

Interpretable Diabetic Retinopathy Classification Using Hybrid ConvNeXt–MLP-Mixer and ConvNeXt–PoolFormer Architectures

Mochammad Arya Jadmika¹, Aris Sugiharto², Dinar Mutiara Kusumo Nugraheni³

¹Master of Information Systems, Postgraduate Program, Diponegoro University, Semarang, Indonesia

^{2,3}Informatics Department, Faculty of Mathematics and Sciences, Diponegoro University, Semarang, Indonesia

Received:

May 27, 2026

Revised:

August 8, 2026

Accepted:

September 25, 2026

Published:

October 24, 2026

Corresponding Author:

Author Name*:

Mochammad Arya Jadmika

Email*:

aryajadmika02@gmail.com

DOI:

10.63158/journalisi.v8i5.1863

© 2026 Journal of Information Systems and Informatics. This open access article is distributed under a (CC-BY License)



Abstract. Diabetic retinopathy (DR) severity grading from retinal fundus images is challenging because retinal abnormalities vary in size, distribution, and severity. This study evaluates two hybrid architectures, ConvNeXt–MLP-Mixer and ConvNeXt–PoolFormer, for five-class DR classification by combining convolutional feature extraction with global feature mixing. Experiments used the APTOS 2019 Blindness Detection dataset from Kaggle. After exact-duplicate filtering, 3,596 images remained: 2,502 for training, 544 for validation, and 550 for held-out testing. The imbalanced test set comprised five DR severity classes. The proposed models were compared with EfficientNet, Vision Transformer, ConvNeXt, MLP-Mixer, and PoolFormer using accuracy, macro- and weighted F1-scores, and macro- and weighted-average one-versus-rest (OvR) ROC-AUC. ConvNeXt–PoolFormer achieved 82.7% test accuracy and 82.7% weighted F1-score, while ConvNeXt–MLP-Mixer achieved 82.4% accuracy and 82.5% weighted F1-score, with the latter obtaining the highest macro F1-score of 67.5%. Moderate DR showed the clearest class-specific improvement, whereas Severe and Proliferative DR remained difficult to classify. Layer-wise Relevance Propagation (LRP) and Grad-CAM generated post-hoc visual explanation maps for representative predictions. These visualizations were treated as exploratory because they were not independently validated using lesion-level annotations or by clinical experts, including ophthalmologists.

Keywords: Diabetic Retinopathy, Fundus Image Classification, ConvNeXt, MLP-Mixer, PoolFormer, Explainable AI, Medical Image Analysis, Multi-class Classification.

1. INTRODUCTION

Diabetic retinopathy (DR), a retinal microvascular complication of diabetes mellitus, remains a major cause of visual impairment and blindness, particularly in the working-age population [1]. As the burden of diabetes continues to grow, scalable methods for retinal assessment are increasingly needed, especially in settings with limited access to ophthalmic services [2]. Automated analysis of color fundus photographs has therefore become an important research area for supporting DR assessment [3]. In this study, the task is defined specifically as five-class severity grading, covering No DR, Mild, Moderate, Severe, and Proliferative DR. Thus, the study evaluates severity classification from fundus images rather than standalone clinical diagnosis or early detection.

Deep learning has enabled retinal images to be analyzed through features learned directly from the data [4], [5]. In DR classification, previous studies have explored conventional CNN architectures such as ResNet and EfficientNet, modern convolutional networks such as ConvNeXt, and Transformer-based approaches including ViT and Swin Transformer. A review covering 94 DR studies found that CNN-based methods remain widely adopted, with APTOS and EyePACS among the most frequently used public datasets [6]. Kumar and Karthikeyan also compared EfficientNet, ResNet, ViT, Swin Transformer, and MLP-Mixer for DR classification, showing that different architectural families provide different predictive and computational characteristics [7]. In the present study, the experimental comparison focuses on EfficientNet, ViT, ConvNeXt, MLP-Mixer, and PoolFormer because these models represent the standalone baselines directly evaluated against the proposed hybrid architectures. Their main characteristics and remaining gaps are summarized in Table 1.

Table 1. Representative Architectures and Explainability Methods

Approach	Representative Reference	Main Mechanism	Relevance to this Study	Remaining Gap
EfficientNet	Tan and Le (2019) [8]; Kumar and Karthikeyan (2021) [7]	Compound model scaling	Efficient CNN architecture used in DR studies	No separate global token-mixing mechanism

Approach	Representative Reference	Main Mechanism	Relevance to this Study	Remaining Gap
Vision Transformer	Dosovitskiy et al. (2021) [9]; Kumar et al. (2023) [10]	Global self-attention	Models interactions among image patches	Relatively high attention complexity
ConvNeXt	Liu et al. (2022) [11]; Devanshi et al. (2023) [12]	Modernized convolution	Used as the convolutional backbone in this study	Long-range interaction is not modeled explicitly through token mixing
MLP-Mixer	Tolstikhin et al. (2021) [13]; Kumar and Karthikeyan (2021) [7]	Token and channel mixing	Provides attention-free interaction across patches	Less studied in DR than established CNNs and Transformers
PoolFormer	Yu et al. (2022) [14]	Pooling-based token mixing	Provides a simple spatial mixing mechanism	Limited evaluation for five-class DR grading
Explainable DR Models	Jiang et al. (2020) [15]; Singh (2022) [16]; Ullah et al. (2022) [17]	Grad-CAM and relevance propagation	Provides visual information associated with predictions	Visual explanations still require clinical validation

Local and broader contextual information are both relevant to DR severity grading. Retinal abnormalities such as microaneurysms, hemorrhages, and exudates may be small and localized, but they can also occur at multiple spatially separated regions across the fundus image. Consequently, distinguishing DR severity levels may depend not only on detecting individual lesion characteristics but also on capturing how abnormal patterns are distributed across a broader retinal area. Previous DR studies have reported long-range relationships among lesion features across fundus images and demonstrated the value of combining local and broader contextual representations [18]. This does not imply that CNNs are unable to represent global information; rather, convolution has a strong locality bias, whereas token-mixing architectures provide a more explicit mechanism for exchanging information across spatial locations.

ConvNeXt was selected as the convolutional backbone because it provides strong hierarchical feature extraction while retaining the inductive properties of modern ConvNets [11]. In the proposed hybrid architectures, the ConvNeXt branch therefore emphasizes local retinal features, while the MLP-Mixer or PoolFormer branch provides a complementary mechanism for integrating information across spatially distributed regions. MLP-Mixer exchanges information across image patches through token-mixing and channel-mixing MLPs [13], allowing spatially separated retinal regions to interact within the learned representation. PoolFormer provides an alternative attention-free mechanism by using pooling-based token mixing within the MetaFormer framework [14], aggregating contextual information across neighboring spatial regions. Pairing either mechanism with ConvNeXt is therefore relevant to DR grading, where localized abnormalities and their distribution across the retina may both influence severity. MLP-Mixer has previously been evaluated for DR classification, but direct comparison between MLP-Mixer and PoolFormer when both are paired with the same ConvNeXt backbone remains limited [7].

Model transparency is another consideration in medical image analysis because predictive performance alone does not reveal which retinal regions influenced a model decision [19], [20]. For this reason, two post-hoc explanation methods were included. LRP traces prediction relevance backward toward the input image [16], [17], while Grad-CAM provides class-related activation maps from convolutional feature representations. Explainable DR studies have previously used visual explanation methods such as Grad-CAM [15]. In this study, LRP and Grad-CAM are used as post-hoc exploratory explanation methods rather than as clinically validated interpretation tools.

The main research gap is the limited evidence on how different global feature-mixing mechanisms perform when combined with the same ConvNeXt backbone for five-class DR severity grading. Accordingly, the objective of this study is to comparatively evaluate MLP-Mixer and PoolFormer as alternative global feature-mixing mechanisms when integrated with ConvNeXt, while keeping the local convolutional backbone and classification setting consistent. The main contribution is therefore the comparative evaluation of ConvNeXt–MLP-Mixer and ConvNeXt–PoolFormer, supported by standalone baseline comparisons and post-hoc explanation analysis. Based on this objective, the research question is: Which global feature-mixing mechanism, MLP-Mixer or PoolFormer,

performs better when combined with ConvNeXt for five-class diabetic retinopathy severity grading?

To answer this research question, ConvNeXt–MLP–Mixer and ConvNeXt–PoolFormer are evaluated on the same five-class DR grading task and compared with EfficientNet, ViT, ConvNeXt, MLP–Mixer, and PoolFormer using overall and class-specific predictive performance, statistical uncertainty, probability calibration, and computational efficiency. LRP and Grad-CAM are additionally used to provide exploratory post-hoc visual explanations of model predictions. All experiments in this study were conducted using the APTOS 2019 dataset, without testing the models on an independent external dataset. Thus, the reported results reflect model performance within this experimental setting and do not establish that the same performance would be obtained across other patient populations, imaging devices, or clinical environments.

2. METHODS

The study followed a sequential experimental procedure, beginning with APTOS 2019 data collection and preparation. After establishing a fixed data partition, the baseline and hybrid architectures were developed and trained, with validation data used for model selection. The selected models were subsequently assessed on the held-out test set, followed by an analysis of their visual explanations. For a consistent architectural comparison, all models used identical training, validation, and test subsets and followed the same preprocessing and evaluation protocol. Figure 1 summarizes the experimental procedure.

2.1. Dataset Acquisition

The experiments used the APTOS 2019 Blindness Detection dataset available through Kaggle [21]. Its original labeled collection contains 3,662 RGB fundus images covering five DR severity categories: No Diabetic Retinopathy (No DR), Mild, Moderate, Severe, and Proliferative Diabetic Retinopathy (PDR). These images formed the source data from which the training, validation, and held-out test subsets were prepared. Only data already available in the public APTOS 2019 collection were analyzed. No participants were recruited, no patients were contacted, and the authors collected no additional clinical data. The experiments relied on the fundus images and diagnostic labels supplied with

the dataset. The data were used for academic research under the conditions of the public source, without attempting to determine the identity of any individual represented in the images.

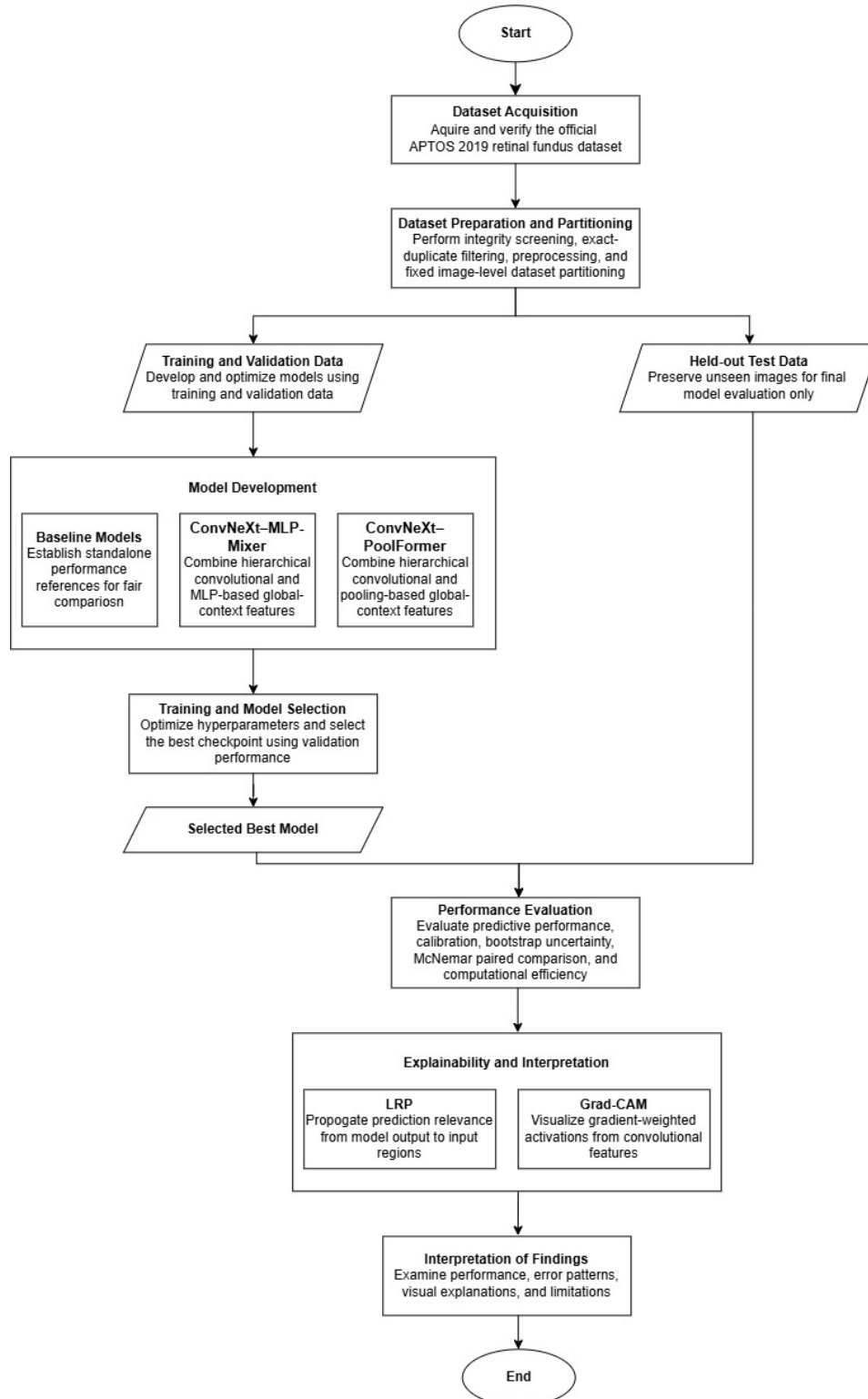


Figure 1. Research Method

2.2. Dataset Preparation and Partitioning

Image availability and label consistency were checked before the models were developed. Exact duplicate images were detected by comparing their SHA-256 hashes. The data were initially divided using a fixed stratified split of approximately 70:15:15. This partition was preserved for every baseline and hybrid model. If the same image was found in more than one subset, a single copy was kept according to the priority order of test, validation, and training. Applying this rule removed 66 cross-subset duplicate entries, leaving no exact-hash matches between the final subsets.

Following duplicate removal, the final dataset contained 3,596 images: 2,502 for training, 544 for validation, and 550 for held-out testing. Model parameters were learned from the training subset. The validation subset supported hyperparameter tuning, early stopping, checkpoint selection, and LRP-parameter selection. In contrast, the 550 test images were kept outside the model-development process and were accessed only for the final evaluation.

Table 2. Dataset Distribution after Exact-duplicate Filtering

Class	Original	Training	Validation	Testing	Final Retained
No DR	1,805	1,259	271	271	1,801
Mild	370	246	55	56	357
Moderate	999	666	146	150	962
Severe	193	134	28	29	191
Proliferative DR	295	197	44	44	285
Total	3,662	2,502	544	550	3,596

All images were converted to RGB, resized once to 224×224 pixels, and stored in a lossless PNG cache. Tensor values were normalized using the ImageNet means [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225]. Training-only augmentation comprised horizontal and vertical flips ($p = 0.5$), rotation within $\pm 20^\circ$, and color jitter with brightness 0.2, contrast 0.2, and saturation 0.1. Validation and test images received only deterministic tensor conversion and normalization. Class imbalance was addressed using weighted cross-entropy during model optimization. The class weights were calculated exclusively from the training subset using the inverse-frequency weighting scheme described in [22], as formalized in Equation (1):

$$w_c = \frac{N}{Cn_c} \quad (1)$$

where w_c denotes the weight assigned to class c , N is the total number of training samples, C is the number of DR severity classes, and n_c is the number of training samples belonging to class c . This weighting assigns greater loss contributions to underrepresented severity classes and lower contributions to more frequent classes, thereby reducing the influence of class imbalance during model optimization. The same class-weighting procedure was applied consistently across all evaluated models.

2.3. Patient-Level Partitioning Limitation

The public APTOS 2019 dataset used in this study does not provide patient identifiers that enable reliable grouping of retinal images at the patient level. Consequently, the stratified training, validation, and test partitioning was performed at the image level, and complete patient-level independence across these subsets could not be verified. To reduce direct image duplication across partitions, SHA-256 hashing was applied to identify and remove exact duplicate images, resulting in zero exact-hash overlap among the final subsets. However, this procedure cannot establish whether different images originated from the same patient. This limitation may introduce residual dependence between subsets and should therefore be considered when interpreting the reported internal test performance. Future evaluations using datasets with explicit patient identifiers should employ patient-level partitioning to provide a stricter assessment of model generalization.

2.4. Model Development

Five standalone architectures, including ConvNeXt, EfficientNet, Vision Transformer, MLP-Mixer, and PoolFormer, were used as baseline models. The proposed models consisted of two parallel dual-branch architectures, namely ConvNeXt-MLP-Mixer and ConvNeXt-PoolFormer. Both architectures were designed to combine local feature extraction and global representation learning within a unified classification framework. In this design, ConvNeXt was selected as the local feature extraction backbone, while MLP-Mixer or PoolFormer was employed as the global feature modeling branch.

In both architectures, the preprocessed retinal fundus image was represented as an RGB tensor of $3 \times 224 \times 224$ and provided independently to the two parallel branches. Thus,

the global branch received the same preprocessed image tensor as the ConvNeXt branch rather than an intermediate feature representation from the local branch. In the ConvNeXt–MLP-Mixer architecture, the MLP-Mixer B/16 branch processed the $3 \times 224 \times 224$ input using non-overlapping 16×16 image patches for token- and channel-mixing operations. In the ConvNeXt–PoolFormer architecture, the PoolFormer-S36 branch received the same $3 \times 224 \times 224$ tensor and processed it through its hierarchical pooling-based stages. The ConvNeXt branch extracted local hierarchical representations associated with retinal structures, whereas the MLP-Mixer or PoolFormer branch modeled broader contextual information through its respective feature-mixing mechanism.

The original classification heads of all backbone networks were removed to allow them to function as feature extractors. The resulting feature representations were subsequently fused and passed through a fully connected classifier to predict the five diabetic retinopathy severity classes. All backbone parameters were optimized using an end-to-end fine-tuning strategy. The overall architecture of the proposed hybrid models is illustrated in Figure 2.

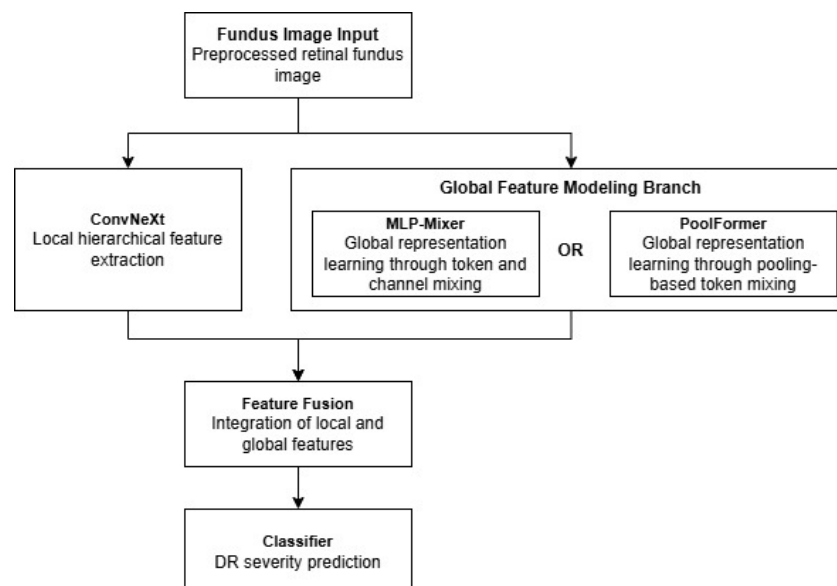


Figure 2. Proposed ConvNeXt–MLP-Mixer and ConvNeXt–PoolFormer Architecture

The two proposed architectures share an identical feature-projection, fusion, and classification strategy, while their primary difference lies in the global feature modeling

branch. ConvNeXt–MLP-Mixer employs token- and channel-mixing operations to model interactions among image patches and feature channels, whereas ConvNeXt–PoolFormer uses pooling-based token mixing within a hierarchical representation. The branch-specific feature representations are independently projected to 256 dimensions and concatenated to form a 512-dimensional fused representation, which is subsequently processed by the shared classifier. The detailed configuration of each component is summarized in Table 3.

Table 3. Configuration of the Proposed Hybrid Models

Component	ConvNeXt–MLP-Mixer	ConvNeXt–PoolFormer
Local Branch	ConvNeXt-Tiny (768-D)	ConvNeXt-Tiny (768-D)
Global Branch	MLP-Mixer B/16 (768-D)	PoolFormer-S36 (512-D)
Branch Projection	256-D per branch	256-D per branch
Fusion	Concatenation (512-D)	Concatenation (512-D)
Classifier	512-256-5	512-256-5
Training	End-to-end fine-tuning	End-to-end fine-tuning

2.5. Training and Model Selection

The experiments were implemented using Python 3.12.13 with PyTorch 2.10.0, timm 1.0.26, and CUDA 12.8 on an NVIDIA Tesla T4 GPU. All backbone networks were initialized using pretrained weights and optimized using an end-to-end fine-tuning strategy. A unified training protocol was applied to all baseline and proposed models to ensure a fair comparison. The models were optimized using AdamW with weighted cross-entropy loss, while mixed-precision training and learning-rate scheduling were applied to improve computational efficiency and training stability.

Hyperparameter optimization was performed using grid search on the proposed hybrid models. Different combinations of learning rate, dropout rate, and architecture depth were tested. Their performance was compared on the validation subset, with macro-F1 used as the main criterion for choosing the best checkpoint. When multiple configurations reached the same macro-F1, validation accuracy was used to break the tie.

The held-out test set was not involved in hyperparameter optimization, early-stopping decisions, checkpoint selection, or explainability parameter tuning. The complete training configuration and hyperparameter search space are summarized in Table 4.

Table 4. Core Training and Hyperparameter Settings

Setting	Value
Optimizer and Loss	AdamW; weighted cross-entropy
Batch Configuration	16; accumulation 2; effective 32
Training Duration	Maximum 10 epochs; patience 3
Scheduler	CosineAnnealingLR
Learning Rates	1×10^{-5} ; 3×10^{-5} ; 1×10^{-4}
Dropout	0.2; 0.3; 0.4
Mixer-block	2; 4; 6
PoolFormer-depth	2; 4; 6
Selection Criterion	Validation macro-F1; accuracy tie-break
Random Seed	42

2.6. Held-out Test Evaluation

The best validation-selected checkpoint from each baseline and hybrid model was evaluated on the same held-out test set. Predictive performance was measured using accuracy; macro and weighted precision, recall, and F1-score; one-versus-rest macro and weighted ROC-AUC; per-class metrics; and raw and normalized confusion matrices. Statistical uncertainty was summarized using 95% bootstrap confidence intervals for accuracy and macro-F1 based on 1,000 resamples. To assess whether the two proposed hybrid models differed significantly in their classification outcomes on the same test samples, an exact two-sided McNemar test was applied to their paired predictions. The test was based on discordant cases in which one hybrid model produced a correct prediction while the other produced an incorrect prediction. Statistical significance was assessed at $\alpha = 0.05$. Predictive confidence and calibration were assessed using softmax probabilities, multiclass Brier score, Expected Calibration Error with 15 bins, and a reliability diagram. Computational efficiency was evaluated using parameter counts, MACs, approximate FLOPs, model size, batch-one latency after 20 warm-up and 100 measured runs, end-to-end inference time, and peak GPU memory.

2.7. Explainability and Interpretation

Post-hoc explanations were generated for representative correct and misclassified test predictions. LRP was implemented with LXT and Zennit Gamma rules to obtain signed pixel relevance, while Grad-CAM used the final depthwise convolutional layer of the ConvNeXt branch. LRP Gamma parameters were selected only from validation images using attribution finiteness, relevance-conservation residual, and normalized total variation; test images were not used for this tuning. Grad-CAM therefore represented the local ConvNeXt contribution rather than the complete dual-branch model. Explanations were interpreted together with confidence and error patterns, with particular attention to Moderate DR misclassifications. Because no clinical specialist independently validated the maps, they were treated as computational explanations rather than definitive clinical evidence.

3. RESULTS AND DISCUSSION

The experimental results are examined from several perspectives rather than through accuracy alone. The analysis covers model selection, overall and class-specific performance, statistical uncertainty, calibration, computational cost, and post-hoc explanations for the five standalone models and the two proposed hybrids. Together, these results provide a broader comparison of model behavior in the five-class DR grading task.

3.1. Hyperparameter Optimization and Model Selection

Hyperparameter optimization produced architecture-dependent configurations, indicating that a single learning rate, regularization, or depth setting was not optimal for all models. The selected configurations are summarized in Table 5. Most architectures selected a learning rate of 1×10^{-4} , whereas MLP-Mixer selected the lower learning rate of 1×10^{-5} . The selected dropout values ranged from 0.2 to 0.4. Six blocks were selected for MLP-Mixer and ConvNeXt-MLP-Mixer, while PoolFormer and ConvNeXt-PoolFormer selected depths of four and six, respectively. The resulting parameter counts also varied substantially, from 4.01 million parameters for EfficientNet to 85.80 million for Vision Transformer, while the two hybrid models contained approximately 58 million parameters each.

Table 5. Selected Hyperparameter Configuration

Model	Best LR	Best Dropout	Best Block/Depth	Epoch	Params (M)	Effective Batch Size
EfficientNet	1×10^{-4}	0.3	-	8	4.01	32
Vision Transformer	1×10^{-4}	0.2	-	9	85.80	32
ConvNeXt	1×10^{-4}	0.3	-	6	27.82	32
MLP-Mixer	1×10^{-5}	0.2	6	9	29.86	32
PoolFormer	1×10^{-4}	0.4	4	9	26.15	32
ConvNeXt-MLP-Mixer	1×10^{-4}	0.2	6	8	58.20	32
ConvNeXt-PoolFormer	1×10^{-4}	0.4	6	9	58.63	32

The validation and test results of the selected configurations are reported separately in Table 6 so that model selection performance can be distinguished from final held-out evaluation. ConvNeXt-PoolFormer achieved the highest validation macro-F1 of 74.0% and the highest validation accuracy of 85.5%. ConvNeXt achieved a validation macro-F1 of 71.9%, followed by ConvNeXt-MLP-Mixer at 71.4%. On the test set, ConvNeXt-PoolFormer achieved the highest accuracy of 82.7%, followed closely by ConvNeXt-MLP-Mixer at 82.4%, while standalone ConvNeXt achieved 80.0%. Thus, the hybrid models improved the test-accuracy point estimate by 2.7 and 2.4 percentage points, respectively, relative to ConvNeXt.

Table 6. Validation and Test Results of the Evaluated Models

Model	Val. Macro-F1	Val. Accuracy	Test Accuracy
EfficientNet	65.2%	78.7%	78.9%
Vision Transformer	47.8%	66.4%	65.1%
ConvNeXt	71.9%	82.9%	80%
MLP-Mixer	63.9%	76.5%	77.8%
PoolFormer	62.1%	75.6%	75.1%
ConvNeXt-MLP-Mixer	71.4%	83.6%	82.4%
ConvNeXt-PoolFormer	74.0%	85.5%	82.7%

The training history of ConvNeXt-MLP-Mixer is shown in Figure 3. Its training loss continued to fall across epochs, but the validation loss dropped mainly during the early

stage and became more variable afterward. A similar divergence appeared in accuracy: training accuracy continued to increase, while validation accuracy was strongest around epoch 8. Validation macro-F1 rose from approximately 0.52 in the first epoch to about 0.71 at epoch 8 and then declined slightly. For this reason, the epoch-8 checkpoint was retained; continuing training beyond this point increased the separation between training and validation behavior without improving class-balanced validation performance.

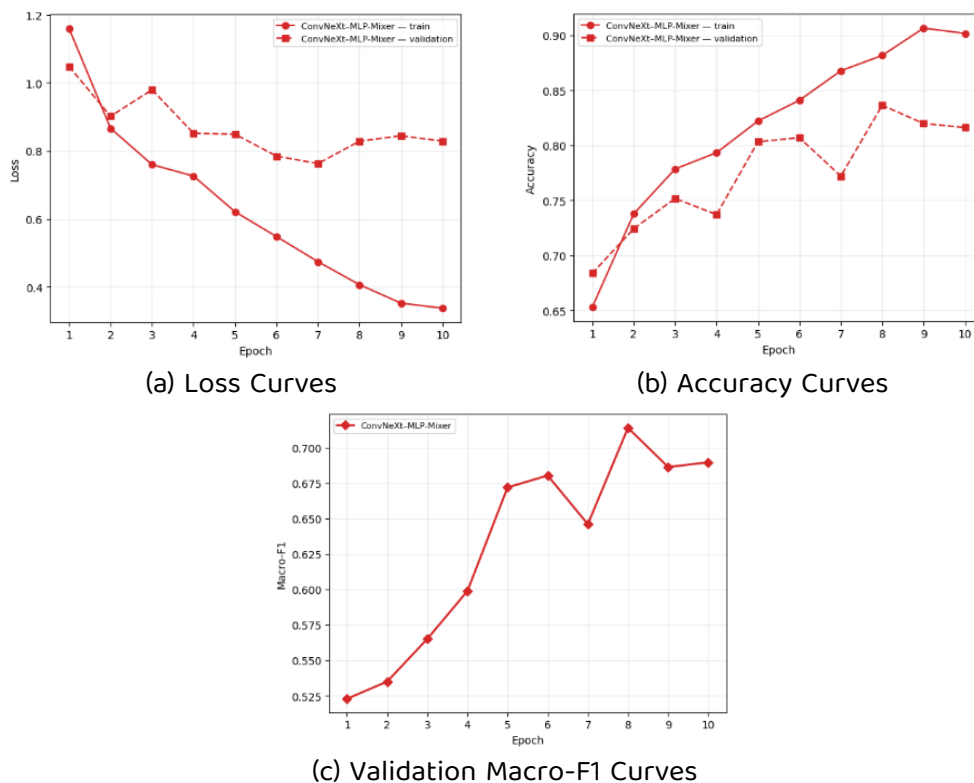
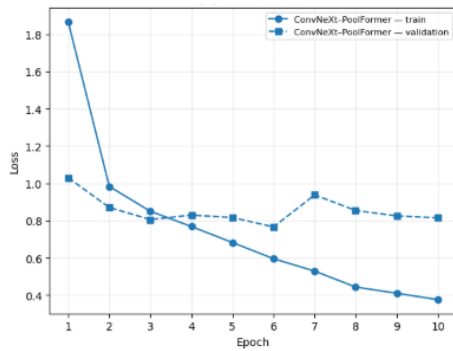
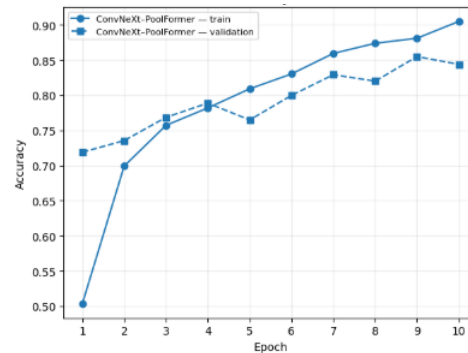


Figure 3. Training History of the ConvNeXt-MLP-Mixer Model

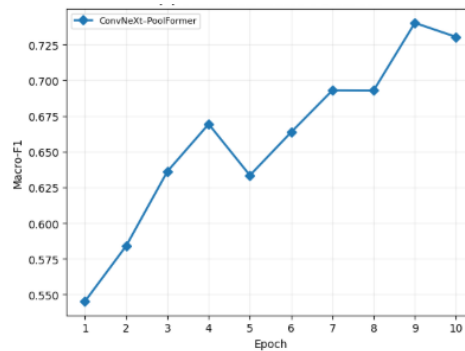
Figure 4 shows the corresponding training behavior of ConvNeXt-PoolFormer. Training loss decreased continuously, while validation loss improved during the early epochs but showed temporary increases later in training. Training accuracy increased to approximately 0.90, whereas validation accuracy reached its highest region around epoch 9 before a slight reduction. Validation macro-F1 increased overall from approximately 0.55 to 0.74 and peaked at epoch 9. The subsequent decrease at epoch 10 supports the epoch-9 checkpoint and indicates that additional training did not improve class-balanced validation performance.



(a) Loss Curves



(b) Accuracy Curves



(c) Validation Macro-F1 Curves

Figure 4. Training History of the ConvNeXt–PoolFormer Model

3.2. Overall Performance and Baseline Comparison

Table 7 compares the overall test performance of the five standalone baselines and the two hybrid architectures using accuracy, macro precision, macro recall, macro-F1, weighted F1, and one-vs-rest ROC-AUC. ConvNeXt PoolFormer achieved the highest accuracy of 82.7% and the highest weighted F1 of 82.7%, followed closely by ConvNeXt-MLP-Mixer with 82.4% accuracy and 82.5% weighted F1. These results indicate a modest overall performance gain from adding a global-processing branch to ConvNeXt.

Table 7. Overall Test Performance of Baseline and Hybrid Models

Model	Test Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1	Macro AUC OvR	Weighted AUC OvR
EfficientNet	78.9%	63.1%	66.3%	63.7%	79.4%	93.6%	96.2%
Vision Transformer	65.1%	49.6%	51.3%	45.2%	65.5%	86.9%	91.6%
ConvNeXt	80%	65.1%	69.5%	66.5%	80.7%	95.5%	97.2%
MLP-Mixer	77.8%	63%	68.2%	63.8%	78.4%	93.6%	96%

Model	Test Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1	Macro AUC OvR	Weighted AUC OvR
PoolFormer	75.1%	58%	61.7%	58.7%	75.7%	92.8%	95.6%
ConvNeXt-MLP-Mixer	82.4%	67.2%	68%	67.5%	82.5%	94.6%	96.9%
ConvNeXt-PoolFormer	82.7%	67.7%	67%	67.1%	82.7%	95%	97%

The ranking of the evaluated models changed according to the performance metric. ConvNeXt-MLP-Mixer achieved the highest macro-F1 of 67.5%, slightly exceeding ConvNeXt-PoolFormer at 67.1% and standalone ConvNeXt at 66.5%. In contrast, ConvNeXt achieved the highest macro recall of 69.5% and the highest macro and weighted one-vs-rest AUC values of 95.5% and 97.2%, respectively. Therefore, ConvNeXt-PoolFormer is best described as having the highest accuracy point estimate rather than as being uniformly superior across all evaluation criteria. The AUC values in Table 6 are explicitly reported as macro-average and weighted-average one-vs-rest measures, which is important because these summaries respond differently to class imbalance.

The standalone ConvNeXt, MLP-Mixer, and PoolFormer results also provide an architecture level comparison with the two hybrids. However, because the architectures were independently optimized, this comparison should not be presented as a controlled tuned versus untuned concatenation ablation. Likewise, LRP is a post-hoc explanation procedure and does not change the classifier output, so a with versus without LRP predictive ablation would not be methodologically meaningful.

3.3. Class-Specific Performance and Error Analysis

Class specific analysis was performed for ConvNeXt and the two proposed hybrid models to determine whether the overall performance gains were distributed consistently across diabetic retinopathy severity levels. Table 8 reports precision, recall, F1-score, and one-vs-rest AUC for No DR, Mild, Moderate, Severe, and Proliferative DR.

Table 8. Class-specific Performance of the Selected Model

Model	Class	Precision	Recall	F1-Score	AUC OvR
ConvNeXt	No DR	98.9%	95.9%	97.4%	99.9%

Model	Class	Precision	Recall	F1-Score	AUC OvR
	Mild	51.9%	73.2%	60.7%	94.3%
	Moderate	81.4%	64%	71.6%	95%
	Severe	42.4%	48.3%	45.2%	93.9%
	Proliferative DR	50.9%	65.9%	57.4%	94.4%
	No DR	98.5%	98.2%	98.3%	99.9%
ConvNeXt-MLP-Mixer	Mild	63%	60.7%	61.8%	95.1%
	Moderate	77.6%	76%	76.8%	94.7%
	Severe	40%	48.3%	43.8%	91.3%
	Proliferative DR	56.8%	56.8%	56.8%	92.2%
	No DR	98.2%	98.2%	98.2%	99.7%
ConvNeXt-PoolFormer	Mild	56.4%	62.5%	59.3%	94.2%
	Moderate	78.8%	79.3%	79.1%	95.3%
	Severe	40.6%	44.8%	42.6%	92.8%
	Proliferative DR	64.7%	50%	56.4%	92.9%
	No DR	98.2%	98.2%	98.2%	99.7%

All three models performed strongly for the No DR class, with F1-scores ranging from 97.4% to 98.3% and one-versus-rest AUC values of at least 99.7%. Performance was substantially lower for the remaining severity levels. For Mild DR, the F1-score was 60.7% for ConvNeXt, 61.8% for ConvNeXt-MLP-Mixer, and 59.3% for ConvNeXt-PoolFormer. Moderate DR showed the clearest benefit from the hybrid architectures, with the F1-score increasing from 71.6% for ConvNeXt to 76.8% for ConvNeXt-MLP-Mixer and 79.1% for ConvNeXt-PoolFormer. The pattern was different for Severe DR, which had the lowest F1-score for each of the three models: 45.2% for ConvNeXt, 43.8% for ConvNeXt-MLP-Mixer, and 42.6% for ConvNeXt-PoolFormer. Proliferative DR was also difficult to distinguish, with corresponding F1-scores of 57.4%, 56.8%, and 56.4%. The variation across classes is important because the high scores obtained for No DR and the improvement observed for Moderate DR are not representative of model behavior at every severity level. For this reason, overall accuracy needs to be considered together with class-specific performance when evaluating the five-class grading task.

Several observations may help explain the weaker results for Severe DR. First, this class had the fewest examples in the data used for the experiments, with 134 images in the training subset and only 29 in the held-out test subset. Weighted cross-entropy gave

greater importance to errors from such underrepresented classes during training, but it could not increase the variety of Severe DR examples available to the model. Second, the confusion matrices indicate that errors involving Severe DR were concentrated mainly in the adjacent Moderate and Proliferative DR grades. The difficulty therefore appears to involve discrimination among neighboring severity levels rather than simply separating DR from No DR. Taken together, the limited Severe DR samples and the errors between adjacent grades provide plausible explanations for its low F1-score. These observations should not be interpreted as a pathological explanation for the errors, however. Lesion-level annotations and independent assessment by clinical experts were not available in this study, so the specific retinal findings responsible for these misclassifications could not be verified.

Moderate DR was the class for which the hybrid models showed the largest gain over standalone ConvNeXt. Its recall increased from 64.0% with ConvNeXt to 76.0% with ConvNeXt-MLP-Mixer and 79.3% with ConvNeXt-PoolFormer. The same pattern appeared in F1-score, which rose from 71.6% to 76.8% and 79.1%, respectively. One possible interpretation is that the global branch adds information from spatially separated retinal regions to the local features learned by ConvNeXt, which may be useful for Moderate cases containing abnormalities at multiple locations. This advantage was class-dependent, however, because comparable gains were not observed for Severe and Proliferative DR. Thus, the addition of a global feature-mixing branch improved selected class-specific outcomes but did not eliminate the challenges associated with minority and adjacent-severity classes.

Figure 5 confirms the class-specific performance pattern. The correct classification rate for Moderate DR increased from 64% with ConvNeXt to 76% with ConvNeXt-MLP-Mixer and 79% with ConvNeXt-PoolFormer. Mild DR was frequently confused with Moderate DR, particularly for ConvNeXt-MLP-Mixer (32%) and ConvNeXt-PoolFormer (29%). Severe DR was also frequently assigned to Moderate or Proliferative DR, while Proliferative DR predictions were distributed between the correct class and neighboring advanced-severity categories. These errors were therefore concentrated primarily between adjacent severity levels rather than between No DR and advanced DR. The pattern further supports the view that class imbalance and similarity between neighboring severity

grades remained important sources of error despite the use of weighted loss and hybrid local-global feature modeling.

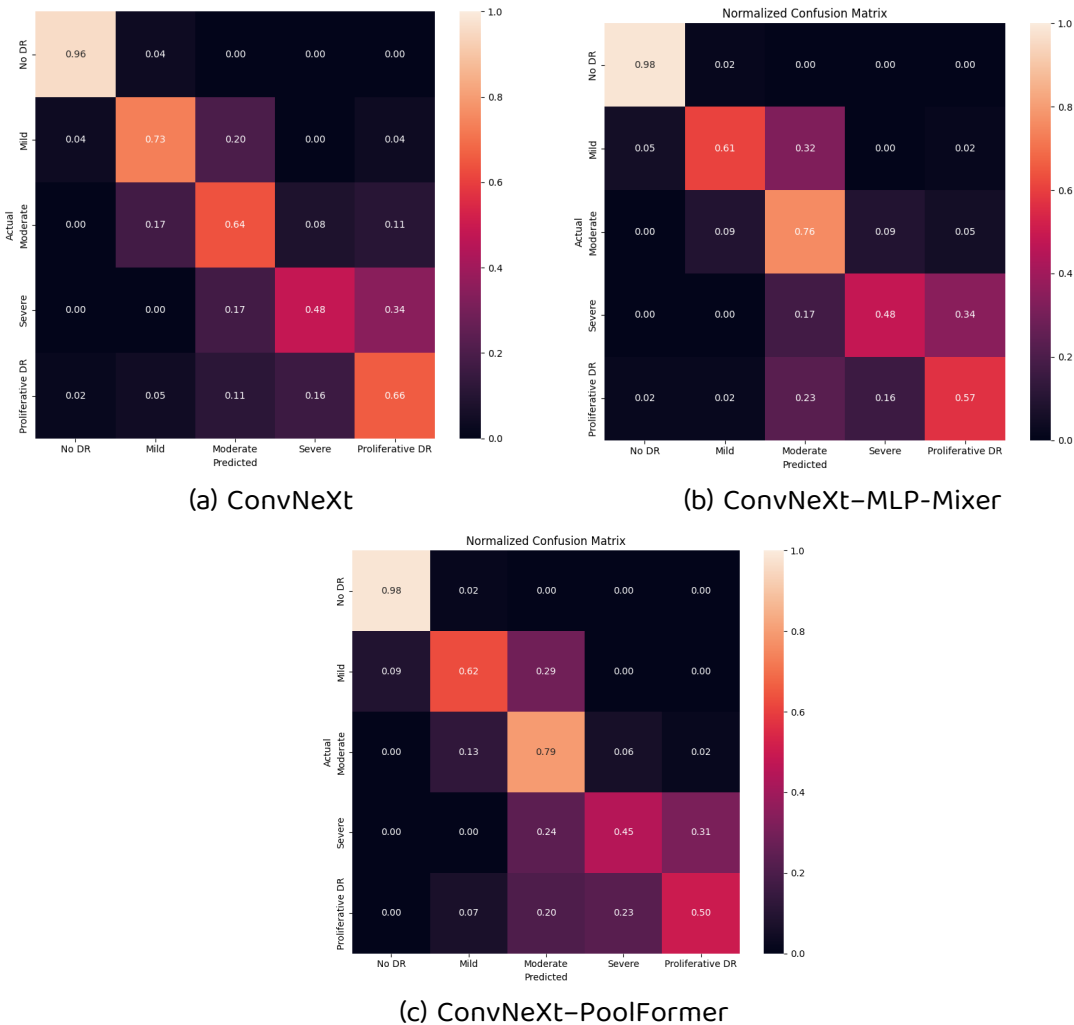


Figure 5. Normalized Confusion Matrices

The relatively strong Moderate DR performance of the hybrid models, together with the persistent confusion involving Severe DR, illustrates that global feature mixing does not benefit every severity level equally. Moderate cases may benefit from the integration of spatially distributed retinal information, whereas Severe cases remain difficult when their learned representations overlap with both Moderate and Proliferative DR. Nevertheless, the present results only demonstrate classification behavior and do not establish the underlying clinical or lesion-level reason for each misclassification.

3.4. Statistical Reliability and Calibration

Point estimates were supplemented with 95% confidence intervals and probability-calibration metrics to characterize sampling uncertainty and the reliability of model confidence. Table 9 reports accuracy and macro-F1 confidence intervals together with the multiclass Brier score and Expected Calibration Error (ECE).

Table 9. Statistical Uncertainty and Calibration Performance

Model	Accuracy	Accuracy 95% CI	Macro F1	Macro F1 95% CI	Brier Score	ECE
EfficientNet	78.9%	75.6-82.2%	63.7%	58.4-68.8%	0.302	0.055
Vision Transformer	65.1%	61.1-68.9%	45.2%	40.6-49.3%	0.415	0.031
ConvNeXt	80%	76.9-83.3%	66.5%	61.3-71.5%	0.265	0.042
MLP-Mixer	77.8%	74.4-81.1%	63.8%	58.7-68.5%	0.314	0.055
PoolFormer	75.1%	71.6-78.4%	58.7%	53.7-63.4%	0.369	0.133
ConvNeXt- MLP-Mixer	82.4%	79.3-85.3%	67.5%	62.1-72.3%	0.255	0.048
ConvNeXt- PoolFormer	82.7%	79.6-85.8%	67.1%	61.4-72.1%	0.251	0.062

ConvNeXt-PoolFormer achieved an accuracy of 82.7% with a 95% confidence interval of 79.6–85.8%, whereas ConvNeXt-MLP-Mixer achieved 82.4% with an interval of 79.3–85.3%. Their macro-F1 estimates were 67.1% (95% CI: 61.4–72.1%) and 67.5% (95% CI: 62.1–72.3%), respectively. The substantial overlap between these confidence intervals indicates considerable uncertainty around the small differences in their point estimates. However, overlapping confidence intervals alone do not constitute a formal paired comparison because both models were evaluated on the same test observations.

To assess the paired difference directly, an exact two-sided McNemar test was applied to the predictions of the two hybrid models on the same held-out test set. Among the discordant cases, ConvNeXt-PoolFormer correctly classified 23 images that were misclassified by ConvNeXt-MLP-Mixer, whereas ConvNeXt-MLP-Mixer correctly classified 20 images that were misclassified by ConvNeXt-PoolFormer. The resulting difference was not statistically significant ($p = 0.7608, \alpha = 0.05$). This result indicates

that the small difference in overall classification outcomes between the two hybrid architectures was not statistically supported on the present test set. Accordingly, the observed accuracy values of 82.7% and 82.4% represent a small numerical difference on this test set, and the paired analysis does not support concluding that one hybrid statistically outperformed the other.

Calibration analysis revealed a different ranking from classification performance. ConvNeXt–PoolFormer achieved the lowest multiclass Brier score of 0.251, narrowly outperforming ConvNeXt–MLP-Mixer at 0.255 and ConvNeXt at 0.265. Because the Brier score evaluates the overall discrepancy between predicted probabilities and observed outcomes, these lower values indicate comparatively better probabilistic prediction quality for the two hybrid models. In contrast, the ECE ranking differed: ConvNeXt achieved an ECE of 0.042, ConvNeXt–MLP-Mixer 0.048, and ConvNeXt–PoolFormer 0.062. Thus, the model with the highest accuracy did not obtain the lowest calibration error.

ViT illustrates particularly well why calibration and classification ability need to be examined separately. Among all evaluated architectures, it obtained the smallest ECE at 0.031. However, the same model achieved only 65.1% accuracy and 45.2% macro-F1, both of which were the lowest results in the comparison. Its confidence estimates can therefore show relatively close agreement with its observed correctness while the model still makes many classification errors. In other words, the low ECE reported for ViT describes one aspect of its confidence behavior and should not be taken as an indication that it provides the strongest predictions.

ConvNeXt–PoolFormer showed the opposite situation: it produced the highest accuracy, but its ECE was higher than those of ConvNeXt and ConvNeXt–MLP-Mixer. This result shows that better classification accuracy was not accompanied by a corresponding improvement in ECE. The Brier score adds another perspective because it evaluates errors across the predicted class-probability distribution, whereas ECE summarizes the discrepancy between confidence and observed accuracy after predictions are grouped into bins. Since these metrics capture different properties of the predicted probabilities, their values should be considered together rather than using either measure alone to judge calibration. ECE is also sensitive to how the confidence bins are defined, which should be considered when interpreting small differences between models.

Considering these results together gives a more complete picture than accuracy alone. The confidence intervals and paired McNemar test indicate that the two hybrids performed similarly on the current test set, without a statistically significant difference in their paired predictions. At the same time, Brier score and ECE did not favor the models in an identical order. The evaluation should therefore consider classification performance, class balance, statistical uncertainty, and calibration as separate but complementary properties of model behavior.

3.5. Computational Efficiency

Table 10 compares the computational characteristics of the standalone and hybrid architectures. The comparison covers parameter count, GFLOPs, model size, training time, inference latency per image, throughput, and peak GPU memory usage. Together, these measurements describe both the theoretical computational requirements and the observed runtime behavior of each model.

Table 10. Computational Efficiency

Model	Params (M)	GFLOPs	Size (MB)	Training (min)	Latency (ms/image)	Throughput (images/s)	Peak GPU (MB)
EfficientNet	4.01	0.83	15.69	1.55	7.59	131.75	74.52
Vision Transformer	85.80	22.57	327.41	5.11	14.58	68.59	1007.88
ConvNeXt	27.82	8.91	106.30	2.15	5.27	189.75	355.33
MLP-Mixer	29.86	12.73	113.95	2.01	5.84	171.23	330.44
PoolFormer	26.15	9.54	100.17	4.09	13.77	72.62	331.54
ConvNeXt-MLP-Mixer	58.20	21.64	222.22	4.37	11.69	85.54	673.47
ConvNeXt-PoolFormer	58.63	18.86	224.27	6.82	23.74	42.12	703.99

Among the evaluated architectures, EfficientNet required the fewest parameters (4.01 M), the lowest operation count (0.83 GFLOPs), and the smallest model storage (15.69 MB). The fastest inference in the experiment was instead obtained by ConvNeXt, which processed an image in 5.27 ms on average and reached a throughput of 189.75 images/s. Vision

Transformer was the largest standalone model, with 85.80 million parameters and a model size of 327.41 MB.

Among the hybrid models, ConvNeXt-PoolFormer required fewer theoretical operations than ConvNeXt-MLP Mixer (18.86 versus 21.64 GFLOPs), representing an approximately 12.8% reduction in GFLOPs. However, the lower operation count did not translate into lower measured runtime or memory usage. ConvNeXt-PoolFormer required 23.74 ms per image and achieved 42.12 images per second, whereas ConvNeXt-MLP-Mixer required 11.69 ms per image and achieved 85.54 images per second. The PoolFormer hybrid also had a longer reported training time, slightly larger model size, and higher peak GPU memory. Therefore, the PoolFormer hybrid can be described as having a lower theoretical operation count, but not lower measured latency or memory consumption under the reported implementation.

From a deployment perspective, the preferred model depends on the operational priority. ConvNeXt-PoolFormer provides the highest accuracy point estimate, whereas ConvNeXt-MLP-Mixer provides a more favorable latency throughput trade-off and a slightly higher macro-F1. Standalone ConvNeXt remains attractive when inference speed is prioritized over the modest accuracy gains of the hybrid architectures.

3.6. Explainability Analysis

LRP and Grad-CAM were applied to both hybrid models to provide complementary post-hoc visualizations of model predictions. LRP propagates relevance from the final hybrid prediction through the fusion mechanism and both model branches to the input, producing fine-grained pixel-level signed relevance patterns. In contrast, Grad-CAM produces broader class-discriminative activation regions from the final convolutional layer of the ConvNeXt branch. Therefore, LRP provides an input-level relevance explanation of the fused hybrid decision, whereas Grad-CAM provides a coarser spatial view of activation within the local convolutional pathway. Correctly classified Severe and Proliferative DR examples are shown first, followed by representative misclassified Moderate DR examples.

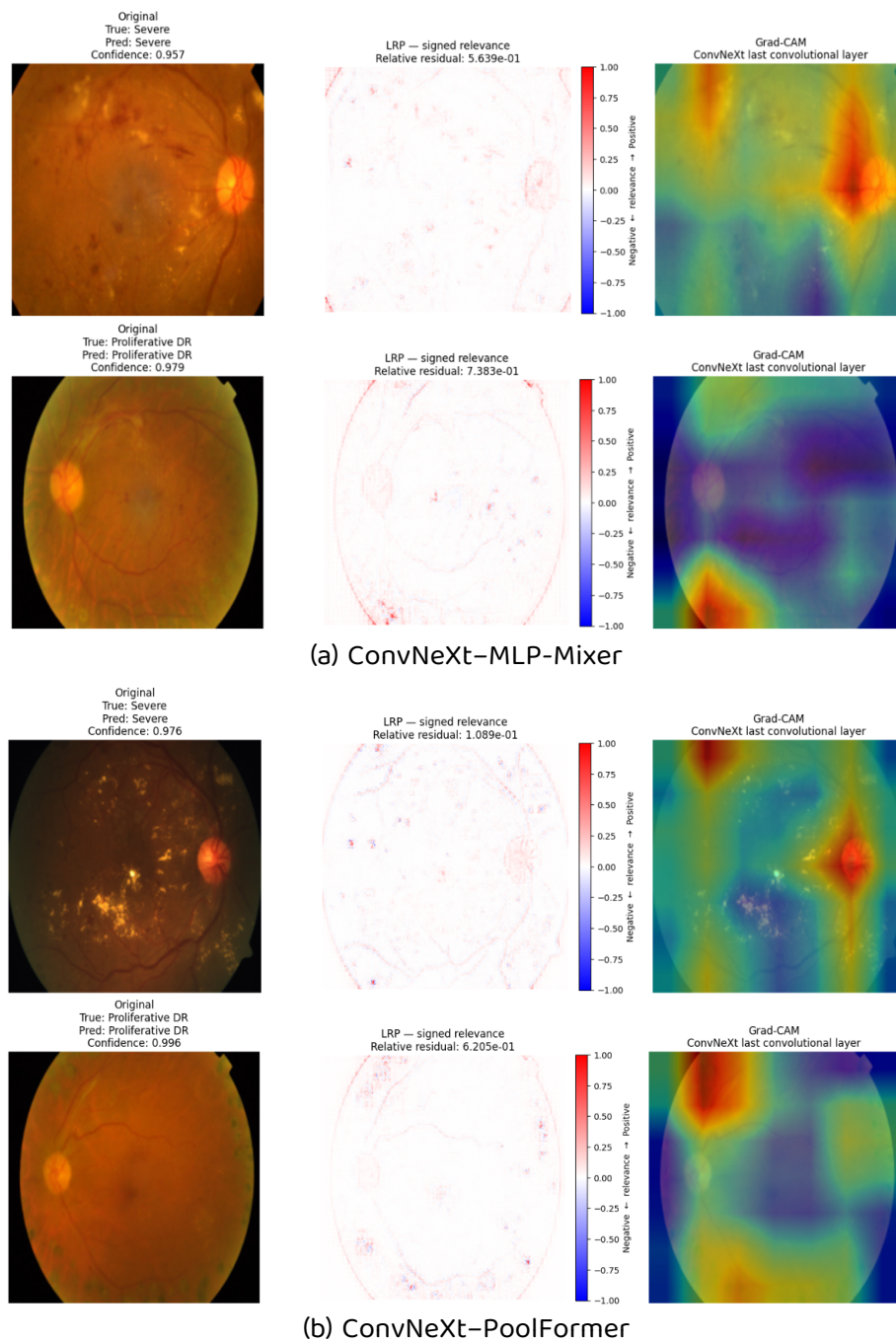


Figure 6. Correct Predictions with LRP and Grad-CAM

For ConvNeXt-MLP-Mixer, the illustrated Severe and Proliferative DR cases were correctly classified with confidence values of 0.957 and 0.979, respectively. For ConvNeXt-PoolFormer, the corresponding illustrated examples were classified correctly with confidence values of 0.976 and 0.996. Consistent with their different explanation mechanisms, the LRP maps provide fine-grained pixel-level relevance patterns across

retinal structures, whereas the Grad-CAM maps highlight broader activation regions associated with the final ConvNeXt convolutional features. In several examples, the Grad-CAM regions overlapped visually abnormal retinal areas, while LRP distributed finer relevance patterns over retinal structures. However, some explanation maps also emphasized the optic disc, peripheral retina, or broad illumination-related regions. These visualizations therefore indicate image regions associated with the model predictions but do not establish that the models relied on clinically valid diagnostic features.

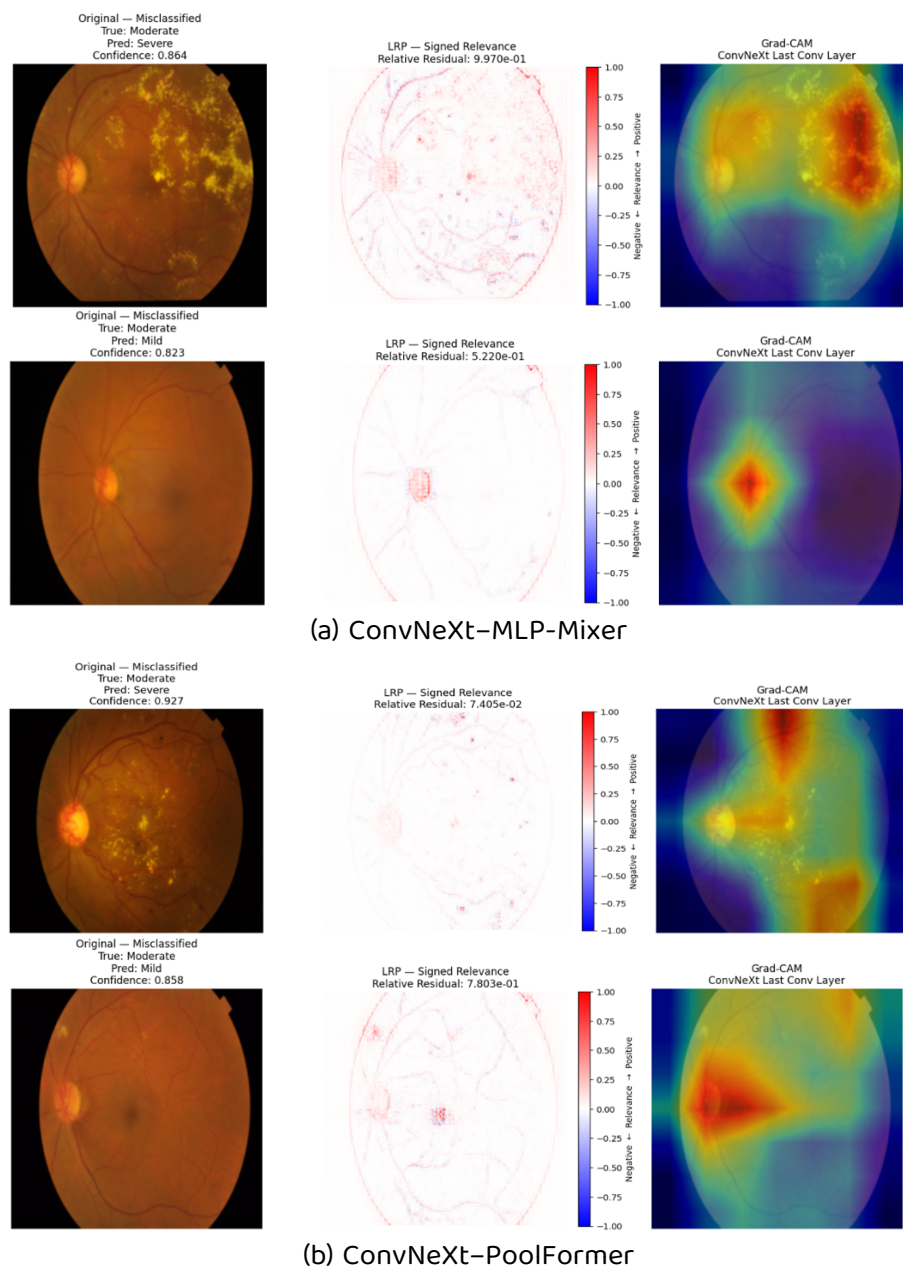


Figure 7. Misclassified Predictions with LRP and Grad-CAM

The misclassified examples illustrate confusion between adjacent severity levels. ConvNeXt-MLP-Mixer predicted one Moderate DR image as Severe with a confidence of 0.864 and another as Mild with a confidence of 0.823. ConvNeXt-PoolFormer showed the same directional error pattern, predicting Moderate DR as Severe with a confidence of 0.927 and as Mild with a confidence of 0.858. The relatively high confidence of these incorrect predictions indicates that the observed errors were not limited to low-confidence boundary cases. The explanation maps again show the complementary spatial characteristics of the two XAI methods: LRP provides fine-grained pixel-level relevance patterns, whereas Grad-CAM identifies broader activation regions from the final convolutional layer. Importantly, neither type of visualization should be interpreted as direct lesion segmentation or as evidence that the highlighted regions correspond to clinically validated pathological features.

Some explanation maps highlighted visible abnormal regions, whereas others concentrated on the optic disc or peripheral retinal areas. Such patterns could reflect pathological context, illumination variation, retinal geometry, or dataset-specific shortcuts. The relative residual values displayed in the LRP panels are diagnostics of the relevance-propagation computation; they are not measures of localization accuracy or expert agreement. Furthermore, because Grad-CAM was generated from the final convolutional layer of the ConvNeXt branch, its activation maps primarily reflect the local convolutional component rather than providing a complete spatial explanation of both branches in the hybrid architecture. In the absence of lesion-level annotations and blinded ophthalmologist or trained-grader assessment, the LRP and Grad-CAM results should therefore be interpreted as exploratory qualitative explanations rather than as evidence of clinically validated interpretability.

3.7. Discussions

The results indicate that adding a global feature-mixing branch to ConvNeXt can improve overall five-class DR grading performance, although the benefit depends on both the global architecture and the evaluation metric. ConvNeXt-PoolFormer achieved the highest accuracy and weighted F1-score, whereas ConvNeXt-MLP-Mixer achieved the highest macro-F1 among the evaluated models. Both hybrid models also substantially improved Moderate DR recall relative to standalone ConvNeXt. These findings suggest that combining local convolutional representations with global feature interactions can

provide complementary information for fundus-image classification. However, the class-specific results show that this benefit was not uniform across severity levels, particularly for Severe and Proliferative DR.

The hybrid architectures involve a trade-off between local feature aggregation and global feature mixing. The ConvNeXt branch provides hierarchical convolutional representations that preserve sensitivity to localized retinal patterns, while the MLP-Mixer and PoolFormer branches introduce broader interactions across spatially distributed image regions through different mixing mechanisms. Global feature mixing may therefore complement convolutional processing when relevant abnormalities are distributed across multiple retinal regions, as suggested by the substantial improvement in Moderate DR recall. Nevertheless, broader contextual modeling does not guarantee improved discrimination for every class. Severe DR remained the lowest-F1 class in both hybrids, indicating that global information did not overcome the limited representation and adjacent-grade confusion observed for this minority class. Thus, local and global representations should be regarded as complementary rather than as universally interchangeable or uniformly beneficial sources of information.

Neither hybrid showed a consistent advantage across all evaluation measures. ConvNeXt–PoolFormer had a 0.3 percentage-point advantage in accuracy, while ConvNeXt–MLP-Mixer was 0.4 percentage points higher in macro-F1. Moreover, the exact McNemar test found no significant difference between their paired predictions ($p = 0.7608$). The calibration results also did not consistently favor the same architecture. Differences were evident in computational behavior as well. PoolFormer performs token mixing through pooling operations, whereas MLP-Mixer uses learned token interactions. Although the PoolFormer hybrid required fewer GFLOPs, it recorded higher latency, longer training time, and greater memory usage in the present experiment. Its lower GFLOP count should therefore be viewed as an advantage in theoretical operation count, not as an indication of lower computational cost across every measured criterion.

Previous studies have demonstrated that combining local feature extraction with global contextual modeling can improve medical image classification performance. Hu et al. [23] showed that CNN-based representations and global modeling mechanisms provide complementary information because CNNs capture local structures, whereas global

modules improve long-range dependency modeling. Similar findings have been reported in diabetic retinopathy classification. Huang et al. [24] combined ConvNeXt with Swin Transformer to integrate local retinal features and global contextual information, while Sekar et al. [25] integrated DenseNet121 and Transformer components for five-class DR grading by combining spatial feature extraction with long-range dependency modeling.

However, these previous approaches mainly relied on Transformer-based mechanisms to obtain global contextual information. In contrast, the present study investigates whether similar local-global feature integration can be achieved through non-attention-based global feature mixing. ConvNeXt-MLP-Mixer and ConvNeXt-PoolFormer provide global interaction through token-channel mixing and pooling-based token mixing, respectively. Therefore, this study extends previous findings by demonstrating that the benefit of hybrid local-global representation learning is not exclusively dependent on Transformer self-attention.

Beyond the architectural comparison with previous Transformer-based approaches, the present study further analyzes how different global mixing mechanisms influence DR severity classification. In ConvNeXt-MLP-Mixer, spatial information is exchanged through token- and channel-mixing operations, whereas ConvNeXt-PoolFormer relies on pooling for token mixing. Both configurations produced improvements over standalone ConvNeXt for some measures, with Moderate DR showing the clearest class-level gain. The improvement did not extend equally to all severity grades, however, and Severe DR remained difficult for both hybrids. These results indicate that global feature mixing can add useful information to the local representation learned by ConvNeXt, but the magnitude of the benefit varies with the mixing architecture and the class being predicted. Absolute performance values should also not be compared directly with previous studies without considering differences in datasets, data partitions, preprocessing, model settings, and evaluation procedures.

Performance also varied considerably across DR grades. No DR reached high F1-scores with each of the three selected models, while the advanced Severe and Proliferative DR classes were harder to classify. Moderate DR benefited most from the hybrid architectures, in contrast to Severe DR, which had the lowest F1-score across the three models. This result needs to be viewed alongside the class distribution and error patterns:

Severe DR had relatively few training examples, and its errors frequently involved the neighboring Moderate and Proliferative DR grades. Weighted cross-entropy and the addition of global feature mixing reduced neither of these difficulties completely. Consequently, aggregate accuracy by itself does not describe how performance is distributed across severity levels; macro-averaged measures, class-wise scores, and normalized confusion matrices provide important complementary information.

The calibration results further demonstrate that predictive discrimination and probability calibration should be considered separately. ConvNeXt–PoolFormer achieved the highest accuracy and lowest Brier score, whereas ConvNeXt–MLP-Mixer achieved a lower ECE than ConvNeXt–PoolFormer. The Vision Transformer provides the clearest counterexample to interpreting low ECE as equivalent to strong classification performance: despite achieving the lowest ECE of 0.031, it produced the lowest accuracy (65.1%) and macro-F1 (45.2%) among the evaluated models. Thus, a low ECE indicates closer agreement between confidence and empirical correctness within the adopted calibration bins, but it does not imply that the underlying class predictions are frequently correct. Conversely, higher classification accuracy does not guarantee superior calibration. For medical decision-support settings, accuracy, class-balanced performance, Brier score, and ECE should therefore be interpreted jointly rather than using any single metric as evidence of model reliability.

Finally, the XAI findings should be interpreted conservatively. LRP and Grad-CAM provide complementary explanations at different levels of spatial granularity and model representation. In the present hybrid architectures, LRP propagates relevance from the final hybrid prediction backward through the fusion mechanism and both the local and global branches to the input, thereby producing fine-grained pixel-level relevance patterns that reflect the fused model decision. In contrast, Grad-CAM derives broader class-discriminative activation regions from the final convolutional layer of the ConvNeXt branch. The Grad-CAM maps therefore characterize spatial evidence represented by the local convolutional pathway rather than providing a complete spatial explanation of both branches. This distinction is important when interpreting Figures 6 and 7, because the two methods provide complementary rather than equivalent views of model behavior. Some explanation maps overlapped visually abnormal retinal regions, whereas others emphasized the optic disc, image boundaries, peripheral retina, or broad illumination

patterns, raising the possibility that the models may also rely on non-lesion or dataset-specific cues. Nevertheless, neither the fine-grained LRP relevance patterns nor the broader Grad-CAM activation regions establish lesion-level correspondence or clinically valid reasoning. Without quantitative lesion localization and blinded expert assessment, these visualizations should therefore be regarded as exploratory analyses of model behavior rather than evidence of clinically validated interpretability.

3.8. Limitations

Several limitations should be considered when interpreting these findings. First, the models were developed and evaluated using a single dataset, APTOS 2019, without independent external validation on a separate retinal-image dataset. Consequently, cross-dataset generalization has not been established, and the reported performance should not be assumed to transfer directly to datasets such as Messidor-2, IDRiD, or EyePACS. Differences in patient populations, imaging devices, acquisition conditions, image quality, class distributions, and preprocessing characteristics may introduce domain shifts that affect model performance. Therefore, the present findings demonstrate performance within the evaluated APTOS 2019 setting rather than generalizable performance across heterogeneous clinical or retinal-imaging datasets. Second, because patient identifiers were not available in the public dataset metadata used in this study, patient-level partitioning could not be verified. Although exact duplicate images were removed before splitting, the possibility of unrecognized patient-level dependence across subsets cannot be completely excluded.

Third, although bootstrap confidence intervals were used to quantify sampling uncertainty and a paired McNemar test was used to compare the two hybrid models on the same held-out test set, the statistical evaluation remains based on a single fixed train-validation-test split. Repeated training or cross-validation was not performed; therefore, variability associated with random initialization, optimization, and alternative data partitions was not quantified. The small performance differences between the two hybrid architectures should consequently be interpreted within this experimental setting rather than as evidence of universal architectural superiority.

Fourth, the explainability analysis remains qualitative and was not validated against lesion-level annotations or blinded ophthalmologist/trained-grader assessments.

Although LRP provides fine-grained pixel-level relevance patterns for the fused hybrid decision and Grad-CAM provides broader activation regions from the ConvNeXt convolutional pathway, neither method establishes that the highlighted regions correspond to clinically validated pathological features. The explanation maps should therefore be regarded as exploratory post-hoc analyses rather than evidence of clinically validated interpretability.

Finally, training time, inference latency, throughput, and GPU-memory measurements are implementation- and hardware-dependent. The reported computational characteristics therefore describe the experimental environment used in this study and should not be assumed to represent performance on clinical workstations, edge devices, or other deployment platforms without device-specific benchmarking.

Future work should prioritize external cross-dataset validation on independent retinal-image datasets such as Messidor-2, IDRiD, or EyePACS to assess robustness under dataset and acquisition shifts. Repeated or cross-validated experiments should also be conducted to quantify training and partition variability. Further evaluation should include quantitative lesion-localization analysis when suitable annotations are available and blinded expert assessment of explanation maps. Such evaluations are necessary before stronger conclusions can be drawn regarding model generalizability, reliability, or clinical interpretability.

4. CONCLUSION

This study evaluated ConvNeXt-MLP-Mixer and ConvNeXt-PoolFormer for five-class diabetic retinopathy grading against standalone EfficientNet, Vision Transformer, ConvNeXt, MLP-Mixer, and PoolFormer architectures. ConvNeXt-PoolFormer achieved the highest test accuracy and weighted F1-score of 82.7%, while ConvNeXt-MLP-Mixer achieved a closely comparable accuracy of 82.4% and the highest macro-F1 of 67.5%. The paired McNemar test indicated no statistically significant difference between the classification outcomes of the two hybrid models, indicating that their small performance difference should be interpreted as metric-dependent within the present experimental setting rather than as evidence of clear architectural superiority. Both hybrid models improved the test-accuracy point estimate over standalone ConvNeXt and showed their

clearest class-specific benefit for Moderate DR, with recall increasing from 64.0% with ConvNeXt to 76.0% with ConvNeXt-MLP-Mixer and 79.3% with ConvNeXt-PoolFormer. However, Severe and Proliferative DR remained comparatively difficult to classify, indicating that the benefit of combining local convolutional representations with global feature mixing was not uniform across severity levels. The computational results also revealed an architecture-dependent trade-off, with ConvNeXt-PoolFormer requiring fewer GFLOPs, whereas ConvNeXt-MLP-Mixer achieved more favorable measured latency and throughput. LRP and Grad-CAM provided complementary but exploratory explanations of model behavior, with LRP producing fine-grained pixel-level relevance patterns for the fused hybrid prediction and Grad-CAM highlighting broader activation regions from the ConvNeXt convolutional pathway; however, these visualizations were not validated using lesion-level annotations or blinded expert assessment and therefore should not be interpreted as evidence of clinically validated model reasoning. The findings are limited to a single public fundus-image dataset and a fixed train-validation-test split, while patient-level separation could not be verified because reliable patient identifiers were unavailable and repeated or cross-validated training was not performed. External cross-dataset validation on independent datasets such as Messidor-2 or IDRiD is therefore necessary to establish generalizability beyond the APTOS 2019 setting. Future work should additionally investigate repeated or cross-validated evaluation, quantitative lesion-level assessment, blinded ophthalmologist or trained-grader evaluation of explanation maps, and deployment-oriented benchmarking under intended clinical hardware and workflow conditions. Overall, the proposed hybrid models represent promising algorithmic approaches for five-class diabetic retinopathy grading within the evaluated experimental setting, but they are not yet ready for clinical deployment.

REFERENCES

- [1] H. Li *et al*, "Current research and future strategies for the management of vision-threatening diabetic retinopathy," *Asia-Pacific J. Ophthalmol*, vol. 13, no. 5, p. 100109, Sep. 2024, doi: 10.1016/j.apjo.2024.100109.
- [2] Y. Meng *et al*, "Global, Regional, and National Burden of Blindness due to Diabetic Retinopathy, 1990–2021," *Ophthalmol. Ther*, vol. 14, no. 10, pp. 2599–2615, Oct. 2025, doi: 10.1007/s40123-025-01230-y.

- [3] T. Zikhali, C. Kalinda, and Z. N. Xulu-Kasaba, "Screening of Diabetic Retinopathy Using Teleophthalmology to Complement Human Resources for Eye Health: A Systematic Review and Meta-Analysis," *Clin. Pract.*, vol. 12, no. 4, pp. 457–467, Aug. 2022, doi: 10.3390/clinpract12040050.
- [4] M. A. Raza, M. Dua, and K. Berwal, "Early Detection of Diabetic Retinopathy Using Pretrained Models: A Focus on Initial Stages," in *2024 1st International Conference on Advanced Computing and Emerging Technologies, ACET 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ACET61898.2024.10730746.
- [5] Y. Miao, S. Tang, P. Du, and Z. Li, "Research on Deep Learning in the Detection and Classification of Diabetic Retinopathy," in *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology, CEI 2021*, Institute of Electrical and Electronics Engineers Inc., Sep. 2021, pp. 107–113. doi: 10.1109/CEI52496.2021.9574509.
- [6] 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.
- [7] N. S. Kumar and B. Ramaswamy Karthikeyan, "Diabetic Retinopathy Detection using CNN, Transformer and MLP based Architectures," in *ISPACS 2021 - International Symposium on Intelligent Signal Processing and Communication Systems: 5G Dream to Reality, Proceeding*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 1–2. doi: 10.1109/ISPACS51563.2021.9651024.
- [8] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, 2019, pp. 6105–6114.
- [9] A. Dosovitskiy *et al*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations*, 2021.
- [10] N. S. Kumar, R. K. Balasubramanian, and M. R. Phirke, "Image Transformers for Diabetic Retinopathy Detection From Fundus Datasets," *Revue d'Intelligence Artificielle*, vol. 37, no. 6, pp. 1617–1627, Dec. 2023, doi: 10.18280/ria.370626.

- [11] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2022, pp. 11966–11976. doi: 10.1109/CVPR52688.2022.01167.
- [12] Devanshi, S. K. Baliarsingh, and P. P. Dev, "An Early diagnosis of diabetic retinopathy using ConvNeXt," in *Proceedings of the 10th International Conference on Signal Processing and Integrated Networks, SPIN 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 739–743. doi: 10.1109/SPIN57001.2023.10116214.
- [13] I. Tolstikhin *et al*, "MLP-Mixer: An all-MLP Architecture for Vision," in *Advances in Neural Information Processing Systems*, 2021, pp. 24261–24272.
- [14] W. Yu *et al*, "MetaFormer is Actually What You Need for Vision," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2022, pp. 10809–10819. doi: 10.1109/CVPR52688.2022.01055.
- [15] H. Jiang *et al*, "A Multi-Label Deep Learning Model with Interpretable Grad-CAM for Diabetic Retinopathy Classification," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2020, pp. 1560–1563. doi: 10.1109/EMBC44109.2020.9175884.
- [16] R. Singh, "Understanding Image Classification Tasks Through Layerwise Relevance Propagation," in *Proceedings - 2022 IEEE 18th International Conference on Intelligent Computer Communication and Processing Conference, ICCP 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 199–203. doi: 10.1109/ICCP56966.2022.10053963.
- [17] I. Ullah, A. Rios, V. Gala, and S. McKeever, "Explaining Deep Learning Models for Tabular Data Using Layer-Wise Relevance Propagation," *Applied Sciences*, vol. 12, no. 1, p. 136, Jan. 2022, doi: 10.3390/app12010136.
- [18] X. Luo *et al*, "A deep convolutional neural network for diabetic retinopathy detection via mining local and long-range dependence," *CAAI Trans. Intell. Technol.*, vol. 9, no. 1, pp. 153–166, 2024, doi: 10.1049/cit2.12155.
- [19] R. Katarya, P. Sharma, N. Soni, and P. Rath, "A Review of Interpretable Deep Learning for Neurological Disease Classification," in *8th International Conference on Advanced Computing and Communication Systems, ICACCS 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 900–906. doi: 10.1109/ICACCS54159.2022.9785321.

- [20] A. M. N. Hirushan and S. P. K. Arachchi, "Integrate Interpretable AI into Chest X ray Diagnosis through a Comparative Analysis of CNN Architectures with Grad CAM Visualizations," in *9th International Conference on Advances in Technology and Computing, ICATC 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICATC64549.2024.11025232.
- [21] Asia Pacific Tele-Ophthalmology Society, "APTOS 2019 blindness detection," Kaggle Competition. Accessed: Aug. 08, 2026. [Online]. Available: <https://www.kaggle.com/competitions/aptos2019-blindness-detection>
- [22] O. Cîmpeanu, I. Mihaela Liliac, M. Mămuleanu, Ștefan-V. Voinea, M. Olteanu, and C.-T. Streba, "Using Convolved Neural Networks in Diagnosing Lung Cancer on Computed Tomography Scans," *Curr. Health Sci. J.*, vol. 51, no. 1, pp. 87–95, 2025, doi: 10.12865/chsj.51.01.09.
- [23] C. Hu, N. Cao, H. Zhou, and B. Guo, "Medical Image Classification with a Hybrid SSM Model Based on CNN and Transformer," *Electronics (Basel)*, vol. 13, no. 15, p. 3094, 2024, doi: 10.3390/electronics13153094.
- [24] X. Huang *et al.*, "A CNN-Transformer fusion network for Diabetic retinopathy image classification," *Comput. Med. Imaging Graph.*, vol. 126, p. 102655, 2025, doi: 10.1016/j.compmedimag.2025.102655.
- [25] P. Sekar, R. Bhoopalan, N. Nagaprasad, T. R. Mamo, S. P. Dhanabal, and R. Krishnaraj, "D-TNet: a hybrid Dense Net-transformer model for robust diabetic retinopathy detection," *Sci. Rep.*, vol. 15, no. 1, p. 39594, 2025, doi: 10.1038/s41598-025-23234-1.