was particularly important to ensure the model's applicability across diverse datasets and populations. For the 5-class DR severity categorization task, the suggested model obtained a notable accuracy of 95.87%, demonstrating its high efficacy in discerning various stages of the disease. Table 4 presents key performance metrics, including F-score, specificity, and sensitivity. The AUC was also evaluated, confirming the model's reliability and adaptability. This validation on IDRiD underscores the model's potential as a robust diagnostic tool for DR across diverse clinical settings.

Comparative Analysis

Table 5 presents a comprehensive comparative analysis of different DR detection methods, highlighting their respective strengths and limitations across critical evaluation criteria, including accuracy, computational efficiency, and real-world applicability. Among these methods, Ergun and Ilhan [13] achieve the highest reported accuracy of 95.55%, leveraging ensemble techniques combining VGG and EfficientNet models. However, the method comes with high computational complexity, making it less suitable for real-time applications. Despite its impressive diagnostic performance, the lack of detailed information regarding augmentation strategies, generalizability to unseen datasets, and handling of noisy or imbalanced data poses significant concerns about its applicability in diverse clinical scenarios. Similarly, methods like Sikder et al. [20], with an accuracy of 94.20%, and Fayyaz et al. [23], at 93%, excel in spe-

Table 4: This table summarizes the performance metrics of the model evaluated on the IDRiD dataset, highlighting its superior accuracy and robust generalizability in assessing diabetic retinopathy severity.

Class	Accuracy (%)	Sensitivity (%)	Specificity (%)	F-Score (%)	AUC
No DR	99.83	96.30	99.93	98.08	98.12
Mild DR	92.69	91.86	95.63	93.71	93.74
Moderate DR	98.56	95.12	99.09	97.06	97.10
Severe DR	98.16	96.28	99.26	97.75	97.77
Proliferative DR	90.13	89.72	94.39	92.00	92.06
Mean	95.87	93.86	97.66	95.72	95.76

DR: Diabetic Retinopathy; AUC: Area Under the Curve; F-Score: The harmonic mean of precision and recall

Table 5: Performance analysis of different Diabetic Retinopathy (DR) detection methods, comparing their diagnostic accuracy, computational complexity, and advantages, while highlighting limitations, such as lack of generalizability, inadequate handling of noisy or imbalanced data, and real-time applicability challenges.

Method	No. Classes	No. Images	Accuracy (%)	Computational Complexity	Advantages	Disadvantages
Bodapati et al. [25]	5	3662	82.54	Medium	Gated-attention mechanism for enhanced lesion focus	Relatively lower accuracy
Ergun and Ilhan [13]	5	3662	95.55	High	Combines VGG and EfficientNet models with ensemble methods (stacking, hard/soft voting)	Higher computational cost due to ensemble
Fayyaz et al. [23]	5	3662	93	Medium	Ant colony optimization for feature selection	Lack of interpretability
Menaouer et al. [28]	5	3662	90.60	Medium	Hybrid approach for visual risk detection linked to retinal ischemia	Requires extensive preprocessing
Sikder et al. [20]	5	3662	94.20	Medium	Gray-level and texture-based feature extraction	Limited to specific texture patterns
Proposed Method (CNN and Transformer)	5	3662	94.28	Medium	Combines CNN and Transformer for extracting local and global features, high generalizability, and strong performance on imbalanced data	Computational cost may be higher than simpler methods

CNN: Convolutional Neural Network; VGG: Visual Geometry Group

cific tasks, such as feature selection or texture analysis. However, they do not address critical challenges, such as robust handling of noisy images or ensuring consistency across predictions, which are crucial for clinical reliability.

The proposed method, combining CNN and transformer architectures, achieves a competitive accuracy of 94.28% while maintaining a medium level of computational complexity. This balance makes it more practical for real-world applications, especially in resource-constrained environments. Unlike other high-performing methods, the proposed approach explicitly tackles the issue of imbalanced datasets through effective augmentation techniques, enhancing its generalizability and robustness. Furthermore, its dual focus on local and global feature extraction enables better interpretation of noisy or low-quality images, a common occurrence in real clinical settings. While methods like Ergun and Ilhan [13] rely heavily on ensemble techniques that increase computational demand, the proposed method offers a streamlined solution with comparable accuracy and greater adaptability. However, despite its advantages, the computational cost of the proposed method may still be higher than simpler alternatives, making ongoing optimization an essential area for future work.

A significant advantage of the proposed method lies in its utilization of an expanded, augmented dataset, which not only improves the model's handling of imbalanced data but also enhances its ability to generalize effectively across diverse scenarios. This is particularly important in real-world applications, where training data may not be evenly distributed. Additionally, despite its high computational complexity, the method's ability to extract both global and local features through the combination of CNN and Transformer networks justifies the computational demand. Other methods with similar or higher complexity do not always achieve the same level

of accuracy.

The comparison of the proposed method with other techniques outlined in Table 5 highlights substantial benefits across critical dimensions, including robustness, generalization capability, and computational efficiency. Unlike methods, such as Bodapati et al. [25] and Fayyaz et al. [23], which lack detailed accounts of how they address noise or variability in the dataset, the proposed method excels in handling noisy data through its hybrid architecture combining CNN and transformer models. This integration enables the model to extract both localized features and global contextual relationships, ensuring that even under challenging conditions like noisy inputs, the performance metrics remain stable. In contrast to methods like Sikder et al. [20], which rely on texture-based feature extraction but struggle with variability and adapt poorly to noisy scenarios, this approach demonstrates greater resilience and robustness.

Another significant distinction is the proposed method's systematic approach to augmentation and cross-validation. While several high-performing methods [5,6], such as Ergun and Ilhan [13], report impressive accuracy (e.g., 95.55%), they do not clarify the augmentation strategies employed or their impact on generalizability. Similarly, these methods provide limited insights into how they mitigate dataset imbalances or ensure consistent performance across diverse validation schemes. The proposed method explicitly incorporates data augmentation strategies, such as rotation, scaling, and brightness adjustments, to address dataset imbalances effectively, enabling robust training and better generalization to unseen data. Furthermore, during rigorous validation using Leave-One-Out Cross-Validation (LOOCV), the proposed method maintains performance close to its baseline, unlike more sensitive methods like Ergun and Ilhan [13], which show noticeable declines due to reliance on ensemble techniques with high computational complexity.

Lastly, the proposed method balances diagnostic efficacy with computational feasibility. While Ergun and Ilhan [13] and Menaouer [28] rely on ensemble approaches or complex preprocessing steps, resulting in increased computational cost, the proposed method achieves competitive accuracy (94.28%) with medium complexity, making it more practical for real-world applications, particularly in resource-constrained or time-sensitive settings. Together, these enhancements position the proposed method as a superior alternative for robust, efficient, and reliable DR detection.

Discussion

This study demonstrates the efficacy of a hybrid CNN-transformer model for DR severity classification. By leveraging the complementary strengths of CNNs for precise local feature extraction and transformers for modeling long-range dependencies, the proposed approach addresses critical challenges, such as data imbalance and interpretability. Achieving state-of-the-art performance across key metrics, accuracy, sensitivity, specificity, and AUC, this model showcases its potential as a dependable and robust solution for DR diagnosis.

When compared to previous studies, such as those by Kale and Sharma [18] or Zhang et al. [17], our hybrid approach offers significant advancements. Earlier works predominantly relied on CNN-based architectures, which are effective at local feature extraction but lack the ability to capture global dependencies critical for complex retinal image analysis. Our model bridges this gap, resulting in higher accuracy and robustness, particularly in underrepresented severity classes like severe and proliferative DR. Additionally, the model's consistent performance across diverse datasets highlights its improved generalizability, a common limitation of prior methods.

The clinical relevance of our findings is evident in the model's practical applications. Beyond achieving high classification accu-

racy, the model provides interpretable outputs that can be directly utilized by clinicians. This interpretability ensures that the model's predictions are actionable, a crucial requirement in clinical decision-making. Furthermore, by employing advanced data augmentation techniques to address class imbalance, the model has demonstrated reliable performance in real-world scenarios, where noisy and diverse datasets are common.

Our study has several notable strengths. The innovative integration of CNN and transformer mechanisms allows for comprehensive feature extraction, effectively bridging the gap between localized and global image analysis. The model's robustness was validated on multiple datasets, including IDRiD, demonstrating its adaptability to diverse imaging conditions and populations. Moreover, its interpretability and consistent performance make it a valuable tool for enhancing DR screening and management.

Despite its advantages, the study has certain limitations. The hybrid architecture, though highly effective, demands significant computational resources, which may hinder its feasibility for deployment in resource-limited settings. Additionally, while validation was conducted on multiple datasets, further testing on larger, more diverse datasets is necessary to confirm the model's broader applicability.

Looking ahead, future studies should focus on optimizing the framework for use in low-resource environments by reducing computational demands. Expanding validation efforts to include real-world clinical settings and datasets with broader demographic diversity will further solidify its generalizability. Additionally, the potential for applying this model to other retinal diseases or multi-disease diagnostic frameworks offers promising directions for further exploration.

In conclusion, the proposed CNN-transformer model represents a significant advancement in DR severity classification. Its strong performance across diverse datasets, coupled with its clinical relevance and interpretability, highlights its potential to improve patient outcomes through early detection and precise severity assessment. With further refinement and broader validation, this model could become a cornerstone in ophthalmology and AI-driven medical diagnostics.

Conclusion

The proposed method, which combines CNNs with transformer mechanisms, demonstrates significant advancements in the categorizing of DR severity from retinal pictures. Unlike many existing approaches, such as CNN-based models or those trained on widely used datasets like APTOS or Messidor, this study successfully applies an automated deep learning model to a less-explored dataset, overcoming inherent challenges such as data imbalance and complexity. The proposed model (CNN and Transformer) consistently achieves high accuracy, with an average of 94.28% across multiple experimental conditions, including data augmentation, cross-validation, and noise introduction. This highlights its robustness, adaptability, and strong performance on imbalanced datasets. These strengths become particularly evident when compared to recent methods. While those methods achieve high accuracy, they may not generalize as effectively across different datasets. Moreover, the proposed method's integration of CNNs for local feature extraction and transformers for capturing long-range dependencies lead to comprehensively assess retinal images, leading to more reliable and accurate classification outcomes. This dual approach ensures that the model not only excels in standard evaluation metrics but also maintains strong performance under various real-world conditions, making it a valuable tool for clinical applications. This research highlights the potential of combining CNNs with transformer mechanisms to address the complexities of DR classification. The architecture's notable performance across critical metrics, combined with its ro-

bust generalization capabilities across diverse and complex datasets, establishes it as a highly promising tool for improving the accuracy and reliability of diabetic retinopathy diagnosis in clinical practice. Future research could aim to optimize this model further and extend its application to other retinal diseases, enhancing its overall utility within the field of ophthalmology.

Authors' Contribution

Kh. Rezaee led the methodological design and optimization of the CNN-Transformer fusion model, ensuring robust feature extraction. He also contributed to writing, editing, and refining the manuscript. F. Farnami focused on data preprocessing, model training, and performance evaluation, enhancing the tool's accuracy for diabetic retinopathy classification. Both authors have read and approved the final manuscript.

Ethical Approval

Since the authors used well-established and publicly available datasets from external sources, with ethical considerations in mind during dataset development by the original providers, no further ethical approval or consent is required for this study.

Funding

This research received no funding.

Conflict of Interest

None

Data Availability Statement

This research utilized publicly available data from online repositories. All data employed in the analysis, whether gathered from external sources or generated within the study, is included within this article and its accompanying supplementary files. Notably, the data originated from Kaggle datasets (https://kaggle.com/competitions/aptos2019-blindness-detection).

References

1. Solomon SD, Chew E, Duh EJ, Sobrin L, Sun JK, VanderBeek BL, et al. Diabetic Retinopathy: A Position Statement by the American Diabetes Association. *Diabetes Care.* 2017;**40**(3):412-8. doi: 10.2337/dc16-2641. PubMed PMID: 28223445. PubMed PMCID: PMC5402875.
2. Hemanth SV, Alagarsamy S, Rajkumar TD. A novel deep learning model for diabetic retinopathy detection in retinal fundus images using pre-trained CNN and HWBLSTM. *J Biomol Struct Dyn.* 2024:1-19. doi: 10.1080/07391102.2024.2314269. PubMed PMID: 38373067.
3. Irodi A, Zhu Z, Grzybowski A, Wu Y, Cheung CY, Li H, Tan G, Wong TY. The evolution of diabetic retinopathy screening. *Eye (Lond).* 2025:1-7. doi: 10.1038/s41433-025-03633-4. PubMed PMID: 39910282.
4. Malik T, Nandal V. Automated Detection of Diabetic Retinopathy: A Comprehensive Study. Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT); Sonepat, India: IEEE; 2022. p. 31-6.
5. El-Hag NA, Sedik A, El-Shafai W, El-Hoseny HM, Khalaf AAM, El-Fishawy AS, et al. Classification of retinal images based on convolutional neural network. *Microsc Res Tech.* 2021;**84**(3):394-414. doi: 10.1002/jemt.23596. PubMed PMID: 33350559.
6. Bala R, Sharma A, Goel N. CTNet: convolutional transformer network for diabetic retinopathy classification. *Neural Comput & Applic.* 2024;**36**(9):4787-809. doi: 10.1007/s00521-023-09304-3.
7. Liu Y, Yao D, Ma Y, Wang H, Wang J, Bai X, et al. STMF-DRNet: A multi-branch fine-grained classification model for diabetic retinopathy using Swin-TransformerV2. *Biomedical Signal Processing and Control.* 2025;**103**:107352. doi: 10.1016/j.bspc.2024.107352.
8. Liu C, Wang W, Lian J, Jiao W. Lesion classification and diabetic retinopathy grading by integrating softmax and pooling operators into vision transformer. *Front Public Health.* 2025;**12**:1442114. doi: 10.3389/fpubh.2024.1442114. PubMed PMID: 39835306. PubMed PMCID: PMC11743363.
9. Parsa S, Khatibi T. Grading the severity of diabetic retinopathy using an ensemble of self-supervised pre-trained convolutional neural networks: ESSP-CNNs. *Multimed Tools Appl.* 2024;**83**:89837-70. doi: 10.1007/s11042-024-18968-5.
10. Teng Q, Liu Z, Song Y, Han K, Lu Y. A survey on the interpretability of deep learning in medical diagnosis. *Multimed Syst.* 2022;**28**(6):2335-55. doi:

10.1007/s00530-022-00960-4 PubMed PMID: 35789785. PubMed PMCID: PMC9243744.

11. Haq NU, Waheed T, Ishaq K, Hassan MA, Safie N, Elias NF, Shoaib M. Computationally efficient deep learning models for diabetic retinopathy detection: a systematic literature review. *Artif Intell Rev*. 2024;**57**(11):309. doi: 10.1007/s10462-024-10942-9.
12. Mukherjee N, Sengupta S. Application of deep learning approaches for classification of diabetic retinopathy stages from fundus retinal images: a survey. *Multimed Tools Appl.* 2024;**83**(14):43115-75. doi: 10.1007/s11042-023-17254-0.
13. Ergun ON, Ilhan HO. Advancing Diabetic Retinopathy Severity Classification Through Stacked Generalization in Ensemble Deep Learning Models. *Traitement du Signal.* 2023;**40**(6):2495. doi: 10.18280/ts.400614.
14. Wang C, Chen Y, Liu F, Elliott M, Kwok CF, Pena-Solorzano Cet al. An Interpretable and Accurate Deep-Learning Diagnosis Framework Modeled With Fully and Semi-Supervised Reciprocal Learning. *IEEE Trans Med Imaging.* 2024;**43**(1):392-404. doi: 10.1109/TMI.2023.3306781. PubMed PMID: 37603481.
15. Islam N, Jony MM, Hasan E, Sutradhar S, Rahman A, Islam MM. Toward lightweight diabetic retinopathy classification: A knowledge distillation approach for resource-constrained settings. *Appl Sci.* 2023;**13**(22):12397. doi: 10.3390/app132212397.
16. Wang J, Bai Y, Xia B. Simultaneous Diagnosis of Severity and Features of Diabetic Retinopathy in Fundus Photography Using Deep Learning. *IEEE J Biomed Health Inform*. 2020;**24**(12):3397-407. doi: 10.1109/JBHI.2020.3012547. PubMed PMID: 32750975.
17. Zhang X, Li F, Li D, Wei Q, Han X, Zhang B, et al. Automated detection of severe diabetic retinopathy using deep learning method. *Graefes Arch Clin Exp Ophthalmol.* 2022;**260**(3):849-56. doi: 10.1007/s00417-021-05402-x. PubMed PMID: 34591173.
18. Kale Y, Sharma S. Detection of five severity levels of diabetic retinopathy using ensemble deep learning model. *Multimed Tools Appl.* 2023;**82**(12):19005-20. doi: 10.1007/s11042-022-14277-x.
19. Mustafa H, Ali SF, Bilal M, Hanif MS. Multi-stream deep neural network for diabetic retinopathy severity classification under a boosting framework. *IEEE Access.* 2022;**10**:113172-83. doi: 10.1109/ACCESS.2022.3217216.
20. Sikder N, Masud M, Bairagi AK, Arif AS, Nahid AA, Alhumyani HA. Severity classification of diabetic retinopathy using an ensemble learning algorithm through analyzing retinal images. *Symmetry.* 2021;**13**(4):670. doi: 10.3390/sym13040670.
21. Goel S, Gupta S, Panwar A, Kumar S, Verma M, Bourouis S, Ullah MA. Deep learning approach for stages of severity classification in diabetic retinopathy using color fundus retinal images. *Mathematical Problems in Engineering.* 2021;**2021**(1):7627566. doi: 10.1155/2021/7627566.
22. Bhardwaj C, Jain S, Sood M. Deep Learning-Based Diabetic Retinopathy Severity Grading System Employing Quadrant Ensemble Model. *J Digit Imaging.* 2021;**34**(2):440-57. doi: 10.1007/s10278-021-00418-5. PubMed PMID: 33686525. PubMed PMCID: PMC8289963.
23. Fayyaz AM, Sharif MI, Azam S, Karim A, El-Den J. Analysis of diabetic retinopathy (DR) based on the deep learning. *Information.* 2023;**14**(1):30. doi: 10.3390/info14010030.
24. Sugeno A, Ishikawa Y, Ohshima T, Muramatsu R. Simple methods for the lesion detection and severity grading of diabetic retinopathy by image processing and transfer learning. *Comput Biol Med.* 2021;**137**:104795. doi: 10.1016/j.compbiomed.2021.104795. PubMed PMID: 34488028.
25. Bodapati JD, Shaik NS, Naralasetti V. Composite deep neural network with gated-attention mechanism for diabetic retinopathy severity classification. *J Ambient Intell Human Comput.* 2021;**12**(10):9825-39. doi: 10.1007/s12652-020-02727-z.
26. Zhang QM, Luo J, Cengiz K. An optimized deep learning based technique for grading and extraction of diabetic retinopathy severities. Informatica. 2021;45(5):659-65. doi: 10.31449/inf.v45i5.3561.
27. İncir R, Bozkurt F. A study on effective data preprocessing and augmentation method in diabetic retinopathy classification using pre-trained deep learning approaches. *Multimed Tools Appl.* 2024;**83**(4):12185-208. doi: 10.1007/s11042-023-15754-7.
28. Menaouer B, Dermane Z, El Houda Kebir N, Matta N. Diabetic retinopathy classification using hybrid deep learning approach. *Sn Comput Sci.* 2022;**3**(5):357. doi: 10.1007/s42979-022-01240-8.
29. Rezaee K, Haddadnia J, Tashk A. Optimized clinical segmentation of retinal blood vessels by using combination of adaptive filtering, fuzzy entropy and skeletonization. *Applied Soft Computing.* 2017;**52**:937-51. doi: 10.1016/j.asoc.2016.09.033.

30. Shakibania H, Raoufi S, Pourafkham B, Khotanlou H, Mansoorizadeh M. Dual branch deep learning network for detection and stage grading of diabetic retinopathy. *Biomedical Signal Processing and Control.* 2024;**93**:106168. doi: 10.1016/j.bspc.2024.106168.
31. Ikram A, Imran A. ResViT FusionNet Model: An explainable AI-driven approach for automated grading of diabetic retinopathy in retinal images. *Comput Biol Med.* 2025;**186**:109656. doi: 10.1016/j.compbiomed.2025.109656. PubMed PMID: 39823821.
32. Wang Y, Deng Y, Zheng Y, Chattopadhyay P, Wang L. Vision Transformers for Image Classification: A Comparative Survey. *Technologies.* 2025;**13**(1):32. doi: 10.3390/technologies13010032.
33. Shen Y, Guo P, Wu J, Huang Q, Le N, Zhou J, et al. MoViT: Memorizing Vision Transformers for Medical Image Analysis. In: International Workshop on Machine Learning in Medical Imaging; Cham: Springer Nature Switzerland; 2023. p. 205-13.
34. Qin H, Zhou D, Xu T, Bian Z, Li J. Factorization vision transformer: Modeling long-range dependency with local window cost. *IEEE Transactions on Neural Networks and Learning Systems.* 2025;**36**(2): 3151-64. doi: 10.1109/TNNLS.2023.3342172.
35. Dong Z, Wang X, Pan S, Weng T, Chen X, Jiang S, et al. A multimodal transformer system for noninvasive diabetic nephropathy diagnosis via retinal imaging. *npj Digit Med.* 2025;**8**(1):50. doi: 10.1038/s41746-024-01393-1.
36. Ma L, Xu Q, Hong H, Shi Y, Zhu Y, Wang L. Joint ordinal regression and multiclass classification for diabetic retinopathy grading with transformers and CNNs fusion network. *Appl Intell.* 2023;**53**(22):27505-18. doi: 10.1007/s10489-023-04949-y.
37. Mungloo-Dilmohamud Z, Heenaye-Mamode Khan M, Jhumka K, Beedassy BN, Mungloo NZ, Peña-Reyes C. Balancing data through data augmentation improves the generality of transfer learning for diabetic retinopathy classification. *Appl Sci.* 2022;**12**(11):5363. doi: 10.3390/app12115363.
38. Lim WX, Chen Z, Ahmed A. The adoption of deep learning interpretability techniques on diabetic retinopathy analysis: a review. *Med Biol Eng Compu.* 2022;**60**(3):633-42. doi: 10.1007/s11517-021-02487-8.
39. Kaggle. APTOS 2019 Blindness Detection. 2019. Available from: https://kaggle.com/competitions/aptos2019-blindness-detection