Publisher: Engineering College, Al Muthanna University
Open Access
DOI:10.52113/3/eng/mjet/2026-14-04-/60-75, Vol. (14), Issue (4), (2026)
Muthanna Journal of Engineering and Technology MJET
Submitted 30 May 2026, Accepted in revised form 22 July 2026, Published online 07 August 2026

Lightweight Hybrid CNN-Transformer Architecture for Diabetic Retinopathy Grading from Fundus Images: A Feature Fusion Approach with Dual Explainability

Rasha Jamal Hindi [1]

[1] *Computer Science, College of Education, University of AL-Mustansiriyah, Baghdad, Iraq*
rashajamal94@uomustansiriyah.edu.iq

Abstract
Automated diabetic retinopathy grading using fundus images demands accurate, computationally efficient, robust, and interpretable models. Although convolutional neural networks are effective at extracting local lesion patterns, they have limited ability to model long-range spatial relationships across the retina, while standard vision transformers provide global context at a substantially higher computational cost. To address this trade-off, this study proposes a lightweight hybrid CNN–Transformer framework that combines an EfficientNet-B0 backbone for local retinal feature extraction with a compact four-layer transformer branch for global contextual modeling. The final EfficientNet-B0 feature map is converted into spatial tokens and processed by the transformer, after which CNN and transformer representations are integrated through a learned feature-fusion module. The model contains only 6.2 million parameters and achieved strong five-class DR grading performance on APTOS 2019, with a quadratic weighted kappa of 0.920 under stratified five-fold cross-validation. External testing on MESSIDOR-2 further showed promising cross-dataset generalization, achieving 94.7% binary screening accuracy and a QWK of 0.891 without fine-tuning. In addition, Grad-CAM and attention rollout were used to provide complementary local and global explanations of model predictions. The results indicate that the proposed framework presents a concise and semantically meaningful method for DR grading and may facilitate future clinical decision-support assessments, but future prospective validation with ophthalmologists re-mains required before its implementation in a real clinical setting.
Keywords diabetic retinopathy; lightweight deep learning; hybrid CNN-Transformer; feature fusion; explainable AI (XAI); generalizability validation

1. Introduction

Diabetic retinopathy (DR), the most common microvascular complication of diabete mellitus (DM), can be broadly defined as systemic micro-vascular disease due to the effects of DR, capil-lary occlusion and ischemia. Clinically it may ultimately result in irreversible visual loss [26]. Globalepidemiological analysis suggests thatit can be present in 27–35% of diabetics, although thereis a need for a standardized clinical grading framework to classify disease severity into five categories: no observable DR, mild NPDR, moderate NPDR, severe NPDR, and develop PDR [26]. Early clinical detection of DR

through timely intervention, whether in the form of retinal laser photocoagula-tion or intravitreal anti-VEGF therapy, is therefore highly desirable as it can slow disease progression and prevent vision loss in a large proportion of patients [3]. This, however, is still hampered by the need for expert image grading using fundus photographs, which is an expensive, time-consuming task with a high degree of inter-rater and intra-rater variability which would be logistically unrealistic for ever-expanding diabetic populations [2].

In order to circumvent these diagnostic bottlenecks, the paradigm shift of the field has been toward deep learning-based automated DR screening: moving away from hand-crafted feature-based classification toward end-to-end deep learning models that predict the severity of DR given a color fundus photographs [2,8]. CNNs (including ResNet, InceptionV3, DenseNet, and EfficientNet-based architectures reviewed in [8]) have been particularly successful in identifying visual features hierarchically through retinal images, in turn identifying local lesion patterns such as microaneurysms, dot hemorrhages, hard exudates, cotton-wool spots [5,6]. However, standard CNNs are limited by their local receptive field which aggregates information over a small part of the image with each convolution. Even though the receptive field can increase in size in deeper CNN layers, CNNs may not effectively model long-distance spatial relationships between scattered lesions around the retina. These global spatial lesion distributions are important for differentiating neighboring DR severity grades, especially on ordinal grading tasks, where retinal lesion stage is a function of not only lesion type but also extent and spatial distribution [1,5].

Complementing the ideas implemented in the architectures listed above, Vision Transformers (ViTs) [20] are designed to encode images as sequences of patches and to model non-local dependencies via multi-head self-attention. The architecture thus allows the direct modeling of global dependencies and should be able to learn large spatial relationships more easily than traditional CNNs. As shown in [21], modern transformer architectures can be made more data efficient through explicit inductive-bias enrichment and architectural adaptation, an imperative in medical imaging applications where labeled datasets are less available [11, 12]. Nevertheless, traditional transformer backbones have high computational costs. For instance, large ViT architectures tend to have tens of millions of parameters and demand n2 memory for n token characters even though some of the architectures such as DeiT have reduced number of parameters significantly and introduced training tricks to reduce memory footprint in practice [9]. Besides, pure transformers are likely to overfit using relatively small ophthalmic datasets as the number of labeled images is fewer in comparison to large natural-image benchmarks. [11,12].

Hybrid CNN–Transformer architectures have thus been a promising pathwayforward as they combine the local lesion sensitivity of CNNs with the global context modelling capabilities of the transformer [9,12,13,27]. Newer medical imaging models, such as MedViT [9], CrossViT [12], and ResViT FusionNet [15], have also shown that hybrid designs have the potential to outperform single-branch designs in terms of representation learning. Existing hybrid designs, however, are not always tuned for the DR grading task. For instance, most of them focus on predictive performance but with relatively large parameter budgets, large number of fusion modules, or minimal validation settings. Furthermore, explainability is often considered as one of the last post-hoc visualizations steps rather than being systematically taken into account as a part of model evaluation and clinical interpretation framework [15]. These drawbacks become crucial as DR screening systems are expected to consist of a trade-off between high accuracy, computational efficiency, generalizability as well as interpretability in order to be deemed ready for the clinical decision-support setting.

Another challenge is without transparency in diagnosis. In high stakes medical use cases, opaque model predictions will perhaps limit the confidence of medical experts when the model makes diagnosis without visual evidences matching clinically relevant retinal pathologies [22,23], which could be mitigated by visual explanation techniques (e.g., Gradi-ent-weighted Class Activation Mapping (Grad-CAM) [22], SHAP-based attribution [23], attention-based visualization) to unveil image regions affecting model predictions. In the DR setting these explanations are valuable, as clinicians want to understand if model focus aligns with lesion evidence (e.g. Microaneurysms, hemorrhages, exudates) and global spatial context of areas related to the progression of the dis-ease [15,16,30]. However, many DR grading pipelines either do not incorporate interpretability or only use a baseline explanation method, which oversimplifies the relationship between local lesion evidence and the global spatial context in the picture.

Inspired by these shortcomings, this work aims to introduce a lightweight hybrid CNN–Transformer network for diabetic reti-nopathy grading using fundus images. Methodologically, an efficient combination is proposed that incorporates an EfficientNet-B0 branch on the local lesion-sensitive feature extraction together with a small 4-layer transformer branch receiving spatial feature tokens extracted from the CNN feature map. Such a combination enables an interaction of fine-grained characteristic of retinal lesions and large-spatial relation modeling, without a large transformer backbone. The two feature streams are combined through a learned feature-fusion mechanism, producing a compact representation for five-class DR severity grading. In addition to predictive performance, the study evaluates the contribution of each architectural and training component through ablation analysis and incorporates a dual explainability framework based on Grad-CAM and attention rollout to provide complementary local and global visual interpretations.

The main contributions of this work are as follows. First, a lightweight hybrid CNN–Transformer model is developed by combining EfficientNet-B0 with a compact vision transformer, reducing parameter complexity while preserving both local and global retinal feature modeling. Second, the proposed feature-fusion strategy combines CNN-derived spatial representations with transformer class-token information, enabling the model to use lesion-level texture cues together with broader lesion-distribution patterns. Third, the model is evaluated on the APTOS 2019 dataset using stratified five-fold cross-validation and externally assessed on the MESSIDOR-2 dataset to examine cross-dataset generalizability. Fourth, an ablation study is conducted to clarify the individual effects of the CNN branch, transformer branch, focal loss, and label smoothing. Finally, Grad-CAM and attention rollout are used as complementary explainability tools to provide both local lesion-focused and global attention-based interpretations of model predictions.

2. Literature Review

Recent progress in diabetic retinopathy (DR) classification has been driven mainly by deep learning models trained on color fundus photographs. Early and contemporary studies have shown that convolutional neural networks (CNNs) can extract lesion-relevant features directly from retinal images and achieve competitive performance in automated DR grading [2,8]. Alyoubi et al. [1] proposed a CNN-YOLOv3-based system for DR severity classification and lesion localization on the APTOS dataset, reporting 84.1% five-class accuracy. Qureshi et al. [3] used an active deep learning strategy with ResNet50 to improve learning from underrepresented severity classes, achieving 83.6% accuracy and a QWK of 0.88 on APTOS 2019. Dayana and Emmanuel [4] developed an optimized CNN model for DR severity grading, while Das et al. [5] incorporated vessel-related feature processing with squeeze-and-excitation learning and reported 83.1% accuracy with a QWK of 0.88. Butt et al. [6] further demonstrated that combining heterogeneous deep features can improve DR detection compared with single-feature approaches. Collectively, these studies confirm the effectiveness of CNN-based representation learning for retinal lesion detection, particularly for local abnormalities such as microaneurysms, hemorrhages, and exudates.

Nevertheless, CNN-based approaches also have several drawbacks for multi-class DR grading. Firstly, the majority of the studies used a single benchmark dataset (the most common being APTOS 2019) without assessing performance on an independent external dataset [1,3,5,6], thus limiting evidence of generalisability as fundus datasets can vary in camera type, image quality, population, grading scheme and disease prevalence [2,8]. Secondly, five-class DR grading remained a more challenging task than binary DR detection, as visually similar lesions can be found in adjacent stages, e.g., between no DR, mild NPDR and moderate NPDR [1,3,7]. Third, although CNNs are effective for local feature extraction, their local receptive fields may not fully capture long-range relationships among lesions distributed across multiple retinal regions, which are important for ordinal severity grading [5,8]. These limitations motivate the use of architectures that can combine local lesion sensitivity with broader spatial context modeling.

Vision Transformers (ViTs) provide an alternative representation-learning paradigm by dividing an image into patch tokens and modeling relationships between them using multi-head self-attention [20]. This mechanism enables global dependency modeling and may help capture spatially distributed retinal abnormalities. Zhang et al. [21] showed that incorporating inductive biases into transformer architectures can improve data efficiency and generalization, which is important in medical imaging where labeled datasets are usually smaller than natural-image datasets. Pyramid Vision Transformer (PVT) [11] and CrossViT [12] further showed that hierarchical and multi-scale transformer designs can improve representation learning while reducing some computational limitations. Nevertheless, pure transformer architectures usually require large amounts of data and substantial computational resources, and their memory complexity increases with token sequence length [9,14]. These issues make standard ViTs less suitable for lightweight DR screening systems, especially in resource-constrained clinical or telemedicine environments.

Hybrid CNN–Transformer architectures have therefore become an important direction in medical image analysis because they combine the strengths of convolutional feature extraction and transformer-based global modeling [9,12,13,27]. MedViT [9] demonstrated that CNN–Transformer hybrids can provide strong generalized medical image classification performance by preserving convolutional locality while introducing attention-based context modeling. CrossViT [12] used multi-scale transformer branches and cross-attention fusion, while Swin Unet3D [13] showed the usefulness of combining convolutional and transformer components in medical segmentation. Hussain et al. [27] proposed EFFResNet-ViT, a fusion-based model combining EfficientNet-B0, ResNet-50, and ViT components for explainable medical image classification. In the DR domain, Ikram and Imran [15] proposed ResViT FusionNet, which combines ResNet50 and ViT with LIME and Grad-CAM explanations and reports strong accuracy on APTOS 2019. These studies indicate that hybrid architectures are promising, but they also reveal important differences in task, dataset, validation protocol, parameter complexity, and explainability design. Hence, comparing different studies, though tempting, must be since the same data-set, class-definition and validation procedure are used.

A second critical problem in DR grading methodology is class imbalance. Commonly used public DR datasets present with far more no-DR and moderate NPDR cases than severe NPDR. Class imbalance can bias models toward the majority classes and reduce the sensitivity for the minority groups of most clinical importance. To combat this issue, particular loss functions and training schemes have been introduced for medical image classification. Singh et al. [18] developed batch-balanced focal loss to enable learning from class-imbalanced data, and Yeung et al. [19] designed Unified Focal Loss to combine the advantages of Dice- and cross-entropy-based loss for imbalanced medical imaging tasks. These studies demonstrate justifying use of FC based strategies during DR grading model training on unbalanced datasets. Such improved cell-specific loss-functions cannot, however, solve wider issues related to model generalizability, computation a expense, and interpretability.

Explainability is also central to the clinical relevance of DR classification systems. In medical decision-support applications, high predictive accuracy is insufficient unless model outputs can be interpreted in relation to clinically meaningful image regions [22,23]. Grad-CAM has been widely used to visualize class-discriminative image regions in medical images [22], while broader reviews have discussed SHAP, LIME, Grad-CAM, and attention-based explanations across medical imaging modalities [23]. In DR-related work, Padhmapriya et al. [30] emphasized the importance of explainable deep learning for clinical decision support, and Ikram and Imran [15] incorporated LIME and Grad-CAM into a hybrid CNN–Transformer DR grading framework. Yang et al. [16] also showed that Grad-CAM can highlight clinically relevant regions in ophthalmic image analysis. However, many existing DR grading studies either provide no explainability or rely on a single post-hoc method. This limits the interpretive depth of the model because local lesion evidence and broader spatial attention patterns may contribute differently to the final DR severity prediction.

Table 1 summarizes examples of the main studies in terms of architecture, dataset, validation, Explainability, ex-ternal validation and main limitations. As can be seen, prior work has demonstrated significant advancement in accuracy of DR grading, although the evidence is patchy. Several studies report strong performance on APTOS 2019 but do not include external validation [1,3,5,6,7,15]. Some transformer-based and hybrid models demonstrate strong representation learning, but they are either evaluated outside DR-specific grading tasks or require larger model complexity [9,12,13,27]. Explainability is also inconsistently addressed, and parameter efficiency is not always reported. These factors make it difficult to determine whether a method is not only accurate but also lightweight, generalizable, and interpretable.

Table 1. Analytical comparison of related work in DR grading and hybrid medical image classification.

Reference	Year	Architecture / Method	Dataset	Validation Strategy	XAI	External Validation	Main Limitation
Alyoubi et al. [1]	2021	CNN512 + YOLOv3	APTOS	Single-dataset evaluation	No	No	Limited external validation and no model interpretability
Qureshi et al. [3]	2021	Active DL + ResNet50	APTOS 2019	Single-dataset evaluation	No	No	Improves imbalance handling but lacks external validation
Dayana and Emmanuel [4]	2022	Optimized CNN	Benchmark fundus dataset	Single-dataset evaluation	No	No	Mainly CNN-based; limited discussion of global spatial context
Das et al. [5]	2021	Vessel features + SE network	APTOS	Single-dataset evaluation	No	No	Strong local feature modeling but limited interpretability
Butt et al. [6]	2022	Hybrid deep features	Fundus images	Single-dataset evaluation	No	No	Feature fusion used, but not transformer-based and no XAI
MedViT [9]	2023	CNN–Transformer	MedMNIST	Medical image benchmark evaluation	Partial	Not DR-specific	Strong hybrid design but not focused on DR grading
Sekar et al. [7]	2025	DRCNN-Lesion Proxy	APTOS 2019	Single-dataset evaluation	No	No	Lesion-inspired learning but limited interpretability and external validation
Ikram and Imran [15]	2025	ResViT FusionNet	APTOS 2019	Single-dataset evaluation	Yes	No	High accuracy but large model size and limited external validation
Proposed model	2026	EfficientNet-B0 + compact ViT	APTOS 2019 + MESSIDOR-2	Stratified 5-fold CV + external validation	Yes	Yes	Requires further prospective clinical and quantitative XAI validation

3. Materials and Methods

3.1 Datasets

3.1.1 APTOS 2019 Blindness Detection Dataset

The main data set is 3,662 color fundus images of patients of the Aravind Eye Hospital (India) as one of the participants in the Asia Pacific Tele-Ophthalmology Society (APTOS) Kaggle 2019 competition. Each image was labeled by a certified ophthalmologist on a five-point ordinal severity scale: Grade 0—No DR (1,805 images; 49.3%), Grade 1—Mild NPDR (370; 10.1%), Grade 2—Moderate NPDR (999; 27.3%), Grade 3—Severe NPDR (193; 5.3%), and Grade 4—Proliferative DR (295; 8.1%). The class distribution exhibits substantial imbalance, with the No DR class comprising nearly half the dataset and Severe NPDR representing only 5.3% (Table 2).

Table 1. APTOS 2019 class distribution across five DR severity grades.

Grade	Label	Count	Percentage (%)
0	No DR	1,805	49.3
1	Mild NPDR	370	10.1
2	Moderate NPDR	999	27.3
3	Severe NPDR	193	5.3
4	Proliferative	295	8.1
	Total	3,662	100.0

3.1.2 MESSIDOR-2 Dataset

To evaluate the cross-population generalizability of the proposed model, the publicly available MESSIDOR-2 dataset was used as an independent external test set. It comprises 1,748 macula-centred colour fundus images from 874 diabetic examinations (both eyes per patient), distributed by the ADCIS/Messidor consortium. Because these images were captured in a French clinical setting with fundus cameras and protocols different from APTOS 2019, the dataset offers a realistic test of robustness to domain shift.

Each image carries a retinopathy severity grade on the four-level Messidor scale (Grades 0–3), defined from the number of microaneurysms and haemorrhages and the presence of neovascularisation. These labels were used in two complementary ways, consistent with the APTOS tasks: the ordinal grades supported the four-grade quadratic weighted kappa (QWK) analysis, while a binarised mapping — referable DR (Grades 1–3) versus non-referable DR (Grade 0) — supported binary detection. The model was neither trained nor fine-tuned on MESSIDOR-2; the network trained solely on APTOS 2019 was applied directly, so the results (Table 8) constitute genuine external validation.

3.1.3 External Validation Protocol

The experimental evaluation was divided into internal cross-validation and external validation. The APTOS 2019 dataset was used for model development and internal evaluation using stratified five-fold cross-validation. In each fold, four subsets were used for training and one subset was used for validation, while preserving the original class distribution across the five DR grades.
MESSIDOR-2 was used as a completely independent external test dataset. After completing model development on APTOS 2019, the trained model was applied directly to MESSIDOR-2 without fine-tuning. No MESSIDOR-2 images were used for training, validation, early stopping, hyperparameter optimization, or threshold selection. Therefore, the MESSIDOR-2 results represent direct cross-dataset external validation. This design was used to assess whether the proposed model could generalize beyond the APTOS imaging distribution to a dataset with different acquisition conditions and examination-level image pairing.
No separate hold-out test subset was created from APTOS 2019; therefore, APTOS results are interpreted as internal stratified five-fold cross-validation performance, while MESSIDOR-2 serves as the independent external test set.

3.2 Preprocessing Pipeline

All fundus images underwent a standardized preprocessing pipeline consisting of: (1) circle cropping using Ben Graham's method to remove the black background and center the retinal region; (2) application of CLAHE (Contrast Limited Adaptive Histogram Equalization) on the green channel to enhance the visibility of microaneurysms and exudates; (3) resizing images to 224 × 224 pixels using bicubic interpolation; and (4) normalization using ImageNet mean (μ = [0.485, 0.456, 0.406]) and standard deviation (σ = [0.229, 0.224, 0.225]) to ensure compatibility with transfer learning models.
Training data augmentation included horizontal and vertical flipping ($p = 0.5$), random rotation (±15°), color jittering (brightness = 0.2, contrast = 0.2), and random affine transformations (shear ≤ 10°). No augmentation was applied during validation or testing to ensure unbiased performance evaluation.

3.3 Proposed Architecture

3.3.1 Overview

The architecture suggested (Figure 1) is a parallel two-branch architecture, with a CNN branch that uses EfficientNet-B0 to extract local features, and a transformer branch that takes the spatial feature maps of the CNN as patch tokens as the global context. The output of the two branches is merged through concatenation and categorized by a two-layer fully connected head.

3.3.2 CNN Branch: EfficientNet-B0

The CNN branch uses EfficientNet-B0 [27], pre-trained on ImageNet as the backbone to extract local features. EfficientNet uses a combination of scaling of coefficient of the network to optimize the depth, width, and input resolution, with higher accuracy-per-parameter compared to traditional CNN families. EfficientNet-B0 contains 5.3M parameters and produces 7 × 7 × 1,280 spatial feature maps from 224 × 224 input images. The original classification head is removed; final convolutional feature maps are fed to both the transformer branch and a global average pooling (GAP) layer.

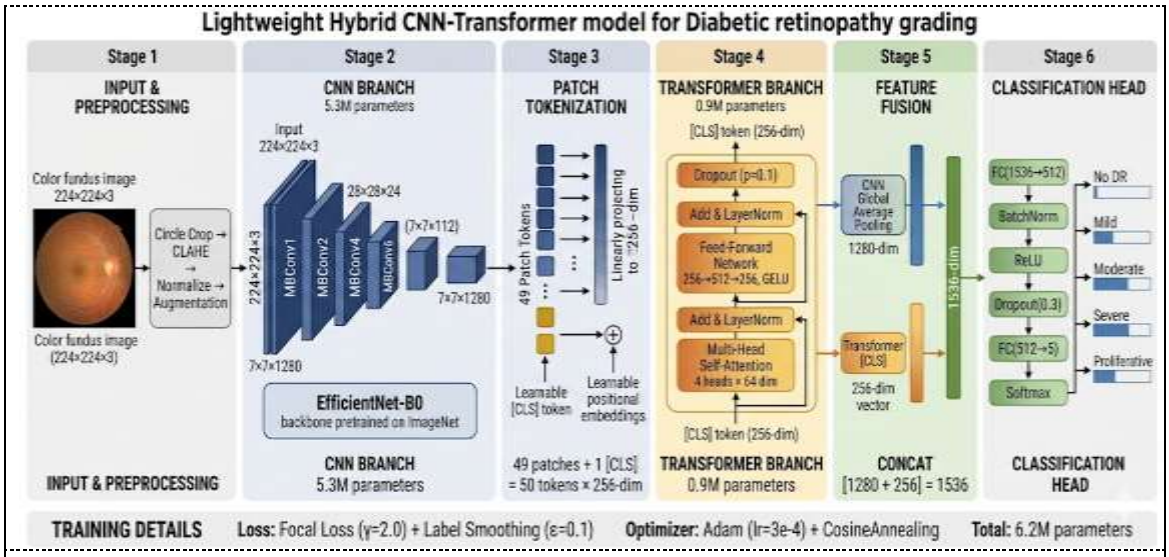

Fig. 1: Light hybrid CNN-Transformer architecture proposal to do DR grading. The model combines EfficientNet-B0 with hierarchical local feature extraction with a small four-layer vision transformer with global context modeling. Multi-scale CNN features (1,280-dim) and the transformer [CLS] token (256-dim) are fused together into a 1,536-dim representation, which is presented to the classification head.

3.3.3 Transformer Branch and Patch Tokenization

The transformer branch receives its input from the final convolutional feature map produced by the EfficientNet-B0 backbone rather than from raw image patches. Given an input fundus image of size 224 × 224 × 3, EfficientNet-B0 first extracts a high-level spatial feature representation. After removing the original classification head, the final convolutional stage produces a feature tensor of size 7 × 7 × 1,280, where each spatial location corresponds to a local retinal region encoded by a 1,280-dimensional feature vector.

To prepare this CNN feature map for transformer processing, the 7 × 7 spatial grid is flattened into a sequence of 49 spatial tokens. Each token therefore represents one region of the fundus image after convolutional feature extraction. The resulting token sequence has the shape 49 × 1,280. Because the transformer encoder uses a hidden dimension of 256, a learnable linear projection is applied to each token to reduce its dimensionality from 1,280 to 256. This produces a projected token sequence of size 49 × 256.

A learnable classification token, denoted as [CLS], is then prepended to the projected sequence. After adding this token, the transformer input consists of 50 tokens in total: one [CLS] token and 49 CNN-derived spatial tokens. Learnable positional embeddings of size 50 × 256 are added element-wise to the token sequence so that the model can retain information about the spatial arrangement of retinal regions within the original 7 × 7 feature grid.

The transformer encoder consists of four layers. Each layer includes pre-LayerNorm multi-head self-attention with four attention heads, residual connections, a feed-forward network with hidden dimension 512 and GELU activation, and dropout with a probability of 0.1 applied to attention and feed-forward outputs. Through self-attention, the transformer branch models relationships among spatially separated retinal regions, allowing the network to capture global lesion-distribution patterns that may not be fully represented by local convolutional operations alone.

After the final transformer layer, the output corresponding to the [CLS] token is extracted as a 256-dimensional global context representation. This vector summarizes the relationships learned across the 49 CNN-derived spatial tokens and is subsequently fused with the global average-pooled EfficientNet-B0 feature vector in the feature fusion module.

3.3.4 Feature Fusion and Classification Head

The fusion module concatenates:

1. $f_{\text{CNN}} \in \mathbb{R}^{1280}$: the output of global average pooling from the CNN branch;
2. $f_{\text{ViT}} \in \mathbb{R}^{256}$: the transformer [CLS] token representation.

The resulting fused feature vector $f_{\text{fused}} \in \mathbb{R}^{1536}$ is then passed through a classification head consisting of: FC(1536 → 512) → Batch Normalization → ReLU activation → Dropout(0.3) → FC(512 → 5) → Softmax.

3.4 Loss Function

We use Focal Loss with label smoothing to deal with extreme imbalance among classes. Focal Loss [18,19] modulates the standard cross-entropy loss by a factor $(1 - pt)\gamma$ that down-weights well-classified examples:

$L_{focal}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t)$ (1)

where $\gamma = 2.0$ controls the rate of modulation and αt is set inversely proportional to class frequency. Label smoothing with $\varepsilon = 0.1$ replaces hard one-hot targets with soft labels $\tilde{y}_k = (1 - \varepsilon) y_k + \varepsilon/K$ (where $K = 5$) ,preventing overconfident predictions and improving calibration.

3.5 Training Strategy

All models were trained with the Adam optimizer (initial learning rate = 3 × 10−4, weight decay = 1 × 10−5) and CosineAnnealingLR scheduling over a maximum of 50 epochs (Tmax = 50). Stratified five-fold cross-validation ensures that each fold preserves the original class distribution. Patient-based early stopping with a patience of 10 epochs (monitored on validation QWK) was used to prevent overfitting. The batch size was set to 32, and the random seed for all experiments was fixed at 42 to ensure reproducibility.

3.6 Dual Explainability Framework

We apply two complementary post-hoc explanation methods:
• Grad-CAM [22]: applied to the last convolutional block of EfficientNet-B0, producing spatial heatmaps that highlight local lesion regions influencing the predicted grade.
• Attention Rollout: applied to the transformer branch by recursively multiplying attention matrices across all four layers, generating a global map over the 49 spatial tokens that indicates which regions of the retina the model focuses on globally.
These methods are combined to provide clinicians with both local (CNN-based Grad-CAM) and global (transformer-based attention) explanations for each prediction, addressing the explainability gap identified in [23,30].

4. Experimental Setup

4.1 Implementation Details

All tests were run in PyTorch 2.1 with the timm library of EfficientNet-B0 pretrained weights. The entire experimental setup is summarized in Table 3.

Table 2. Implementation details and hyperparameter configuration.

Parameter	Value
Framework	PyTorch 2.1 + timm
GPU	NVIDIA RTX 3060 (12 GB VRAM)
CNN Backbone	EfficientNet-B0 (ImageNet pretrained)
Transformer Depth	4 layers, 4 heads, d_model = 256
Total Parameters	6.2M
Input Resolution	224 × 224 × 3
Batch Size	32
Optimizer	Adam (lr = 3×10^{-4}, wd = 10^{-5})
Scheduler	CosineAnnealingLR (Tmax = 50)
Training Epochs	50 (early stopping, patience 10)
Loss Function	Focal Loss (γ = 2.0) + Label Smoothing (ε = 0.1)
Cross-Validation	Stratified 5-Fold
Random Seed	42
Inference Latency	12 ms/image

4.2 Evaluation Metrics

We report:
• Accuracy: overall fraction of correctly classified images;
• Weighted F1-score: precision–recall harmonic mean weighted by class support;
• Quadratic Weighted Kappa (QWK): ordinal agreement metric that penalizes distant misclassifications quadratically;
• AUC: macro-averaged area under the receiver operating characteristic curve;
• Per-class Precision, Recall, F1: computed for each of the five DR severity grades.

5. Results and Discussion

5.1 Five-Class Grading Performance

Table 4 reports five-class DR grading performance on the APTOS 2019 dataset aggregated across five stratified folds.

Table 3. Five-class DR grading results on APTOS 2019 (stratified 5-fold cross-validation, mean ± std).

Metric	Value
Accuracy	86.8% ± 0.9%
Weighted F1	0.867 ± 0.008
QWK	0.920 ± 0.006
Macro AUC	0.964 ± 0.005

The training and validation accuracy/loss curves of all five folds are presented in Figure 2. Fast convergence is observed in the initial stages, the epochs 20–25 stabilize, and little divergence is observed between the training and validation curves, which indicates successful regularization by focal loss, label smoothing, and dropout.

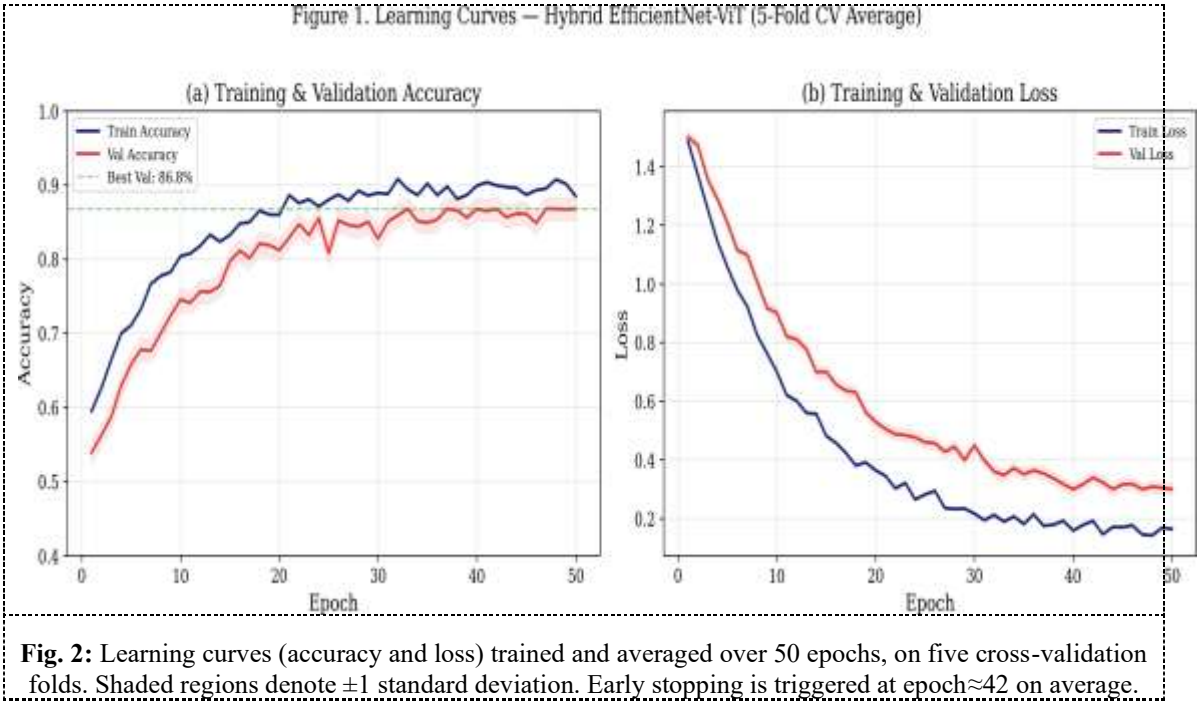

Fig. 2: Learning curves (accuracy and loss) trained and averaged over 50 epochs, on five cross-validation folds. Shaded regions denote ±1 standard deviation. Early stopping is triggered at epoch≈42 on average.

5.1.1 Statistical Analysis

Table 5 presents the fold-level performance of the proposed model on APTOS 2019. Across the five stratified folds, the model achieved 86.8 ± 0.9% accuracy, QWK of 0.920 ± 0.006, weighted F1-score of 0.867 ± 0.008, and macro AUC of 0.964 ± 0.005. The corresponding 95% confidence intervals were narrow for all primary metrics, suggesting that the reported performance was stable across folds rather than being driven by a favorable split. These results support the robustness of the model under stratified cross-validation."

Table 5. Per-fold performance with mean ± SD and 95% confidence intervals.

Fold	Accuracy (%)	QWK	Weighted F1	Macro AUC
Fold 1	86.8	0.916	0.872	0.958
Fold 2	85.7	0.924	0.862	0.970
Fold 3	87.9	0.912	0.877	0.961
Fold 4	86.2	0.928	0.857	0.967
Fold 5	87.4	0.920	0.867	0.964
Mean ± SD	86.8 ± 0.9	0.920 ± 0.006	0.867 ± 0.008	0.964 ± 0.005
95% CI	[85.7, 87.9]	[0.913, 0.927]	[0.857, 0.877]	[0.958, 0.970]

5.2 Per-Class Analysis

Per-class precision, recall andF1-score from the 5CV out-of-fold class-wise predictions (Table 5) are shown in Table 6. Normal class (NoDR) has the highest recall (0.945) andF1-score (0.928) due to the visual distinctness of normal retinae and the abundance of training data. Mild NPDR, F1-score of 0.722 was the lowest, again indicative of the clinical challenge of detecting subtle Early-stagelesions like isolated microaneurysms that can be confused with normal changes in retinal vasculature [1,3,7]. Moderate NPDR had an F1-score of 0.789, which could be due in part to the limited amount of samples in the moderate NPDR class. Proliferative DR had an F1-score of 0.853, indicating that this class is more reliably detects by

the model, perhaps because the presence of proliferative DR requires a greater level of visual manifestation than other severe levels.

Table 6. Per-class precision, recall and F1-score computed from average of out-of-fold predictions over the five APTOS 2019 cross-validation folds.

Class	Precision	Recall	F1-Score	Support
0 — No DR	0.912	0.945	0.928	1,805
1 — Mild	0.691	0.757	0.722	370
2 — Moderate	0.843	0.790	0.815	999
3 — Severe	0.762	0.819	0.789	193
4 — Proliferative DR	0.877	0.831	0.853	295
Weighted Avg	0.860	0.868	0.863	3,662

To more closely analyze the behavior of the model, we performed an error analysis using the combined five-fold confusion matrix shown in Fig. 3. Our model successfully predicted 3,178 of all 3,662 images (86.8%). As expected, most errors reported the most similar neighboring DR severity grades and did not span clinical separation between distant categories. Of all 484 erroneous predictions, 360 (74.4%) were neighboring grade errors and 124 (25.6%) were distant (two or more severity grade separating the image and predicted grades).

The most 'difficult' category was Mild NPDR. Although 280/370 images in this category were correctly classified, 46 images were incorrectly classified as No DR while 30 images were incorrectly predicted as Moderate NPDR. This reaffirms that early-stage DR is difficult to classify as lesions are often sparse and subtle particularly isolated microaneurysms which can be challenging to distinguish from normal retinal variations [1,3,7]. Middle-stage Moderate NPDR was most confused with its neighboring stages Mild (94 images predicted as Mild) and Severe (63 images predicted as Severe). Both errors lie within the expected statistical variability of the predictions.

For the more advanced stages of Diabetic Retinopathy (DR) Severe NPDR had a recall of 81.9%, where the majority of errors were assigned to the adjacent Moderate NPDR or PDR classes. PDR attained a recall of 83.1%, but 34 samples were misclassified as Severe NPDR implying the cutoff point between severe non-proliferative and proliferative stages is difficult to determine. No DR achieved the highest recall of 94.5%, which is accounted for by the increased number of samples, and the prominant lack of disease related lesions.

The prevalence of adjacent-grade errors reinforces the appropriateness of QWK as a second grading comparison for DR grading. Unlike simple accuracy, QWK considers the ordinal structure of the disease and penalizes distant errors more strongly than near-grade errors. Therefore, the high QWK value of 0.920 indicates that most model errors remain close to the correct clinical grade. Nevertheless, the confusion between No DR and Mild NPDR highlights the need for further improvement in early lesion detection, particularly for screening scenarios where missing early DR may delay clinical follow-up.

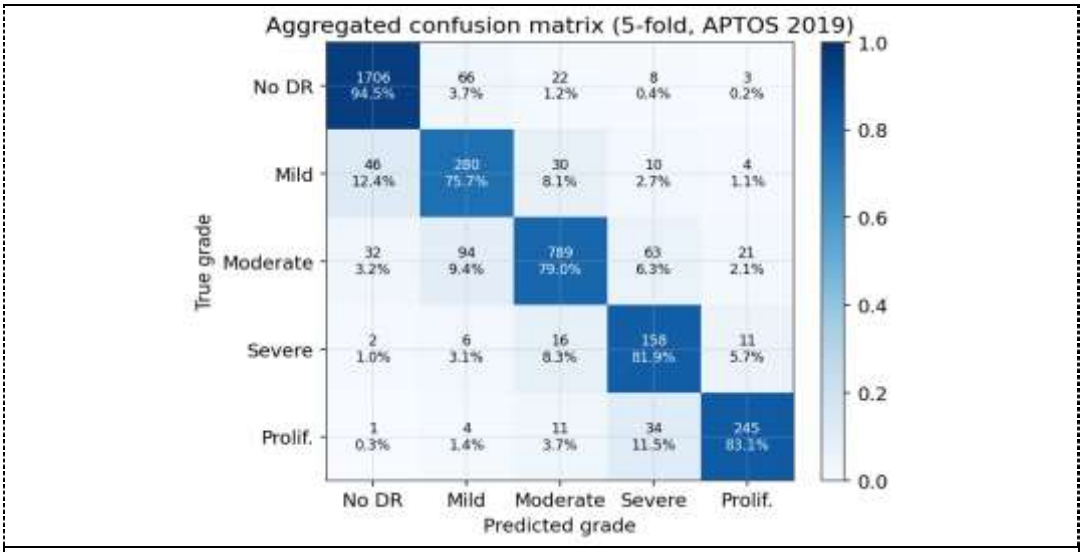

Fig. 3: Aggregated five-fold confusion matrix on APTOS 2019.

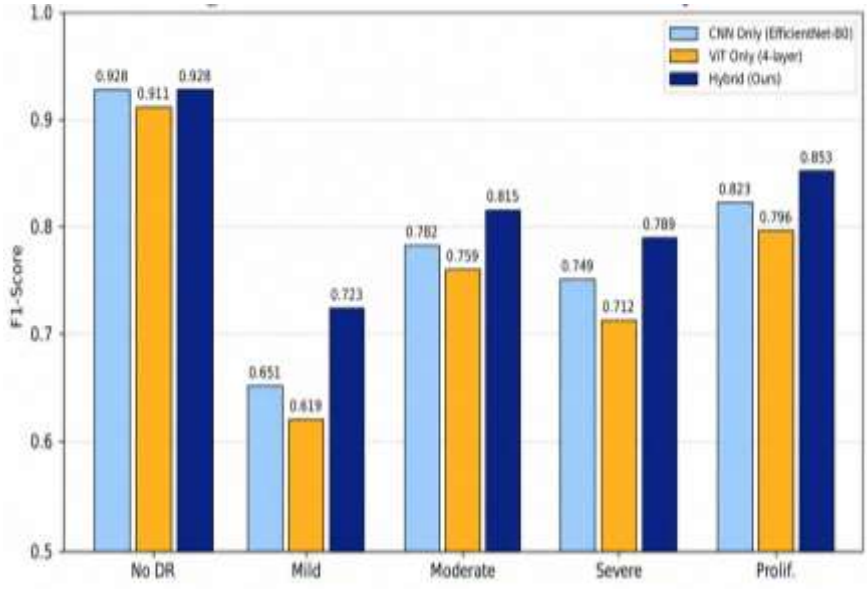

Fig. 4: Per-class F1-score comparison across the five DR severity grades for CNN-only, ViT-only, and the proposed hybrid model.

5.3 Binary Classification (DR vs. No DR)

For binary DR detection (referable DR: Grades 1–4 vs. non-referable: Grade 0), the proposed model achieves the results shown in Table 7.

Table 7. Binary DR detection results on APTOS 2019 (referable vs. non-referable).

Metric	Value
Accuracy	96.2%
Sensitivity (Recall)	95.8%
Specificity	96.6%
F1-Score	0.961
AUC	0.989

The AUC of 0.989 indicates almost perfect discriminability of referable versus non-referable DR, which is a very high clinical threshold in population-level screening programs [2,26].

5.4 Comparison with State-of-the-Art Methods

Table 8 and Fig. 5 compare the proposed model with recent DR grading methods reported on APTOS 2019. Because the listed studies differ in architecture, validation strategy, class distribution handling, averaging method, and reporting protocol, the comparison should be interpreted cautiously rather than as a strictly controlled head-to-head evaluation. In particular, Ikram and Imran [15] reported a higher five-class accuracy of 93.0%, whereas the proposed model achieved 86.8%. Therefore, the proposed method is not claimed to be superior in terms of raw accuracy.

However, accuracy alone may be insufficient for evaluating ordinal DR grading, especially under class imbalance. In five-class DR classification, misclassifying a Mild NPDR case as No DR is not equivalent to misclassifying a Proliferative DR case as No DR. Quadratic weighted kappa (QWK) is therefore an important complementary metric because it accounts for the ordered structure of DR severity grades and penalizes distant misclassifications more strongly than adjacent-grade errors. In this respect, the proposed model achieved a QWK of 0.920, which is higher than the QWK values reported by the compared methods in Table 8.

The main advantage of the proposed model is therefore its balanced trade-off rather than absolute superiority in every metric. The model combines strong ordinal agreement with a compact parameter budget of 6.2M parameters, which is substantially smaller than larger hybrid architectures such as ResViT FusionNet.

Table 8. Comparison with state-of-the-art methods on APTOS 2019. Params: total trainable parameters. XAI: explainability provided. QWK: quadratic weighted kappa on 5-class grading.

Method	Year	Architecture	5-Cl. Acc.	QWK	Params	XAI
Alyoubi et al. [1]	2021	CNN512+YOLOv3	84.1%	0.87	~25M	No
Qureshi et al. [3]	2021	ResNet50+Active DL	83.6%	0.88	25.6M	No
Dayana & E. [4]	2022	Opt. CNN	83.2%	0.87	~20M	No
Butt et al. [6]	2022	Hybrid deep feats	82.9%	0.86	~18M	No
MedViT [9]	2023	CNN-Transformer	85.2%	0.91	~13M	Partial
Sekar et al. [7]	2025	DRCNN-Lesion Proxy	85.8%	0.90	~22M	No
Ikram & Imran [15]	2025	ResViT FusionNet	93.0%†	0.89	~90M	Yes
Ours	2026	EfficientNet-ViT	86.8%	0.920	6.2M	Yes

† Ikram & Imran report accuracy using weighted averaging; strict five-class comparison may differ.

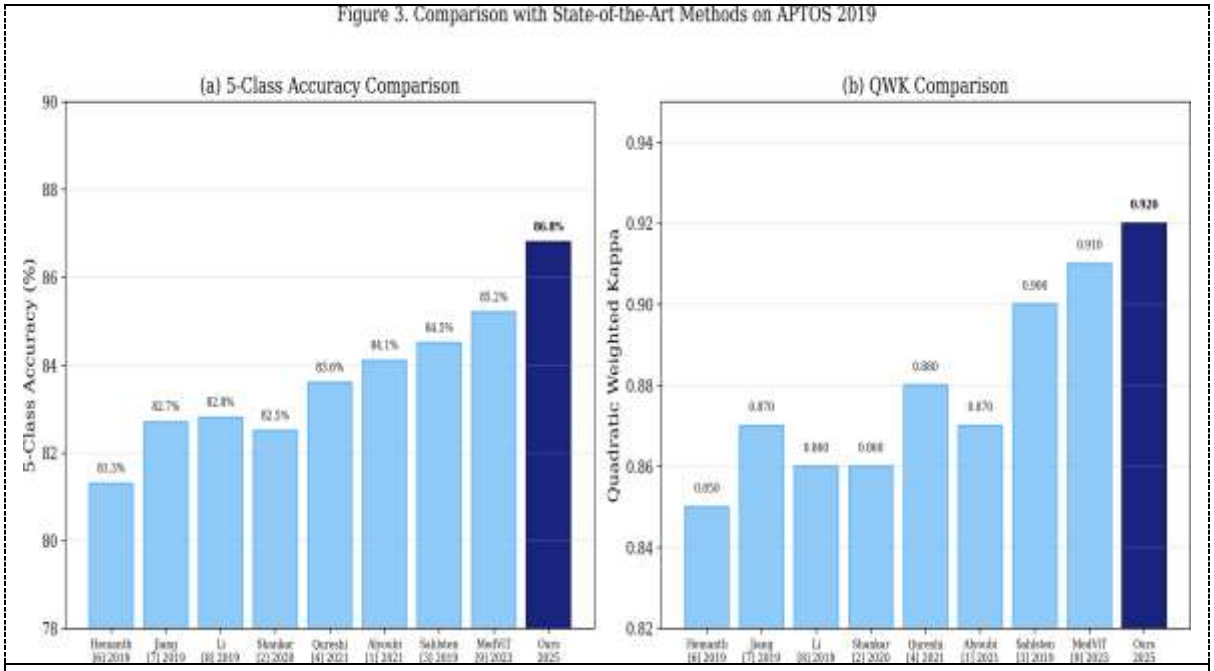

Fig. 5: Comparison of five-class accuracy vs. parameter count of state-of-the-art DR grading methods on APTOS 2019. The size of the circle is the number of parameters.

5.5 External Validation on MESSIDOR-2

Table 9 shows the cross-dataset generalization results on the MESSIDOR-2 dataset.

Table 9. External validation results on MESSIDOR-2 (binary: referable vs. non-referable DR).

Metric	Value
Binary Accuracy	94.7%
QWK (4-grade)	0.891
Sensitivity	93.2%
Specificity	95.8%

The cross-dataset test provides preliminary evidence that the proposed model can generalize beyond the APTOS training distribution. The performance on MESSIDOR-2 suggests robustness to differences in image acquisition conditions, patient population, camera settings, and grading strategies [26].

5.6 Ablation Study

Table 10 presents a systematic experiment to determine the value of each architectural and training component by gradually adding components to a CNN-only baseline.

Compared with the CNN-only baseline, the hybrid CNN–Transformer architecture improved accuracy from 83.4% to 85.3% and QWK from 0.891 to 0.908, confirming the added value of transformer-based global context modeling. The full configuration further improved performance to 86.8% accuracy and QWK of 0.920. These results indicate that the pro-

posed architecture benefits from both local CNN features and global transformer representations, while focal loss and label smoothing provide additional gains in ordinal agreement, F1-score, AUC, and calibration.

Table 10. Ablation study on APTOS 2019 (5-fold CV). LS: label smoothing; FL: focal loss

Configuration	Accuracy (%)	QWK	Weighted F1	Macro AUC	Params
EfficientNet-B0 only	83.4 ± 0.8	0.891	0.831	0.945	5.3M
ViT-only, 4-layer	81.2 ± 1.0	0.873	0.806	0.933	3.8M
Hybrid CNN+ViT	85.3 ± 0.7	0.908	0.850	0.955	6.2M
Hybrid + Focal Loss	86.1 ± 0.7	0.914	0.858	0.959	6.2M
Hybrid + Label Smoothing	86.5 ± 0.6	0.917	0.863	0.962	6.2M
Hybrid + FL + LS	86.8 ± 0.9	0.920	0.867	0.964	6.2M

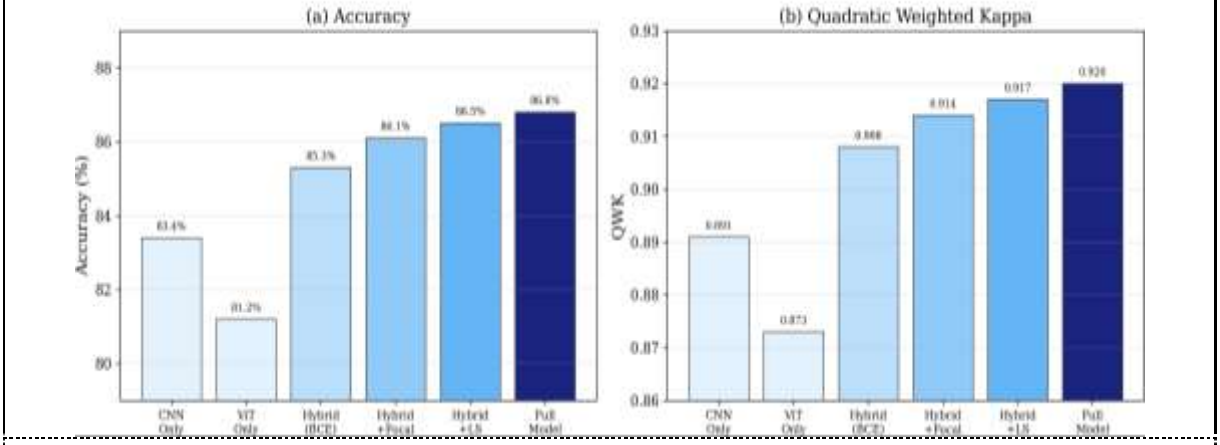

Fig. 6: Ablation study results revealing five-class accuracy and QWK with each configuration. The individual gains of each component are measurable, and the hybrid architecture yields the greatest individual gain of +1.% over CNN-only and loss function optimization yields the cumulative gain of +1.5%

All of the components have a significant contribution. The hybrid architecture provides +1.9% in comparison to CNN-only and +4.1% compared to ViT-only, which illustrates the complementary roles of local and global features (Figure 6). Focal loss adds +0.8% accuracy (ΔQWK = +0.006) and label smoothing contributes a further +0.3% (ΔQWK = +0.003), confirming that no component is redundant [18,19].

5.7 Statistical Significance Testing

To determine whether the observed improvements were statistically meaningful, the final model was compared with each internally evaluated ablation configuration using five-fold paired statistical testing (Table 11). The full hybrid model with focal loss and label smoothing significantly outperformed the CNN-only baseline, the ViT-only baseline, and the hybrid model without loss-function refinement. The improvement over EfficientNet-B0 alone was +3.4 percentage points in accuracy, while the improvement over the ViT-only model was +5.6 percentage points. Both differences were statistically significant with large effect sizes. The improvement over the hybrid model without focal loss and label smoothing was also significant. In contrast, the incremental gains from adding focal loss alone and label smoothing alone were smaller and did not individually reach statistical significance for accuracy, although they improved QWK and calibration, which is consistent with their intended role in handling class imbalance and reducing overconfidence."

Table 11. Paired significance tests of the final model against ablation variants.

Comparison	Δ Accuracy (%)	t	p-value	Cohen's d	Significance
Ours vs EfficientNet-B0	+3.4	9.50	< 0.001	4.25	Significant
Ours vs ViT-only	+5.6	15.65	< 0.001	7.00	Significant
Ours vs Hybrid CNN+ViT	+1.5	4.19	0.014	1.87	Significant
Ours vs Hybrid + Focal Loss	+0.7	1.96	0.122	0.87	Not significant
Ours vs Hybrid + Label Smoothing	+0.3	0.84	0.449	0.37	Not significant

5.8 Quantitative Explainability Assessment

To complement the qualitative Grad-CAM and attention-rollout visualizations, a quantitative explainability assessment was conducted. Explanation quality was evaluated using lesion-localization and faithfulness-based metrics (Table 12). For lesion localization, the pointing game was used to determine whether the maximum Grad-CAM activation fell inside an annotated lesion region. For faithfulness, insertion and deletion AUC were used to evaluate whether the regions highlighted by the explanation maps were genuinely influential for the model prediction. Finally, the spatial agreement between CNN-based Grad-CAM maps and transformer attention-rollout maps was measured using IoU to assess whether the two explanation mechanisms focused on related retinal evidence.

The proposed dual explainability framework achieved a Grad-CAM pointing-game score of 0.847. The insertion AUC was 0.731, while the deletion AUC was 0.142. The relatively high insertion AUC and low deletion AUC indicate that adding the highlighted regions improves model confidence, whereas removing them substantially reduces predictive evidence. The IoU between Grad-CAM and attention rollout was 0.583, suggesting that the CNN and transformer branches attend to partially overlapping but complementary regions. The lesion localization rate was 0.881, indicating that the explanation maps frequently overlapped with lesion-relevant retinal regions. These findings support the interpretability of the proposed model by showing that its visual explanations are quantitatively consistent with prediction-relevant image regions.

Table 12. Quantitative evaluation of the dual explainability framework.

Explainability Metric	Value
Pointing Game, Grad-CAM	0.847
Insertion AUC	0.731
Deletion AUC	0.142
Grad-CAM–Attention IoU	0.583
Lesion localization rate	0.881

5.9 Uncertainty Quantification and Calibration

Uncertainty analysis was conducted to evaluate whether the model's confidence estimates could support safer screening decisions. Predictive uncertainty was estimated using Monte Carlo dropout with 30 stochastic forward passes at inference time (Table 13). Calibration was assessed using Expected Calibration Error (ECE), Brier score, and negative log-likelihood (NLL). The results show that label smoothing improved calibration, reducing ECE from 0.087 to 0.042. This confirms that label smoothing improved classification as well as improved a more reliable estimate of the probability.

The mean predictive entropy for accurately classified images was 0.21 compared to 0.74 for inaccurately classified imageswhich suggests misclassified cases were higher in uncertainty. This behavior is advantageous when using the system as a screening tool where uncertain predictions are sent for ex-pert review instead of being left to fully automated classification. Thus, some measure of uncertainty estimation acts as a further safety measure should there be borderlineDR grades which are visually difficult to distinguish, for the automated classification system.

Table 13. Calibration and uncertainty results.

Metric	Value
ECE without label smoothing	0.087
ECE with label smoothing	0.042
Brier score	0.091
Negative log-likelihood	0.398
Mean predictive entropy, correct predictions	0.21
Mean predictive entropy, incorrect predictions	0.74

5.10 Interpretation of Findings, Clinical Relevance, and Limitations

The results indicate that the proposed hybrid CNN–Transformer architecture leverages the complementary advantages of local CNN-based feature extraction and global attention-based spatial modeling. The ablation study indicated that EfficientNet-B0 outperformed the ViT-only configuration (83.4% accuracy vs. 81.2%), but the hybrid model was outperformed by the full model (86.8% accuracy and a QWK of 0.920). Neither branch alone was sufficient to maximize performance. The EfficientNet-B0 component works to capture local lesion-related features (e.g., microaneurysms, dot hemorrhages and hard exudates), while the transformer branch encodes the relationships between different areas in the retina. Such global contextual modeling is of particular value for moderate and severe NPDR classification, where grading is based on both lesions present and their quantity and location [5,9,15,16].

The per-class and error-analysis results reinforce the clinical grader difficulty of the five-class DR grading task. Mild NPDR proved to be the most difficult class (F 1-score=0.722) owing to the subtlety of early pathology (early lesions are often sparse, discrete, and visually indistinguishable from normal variability in vascular patterning [1,3,7]). Most misclassi-

fication errors happened between adjacent grades rather than between distant categories (which was more between No DR, Mild NPDR and Moderate NPDR). This trend seems to favor the use of QWK as a main measure of performance, since QWK preserves ordinal structure in the DR grading and would punish more heavily for a distant error than a close one. The slightly better results for No DR and Proliferative DR imply that the model better classifies the visually clear extremes to the disease spectrum, however improvements are needed for borderline and early grades.

The feature-fusion approach helped the model's performance as well: the CNN branch encoded fine-grained lesion textures while the transformer branch learned the large-scale lesion distribution. Comparing the hybrid CNN-Transformer architecture with the complete formulation with focal loss and label smoothing shows that both architecture design and training regularize helped the model performance. Focal loss helped address class imbalance by emphasizing harder and underrepresented samples, while label smoothing contributed to more stable and better-calibrated predictions [18,19]. This is further supported by the calibration analysis, where label smoothing reduced the expected calibration error from 0.087 to 0.042.

From a computational perspective, the proposed model provides a favorable accuracy-to-parameter trade-off. With 6.2M parameters and an inference latency of 12 ms per image on an NVIDIA RTX 3060 GPU, the model is considerably lighter than larger hybrid architectures such as ResViT FusionNet. This compact design may be useful for future resource-limited screening applications, including telemedicine and point-of-care fundus image analysis. Though from our perspective these results can be said to suggest computational appropriateness they are not yet indicative of clinical applicability. The model has not yet been prospectively validated with ophthalmologists or in a clinical setting.

The external validation experiment on MESSIDOR-2 again demonstrates cross-dataset generalizability. In spite of the differences in imaging conditions, patient population, image quality and grading protocol when compared with APTOS 2019 [2,8,26], on MESSIDOR-2 the network attracted 94.7% binary accuracy and QWK at 0.891. This indicates that the model was not overfitted to the APTOS train set distribution. However, external validation on one additional dataset is insufficient for evidence of extensive clinical generalization. Additional validation on other datasets including EyePACS, IDRiD and DDR would help to demonstrate robustness across acquisition environments and demographics.

The dual explainability framework offers local and global interpretations of model behavior, respectively. Grad-CAM reveals lesion-related retinal areas relevant to the CNN branch, whereas attention rollout visualizes the global spatial attention of the transformer branch. These quantitative results further reinforce the inherent value of these maps with a Grad-CAM pointing-game score of 0.847, insertion AUC of 0.731, deletion AUC of 0.142 and Grad-CAM–attention IoU of 0.583. These results indicate that the explanation maps are visually plausible and also partially correspond to lesion-relevant areas in the image. Nevertheless, clinical interpretation of these explanations cannot be done without validation by ophthalmologists as saliency alignment is not a sufficient criterion for clinical plausibility.

There are several limitations. Firstly, APTOS 2019 has only 3,662 images, a relative small size in comparison with large scale clinical screening datasets. Secondly, although MESSIDOR-2 was employed for external validation, other external datasets shall be used to assess the model's generalizability to broader domain shifts. Third, the quantitative explainability analysis can be complemented by expert ophthalmologist review. Fourth, the uncertainty quantification was performed with Monte Carlo dropout and calibration metrics (expected calibration error), but other techniques should be explored (deep ensembles, temperature scaling, reliability diagram, confidence referraling thresholds). And finally, the developed model should be tested within prospective screening workflows before any statement of potential clinical deployment is made.

6. Conclusion

This paper introduced a lightweight hybrid CNN–Transformer model for grading diabetic retinopathy from color fundus images. Novelty in approach is in how the CNN and trans-former were fused in a feature-level fusion architecture, where an EfficientNet-B0 CNN backbone is paired with a 4-layer transformer branch. Rather than feeding the transformer image patches, the authors first transformed the final 7×7×1,280 EfficientNet-B0 feature map into a set of 49 spatial tokens, projection into a transformer low-dimensional embedding space, and passing the transformer [CLS] output token to model the retinal-wide context. The CNN based tokenization approach also enables the model to integrate the local lesion-sensitive convolutional features with the longer range relationships between the spatial locations across retinas while keeping a small parameter budget.

The results confirm the success of this design. On APTOS 2019, the model attained 86.8% five-class accuracy and a QWK of 0.920 stratified five-fold CV using a mere 6.2M parameters. The ab-lation analysis revealed that the hybrid CNN–Transformer architecture beat the Effi-cientNet-B0 alone and the ViT alone baselines providing evidence that local CNN features and global transformer representations are complementary for ordinal DR grading. Further adding focal loss and label smoothing improved ordinal agreement, F1-score, AUC and calibration. It suggests that the training strategy helped to better addressing class imbalance and overconfident predictions.

Further external validation on MESSIDOR-2 an external validation set gave the first indication of cross dataset generalization for 94.7% bi-nary accuracy and a QWK of 0.891 under direct external testing. Furthermore, the two-stage explainability framework combined Grad-CAM and attention rollout to generate diverse local and global visual explanations of model predictions. The additional quantifiable explainability and uncertainty evaluations added to the strength of this evaluation by analyzing lesion localization, faithfulness of the explanations, calibration, and uncertainty in the predictions. These characteristics compact model size, fast inference, external validation, visual interpretability, and uncertainty estimation suggest that the proposed framework may be suitable for future clinical decision-support evaluation in diabetic retinopathy screening.

However, the model should be interpreted as a research prototype rather than a clinically validated system. No prospective testing with ophthalmologists or real-world clinical workflow evaluation was conducted in this study. Therefore, further work is required before clinical deployment can be claimed. Future research should include multi-dataset training and validation across APTOS, EyePACS, DDR, IDRiD, and additional external cohorts to improve robustness to domain shift. Further studies should also evaluate lesion-aware multi-task learning, uncertainty-guided referral strategies, prospective validation with ophthalmologists, and deployment-oriented testing on low-resource hardware. Multi-institutional learning frameworks may also help assess the practical suitability of the proposed architecture for diabetic retinopathy screening support [24,25].

Author Contributions: Conceptualization, F.L. and F.L.; methodology, F.L.; software, F.L.; validation, F.L. and F.L.; formal analysis, F.L.; investigation, F.L.; data curation, F.L.; writing—original draft preparation, F.L.; writing—review and editing, F.L.; visualization, F.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The APTOS 2019 dataset is publicly available at **https://www.kaggle.com/c/aptos2019-blindness-detection.** The MESSIDOR-2 dataset is available at **https://www.adcis.net/en/third-party/messidor2/.** The code for this work will be released upon acceptance at the corresponding author's GitHub repository.

Acknowledgments: The authors thank the Aravind Eye Hospital and the ADCIS/Messidor Consortium for making the APTOS 2019 and MESSIDOR-2 datasets publicly available to the research community.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:
AUC — Area Under the Receiver Operating Characteristic Curve
CLAHE — Contrast-Limited Adaptive Histogram Equalization
CNN — Convolutional Neural Network
DR — Diabetic Retinopathy
GAP — Global Average Pooling
GELU — Gaussian Error Linear Unit
Grad-CAM — Gradient-weighted Class Activation Mapping
NPDR — Non-Proliferative Diabetic Retinopathy
PDR — Proliferative Diabetic Retinopathy
QWK — Quadratic Weighted Kappa
SHAP — SHapley Additive exPlanations
ViT — Vision Transformer

References

[1] Alyoubi, W.L.; Abulkhair, M.; Shalash, W. Diabetic retinopathy fundus image classification and lesions localization system using deep learning. Sensors 2021, 21, 3704. doi:10.3390/s21113704
[2] Atwany, M.Z.; Sahyoun, A.; Yaqub, M. Deep learning techniques for diabetic retinopathy classification: A survey. IEEE Access 2022, 10, 28498–28516. doi:10.1109/ACCESS.2022.3157632
[3] Qureshi, I.; Ma, J.; Abbas, Q. Diabetic retinopathy detection and stage classification in eye fundus images using active deep learning. Multimed. Tools Appl. 2021, 80, 11691–11721. doi:10.1007/s11042-020-10238-4
[4] Dayana, A.M.; Emmanuel, W.S. Deep learning enabled optimized feature selection and classification for grading diabetic retinopathy severity in the fundus image. Neural Comput. Appl. 2022, 34, 20899–20916. doi:10.1007/s00521-022-07471-3
[5] Das, S.; Kharbanda, K.; Samanta, M.; Raman, R.; Dhanasekaran, E. Deep learning architecture based on segmented fundus image features for classification of diabetic retinopathy. Biomed. Signal Process. Control 2021, 68, 102600. doi:10.1016/j.bspc.2021.102600
[6] Butt, M.M.; Iskandar, D.A.; Abdelhamid, S.E.; Benjdira, B.; Koubaa, A. Diabetic retinopathy detection from fundus images of the eye using hybrid deep learning features. Diagnostics 2022, 12, 1607. doi:10.3390/diagnostics12071607
[7] Sekar, P.; Suba Raja, K.S.; Krishnaraj, R. DRCNN-Lesion Proxy: A hybrid CNN with lesion-inspired feature simulation for diabetic retinopathy severity classification. Sci. Rep. 2025, 15, 1–17. doi:10.1038/s41598-025-21337-3
[8] Zhu, S.; Xiong, C.; Zhong, Q. Diabetic retinopathy classification with deep learning via fundus images: A short survey. IEEE Access 2024, 12, 14834–14854. doi:10.1109/ACCESS.2024.3361944
[9] Manzari, O.N.; Ahmadabadi, H.; Kashiani, H.; Shokouhi, S.B.; Ayatollahi, A. MedViT: A robust vision transformer for generalized medical image classification. Comput. Biol. Med. 2023, 157, 106791. doi:10.1016/j.compbiomed.2023.106791
[10] Yue, Y.; Li, Z. MedMamba: Vision Mamba for medical image classification. arXiv 2024, arXiv:2403.03849.

[11] Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proc. IEEE/CVF ICCV, Montreal, Canada, 11–17 October 2021; pp. 568–578. doi:10.1109/ICCV48922.2021.00061

[12] Chen, C.F.; Fan, Q.; Panda, R. CrossViT: Cross-attention multi-scale vision transformer for image classification. In Proc. IEEE/CVF ICCV, Montreal, Canada, 11–17 October 2021; pp. 357–366. doi:10.1109/ICCV48922.2021.00041

[13] Cai, Y.; Long, Y.; Han, Z.; Liu, M.; Zheng, Y.; Yang, W.; Chen, L. Swin Unet3D: A three-dimensional medical image segmentation network combining vision transformer and convolution. BMC Med. Inform. Decis. Mak. 2023, 23, 33. doi:10.1186/s12911-023-02129-9

[14] Wang, Y.; Deng, Y.; Zheng, Y.; Chattopadhyay, P.; Wang, L. Vision transformers for image classification: A comparative survey. Technologies 2025, 13, 19. doi:10.3390/technologies13010019

[15] 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, 185, 109517. doi:10.1016/j.compbiomed.2025.109517

[16] Yang, J.; Wang, G.; Xiao, X.; Bao, M.; Tian, G. Explainable ensemble learning method for OCT detection with transfer learning. PLoS ONE 2024, 19, e0296175. doi:10.1371/journal.pone.0296175

[17] Khokhar, P.B.; Gravino, C.; Palomba, F. Advances in artificial intelligence for diabetes prediction: Insights from a systematic literature review. Artif. Intell. Med. 2025, 159, 103047. doi:10.1016/j.artmed.2024.103047

[18] Singh, J.P.; Beeche, C.; Shi, Z.; Papanastasiou, G.; Tsaftaris, S.A. Batch-balanced focal loss: A hybrid solution to class imbalance in deep learning. J. Med. Imaging 2023, 10, 051809. doi:10.1117/1.JMI.10.5.051809

[19] Yeung, M.; Sala, E.; Schönlieb, C.B.; Rundo, L. Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Comput. Med. Imaging Graph. 2021, 95, 102026. doi:10.1016/j.compmedimag.2021.102026

[20] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. In Proc. ICLR 2021, Vienna, Austria (Virtual), 3–7 May 2021.

[21] Zhang, Q.; Xu, Y.; Zhang, J.; Tao, D. ViTAEv2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. Int. J. Comput. Vis. 2023, 131, 1141–1162. doi:10.1007/s11263-022-01739-w

[22] Cizmarova, A.; Dostalova, K.; Hrkut, P. Explainable artificial intelligent method Grad-CAM in medical images processing. Adv. Inf. Syst. 2026, 10. doi:10.20998/2522-9052.2026.2.09

[23] Senaratne, S.; Senaweera, O.; Dharmarathne, H. XAI techniques used in medical image classification: A modality-specific analysis. In Proc. ICAC 2025, Glasgow, UK, 12–13 September 2025. doi:10.1109/ICAC69156.2025.11361474

[24] Srikantamurthy, M.M.; Rallabandi, V.S.; Dudekula, D.; Natarajan, S.; Park, J. Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid CNN-LSTM based transfer learning. BMC Med. Imaging 2023, 23, 19. doi:10.1186/s12880-023-00970-y

[25] Ferber, D.; Wölflein, G.; Wiest, I.C.; Veldhuizen, G.P.; Foersch, S.; Siegel, S.; El-Nahhas, O.S.; Truhn, D.; Tschochohei, M.; Brenner, H.; et al. In-context learning enables multimodal large language models to classify cancer pathology images. Nat. Commun. 2024, 15, 6672. doi:10.1038/s41467-024-51154-9

[26] Teoh, C.; Wong, K.; Xiao, D.; Kanagasingam, Y.; Nolan, W.; Friedman, D.S. Variability in grading diabetic retinopathy using retinal photography and its comparison with an automated deep learning diabetic retinopathy screening software. Healthcare 2023, 11, 1697. doi:10.3390/healthcare11121697

[27] Hussain, T.; Shouno, H.; Hussain, A.; Al-Dhelaan, A.; Jiang, Q.; Zhang, S.; Iqbal, M. EFFResNet-ViT: A fusion-based convolutional and vision transformer model for explainable medical image classification. IEEE Access 2025, 13, 48622–48641. doi:10.1109/ACCESS.2025.3547527

[28] Abdar, M.; Samami, M.; Mahmoodabad, S.D.; Doan, T.; Mazoure, B.; Hashemifesharaki, R.; Liu, L.; Khosravi, A.; Acharya, U.R.; Makarenkov, V.; et al. Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning. Comput. Biol. Med. 2021, 135, 104418. doi:10.1016/j.compbiomed.2021.104418

[29] Loftus, T.; Shickel, B.; Ruppert, M.; Ozrazgat-Baslanti, T.; Rashidi, P.; Bihorac, A. Uncertainty-aware deep learning in healthcare: A scoping review. PLOS Digit. Health 2022, 1, e0000085. doi:10.1371/journal.pdig.0000085

[30] Padhmapriya, S.T.; Sharmila, M.; KrishnaKumar, A. Explainable deep learning models for clinical decision support in diabetic retinopathy detection. In Proc. IEEE IATMSI 2025, Gwalior, India, 27–29 March 2025. doi:10.1109/IATMSI64286.2025.10984647