VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2026, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 14, Number 1, January-March 2026 pp:124-136

Hybrid Convolutional Transformer Learning Utilizing Ordinal Sensitive Loss for Automated Grading of Diabetic Retinopathy

Muhammad Suleman Memon[1*], **Mumtaz Qabulio**[2], **Nazish Basir**[2], **Asia Khatoon Soomro**[3], **Syeda Hira Fatima Naqvi**[3]

[1]Department of Information Technology, Dadu Campus University of Sindh, Dadu, Pakistan; [2]Department of Software Engineering Faculty of Engineering & Technology,University of Sindh, Jamshoro, Pakistan; [3]Institute of Mathematics & Computer Science, University of Sindh, Jamshoro, Pakistan

Keywords: *Diabetic Retinopathy, Hybrid Deep Learning, Vision Transformer, Ordinal Classification, Medical Image Analysis, Fundus Image Grading.*

Journal Info:
Submitted:
February 01, 2026
Accepted:
March 02, 2026
Published:
March 08, 2026

Abstract Diabetic retinopathy (DR) is a major cause of preventable blindness globally requiring precise and dependable automated grading systems to facilitate extensive screening initiatives. Recent deep learning techniques utilizing convolutional neural networks (CNNs) have yielded encouraging outcomes however, they predominantly concentrate on local lesion identification and frequently neglect to encompass the global retinal context. Furthermore, the majority of current methodologies regard diabetic retinopathy grading as a conventional multi-class classification issue, neglecting the ordinal characteristics of disease severity and the significant class imbalance present in clinical datasets. In this paper, we propose a hybrid convolutional-transformer learning framework with ordinal-sensitive loss for automated diabetic retinopathy grading. The suggested model combines a deep CNN backbone for strong local feature extraction with a lightweight transformer encoder that works on convolutionally down sampled feature maps to model long-range dependencies in a computationally efficient manner. This design reduces the quadratic complexity of self-attention while keeping the global context information needed to figure out how bad something is. To deal with class imbalance and penalties for misclassifying ordinal data, an ordinal-sensitive focal loss is used to make the model focus on clinically important mistakes. We test the framework on publicly available fundus image datasets using a wide range of performance metrics, such as accuracy, macro-F1 score, balanced accuracy, area under the ROC curve (AUC), and quadratic Cohen's kappa. Experimental results show that the suggested method consistently beats CNN-only baselines and standard cross-entropy-based training, getting better accuracy of 85%.

***Correspondence author email address:** msuleman@usindh.edu.pk
DOI: 10.21015/vtse.v14i1.2344

1 Introduction

Diabetic retinopathy (DR) is a significant contributor to vision impairment. Accurate assessment of diabetic retinopathy is very important to guarantee immediate and suitable intervention. DR is classified into two primary stages non-proliferative diabetic retinopathy (NPDR) and proliferative diabetic retinopathy (PDR) [1]. Diabetic retinopathy (DR) is a long-term problem with small blood vessels that can happen in people with diabetes mellitus [2]. It is still one of the main causes of

vision loss that can be avoided around the world [3–5].

The research conducted by the International Diabetes Federation indicates a daily rise in the prevalence of diabetic retinopathy (DR). Diabetic retinopathy (DR) is regarded as the foremost disease jeopardizing human optical health. Symptoms may not be present in diabetic retinopathy. The symptoms include impaired night vision, fluctuating eyesight and an inability to distinguish colors [6].

Early detection [7, 8] and accurate evaluation of diabetic retinopathy (DR) severity are crucial for timely clinical intervention and the avoidance of permanent retinal damage. The eye doctor's manual evaluation of retinal fundus images is time-consuming, subjective and difficult to scale particularly in regions with a limited number of specialists.

These constraints have catalyzed considerable research interest in automated computer-aided diagnostic systems for diabetic retinopathy screening and severity assessment [9, 10].

In the last few years deep learning methods especially convolutional neural networks (CNNs) because of the automatic feature extraction have demonstrated superior performance of automatically finding and classifying DR from retinal fundus images [11]. CNN-based methods are good at learning the unique local features that are linked to pathological retinal lesions like microaneurysms, hemorrhages, and exudates. Localized features are not the only thing that determines the severity of DR. The overall retinal context and the spatial relationships between lesions are also very important for telling apart adjacent levels of disease severity. Because their receptive fields are limited to one area, purely convolutional architectures can't model long-range dependencies between retinal images.

Transformer-based models have gotten a lot of attention because their self-attention mechanisms let them model long-range dependencies and get global contextual information [12, 13] Recently researchers have started using vision transformer and CNN transformer hybrid networks for DR grading. These networks perform better than regular CNNs because they better represent global features [14]. However, directly using transformers on high-resolution fundus images requires a lot of processing power and memory, which makes them less useful in clinical settings. Most current DR grading methods also treat the task as a standard multi-class classification problem which ignores the fact that disease severity is ordinal. For example, misclassifying a severe case as mild is worse than mixing up adjacent grades.

This study proposed a hybrid convolutional-transformer learning framework with an ordinal-sensitive loss function for automated DR grading to deal with these problems. The proposed model combines the strengths of CNNs for extracting local features with the strengths of transformers for modeling global context. Also, an ordinal- sensitive loss function is added to explicitly model the ordered structure of DR severity and reduce class imbalance. This joint optimization strategy makes grading more accurate, makes clinical results more reliable, and makes expert annotations more consistent.

2 Literature Review

To improve DR grading accuracy the author in [1] proposed a deep learning model using frequency attention modules and frequency domain spatial modules. The proposed model was trained on binary dataset. The author in [3] investigated various advanced methods for the automatic detection of DR. The author in [4] proposed self-attention mechanism using swin transformer as the backbone. The author in [5] implemented a hybrid approach for the detection DR. The multilayer perceptive learning employed in [6] for the detection of DR. The suggested technique used different image preprocessing methods at different hidden layers. The author in [7] proposed hybrid model which uses CNN and RNN achieving accuracy of 97.5% on eyepacs dataset. The author in [8] conducted a short survey which includes the articles from 2018 to 2023.

In most articles, the transfer learning was chosen as a good strategy for DR. The datasets included APTOS and EyePACS. The author in [9] used different pre-trained models. The author in [10] proposed a hybrid technique which involved Multi-Scale Discriminative Robust Local Binary Pattern (MS-DRLBP) features integrated with a hybrid Convolutional Neural Network-Radial Basis Function (CNN-RBF). Author employed a binary class dataset for classification. Different preprocessing tech-

niques were used by the author which include noise reduction and morphological operations. The proposed study uses two classes normal and diabetic retinopathy. The model achieved an accuracy of 96.10%. The paper [11] proposes an explainable deep learning framework for automatic diabetic retinopathy (DR) grading using a modified ResNet-50 architecture combined with transfer learning, fine-tuning and regularization techniques. To enhance transparency the authors integrate SHAP to provide visual explanations of the model's predictions.

The model was trained on the APTOS-2019 dataset and validated across four additional public datasets demonstrating strong generalization performance. Importantly the SHAP analysis revealed clinically meaningful insights highlighting retinal vasculature changes and peripheral lesions as significant indicators of DR progression. The author of [12] proposed a vision transformer model to detect blindness from IDRiD dataset achieving an accuracy of 0.825. To detect DR the author in [13] proposed a CNN architecture with mamba. The proposed model uses mamba for resource constraint devices.

The vision transformers [14] show a great potential for detecting DR. They are used for various problems such as vessel segmentation and fovea localization. The author in [15] proposed MobileNetV2 and Graph Convolution Network. The study involved dataset from APTOS 2019 which contains five classes. The proposed study achieved a validation accuracy of 82.5%. The author in [16] proposed a hybrid approach using UNet++ and Capsule Network for the detection of diabetic retinopathy. A small-scale retinal dataset was trained using the hybrid approach. Different preprocessing techniques were used including CLAHE and histogram equalization to improve the quality of the image. The proposed methodology achieved an accuracy of 97.7% in image classification and dice coefficient of 0.92 and IOU score of 0.85 in image segmentation. The author in [17] worked on pretrained model DenseNet121 with Bayesian approach with approximation methods like MC dropout, MFVI, and Deterministic for the classification of diabetic retinopathy.

The suggested study used APTOS dataset which is publicly available on Kaggle. The second dataset selected for the training was DDR which include 13,673 fundus images. The proposed study combined both the datasets. The author in [18] proposes a deep learning based framework (RSG-Net) for automatic diabetic retinopathy (DR) grading using color fundus images. The model performs both 4-class severity grading (Normal, Mild, Moderate, Proliferative) and binary classification (DR vs. No-DR). The experiments are conducted on the Messidor-1 dataset (1200 images). The authors design a custom CNN architecture from scratch consisting of four convolutional layers, max-pooling, batch normalization, dropout and fully connected layers. Extensive image preprocessing is applied including bounding-box, cropping to remove black background, Gaussian blur for denoising, Histogram Equalization (HE) in YUV color space and image resizing to 200×200.

To address class imbalance, aggressive data augmentation is used for increasing the dataset size to 8304 images (4-class) and 4800 images (binary). The model is trained using a 70:10:20 split. The author in [19] uses a transfer learning method using DenseNet and integrated it with raspberry pi. The suggested study was trained on Messidor-1 dataset. The proposed approach was trained solely on Messidor-1 by applying offline data augmentation and forming new dataset with much balanced classes. The proposed method achieved an accuracy of 88%. The author in [20] proposed a model using CNN. The model was trained on two datasets APTOS and DDR. Initially the data processing and data augmentation techniques were used. In data processing the CLAHE was used. The proposed strategy achieved an accuracy of 71%.

The author in [21] proposed a model DRNet13. The proposed model contains three sets of convolutional layers and pooling layers, a normalization layer, two fully connected layers (dense layers), a dropout layer, and the output layer. The author applied data augmentation techniques to expand the dataset. Initially the size of the dataset was 3662 which was increased to 7500. The proposed strategy was applied to a balanced class.

A summary of the key studies, datasets, models, and their limitations is presented in Table 1.

2.1 Transformer-Based Methods for Diabetic Retinopathy

Recent research has increasingly investigated transformer-based architectures for the classification of diabetic

Table 1. Summary of the Literature

Ref.	Dataset	Model	Task	Acc.	Limitations
[9]	Binary Fundus Dataset	VGG16, ResNet50, InceptionV3, MobileNet, DenseNet121	Classification	93.0%	Binary classification only; disease severity levels not considered.
[10]	Binary Class	MS-DRLBP, CNN-RBF, Otsu thresholding	Classification	96.10%	The author trained on binary dataset only.
[15]	Multiclass (APTOS)	DenseNet201, ResNet50, VGG19, MobileNetV2 and Ensemble Method	Classification	82.5%	The study only utilizes the APTOS 2019 Blindness Detection dataset lacking external validation against other prevalent benchmarks like EyePACS, Messidor, IDRiD or DIARETDB1.
[16]	Small Retinal Dataset	UNet++, Capsule Network	Classification and Segmentation	97.7%	The study employs a small dataset this can cause issues for generalizability.
[17]	APTOS+DDR	DenseNet121, Bayesian	Classification	97.68%	The study employs training on a combined dataset with much more balanced dataset. The study was not on evaluated on APTOS dataset solely.
[18]	Messidor-1	custom CNN architecture	Classification	99.36%	There is no testing on EyePACS, APTOS, DDR, or IDRiD, making the results dataset-specific and limiting clinical applicability. The paper states that augmentation increased the dataset size and then a 70:10:20 split was applied. This strongly suggests that augmented versions of the same image may appear in both training and testing sets.
[19]	Augmented APTOS Dataset	DenseNet	Classification	88%	The author used a balanced dataset after augmentation. The model was not evaluated on APTOS data, EyePACS and other datasets.
[20]	APTOS, DDR	CNN	Classification	71%	The proposed model achieved a less accuracy.

retinopathy (DR) owing to their capacity to model long-range dependencies. Vision Transformer (ViT) models have been utilized on fundus images by segmenting them into fixed-size patches and acquiring global contextual representations. ViT-based models showed promise in terms of performance, but they often need big datasets and a lot of computing power which makes them hard to use in clinical settings.

It has been suggested that hybrid CNN Transformer architectures could combine the strong ability of CNNs to extract local features with the strong ability of transformers to model the whole picture. For instance, some studies used CNN backbones like ResNet or EfficientNet to get feature maps, and then transformer encoders to find global relationships between different parts of the retina. Even though these methods got better at making predictions, most of them didn't take into account the fact that DR severity levels are ordered and didn't include a full explainability analysis.

The proposed model is different from other methods because it has a lightweight transformer encoder that works on tokens that are generated on the fly. This makes it possible to model global context efficiently while keeping the computational complexity low.

3 Methodology

3.1 Dataset Details

The APTOS dataset used in the study was downloaded from a Kaggle. The dataset contains total five classes from 0 to 4. Where 0 for No DR, 1 for Moderate, 2 for Mild, 3 for Proliferate DR and 4 for Severe. The APTOS dataset normally is much imbalanced dataset.

Figure 1 shows different sample images and Figure 2 shows the visual graph of different classes.

3.2 Proposed Framework

Deep learning plays a key role in medical image segmentation such detecting tumors from the brain images [22].

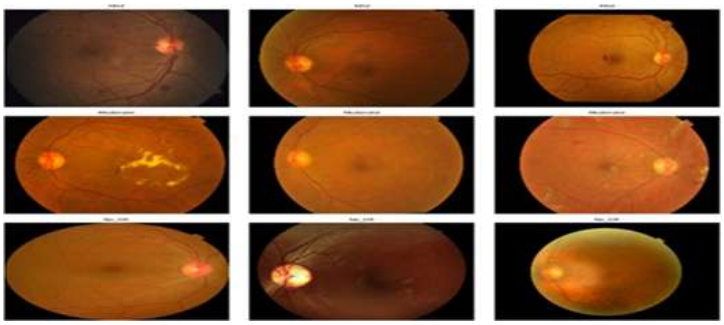

Figure 1. Sample Images

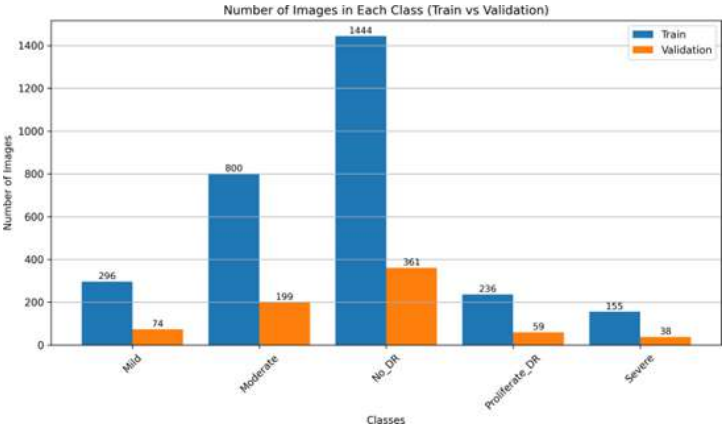

Figure 2. Train and Test Image Distribution

The proposed framework has three main parts a backbone for extracting features using convolution, a module for modeling global context using a transformer and a head for learning and classifying that is sensitive to ordinal data. The proposed framework presents an attention-refined tokenization technique in contrast to traditional hybrid architectures that tokenize backbone feature maps directly. Convolutional Block Attention Modules (CBAM) are specifically used to improve convolutional features that were taken from the EfficientNetB5 backbone.

This allows for spatial–channel recalibration before token formation.Additionally, in order to reduce computational complexity while maintaining contextual modeling capacity, the transformer branch is designed as a lightweight encoder with reduced depth. Complementary representation learning is ensured by combining transformer-encoded contextual tokens with globally pooled CNN descriptors.Furthermore, inter-class severity relationships which are generally disregarded in categorical cross-entropy optimization are explicitly encoded through the integration of an ordinal-sensitive loss formulation.

Figure 3 shows an overview of the architecture.

3.2.1 Convolutional Feature Extraction

Let the input retinal fundus image be defined as:

$$I \in \mathbb{R}^{224\times224\times3} \tag{1}$$

A pretrained EfficientNet backbone extracts spatial features:

$$F_c = \phi_{\text{CNN}}(I) \tag{2}$$

where ϕ_{CNN} represents the convolutional feature extractor and

$$F_c \in \mathbb{R}^{H\times W\times C}.$$

Global Average Pooling (GAP) is then applied:

$$f_c = \text{GAP}(F_c) \tag{3}$$

$$f_c \in \mathbb{R}^{2048} \tag{4}$$

This produces the local spatial feature representation.

3.2.2 Transformer-Based Context Modeling

To capture global contextual relationships, the convolutional feature map is tokenized.

The spatial features are flattened into tokens

$$T = \text{Flatten}(F_c) \tag{5}$$

$$T \in \mathbb{R}^{N\times d} \tag{6}$$

where N denotes the number of tokens and d represents the embedding dimension.

The self-attention mechanism is defined as:

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

where

$$Q = TW_Q, \quad K = TW_K, \quad V = TW_V.$$

The transformer encoder output is expressed as:

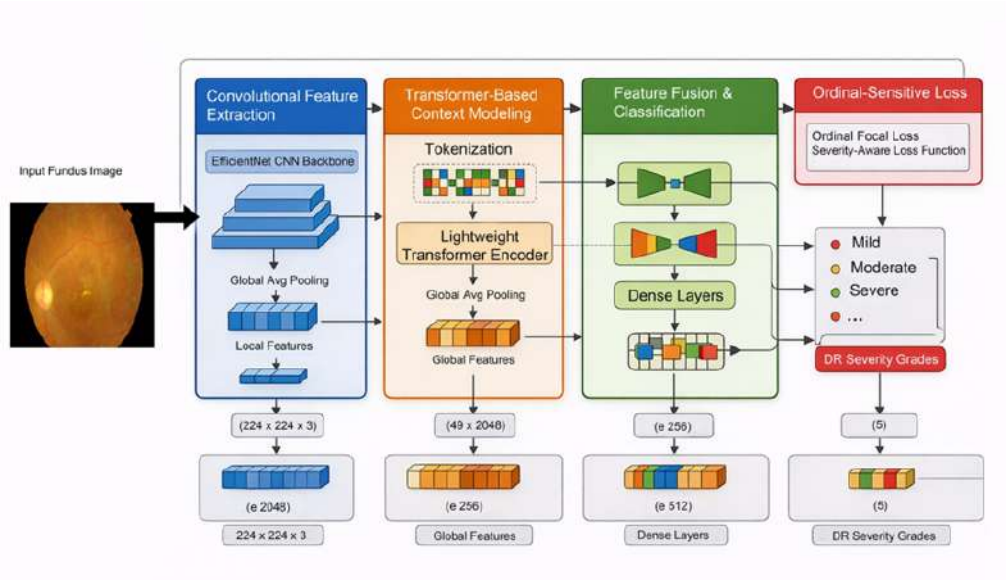

Figure 3. Proposed Model

$$F_t = \phi_{\text{Transformer}}(T) \tag{8}$$

After global pooling:

$$f_t = \text{GAP}(F_t) \tag{9}$$

$$f_t \in \mathbb{R}^{256} \tag{10}$$

3.2.3 Feature Fusion and Classification

The CNN and transformer features are concatenated:

$$f_{\text{fusion}} = [f_c; f_t] \tag{11}$$

$$f_{\text{fusion}} \in \mathbb{R}^{2304} \tag{12}$$

The fused feature vector is passed through fully connected layers:

$$z = W_1 f_{\text{fusion}} + b_1 \tag{13}$$

$$z = \sigma(z) \tag{14}$$

The final output logits are:

$$y = W_2 z + b_2 \tag{15}$$

Softmax classification is defined as:

$$\hat{y}_i = \frac{e^{y_i}}{\sum_{j=1}^{5} e^{y_j}} \tag{16}$$

$$\hat{y} \in \mathbb{R}^5 \tag{17}$$

3.2.4 Ordinal-Sensitive Focal Loss

Since diabetic retinopathy severity follows an ordinal progression (0 < 1 < 2 < 3 < 4), an ordinal penalty is introduced.

The standard cross-entropy loss is defined as:

$$L_{\text{CE}} = -\sum_{i=1}^{5} y_i \log(\hat{y}_i) \tag{18}$$

The focal loss is given by:

$$L_{\text{FL}} = -\alpha(1 - \hat{y}_t)^{\gamma} \log(\hat{y}_t) \tag{19}$$

The ordinal distance penalty is defined as:

$$d = \left| y_{\text{true}} - y_{\text{pred}} \right| \tag{20}$$

The final loss function is expressed as:

$$L = L_{\text{FL}} + \lambda d \tag{21}$$

where λ controls the severity penalty.

3.2.5 EfficientNetB5 Backbone Configuration

The backbone network is EfficientNetB5 because it works better when you scale the depth, width and resolution all at once. The network starts with ImageNet pre-trained weights and works with fundus images that have been resized to 224 × 224 pixels. The last convolutional block makes a feature map that is 7 × 7 × 2048 pixels and has a lot of semantic information. To improve feature representation even more a Convolutional Block Attention Module (CBAM) is used to highlight areas that are important for diagnosis.

3.2.6 Lightweight Transformer Encoder

The transformer encoder operates on dynamically generated tokens obtained by reshaping the convolutional feature map into a sequence of 49 tokens, each corresponding to a spatial region of the retina. A multi-head self-attention mechanism with four attention heads is employed, where each head uses a key dimension of 64. This allows the model to learn global contextual relationships between retinal regions. A feed-forward network with a hidden dimension of 512 follows the attention layer enabling nonlinear feature transformation while maintaining computational efficiency. The lightweight transformer encoder has two encoder layers stacked on top of each other. Each layer is designed to achieve a balance between visual capacity and computational efficiency.

The following make up each encoder layer.

a) Multi-Head Self-Attention: There are four attention heads.
b) Embedding Dimension: 2048 (this comes from the depth of the backbone channel)
c) The Key Dimension is 64 per head.
d) The feed-forward network expansion ratio is 4× channel compression (2048 $\rightarrow$ 1 512 $\rightarrow$ 1 2048).
e) The dropout rate is 0.1.

After both the attention and feed-forward sublayers residual skip connections and layer normalization are used to keep the gradient from getting out of control. Since the token sequence comes from convolutional grids that are arranged in space explicit positional encoding was not added. This kept the number of parameters low while still keeping the spatial locality that convolutional receptive fields implicitly encode.

3.2.7 Tokenization Strategy

The lightweight transformer encoder is composed of two stacked encoder layers, each designed to balance representational capacity and computational efficiency. Residual skip connections and layer normalization are employed after both the attention and feed-forward sublayers to stabilize gradient propagation. As the token sequence is derived from spatially ordered convolutional grids, explicit positional encoding was not introduced, thereby reducing parameter overhead while preserving spatial locality implicitly encoded by convolutional receptive fields. The detailed configuration of the tokenization strategy is presented in Table 2.

Table 2. Tokenization Strategy

Tokenization Strategy.	**Value**
Multi-Head Self-Attention	4 attention heads
Embedding Dimension	2048 (inherited from backbone channel depth)
Key Dimensions	64 per head
Feed-Forward Network Expansion Ratio	4× channel compression (2048 $\rightarrow$ 1 512 $\rightarrow$ 1 2048)
Dropout	0.1

3.2.8 Extracting Convolutional Features

We use a pretrained EfficientNetB5-based CNN as the backbone to get high-level spatial features from input fundus images. CNN takes resized images of the retina and makes deep feature maps that show local lesion patterns that are important for diabetic retinopathy. Global average pooling is used to make small CNN feature representations, which mitigates overfitting and the amount of work that needs to be done.

3.2.9 Global Context Modeling Based on Transformers

A lightweight transformer encoder is used on the convolutionally down sampled feature maps instead of the raw image patches to capture long-range dependencies and the overall structure of the retina. The feature maps are turned into a sequence of tokens, which reduces the number of tokens by a lot and stops the quadratic memory growth that happens with regular self-attention. We use multi-head self-attention to model how different

parts of space interact with each other. Then we use residual connections and layer normalization to make sure the training is stable. Global average pooling combines the output of the transformer to make a global contextual feature vector.

3.2.10 Combining and Classifying Features

The local features from the CNN and the global features from the transformer are combined and sent through fully connected layers so that they can learn together. This fusion strategy lets the model use both lesion-level and context-level information to grade DR in a strong way. The last layer of classification gives class probabilities that match the ordered levels of DR severity.

3.2.11 Loss Function That Is Ordinal-Sensitive

An ordinal-sensitive focal loss is used to deal with a big class imbalance and the fact that DR grading is ordinal. This loss punishes mistakes based on how serious they are giving more weight to clinically important mistakes while reducing bias toward majority classes. This design helps the model learn more useful severity boundaries than standard categorical cross-entropy.

3.2.12 Study of Ablation and Justification of Metrics

Comprehensive ablation studies are performed to evaluate the efficacy of each component. To see how well it works, a baseline with only CNN is tested first. Next the transformer module is added to see how it helps with global context modeling. Finally, the ordinal-sensitive loss replaces the standard cross-entropy to see how it affects learning that is aware of severity. The results show that the hybrid CNN–Transformer architecture always improves the macro-F1 score, balanced accuracy, and quadratic Cohen's kappa, which means that it agrees more with expert grading.

Because DR datasets are very unbalanced and ordinal, relying only on accuracy can be misleading. The macro-F1 score and balanced accuracy give a fair picture of how well all classes are doing while the AUC score shows how well the model can tell the difference between different classes. Quadratic Cohen's kappa is said to measure how much agreement there is with eye doctors while punishing big differences in severity. These metrics, when used together, give a full and clinically useful assessment of the proposed framework.

4 Results

To fix the class imbalance and punish distant grade misclassifications the proposed hybrid convolutional transformer model was trained using focal loss with ordinal sensitivity. Figure 4 shows the curves for training and validation accuracy over time. The model shows stable and monotonic convergence with training accuracy steadily rising and validation accuracy closely following the training trend. This shows that the model can generalize well and is not overfitting. By the end of the last epoch the model had a validation accuracy of over 81.94%. This showed that combining CNN-based local feature extraction with transformer-based global context modeling worked well. The small difference between training and validation accuracy shows that the learning framework is strong and that the lightweight transformer and dropout layers have a regularizing effect.

Table 3. Hyperparameter Tuning Strategy

Parameter	Value
Learning Rate	$1e^{-3}$
Batch Size	8
Dropout	0.5
Attention Heads	4
Transformer Layers	2

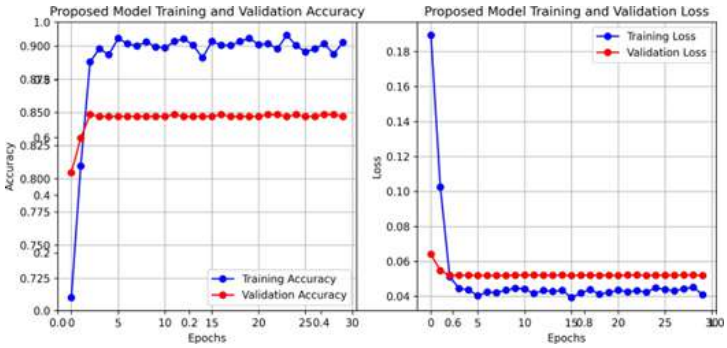

Figure 4. Proposed Model Training and Validation Accuracy

Figure 4 illustrates the training and validation accuracy of the proposed model across different training epochs.

The quadratic weighted kappa score shows that the predictions of the model and the experts notes are almost the same. The high kappa value means that most of the mistakes in predictions happen between DR severity levels that are close to each other not between levels that are very different. This is because kappa looks at the ordinal distance between grades that were incorrectly classified. This behavior is good for the patient because mixing up adjacent grades is not as bad as mixing up severe disease with mild disease, Figure 5 illustrates the kappa curve.

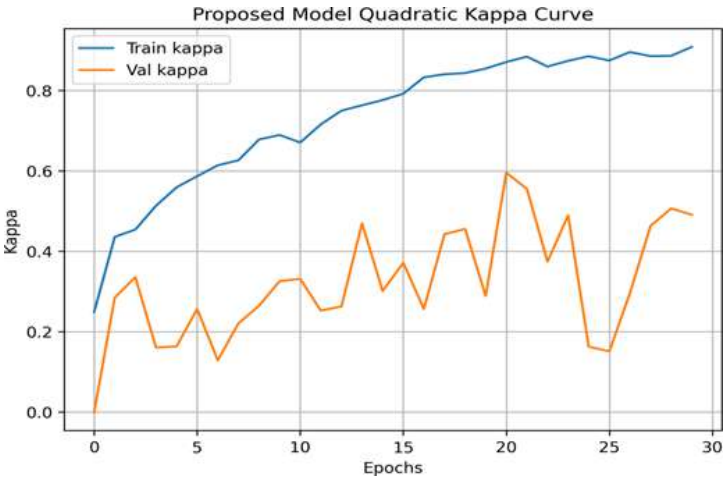

Figure 5. Proposed Model Kappa Curve

The balanced accuracy shows that the model is still good even when there is an uneven number of classes. This means that classes that are not as common, like proliferative DR won't be overpowered by classes that are more common. The high top-2 accuracy also shows that the representations learned are reliable because the right grade is often ranked among the top two predictions, even when the top prediction is wrong. The balanced accuracy of the proposed model is illustrated in Figure 6.

Figure 7 shows the confusion matrix.

The normalized confusion matrix for grading diabetic retinopathy into five classes can be seen in Figure 8. The diagonal in the matrix is very strong which means that most of the predictions are right at all levels of severity. Most mistakes happen when grades are close to each other like mild versus moderate or moderate versus severe DR. Even professional graders don't always know what is necessary like this shows.

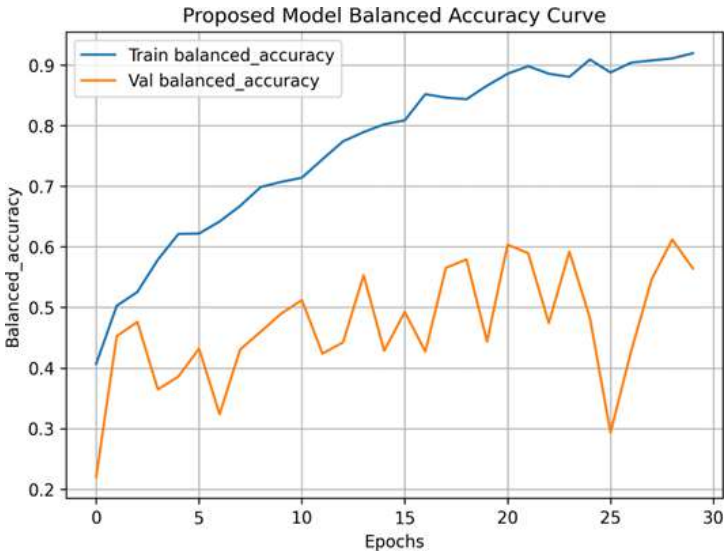

Figure 6. Proposed Model Balanced Accuracy

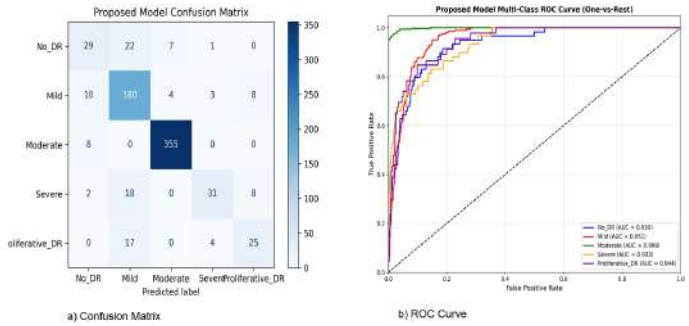

Figure 7. Proposed Model Confusion Matrix and ROC

It is very rare for professionals to make big mistakes when grading like calling proliferative DR no DR or mild DR. This result shows that the ordinal-sensitive loss function works to stop big ordinal deviations and improve clinically important decision-making.

The receiver operating characteristic (ROC) curves for each DR grade are shown in Figure 7. The model always gets high AUC values for all classes, which means it can tell the difference between them very well. The macro-averaged AUC shows once more that the model can tell the difference between different levels of severity, even when the classes aren't evenly split.

Classes that match advanced DR stages are very easy to tell apart, which is important for getting people to the right place for treatment quickly. The ROC curves that are smooth and well-separated show that the outputs

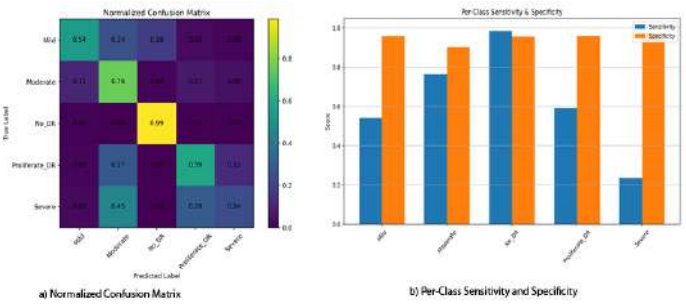

Figure 8. Normalized Confusion Matrix and Specificity

are stable and that the confidence estimation is correct. These are both important for real-world screening systems.

Table 4 shows the results of the quantitative evaluation in terms of precision, recall, F1-score. The proposed model has a high macro-averaged F1-score because it has high precision and recall for all classes. This means that the model does a good job of balancing false positives and false negatives which is very important for clinical screening applications. The visual representation of precision, recall, and F1-score for each class is illustrated in Figure 9.

Table 4. Summary of the results of Precision, Recall and F1 Score

Class	Precision	Recall	F-Score	Support
No DR	0.59	0.49	0.54	59
Mild	0.76	0.88	0.81	205
Moderate	0.97	0.98	0.97	363
Severe	0.79	0.53	0.63	59
Proliferative DR	0.61	0.54	0.57	46

The performance comparison with ResNetRS50 is illustrated in Figures 10 and 11.

The comparative performance of the proposed model with baseline architectures is presented in Table 5.

Table 5. Result Comparison

Model	Accuracy
ConvexNextBase	61%
ResNetRS50	81%
Proposed	85%

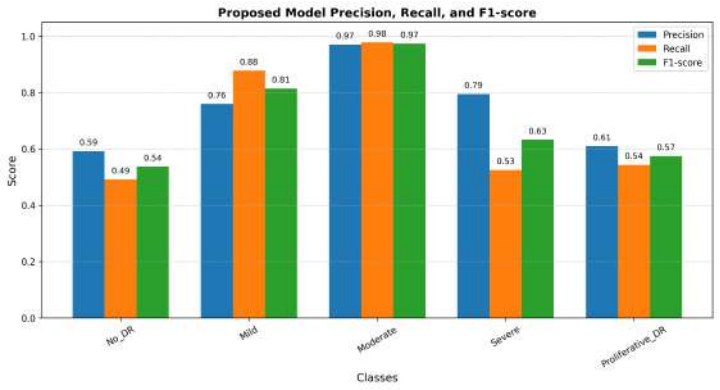

Figure 9. Precision Recall F1 Score

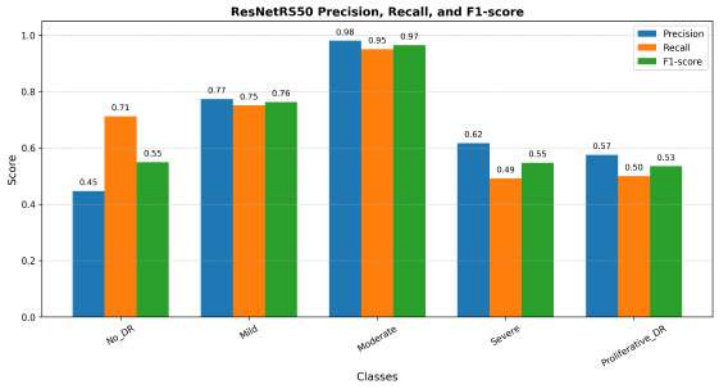

Figure 10. ResNetRS50 Precision Recall F1 Score

5 Discussion and conclusions

The experimental results clearly indicate that the proposed hybrid architecture exceeds traditional exclusive reliance on CNN methodologies by integrating global contextual reasoning through transformer-based self-attention. Microaneurysms and hemorrhages are two small problems in the body that CNNs are good at finding. The transformer section on the other hand lets you model extended spatial relationships across the retina. This is important for getting the severity grade precise.

Also treating diabetic retinopathy grading as an ordinal learning problem instead of a flat multi-class task makes it much more useful in a clinical setting. Adding ordinal-sensitive loss improves kappa scores and makes error distributions more useful which makes the model behave more like how experts grade.

The proposed framework is a good choice for large-scale automated diabetic retinopathy screening and de-

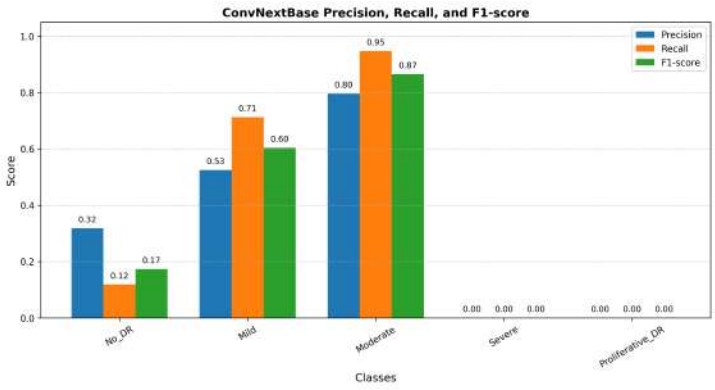

Figure 11. ResNetRS50 Precision Recall F1 Score

cision support in clinical workflows because it has high accuracy, strong ordinal agreement and strong generalization. There are three main reasons why the proposed model works better than other models. First the EfficientNetB5 backbone uses compound scaling to accurately capture small details in the retina like microaneurysms and hemorrhages.

The lightweight transformer encoder also lets you combine information from different parts of the world which is important for modeling how disease severity changes over time.The proposed hybrid architecture shows better generalization and robustness than pure CNN-based models as shown by higher Quadratic Kappa and balanced accuracy scores. The use of ordinal-sensitive loss also makes the optimization process fit with the clinical grading system for diabetic retinopathy which leads to predictions that are more useful in a clinical setting.

The proposed method effectively overcomes the limitations of architectures that only use convolutional or transformer networks by combining the strong local feature extraction ability of convolutional neural networks with the global contextual modeling ability of transformer self-attention. The hybrid design allows for strong representation of both small pathological lesions and long-range spatial dependencies which are important for accurately assessing the severity of DR.

A significant contribution of this study is the explicit modeling of the ordinal characteristics of diabetic retinopathy grades. The proposed ordinal-sensitive loss is different from standard multi-class classification methods because it punishes grade errors that are far away more severely while still allowing for severe class imbalance. This design choice makes the clinical relevance better as shown by steady improvements in ordinal-aware evaluation metrics like quadratic weighted kappa and balanced accuracy as well as standard accuracy and AUC measures.

A comprehensive experimental evaluation demonstrates that the proposed framework outperforms and exhibits greater consistency than models that utilize solely CNNs or transformers. The ablation study further confirms that the transformer branch, feature-level fusion strategy and ordinal-aware optimization all work together to make things better. The model is important because it can be used in the real world without using too much memory or processing power which is a problem with full vision transformer architectures.

From a clinical perspective the proposed method enhances grading reliability and alignment with expert annotations thereby improving its potential effectiveness as a computer-assisted diagnostic tool for comprehensive diabetic retinopathy screening programs. Future research will explore multi-scale tokenization, uncertainty-aware ordinal learning and external validation across diverse datasets and imaging devices to improve robustness and generalizability. This study advances the domain of automated diabetic retinopathy grading by offering a clinically relevant, interpretable and computationally efficient hybrid learning framework.

Author Contributions

Muhammad Suleman Memon: Conceptualization, Methodology, Software,Supervision **Mumtaz Qabulio**: Data curation, Writing- Original draft preparation. **Nazish Basir**: Visualization, Investigation. **Asia Khatoon Soomro**: Software, Validation. **Hira Naqvi**: Reviewing and Editing.

Compliance with Ethical Standards

It is declared that all authors do not have any conflict of interest. Furthermore, informed consent was obtained from all individual participants included in the study.

Funding Information

No external funding was received for this study.

References

[1] C. Yu et al., "FF-ResNet-DR model: a deep learning model for diabetic retinopathy grading by frequency domain attention," Electronic Research Archive, vol. 33, no. 2, pp. 725–743, 2025, doi: 10.3934/era.2025033.

[2] M. S. Khan, M. G. Hyder Talpur, and M. Aslam, "Comparative Analysis of Time Series Forecasting using ARIMA, and GRNNs Models: A Case Study of Death Rate of Diabetic Mellitus in Canada," *VFAST Transactions on Mathematics*, vol. 12, no. 1, pp. 415–423, 2024, doi: 10.21015/vtm.v12i1.1894.

[3] D. Bhulakshmi and D. S. Rajput, "A systematic review on diabetic retinopathy detection and classification based on deep learning techniques using fundus images," *PeerJ Comput. Sci.*, vol. 10, 2024, doi: 10.7717/PEERJ-CS.1947.

[4] D. Badar, J. Abbas, R. Alsini, T. Abbas, W. ChengLiang, and A. Daud, "Transformer attention fusion for fine grained medical image classification," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-07561-x.

[5] A. Bilal, M. Shafiq, W. J. Obidallah, Y. A. Alduraywish, A. Tahir, and H. Long, "Quantum chimp-enanced SqueezeNet for precise diabetic retinopathy classification," Sci. Rep., vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-97686-w.

[6] D. Muthusamy and P. Palani, "Deep learning model using classification for diabetic retinopathy detection: an overview," Artif. Intell. Rev., vol. 57, no. 7, Jul. 2024, doi: 10.1007/s10462-024-10806-2.

[7] M. Sushith, A. Sathiya, V. Kalaipoonguzhali, and V. Sathya, "A hybrid deep learning framework for early detection of diabetic retinopathy using retinal fundus images," Sci. Rep., vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-99309-w.

[8] S. Zhu, C. Xiong, Q. Zhong, and Y. Yao, "Diabetic Retinopathy Classification with Deep Learning via Fundus Images: A Short Survey," IEEE Access, vol. 12, pp. 20540–20558, 2024, doi: 10.1109/ACCESS.2024.3361944.

[9] A. Mofreh, A. Adel, N. Elnady, and M. Khaled, "Detection of Diabetic Retinopathy Using Deep Learning Techniques", doi: 10.21608/erurj.2025.369875.1238.

[10] A. M. Mutawa, G. R. Hemalakshmi, N. B. Prakash, and M. Murugappan, "Randomization-Driven Hybrid Deep Learning for Diabetic Retinopathy Detection," IEEE Access, 2025, doi: 10.1109/ACCESS.2025.3546359.

[11] M. Herrero-Tudela, R. Romero-Oraá, R. Hornero, G. C. Gutiérrez Tobal, M. I. López, and M. García, "An explainable deep-learning model reveals clinical clues in diabetic retinopathy through SHAP," Biomed. Signal Process. Control, vol. 102, Apr. 2025, doi: 10.1016/j.bspc.2024.107328.

[12] W. Nazih, A. O. Aseeri, O. Y. Atallah, and S. El-Sappagh, "Vision Transformer Model for Predicting the Severity of Diabetic Retinopathy in Fundus Photography-Based Retina Images," IEEE Access, vol. 11, pp. 117546–117561, 2023, doi: 10.1109/ACCESS.2023.3326528.

[13] Z. Liu, A. Gao, H. Sheng, and X. Wang, "Identification of diabetic retinopathy lesions in fundus images by integrating CNN and vision mamba models," PLoS One, vol. 20, no. 1 January, Jan. 2025, doi: 10.1371/journal.pone.0318264.

[14] E. Z. Ye, J. Ye, and E. H. Ye, "Applications of Vision Transformers in Retinal Imaging: A Systematic Review," Feb. 01, 2023. doi: 10.22541/au.167528318.80645903/v1.

[15] F. Mostafa, H. Khan, F. Farhana, and M. A. H. Miah, "Application of Deep Learning Framework for Early Prediction of Diabetic Retinopathy," AppliedMath, vol. 5, no. 1, Mar. 2025, doi: 10.3390/appliedmath5010011.

[16] G. I et al., "Enhanced diabetic retinopathy detection using U-shaped network and capsule network-driven deep learning," MethodsX, vol. 14, Jun. 2025, doi: 10.1016/j.mex.2024.103052.

[17] M. Akram et al., "Uncertainty-aware diabetic retinopathy detection using deep learning enhanced by Bayesian approaches," Sci. Rep., vol. 15, no. 1, p. 1342, Dec. 2025, doi: 10.1038/s41598-024-84478-x.

[18] S. Akhtar et al., "A deep learning based model for diabetic retinopathy grading," Sci. Rep., vol. 15, no. 1, p. 3763, Dec. 2025, doi: 10.1038/s41598-025-87171-9.

[19] S. Ajith Kumar, J. S. Kumar, and S. C. Bharadwaj Mahabaleswara, "Efficient diabetic retinopathy detection using deep learning approaches and Raspberry Pi 4," Bulletin of Electrical Engineering and Informatics, vol. 14, no. 2, pp. 1063–1072, Apr. 2025, doi: 10.11591/eei.v14i2.8248.

[20] A. M. Mutawa, K. Al-Sabti, S. Raizada, and S. Sruthi, "A Deep Learning Model for Detecting Diabetic Retinopathy Stages with Discrete Wavelet Transform," Applied Sciences (Switzerland), vol. 14, no. 11, Jun. 2024, doi: 10.3390/app14114428.

[21] F. M. J. Mehedi Shamrat et al., “An advanced deep neural network for fundus image analysis and enhancing diabetic retinopathy detection,” Healthcare Analytics, vol. 5, Jun. 2024, doi: 10.1016/j.health.2024.100303.

[22] W. Yaseen, A. U. Khan, and M. Sajid, “Brain Tumor Segmentation Using Deep Learning,” VFAST Transactions on Software Engineering, vol. 11, no. 2, pp. 113–123, 2023, doi: 10.21015/vtse.v11i2.1533.