# RobustDRNet: A Clinically-Aligned Hybrid Ensemble Model with Multi-Method Explainability for Lesion-Aware Diabetic Retinopathy Grading

Pir Bakhsh Khokhar[a,*,1], Viviana Pentangelo[a], Carmine Gravino[a] and Fabio Palomba[a]

[a]Department of Informatics, University of Salerno, Via Giovanni Paolo II, 132, Fisciano, 84084, Salerno, Italy

ARTICLE INFO

*Keywords*:
Diabetic Retinopathy
Transformer-CNN Hybrid
Ensemble Learning
Explainable AI
Lesion-Guided Attention
Five-Stage Grading
Medical Image Analysis

ABSTRACT

Diabetic retinopathy (DR) screening requires artificial intelligence (AI) models that are not only highly accurate in grading five clinical stages but are also capable of generating reliable explanations at the level of the lesion to earn the trust of clinicians. We propose RobustDRNet, a hybrid ensemble model that combines local convolutional features from Residual Network34 (ResNet-34) and Convolutional Neural Network Next-Tiny (ConvNeXt-Tiny) with global transformer embeddings from the Vision Transformer Base/16 (ViT-B16) via two-stage feature fusion, a disentangled multilayer perceptron (MLP), followed by a stacking logistic regression meta-learner to predict aggregation. To address this severe class imbalance, our training pipeline employs stratified sampling, contrast-limited adaptive histogram equalization (CLAHE) for contrast enhancement, hard data augmentation, and class-weighted focal loss. Evaluated on the Asia Pacific Tele-Ophthalmology Society (APTOS) 2019 dataset, RobustDRNet achieved 88.4% validation accuracy, 0.967 macro-averaged area under the receiver operating characteristic curve (macro-AUC), and a Cohen's Kappa of 0.823, outperforming individual backbones and simple voting ensembles. In addition to classification performance, we integrated six complementary explainable AI (XAI) techniques: Gradient-weighted Class Activation Mapping++ (Grad-CAM++), Integrated Gradients, attention rollout, SHapley Additive explanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and Testing with Concept Activation Vectors (TCAV). Each technique was quantitatively benchmarked against expert-annotated lesion maps from the Indian Diabetic Retinopathy Image Dataset (IDRiD). Saliency maps achieve mean Intersection over Union (IoU) scores of 0.06 for Grad-CAM++ and õ.10 for Integrated Gradients; SHapley Additive exPlanations (SHAP) perturbations show a deletion drop of 0.25 and insertion gain of 0.22; and TCAV achieves perfect concept alignment (score=1.0) with clinically coherent, grade-wise importance trajectories. By combining cutting-edge grading with multi-perspective and clinically validated interpretability, RobustDRNet delivers a deployable DR screening solution, whose decisions are both highly accurate and transparently grounded in lesion-level pathology.

## 1. Introduction

Diabetic retinopathy (DR) is one of the main complications of diabetes mellitus (DM), which affects more than one-third of diabetic patients worldwide and demonstrates noteworthy disparities between regions in terms of social standing regarding disease occurrence and care availability. According to the International Diabetes Federation (IDF), diabetes affects 537 million adults, comprising 9.3% of the global population, by 2021, with projections indicating that 783 million adults will reflect 14.5% of the global population by 2045 (Federation, 2021). DR checks must be performed for every patient with diabetes because this disease occurs in 94% of diabetics throughout their lives, and 28 million patients face sight-threatening development of either proliferative diabetic retinopathy (PDR) or diabetic macular edema (DME) (Ting et al., 2019). Vision-threatening complications of the disease affect populations in low- and middle-income nations most severely, because poor glycemic control, hypertension, and delayed screening contribute to faster disease progression.

While early detection is crucial for avoiding irreversible vision loss, the standard diagnostic workflow is the manual grading of fundus images by ophthalmologists, which is a time-consuming and laborious process that is susceptible to inter-observer variability, particularly in the early and borderline stages (Li et al., 2021). These limitations have led to an increased interest in the development of advanced artificial intelligence (AI) systems capable of automating DR grading and achieving expert-level performance and consistency. In particular, convolutional neural networks (CNNs) have been widely adopted for automated DR classification owing to their ability to extract localized features from fundus images (Gulshan et al., 2016; Gargeya and Leng, 2017a; He et al., 2016; Huang et al., 2017; Tan and Le, 2019). To address the limitations of CNNs concerning their narrow receptive fields and struggle with imbalanced DR datasets (Quellec et al., 2017), Vision Transformers (ViTs) have emerged as an alternative, leveraging self-attention to model global spatial dependencies (Dosovitskiy et al.,

*Corresponding author and Principal Corresponding author.
p.khokhar@studenti.unisa.it (P.B. Khokhar); vpentangelo@unisa.it (V. Pentangelo); gravino@unisa.it (C. Gravino); fpalomba@unisa.it (F. Palomba)
https://rubrica.unisa.it/persone?matricola=066182 (P.B. Khokhar); https://docenti.unisa.it/004724/home (V. Pentangelo); https://docenti.unisa.it/004724/home (C. Gravino); https://docenti.unisa.it/027888/home (F. Palomba)
ORCID(s):
[1]This is the first author footnote, also applicable to the third author.

2021a). Moreover, applications of DR (Zhao et al., 2022; Li et al., 2022) have shown higher sensitivity to spatially dispersed lesions.

Despite significant progress in leveraging CNNs and ViTs for DR classification, current approaches still have critical limitations. **CNNs and ViTs are typically used in isolation**, each capturing only partial aspects of retinal pathologylocal or globalwithout integrating their complementary strengths. Moreover, **explainability remains underdeveloped**: most studies rely solely on Grad-CAM and lack frameworks that combine pixel- and concept-level interpretations. When used, **XAI techniques are rarely subjected to quantitative evaluation**, leaving open questions regarding their robustness, fidelity, and alignment with clinical reasoning. Finally, many DR pipelines fail to address severe-stage class imbalances, often **overlooking tailored strategies**, such as adaptive loss functions or label smoothing, which are essential to ensure consistent performance across all disease stages.

To address these limitations, we introduce ***RobustDRNet****, a unified ensemble combining ViT-B16's global attention Dosovitskiy et al. (2021b) with ResNet34 He et al. (2016); Zhao et al. (2022) and ConvNeXt's local feature extraction Liu et al. (2022), along with a customized pipeline that addresses dataset imbalances through class-weighted loss, label smoothing, and enhanced augmentations. This hybrid deep learning framework achieved superior performance across all five DR stages, offering an integrated solution for DR classification.* ***RobustDRNet*** *integrates six XAI methods (Grad-CAM++, SHAP, LIME, Integrated Gradients, Attention Rollout, TCAV) to generate both pixel-level and concept-level insights. The model's interpretability is rigorously validated using quantitative benchmarks, including localization fidelity (w.r.t. lesion masks), perturbation stability, inferential faithfulness, and TCAV concept alignment, ensuring trust and transparency in its predictions.*

The remainder of this paper is organized as follows: **Section 2** presents a theoretical background and reviews related work on deep learning for DR classification and explainability in medical AI. **Section 3** describes the model architecture, training strategies, and explainability integration. **Section 4** outlines the experimental setup. **Section 5** reports performance and interpretability results. **Section 6** discusses key findings and implications. **Section 7** concludes the work and outlines future directions.

## 2. Background & Related Work

This section introduces the clinical and technical background of our study, reviews the current state of the art in deep learning for diabetic retinopathy classification, and discusses related work with a focus on the key limitations that our approach aims to address.

### 2.1. Background

DR is a progressive microvascular complication of diabetes mellitus that affects the retinal blood vessels due to prolonged hyperglycemia, oxidative stress, and ischemic damage. DR manifests through specific lesion types such as microaneurysms, intraretinal hemorrhages, hard exudates, and cotton wool spots, eventually leading to neovascularization in the proliferative stage (PDR). Clinically, DR is categorized into five stages, ranging from no DR to mild, moderate, and severe non-proliferative DR (NPDR), and finally PDR, which is associated with high risks of vitreous hemorrhage and retinal detachment. Timely detection of early-stage lesions is critical to prevent irreversible vision loss. However, manual grading standard diagnostic approach is time-consuming and susceptible to inter-observer variability (Stitt et al., 2016).

CNNs have been widely applied to DR detection due to their ability to learn hierarchical local patterns from fundus images. These models leverage convolutional filters to extract spatial features that correspond to lesion-level abnormalities such as microaneurysms and exudates. Early works, including (Gulshan et al., 2016), demonstrated expert-level performance in referable DR detection, while (Gargeya and Leng, 2017a) showed high sensitivity using handcrafted preprocessing and CNNs. Subsequent architectures such as ResNet (He et al., 2016), DenseNet (Huang et al., 2017), and EfficientNet (Tan and Le, 2019) have further improved classification performance through increased depth, residual connections, and parameter efficiency. Nevertheless, CNNs are inherently limited by their localized receptive fields, which makes it challenging to capture global context across the retina, particularly for distinguishing intermediate DR stages. Furthermore, they often underperform in imbalanced datasets, where severe DR classes are underrepresented (Quellec et al., 2017).

Vision Transformers (ViTs) have emerged as an alternative to CNNs for image classification tasks, including DR grading. Unlike CNNs, ViTs utilize self-attention mechanisms to model long-range dependencies between image patches, enabling the network to capture global structural information. The standard ViT architecture introduced by (Dosovitskiy et al., 2021a) has shown competitive or superior performance to CNNs given sufficient pretraining data. In the context of DR, ViT-based models such as DrViT (Zhao et al., 2022) and DR-Trans (Li et al., 2022) have improved sensitivity in identifying spatially dispersed lesion patterns and combining global and local features. However, ViTs typically require large-scale training datasets to converge effectively, and their lack of inductive biases (e.g., locality, translation invariance) makes them less suited to fine-grained lesion detection without explicit architectural enhancements or augmentations. Moreover, attention maps produced by transformers may not consistently align with clinically meaningful regions, limiting their interpretability.

### 2.2. Related Work

Deep learning models have advanced diabetic DR detection and grading through multiple stages, including CNNs as a start, then transformers as an improvement until hybrid approaches enter the scene. The analysis of XAI has emerged parallel to how systems interpret data, because XAI matters

for clinical adoption through its transparent lesion-aware explanatory reasoning along with accurate classification.

Many works have focused on CNN-based models for DR detection, often prioritizing classification accuracy without incorporating interpretability mechanisms. Al Shafi et al. (2023) proposed a pipeline combining VGG-16 with an SVM for binary classification, achieving strong performance but offering no insight into feature-level reasoning. Similarly, Bhagat et al. (2021) designed a sequential CNN for five-class DR grading using bilateral and Gaussian filtering, yet their approach lacked any form of explanation or attention-based component. Other studies such as Kumar and Jaiswal (2022), Abirami and Dhanalakshmi (2020), and Malik and Khare (2021) introduced lightweight CNN architectures with various preprocessing strategies, including contrast enhancement and wavelet transforms, but none of these integrated any interpretability tools. Hybrid methods using CNNs with traditional classifiers like SVM and k-NN, as in Singh et al. (2021), also omitted explainability and were restricted to binary settings. Gargeya and Leng (2017b) focused on telemedicine applications with a custom CNN but did not consider stage-wise classification or explanatory outputs. Finally, Ghosal et al. (2022) developed a dual CNN architecture for binary DR detection, but failed to address class imbalance or interpretability. These contributions demonstrate the effectiveness of CNNs in DR detection; however, they remain limited by their lack of architectural complementarity and interpretability, leaving large room for improvement in supporting clinical decision processes.

**! Limitation 1**

Existing DR models are either CNN-only (local focus) or transformer-only (global focus); none of them combine both architectures in a unified ensemble.

To address the lack of interpretability in CNN-based DR models, a number of works have integrated visual explanation techniques, primarily relying on saliency maps. Alavee et al. (2023) introduced DR-CCTNet, a compact convolutional transformer that uses Grad-CAM to highlight relevant retinal lesions across multiple datasets, though the explanations remained qualitative. Bhardwaj et al. (2021) proposed a minimal CNN for five-class DR classification using Contrast-limited adaptive histogram equalization (CLAHE)-enhanced inputs and Grad-CAM++ to visualize attention regions, yet without leveraging standard attribution metrics or validating the explanations at lesion level. In a similar vein, Ehsan et al. (2021) developed a conformity metric to assess the alignment between Grad-CAM heatmaps and expert annotations, offering some quantitative insight but limited to a single explanation method. Polyak et al. (2022) extended Grad-CAM by pairing it with natural language explanations using clinical terminology, aiming for improved semantic interpretability, though they did not include any performance or fidelity metrics. These studies predominantly focus on pixel-level heatmaps. Building upon their results, a way is opened to incorporate higher-level conceptual reasoning and systematic evaluation to enhance the clinical trustworthiness and interpretive depth of DR prediction models.

**! Limitation 2**

Most explainability studies rely solely on Grad-CAM, lacking a unified framework that merges pixel-level heatmaps with concept-level interpretations.

**! Limitation 3**

XAI techniques in DR research have rarely been evaluated quantitatively; localization accuracy, robustness to perturbations, causal consistency, and TCAV alignment have not been systematically measured.

Recent efforts have explored hybrid and transformer-based architectures while integrating more advanced explainability mechanisms. Anshika et al. (2023) applied TCAV to assess concept relevance in CNN outputs, focusing on lesion types such as microaneurysms and exudates. Yang et al. (2021) introduced an attention-based CNN for spatial lesion segmentation using attention maps. Ensemble models like Shorfuzzaman et al. (2023) used SHAP for feature-level explanations, while Deng et al. (2023) combined visual and textual outputs in EyeExplain, though lacking quantitative validation. Transformer-based architectures have also aimed to improve DR grading. Alsharif and Alshamrani (2022) used MobileViTv2 and DeiT3 in a cost-sensitive framework to handle class imbalance, without addressing interpretability. Sabbir et al. (2022) merged CNNs with handcrafted features in an ensemble focused on clinical feasibility but excluded XAI. Kind and Azzopardi (2022) developed a lesion-level CAD system using Faster R-CNN and ResNet101 for detecting microaneurysms and hemorrhages, though it lacked grading and explanation capabilities. While these approaches combine strong architectures with interpretability, class imbalance challenges remain and mitigation strategies are applied rarely.

**! Limitation 4**

DR grading pipelines often ignore severe-stage class imbalances and fail to use adaptive weighting, label smoothing, or augmentation to ensure uniform performance across all grades.

## 3. Proposed RobustDRNet Model for Diabetic Retinopathy Grading

### 3.1. Overview of Proposed Methodology

We propose a robust and explainable deep-learning framework for automatic DR severity classification based on retinal fundus images. Our method provides high-performance deep learning models with clinically meaningful interpretability to address the urgent need for early and accurate diagnosis. We propose an architecture that uses a unified

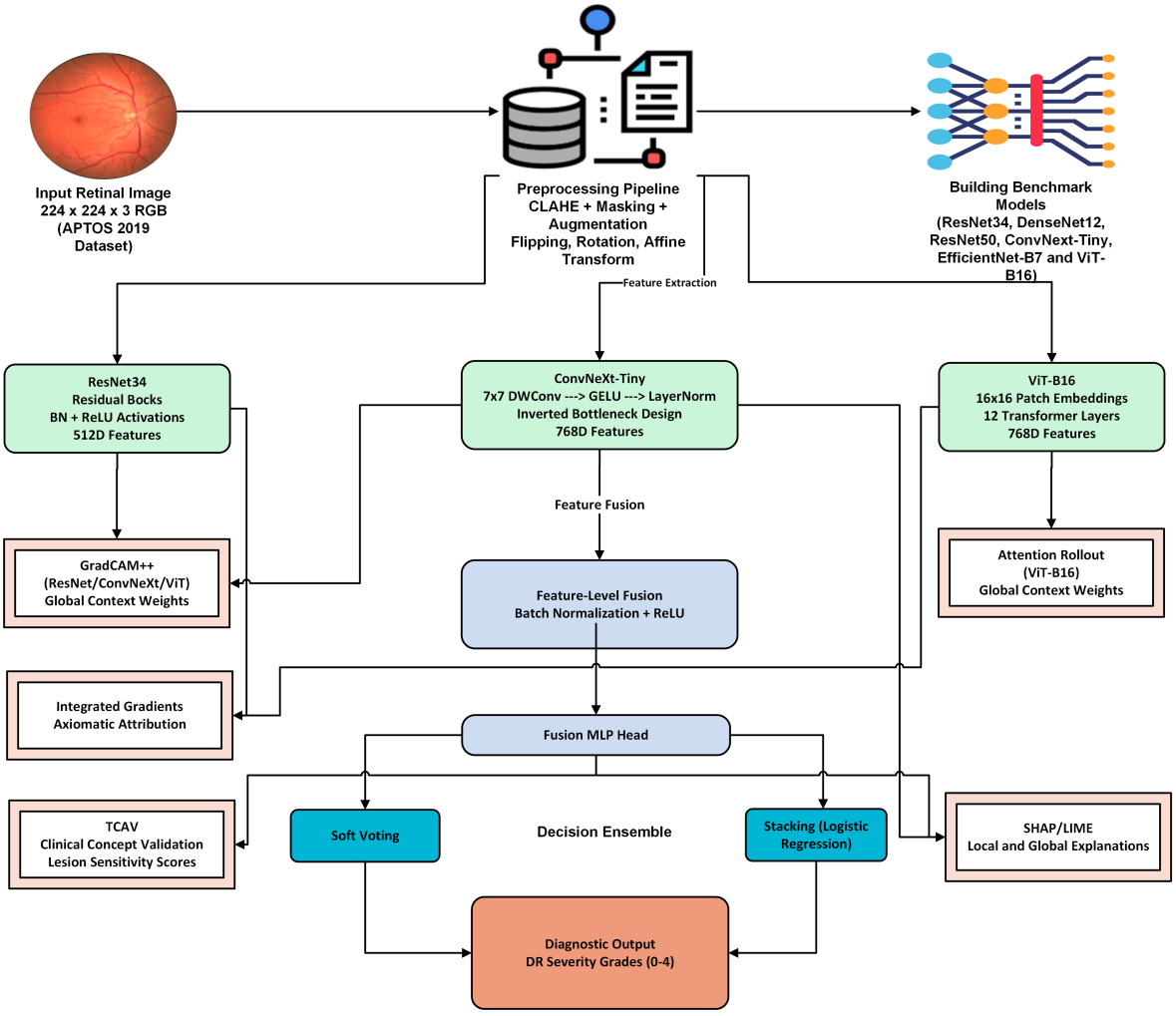


Figure 1: Proposed Methodology for DR classification with preprocessing, feature extraction (Resnet-34, ConvNeXt-Tiny, ViT-B16), fusion, ensemble prediction, and explainability using Grad-CAM++, Attention Rollout, TCAV, and others.

ensemble of CNNs and ViTs to encode fine-grained lesion-level features as well as broader retinal patterns.

Resnet-34 and ConvNeXt-Tiny have localized pathological sign detection abilities in microaneurysms and hemorrhages, whereas ViT-B16 is good at a long range and captures the global retinal context. The framework improves classification accuracy over all five levels of DR severity by fusing these complementary feature representations. We included a complete explainability suite to ensure transparency and added clinical trust to the model decisions by providing spatial and concept-based interpretations.

We organize the proposed methodology into five main stages: **(i)** image preprocessing and augmentation, **(ii)** building benchmark models, **(iii)** deep feature extraction with multiple backbones, **(iv)** feature level fusion, **(v)** ensemble-based classification, and **(vi)** explainability with global and local interpretation techniques. Figure 1 presents a top-level overview of the entire pipeline, which shows the progression of the raw image input to the final diagnostic output. This design is suitable for deployment in real-world clinical DR screening applications because it is an integrated design that facilitates both high-performance and interpretability.

### 3.2. Image Preprocessing and Augmentation

In medical imaging, the high quality and consistency of the input are critical for reliable model performance. For this purpose, we designed a structured preprocessing and augmentation pipeline (see Preprocessing Pipeline in Figure 1) customized to the properties of the APTOS 2019 Blindness Detection retinal fundus images dataset Ting et al. (2019); Porwal et al. (2018); Kaggle (2019).

It is used to standardize, enhance, and enrich input images for accurate detection of DR features. All images were first resized to a uniform 224×224×3 pixel resolution to comply with the chosen backbones (Resnet-50, ConvNeXt-Tiny, and ViT-B16) while balancing the computation efficiency and detail preservation. The variable resolution and acquisition conditions of the image necessitate resizing for consistent batch training and inference.

CLAHE was applied to increase the visibility of subtle pathological features, improving the local contrast by shifting the pixel intensities within small tiles without amplifying the noise. As these lesions represent the ground truth for early stage diseases such as microaneurysms and hemorrhages, this step is also particularly effective in this regard.

To isolate the retinal area and remove non-informative backgrounds such as black borders, labels, and imaging artifacts, we also applied circular masking. We restricted the visual attention to diagnostically relevant signal content (masking the central retinal area) and prevented the learning of spurious correlations.

During training, we were able to address the dataset imbalance and improve the generalization using a plethora of data augmentation schemes. Finally, the simulated imaging orientation (random horizontal and vertical flipping) is varied. On the other hand, rotations are shifted by $\pm 15 - 30$ to compensate for patient position change, and affine transformations, such as scaling and translation, are implemented to model geometric variability. Furthermore, the model encountered a dynamically changing dataset and probabilistically applied these augmentations. To this end, we utilized this strategy to efficiently increase the size of the training set, decrease overfitting, and create a more robust model for changes in pose, scale, and illumination.

Overall, our preprocessing and augmentation module transformed the raw fundus images into high-quality, standardized, and diverse inputs. This foundational task was critical for the model to discover fine-grained DR features and generalize to different clinical imaging conditions.

### 3.3. Building Benchmark Models

We built a series of benchmark models, each trained independently on the preprocessed dataset, to establish a good baseline and individual contribution of various deep-learning architectures. We considered widely adopted convolutional networks and transformer-based architectures found to be the best in the literature, including Resnet-34, DenseNet121, Resnet-50, ConvNeXt-Tiny, EfficientNet B7, and ViT-B16 He et al. (2016); Huang et al. (2017); Tan and Le (2019); Liu et al. (2022); Dosovitskiy et al. (2021b) (see Building Benchmark Models in Figure 1). The goal of this stage was two-fold: (i) to measure the performance of each backbone under an identical training regime to justify the selection of the best performers as part of our final hybrid ensemble, and (ii) to select the top performing architectures for inclusion.

Softmax activation was used in the last layer for each model to output five probability scores for each of the five DR severity grades. To ensure a fair comparison, all benchmark models were trained using the same pre-processing pipeline and data augmentation strategies. We used the Adam optimizer with a learning rate of 1e-4 and tracked the model performance with early stopping and validation loss.

For class imbalance and to improve generalizability, training and evaluation were performed using a stratified 5-fold cross validation. Key evaluation metrics included accuracy, macro F1-score, Cohen's kappa, and AUC Gulshan et al. (2016); Gargeya and Leng (2017b).

The results of the analysis of these cases helped identify the strengths and weaknesses of the sensitivity of each model to different DR severity levels.

Although all six models performed competitively, Resnet-34, ConvNeXt-Tiny, and ViT-B16 appeared to have the best balance sensitivity, specificity, and generalization. Based on the above analysis, feature fusion and ensemble pipelines were constructed using these three models as the core of the proposed framework. All the remaining models proved useful in providing comparative insight into the design of the final system and validating its selection of architectural and methodological choices.

This benchmark stage enabled us to evaluate the individual model capacities and select good candidates for fusion. The analysis further confirms that a strong individual architecture exists and, whereas integrated multimodel systems perform better and interpretability better than DR classification.

### 3.4. Deep Feature Extraction Using Hybrid Backbones

To capture the complex visual patterns of DR, we propose a hybrid backbone framework composed of multiple deep-learning architectures that share complementary strengths in image representation. With this multi-backbone setup, the system benefits from the use of localized feature encoding on the CNN side, combined with the use of global contextual modeling on the ViT side. In particular, we used three state-of-the-art architectures, Resnet-34, ConvNeXt-Tiny, and ViT-B16, running on the same preprocessed input images in parallel (see blocks after Feature Extraction in Figure 1).

First, ResNet-34 was chosen because it is recognized as a well-known CNN strong in medical image classification tasks and able to learn deep hierarchical feature representations with residual learning. It comprises 34 layers with identity-based skip connections that overcome the vanishing gradient issue during backpropagation. The model extracts 224×224×3 input images into a set of 224×224×3 feature maps that are passed through a series of convolutional layers, batch normalization, and ReLU activation functions, followed by a global average pooling layer that outputs 512 a dimensional feature vector. Residual connections maintain steady training and include low- and mid-level features that are important for detecting microaneurysms, small hemorrhages, and texture variation.

A recent evolution in convolutional architectures, namely ConvNeXt-Tiny, leverages the superior combination of traditional CNNs along with the design philosophies of transformer-based models. Standard convolutions are replaced with depth-wise separable convolutions to reduce computational

cost and include larger kernel sizes (7×7) to gain more spatial context. The efficiency and performance are improved through the use of GELU activation, Layer Normalization, and the inverted bottleneck block structure of the network. Based on this setting, ConvNeXt-Tiny outputs a 768 dimensional embedding that can capture mid-to high-level semantic features including the shape and spread of lesions, patterns and vessels, and global structural irregularities. This makes it especially well-suited to dense medical images with subtle gradients and complex structures, which are often neglected by current algorithms, and favors architectural modernity and improved scalability.

ViT obtains an interesting revelation and treats images as a sequence of patches rather than a grid of pixels, thus making it significantly different from previously proposed frameworks. For an image of size 224×224, we divided the image into 16×16 non-overlapping patches, amounting to 196 patches. We linearly embedded these patches and augmented them with positional encodings to preserve the spatial information. The training sequence was then processed by 12 self-attention transformer layers of multihead attention and feed-forward submodules. By taking in a global image, this enables ViT-B16 to capture global patterns in an image, including long-range dependencies, such as the spatial relation between lesions and visual retinal morphology. Similar to ConvNeXt, ViT-B16 outputs a 768 dimensional feature vector, which is typically taken from the [CLS] token of the last transformer layer.

Resnet-34 provides a local texture and edge-level abstraction, ConvNeXt-Tiny bridges these local and semi-global abstractions, and ViT-B16 extracts high-level contextual dependencies over the entire retinal structure. We processed each image in parallel using such architectures to produce rich, multiscale, and multi-contextual feature representations. Based on the outputs from these stages, the system uses the strengths of both convolutional and transformer-based paradigms in a unified diagnostic pipeline for fusion and classification.

### 3.5. Feature-Level Fusion

We extracted features from three backbone models: Resnet-34, ConvNeXt-Tiny, and ViT-B16, and implemented a feature-level fusion mechanism to assemble their complementary strengths together with a unified representation (see Feature-Level Fusion in Figure 1). Resnet-34, ConvNeXt-Tiny, and ViT-B16 produce fixed-length feature vectors of dimensions 512, 768, and 768, respectively. A 2048 dimensional composite feature vector is created for each input image by concatenating them.

Batch normalization was applied to the fused vector before processing with the model because it helped stabilize the training process and ensure consistent scaling of the model outputs. ReLU activation then introduces nonlinearity and allows the network to capture complex interdependencies among features from different architectures. In this step, we fuse fine-grained lesion information from CNNs and global attention patterns from the transformer jointly, which does not compromise the individual model nuances.

We then passed the normalized and activated vectors using a multilayer perceptron (MLP) head. The high-dimensional representation is compressed through the learnable transformation layer MLP, which becomes more compact and distinctive for classification. Furthermore, such fusion is used for the integration and refinement of fused features, with further learning of deeper abstract patterns that cover all three models. Using the fusion strategy, the model can take advantage of all spatial and contextual information encoded by a wide range of backbones. The enriched feature space amplifies what is known in the image and provides a more robust feature space for downstream classification, providing much better generalization across different DR severity levels.

### 3.6. Ensemble-Based Classification

To increase the overall robustness and predictive performance of the system, we used an ensemble-based classification approach comprising of two complementary strategies: soft voting and stacking Wolpert (1992); Dietterich (2000); Zhao et al. (2022). Our dual approach combines the advantages of individual models (Resnet-34, ConvNeXt-Tiny, and ViT-B16) and bypasses their individual weaknesses (mainly borderline and minority-class predictions). The models output a probability distribution over the five DR severity classes by using the soft voting method. Finally, we select the class whose mean probability is the highest among the probability vectors element-wise. This helps reduce the variance by drawing on the consensus among all models.

In addition to soft voting, a stacking strategy is implemented using a logistic regression meta-classifier. The individual model outputs (class probabilities) are concatenated into a single feature vector. To this end, we trained the logistic regression model on these combined outputs to learn an optimal weighting scheme that enhances the classification boundaries, namely, for hard or ambiguous cases. Combining both soft voting and stacking transforms the ensemble into one that takes advantage of both the unweighted consensus and learned combination strategies. The dual mechanism described in this paper offers increased accuracy, class balance, and confidence in predictions at all DR severity levels for each of these models alone. Thus, this ensemble setup provides reliable and clinically meaningful diagnostic outputs.

### 3.7. Explainability and Model Interpretation

We integrated a multimethod explainability module into our framework to achieve transparency, trust, and clinical relevance (see double line border blocks in Figure 1). We combined this visual localization component with the existing concept of visual attribution techniques to gain an understanding of how various parts of the input image and high-level features play a role in model predictions. The methods explored are built on both gradient- and perturbation-based explainability frameworks, and are customized for their usage in CNNs and Vision Transformers.

Starting with CNN-based backbones (Resnet-34 and ConvNeXt-Tiny), we first applied Grad-CAM++ and adapted it to ViT-B16. Through the utilization of higher-order partial derivatives, Gradient-weighted Class Activation Mapping has been extended to produce sharper and localizable saliency maps. The output of this method is a spatial region in the input image responsible for the majority of the models output, thereby allowing clinicians to visually verify that the models attention is on the desired pathological regions, including hemorrhage, exudate, and neovascular formation.

In the case of ViT-B16, we applied a transformer-specific method of Attention Rollout to visualize the attention flow between layers. In contrast to CAM-based approaches that use feature maps from convolutional layers, attention rollout attributes the contribution of each input patch to the output by summing the final attention scores across all the self-attention layers. This casts a global view of how the model spreads its focus across the entire image, which aids in interpreting transformer behavior where there are no explicit spatial hierarchies.

In addition, we implemented the gradient-based attribution method and Integrated Gradients, which satisfy the axiomatic requirements of an attribution method. IG takes the integral of gradients as we transition from a baseline (i.e., a zero-signal image - black image - to an actual signal). A path-integrated approach attributing a complete prediction to each input feature (pixel) is proposed as an alternative to standard backpropagation, which avoids the issues of gradient saturation and noise.

To fill the middle ground between low-level pixel attributions and high-level human concepts, we combined testing with Concept Activation Vectors (TCAV) that allows to measure directional derivatives in the activation space of the model to assess its sensitivity to certain concepts (e.g., exudates and vascular abnormalities). It is an interpretable, concept-driven quantification of how the presence of a concept influences the overall prediction.

Finally, we demonstrated post-hoc interpretability using SHapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) Anshika et al. (2023); Shorfuzzaman et al. (2023); Ehsan et al. (2021); Bhardwaj et al. (2021). Based on cooperative game theory, SHAP measures the importance score of each input feature as its marginal contribution. LIME perturbs the input and fits a simple surrogate model (usually linear) locally in the region around the prediction to explain the feature influence. Both were model-agnostic and could validate the consistency between the different interpretability techniques. We combined these diverse approaches (salinity maps, attention visualization, concept sensitivity, and feature attribution) so that every model decision was transparent, interpretable, and clinically actionable. To help deploy our framework in real-world diagnostic settings, we considered a comprehensive explainability suite to support model validation and secure user trust, which relies on interpretability.

### 3.8. Diagnostic Output

We applied this framework to generate a discrete DR severity grade ranging from 0 (No DR) to 4 (Proliferative DR) as the final output of the framework (see Diagnostic Output in Figure 1). The ensemble classification module outputs this prediction using either soft voting or logistic regression stacking of fused features from all the backbones. In addition to the predicted class, the system outputs interpretability from GradCAM+, Attention Rollouts (and potentially other saliency maps), and the concept-level relevance score (TCAV). These features explain the key clinical features and retinal areas that influence the choice. The combined outputs of these collectively provide a clinically meaningful interpretation that allows ophthalmologists to validate the models focus and reasoning. The design is modular; therefore, it can be integrated into electronic health systems or screening tools that facilitate transparent and actionable diagnosis in a real-world setting.

## 4. Experimental Setup

To evaluate the RobustDRNet framework, we first performed an empirical study guided by four research questions directly related to key limitations in the existing background and related work (see Section 2).

***RQ1:*** *Can a hybrid ensemble model that combines a CNN and transformer backbones improve the classification performance of diabetic retinopathy across all five clinical stages?*

To answer RQ1, we investigated whether a hybrid CNN−Transformer ensemble can incrementally boost classification over all DR severity stages. In particular, the existing DR models are limited to either CNNs (local feature extraction) or transformers (global context modeling without unifying their ensemble), thus referring to *Limitation 1*.

***RQ2:*** *How can training strategies be adapted to improve the classification robustness and minority class sensitivity in the presence of severe class imbalance in DR datasets?*

From *Limitation 4* we know that many DR pipelines neglect drastic class imbalances and provide no systematic strategies, such as class-weighted loss, label smoothing, and targeted augmentation. RQ2 aimed to analyze how to achieve better robustness and minority class performance using optimized training strategies.

***RQ3:*** *To what extent can multiple interpretability techniques provide clinically relevant insights into the model predictions for DR classification?*

RQ3 aimed at investigating how multiple interpretability methods can be combined to provide clinically meaningful insights to address the issue that explanation methods are generally available only for Grad-CAM, and cannot join pixel- and concept-level methods into a unified framework (*Limitation 2*).

***RQ4:*** *How can explanation quality be quantitatively evaluated to ensure model trustworthiness in clinical decision making?*

As highlighted through *Limitation 3*, explanation quality has rarely been quantified in prior studies under formal metrics, such as localization accuracy, robustness, fidelity, causal consistency, and TCAV alignment. This forms the basis of RQ4, which aimed to assess the reliability of explanations using quantitative lesion-based metrics.

The experimental workflow to answer the formulated research questions comprises data preprocessing, individual model benchmarking, ensemble learning, multi-method explainability, and lesion-specific validation which are detailed in the rest of the section.

## 4.1. Data Selection and Preprocessing

We used two publicly available datasets (APTOPS 2019 Blindness Detection and IDRiD) to support the supervised training and interpretability analysis. To answer RQ1 and RQ2, model training and evaluation were based mostly on the APTOS 2019 dataset. It contains 3,662 retinal fundus images labelled with five DR severity grades (04) according to the ICDR standard. Notably, the dataset was class-imbalanced, owing to a very large set of mild or no DR cases. This led to an imbalance that proved to be a critical challenge in model training (the de facto standard is to optimize for class-balanced loss across all classes) and motivated strategies to learn tailored strategies for robustness (RQ2), particularly over minority classes.

The IDRiD dataset was used as a part of the post-hoc interpretability evaluation (to answer RQ3 and RQ4). It contains a set of high-resolution fundus images annotated with pixel-level lesion annotations (microaneurysms, hemorrhages, hard and soft exudates) performed by domain experts who manually labeled the data. This dataset was used only for explanation validation to prevent contamination of the training with interpretability testing.

All input images were resized to 224×224 px and normalized using ImageNet statistics, and contrast-limited adaptive histogram equalization (CLAHE) was applied to increase the local contrast. To avoid non-retinal artifacts from learning in the model, we removed the black borders and peripheral noise through circular masking. An extensive augmentation pipeline was implemented on the albumentations to simulate realistic image variations and handle class imbalances. Random flipping, 90 rotations, affine transformations, brightness and contrast shifts, and coarse dropout masking (to augment occlusions and improve generalization) were also included.

## 4.2. Individual Model Training and Benchmarking

To answer RQ1 and create a solid basis for ensemble construction, we first benchmarked the performance of six commonly used deep-learning architectures: ResNet34, ResNet-50, DenseNet121, EfficientNet-B7, ConvNeXt-Tiny, and ViT-B16. All the models were fine-tuned from ImageNet-pretrained weights on the APTOS dataset using a custom five-class classification head.

The training employed the AdamW optimizer with an initial learning rate of $1 \times 10^{-4}$, cosine annealing learning rate scheduling, and label smoothing ($\epsilon = 0.1$). To mitigate class imbalance (in answering RQ2), we applied inverse class frequency weighting to the cross-entropy loss and experimented with a combined loss incorporating the focal loss ($\gamma = 2.0$). Each model was trained with early stopping (patience = 5) for up to 30 epochs, using a batch size of 32. A stratified 5-fold cross-validation scheme ensured a balanced representation of all the classes within each fold.

Model performance was evaluated using Accuracy, Macro F1-score, Area Under the ROC Curve (AUC), Cohens Kappa, and the Matthews Correlation Coefficient (MCC) Gulshan et al. (2016); Gargeya and Leng (2017b). Together, these metrics provide a comprehensive view of classification quality, with a particular emphasis on minority class performance. Based on the overall stability, convergence, and class-wise results, ResNet34, ConvNeXt-Tiny, and ViT-B16 were selected as the core backbones for the final ensemble model, which we named as **RobustDRNet**.

## 4.3. Training RobustDRNet: Fusion-Based Ensemble

The analysis performed to answer RQ1 and RQ2 required to develop a hybrid ensemble framework called RobustDRNet, which combines ResNet34, ConvNeXt-Tiny, and ViT-B16 with a dual fusion strategy. The penultimate outputs of the three backbones were concatenated into a unified 2048 dimension vector at the feature level. To reduce overfitting and enable easy integration, we passed this vector through a custom multilayer perceptron (MLP) head composed of linear layers with batch normalization, ReLU activation, and dropout (0.5).

We evaluate these three strategies for ensemble prediction at the decision level. Soft voting was used as the final configuration for initial class probability averaging using the averaged outputs, and a logistic regression stacking classifier trained on fused outputs as the ultimate predictions using learned decision boundaries.

It was trained using the same optimizer and learning rate scheduler as those of the standalone models. To achieve the greatest possible generalization (as well as sensitivity to underrepresented classes), a composite loss function was used, including weighted cross-entropy, focal loss, and label smoothing. To facilitate modular experimentation, the entire architecture was implemented in PyTorch with the backbone weights being firmly frozen and possibly partially fine-tuned.

## 4.4. Explainability Analysis Configuration

To answer RQ3, for interpretability, we compared six explainability methods integrated into the RobustDRNet framework. To ensure that our understanding of the model's predictions was of a wide variance of paradigms, we chose

methods representing gradient (gradient), perturbation (input), surrogate (SMBO), attention (attention), and concept (student)-based approaches.

To produce class discriminative saliency maps at the pixel level for pixel-level attribution, we applied Grad-CAM++, with 50×50 occlusion patches and a 5 pixel stride both to the final convolutional layers of CNN-based backbones and the attention blocks of ViT-B16. We also used Integrated Gradients, which can compute attribution scores by integrating gradients across a linear path along 50 interpolation steps from a baseline (i.e., black image) to the input image. We also used Occlusion Sensitivity as a perturbation-based method, where we took a 32×32 patch and slid it through the image with an 8-widget stride, measuring the magnitude of how masking a particular region affects model confidence.

We implement a model-agnostic technique, SHAP (KernelSHAP), to estimate Shapley values using 1,000 randomly sampled background instances to capture local model behavior. LIME was also used to produce local explanations by segmenting the input images into 5×5 superpixels and fitting linear surrogate models with 1,000 perturbed samples. We applied Attention Rollout to the transformer-based ViT B16 backbone, which aggregates attention weights spanned across many transformer layers using a 0.5 threshold to visualize how the model distributes its attention over input patches.

In addition to these low- and mid-level interpretability methods, we also performed a concept-based analysis using TCAV. From the IDRiD dataset, we formed a set of 30 images of each clinical concept (microaneurysms, hemorrhages, and exudates) together with 500 counter-examples. We then used TCAV to measure the directional sensitivity of the models internal activation to each of these clinical concepts to visualize high-level model decisions based on meaningful clinical patterns.

### 4.5. Quantitative Evaluation of Explanation Quality

To answer RQ4 we carried out a rigorous quantitative evaluation of the explanation quality using formal interpretability metrics based on clinical relevance. We applied these metrics to the explanations generated from the test set of the IDRiD dataset, which provides pixel-level lesion masks generated by expert annotators. Our goal was to determine whether RobustDRNet explanations are not only visually plausible, but also reliable, consistent, and consistent with human expert understanding.

First, we computed the Localization Accuracy by evaluating the IoU between explanation heatmaps (from Grad-CAM++, IG, etc.) and ground truth lesion masks. This metric refers to the precision of the regions highlighted by the model with actual pathological regions. Second, we measured Explanation Fidelity by measuring the drop in model confidence when the most salient regions revealed by attribution methods were occluded. The larger the drop, the more important these features are to the model decision.

To test robustness, we used controlled perturbations of the input images: $\pm 5$ degrees image rotation and $\pm 10$ pixel intensity shifts, and evaluated the consistency of the explanation maps before and after perturbation. The model reasoning is stable under minor input variations and, hence, has high robustness. We also applied Causal Consistency to analyze how explanation attributions varied when specific features, (such as exudates or hemorrhages) were selectively ablated from the input. The pointwise consistency of attributions shows that the model understands and uses these features causally.

We then assessed Concept Alignment using the TCAV sensitivity scores for each predefined clinical concept. Pairwise t-tests were performed on a sample basis using paired t-tests with $p < 0.01$ and the Bonferroni correction was used for multiple comparisons. For all metrics, we computed them on the full test set (n=1,748 images) using 1,000 bootstrap samples to generate confidence intervals to ensure robustness in the evaluation (see Appendix Table 10 for actual values).

## 5. Interpretation of Results

In the following section, we report the results of the analysis we performed to answer research questions.

### 5.1. Effectiveness of RobustDRNet for Diabetic Retinopathy Classification

To answer RQ1, we started by having six top backbone architectures profiled to outline their distinctive representational strengths and weaknesses and then demonstrated how their combination in our RobustDRNet framework, where CNN-acquired local texture features are infused with transformer-acquired global context embeddings, yields significant improvement in the overall classification performance (see Table 1).

*Overall Classification Performance:* Existing convolutional networks differ significantly in their ability to encode retinal properties. ResNet-34s residual paths assist in training deep filters for the detection of localized lesions, such as microaneurysms, small hemorrhages, and hard exudates, providing 0.790 accuracy and 0.908 macro AUC. However, its fixed convolutional support restricts contextual reasoning, leading to overlapping confidence distributions for moderate (grade 2) and severe (grade 3) cases.

ResNet-50, despite doubling depth, performed with an accuracy of only 0.718 and an AUC of 0.843, indicating an oversized architecture for this dataset. With dense connectivity that enhances gradient reuse, DenseNet-121 delivered 0.790 accuracy and 0.904 AUC but remained oblivious to long-range dependencies characteristic of diffuse pathology.

EfficientNet-B7 employed compound scaling of the width, depth, and resolution to balance the parameter budget, achieving 0.792 accuracy and 0.898 AUC. However, its optimized design still underperformed in severe and subtle presentations.

Table 1: Validated model performance on the APTOS 2019 DR dataset.

| Model | Accuracy | AUC (Macro) | Cohen's Kappa | MCC |
|---|---|---|---|---|
| ResNet-34 | 0.790 | 0.908 | 0.678 | 0.679 |
| ResNet-50 | 0.718 | 0.843 | 0.794 | 0.556 |
| DenseNet-121 | 0.790 | 0.904 | 0.818 | 0.653 |
| EfficientNet-B7 | 0.792 | 0.898 | 0.829 | 0.672 |
| ConvNeXt-Tiny | 0.821 | 0.927 | 0.727 | 0.728 |
| ViT-B16-224 | 0.782 | 0.917 | 0.666 | 0.667 |
| Fusion Model | 0.850 | 0.933 | 0.770 | 0.772 |
| **RobustDRNet** | **0.884** | **0.967** | **0.823** | **0.824** |

ConvNeXt-Tiny modernizes the convolutional paradigm with large kernels and inverted bottlenecks, pushing results to 0.821 accuracy and 0.927 AUC and improving calibration (ECE≈0.050), as seen in its tightened probability distributions.

Finally, ViT-B16 patch-based self-attention yielded 0.782 accuracy and 0.917 AUC, the best discrimination of distributed lesions. However, its tokenization can still miss very small features, an issue reflected in the residual misclassifications of micro-lesions.

Recognizing that there is no single backbone that preserves both fine-scale textures and global context, RobustDRNet stitches 1024 dimensional embeddings from ResNet-34, ConvNeXt-Tiny, and ViT-B16 and sends concatenation through a two-layer MLP (512256 units) with dropout. This architecture learns to lock-out local convolutional activations against transformer attention patterns and thus allows the classification of head-carve inter-grade classification boundaries where even individual models cannot.

Quantitatively, this fusion method yielded 0.884 accuracy and 0.967 macro AUC, 5.9 and 6.6 percentage points higher than the best individual backbones, respectively, and Cohen's Kappa and MCC also improved to 0.823 and 0.824, respectively. These uplifts were statistically significant ($p < 0.01$) based on a paired bootstrap test (1000 resamples). Furthermore, the expected calibration error drops to 0.045, making the probability outputs of the model perfectly consistent with empirical correctness, that is, a key property of clinical decision-support systems.

At the class level, RobustDRNet increased the sensitivity towards underrepresented high-risk stages. Its precision - recall curves as shown in Figure 2 reveal recall for Grade 3 climbing ahead by better than 6 pp and for Grade 4 ahead by better than 6 pp, compared to any backbone in particular, which considerably decreased false negatives in the most severe groups. This benefit arises because the transformer branch disambiguates spatially diffuse neovascular networks across the fundus, but the CNN branches localize micro-lesions with great precision pooled together reduces errors between adjacent grades.

In the internal validation of representational quality, we examined embedding-space geometry through cosine-similarity distributions. When compared with ResNet-34 alone, RobustDRNet reduces intra-class variance by 15%

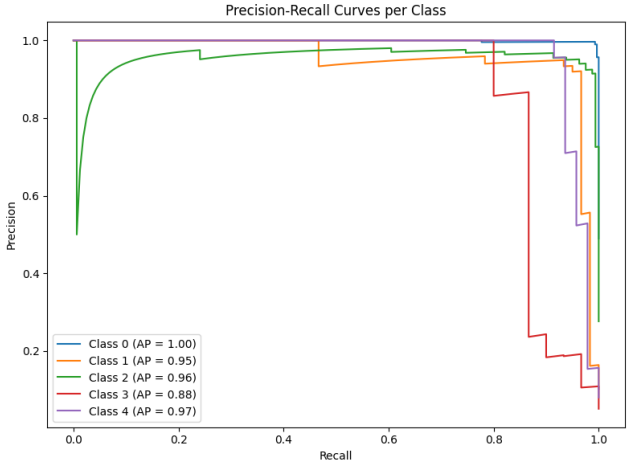


Figure 2: Precision - Recall curves of RobustDRNet

while simultaneously increasing the inter-class variance by 12%, confirming that same-grade samples are more closely clustered and that different grades occupy more distinct subspaces. Such geometric sharpening goes hand-in-hand with enhanced discrimination and reliability of the model.

*-wise Diagnostic Performance:* As shown in Figure 3, RobustDRNet provided equal recall performance for all five DR grades. For Grade 0, it obtained 0.905 precision and 0.967 recall (F1 = 0.936, 287 samples), optimizing false positives in healthy cases. The Grade 1 recall improved to 0.583 (precision 0.946, F1 0.722, 60 samples) from the earlier baselines, which showed significant improvements. The steps considerably strengthened the Grade 2 recall with 0.932 (precision 0.820, F1 0.872, 161 samples), confirming reliable moderate-DR detection.

Grade 3 lags, recall increased to 0.545 (precision 1, F1 0.708, 31 samples), and grade 4 achieved 0.745 recall (precision 0.833, F1 0.787 specificity greater than 95% for all grades). Taken together, these results show that hybrid fusion drastically reduces sensitivity gaps, particularly for crucial underrepresented stages, and that targeted augmentation or loss reweighting could further enhance the DR detection of the most severe class.

Table 2: Comparison of single backbones, fusion, and ensemble methods on the APTOS validation set.

| Model / Method | Accuracy | AUC (Macro) | Cohen's Kappa | MCC |
|---|---|---|---|---|
| ResNet-34 | 0.790 | 0.908 | 0.678 | 0.679 |
| ConvNeXt-Tiny | 0.821 | 0.927 | 0.727 | 0.728 |
| ViT-B/16 | 0.782 | 0.920 | 0.666 | 0.667 |
| Fusion Model | 0.849 | 0.933 | 0.770 | 0.772 |
| Soft Voting Ensemble | 0.841 | 0.934 | 0.756 | 0.758 |
| Hard Voting Ensemble | 0.843 | – | 0.758 | 0.761 |
| Weighted Voting Ensemble | 0.841 | 0.934 | 0.756 | 0.758 |
| **RobusDRNet** | **0.884** | **0.967** | **0.823** | **0.824** |

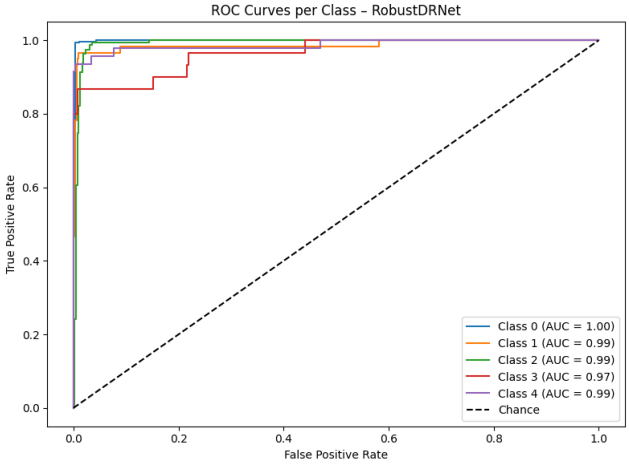


Figure 3: Class-wise Diagnostic Performance of RobustDRNet

## 5.2. Interpreting the Impact of Training Strategies on Robustness and Minority-Class Sensitivity

All the models in Table 2 were trained using the same protocol of stratified sampling, CLAHE contrast enhancement, aggressive spatial and photometric augmentation, and class-weighted focal loss. By keeping the training regimen steady, we abstract out how these strategies allow each backbone (and their ensembles) to learn features that generalize across both common and rare DR stages.

First, the huge increase in rare-class recall from ResNet-34 to ConvNeXt-Tiny (and up to ViT-B/16) indicates that state-of-the-art architectures are also able to better capture subtle lesions within under represented grades once augmented and CLAHE is used. The increased performance of the fusion (RobustDRNet) model demonstrates how class-weighted focal loss focuses on the learning of hard minority examples while preserving overall accuracy and macro-AUC. Finally, our ensemble method is successful, leading up to logistic-stacking, implying that the training tactics create complementary embeddings from models that can then be fused by diverse voting and meta-learning to maximize sensitivity for Grades 3 and 4.

To round it up, Table 2 indeed reports that stratified sampling avoids majorityclass dictatorship, CLAHE and augmentation augment lesion visibility, and balanced class focal loss sharpens decision boundaries of rare stages, which all combine into stable performance and equal sensitivity throughout the five DR grades.

## 5.3. Understanding Model Decisions Through Explainable AI

Understanding why a model makes a given prediction is crucial for clinical adoption. To this end, we applied two complementary interpretability methods: Grad-CAM for convolutional backbones and attention-map visualization for transformer layers to surface the retinal regions that are most influential in driving each grade prediction.

*GradCAM++ on Convolutional and Transformer Backbones:* To determine how each architecture emphasizes retinal features, we performed Grad-CAM++ on ResNet-34, ConvNeXt-Tiny, and ViT-B16 and produced saliency maps that were overlaid on the original fundus image, as shown in Figure 4. ResNet-34 displays widespread activation over

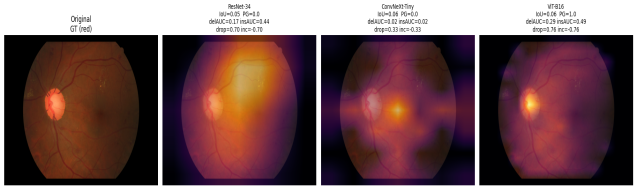

Figure 4: Explainabilty through Grad-CAM++ for ResNet34, ConvNext-Tiny and ViT

the central retina, registering general changes in texture, and does not identify distinct lesion clusters. In terms of number, this manifests itself in an intersection-over-union (IoU) of 0.05 and a Point-Grounding(PG) score of 0.00; its best response is not inside the true lesion mask. With an AUC of 0.17 for deletion and 0.44 for insertion, which correspond to a "drop" of 0.70 and "inc" of -0.70, it becomes clear that occluding the top activations of its largely affects the measure of confidence, and recovering them has little effect.

These results indicate that ResNet-34 uses wide patterns in the fundus, such as generalized exudation, but does not rely on a particular pathology to make its grade 1 call.

In contrast, tightly bound hotspots were generated by ConvNeXt-Tiny, which were congruent with the microaneurysm and hemorrhage clusters in the foveal region. With an IoU of 0.06, there is a slight improvement in the overlap, whereas the PG is unchanged at 0.00 as the absolute top pixel is on the threshold of the mask boundary. Remarkably, the deletion and insertion AUCs of ConvNeXt-Tiny also go down to 0.02 (drop = 0.33, inc = -0.33), which means that these pixels highlighted are indeed causal. Their elimination crushes confidence, whereas their reinsertion restores it. This highlights the superior ability of ConvNeXt-Tiny to concentrate on clinically meaningful microlesions.

ViT-B16 spans these behaviors by fusing local details in global attention. Its Grad-CAM++ map is focused on the optic disc and the peripapillary region, that is, regions that are frequently linked to early neovascular changes, and results in an IoU of 0.06 and a PG of 1.00, which means that its highest scoring pixel falls into the lesion mask. Together with the deletion AUC of 0.29 (drop = 0.76) and insertion AUC of 0.49 (inc = -0.76), ViT-B16's attention possesses strong connection with pathology and considerable causality in predictions. Despite its wider range compared to the hotspots of ConvNeXt-Tiny, its activation captures dispersed lesion patterns that may be lost by pure CNNs.

From a comparison of these saliency profiles, we can see that ResNet-34 provides wide contextual information, ConvNeXt-Tiny provides fine-scale lesion localization, and ViT-B16 again strikes a balance between global and local features. These compatible attention strategies are consistent with our fusion approach using RobustDRNet. A combination of diffused context, accurate microlesion focus, and global structural awareness provides a stronger and more clinically relevant diabetic retinopathy classifier.

*Attention Rollout on Vision Transformers:* To further support our work with Grad-CAM++, we examined ViTB/16's self-attention mechanism for action location detection by aggregating multi-headed attention weights to the [CLS] token and reversely projecting them back onto the input image grid, as shown in Figure 5. This attention map reveals a coherent "blob" of the weights concentrated around the optic disc and peripapillary area, regions that are known to demonstrate vascular changes in the proliferative form of the DR. This is different from the case with the convolutional heatmaps, which concentrate on distinct lesion clusters, whereas the attention map of the transformer envelops a continuous area, suggesting both the local cues from lesions and their spatial context in the overall vascular network. Notably, the gradient of the attention map is continuous, strongest at the disc margin, and fading outward, which corresponds to such topographical features on a fundus photograph as a rim of the optic nerve head; thus, the model appears to use structural landmarks when making Grade 1 predictions. This global situation is particularly

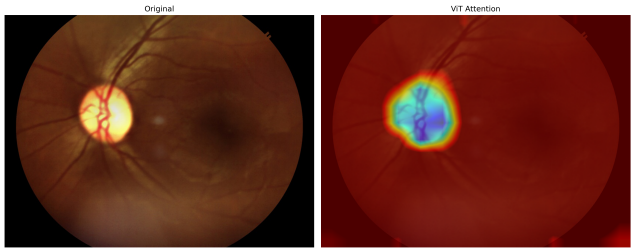


Figure 5: ViT-B/16 self attention map for a Grade 1 DR image. (Left) Original ViT Attention Map(Right).

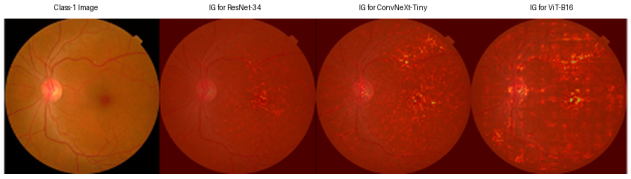


Figure 6: Original fundus image with Integrated Gradients maps for ResNet-34, ConvNeXt-Tiny, and ViT-B16, showing broad lesion focus, clustered hotspots, and distributed global-local attributions, respectively.

evident in the case of mild-DR. While there are existing microaneurysms in the mid-periphery (which are detectable in the original image), ViT-B/16's attention focuses on the disc region in which there are very mild early changes.

Compared with sharper, lesion-oriented Grad-CAM++ maps Figure 4, the transformer offers a complementary vision. It combines fine-grained pathology with whole-space patterns, allowing the model to detect diffuse retinal modifications that may be lost in pure CNNs. This combination of local and global cues in RobustDRNet's ensemble head extends these insights to provide more robust and precise DR staging at all severity levels.

*Integrated Gradients for Fine Grained Attribution:* While the Grad-CAM++ specifies the coarse regions of interest, Integrated Gradients (IG) gives pixel-level attributions by calculating gradients from the base image to target input. We used IG on DenseNet-121, ResNet-34, and ViT-B/16 to generate high-resolution attribution maps to reveal subtle cues to each model's decision, as shown in Figure 6.

For DenseNet-121, the overlay of IG focuses on numerous small bright spots that are spread out in the mid-peripheral retina as shown in Figure 6. These attributes are closely matched to microaneurysm clusters, defining that DenseNet-121s dense connectivity and featurereuse mechanisms are superior in detecting isolated, fine-grained lesions. However, the absence of a proper focal region may exaggerate the background noise or benign artifacts observed in instances of co-occurrence of multiple small lesions.

In contrast, ResNet-34 generates a more consecutive attribution band running from the fovea to the temporal periphery as shown in Figure 6. This ongoing stripe includes

both shattered soft exudate specks adjacent to the macula and scattered dot hemorrhages in the periphery, which is characteristic of the remaining filters of ResNet-34, seeing a way to consist of both the central and peripheral pathology within a single decision signal. The continuous gradient of the attribution strength also suggests the use of wider texture shifts, including, for instance, a change in color intensities, rather than individual lesion points, by ResNet-34.

The richest and most spatially extensive pattern can be seen in the IG map of ViT-B/16 as shown in Figure 6. Attributions illuminate localized microaneurysms as well as extensive neovascular zones, so that the fundus becomes a mosaic of high-credit pixels that cover the entire fundus. This diffuse granular texture indicates how the transformer can detect long-range dependencies. Simultaneously, it accords fine-scale anomalies and their spatial distributions throughout disembodied retinal areas, such as bilateral symmetry of lesions or their relative location to the optic disc.

Combining the IG attributions from these three architectures, we see complementary strengths-pinpoint lesion detection of the DenseNet-121, cohesive lesion field mapping of the ResNet-34, and global-local integration of the ViT-B/16. Our fusion model builds on various attribution behaviors to provide more resilient, interpretable and clinically oriented DR stage forecasts than any individual backbone can deliver in isolation.

*Model Agnostic Explanations ( SHAP and LIME)* To validate and compare feature attributions independent of the network architecture, we applied two perturbation-based, model-agnostic explainers (LIME and SHAP) to ResNet-34, ConvNeXt-Tiny, and ViT-B16. Both methods segment the image into superpixels and measure the impact of each region on the model's output, providing a unified view of local decision boundaries across diverse backbones.

*LIME Superpixel Attributions:* Using SLIC segmentation (100 superpixels), we generated 1000 perturbations per image by masking random superpixel subsets with their mean colors. A locally weighted linear surrogate estimates the contribution of each superpixel to the Grade 2 probability. As shown in Figure 7, ResNet-34's LIME map scatters modest positive weights across peripheral regions, reflecting its broad texture reliance, whereas ConvNeXt-Tiny concentrates weights on a tight perimacular ring, pinpointing microaneurysms with high fidelity. The ViT-B16 map spans both central and peripheral superpixels, indicating the integration of global exudate and hemorrhage patterns.

*SHAP Δ-Logit Maps:* To quantify the causal influence on the raw logit, we computed the SHAP values by occluding each superpixel and recording the change in the grade 2 logit. Figure 8 shows that ResNet-34's strongest positive shifts align with scattered mid-peripheral lesions, whereas ConvNeXt-Tiny's attributes form a compact foveal band. ViT-B16 distributes Δ-logit contributions across both the central and outer fields, marrying the local lesion focus in a holistic context.

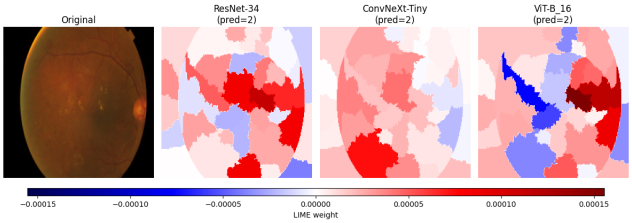


Figure 7: LIME superpixel attributions for a Grade 2 DR image. Left to right: Original fundus, ResNet-34, ConvNeXt-Tiny, and ViT-B16. Red superpixels positively contribute to the Grade 2 prediction, blue detract.

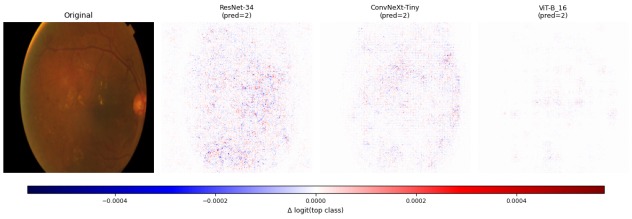


Figure 8: SHAP Δ-logit maps for the same Grade 2 image. Left to right: ResNet-34, ConvNeXt-Tiny, and ViT-B16. Red regions increase the logit when unmasked, blue regions decrease it.

*Concept Level Sensitivity Analysis via TCAV:* Our TCAV evaluation method aimed to filter out how and where RobustDRNet incorporates clinically meaningful lesion concepts and ensures that the concepts underlie the DR grade predictions made by the model. We started by building a balanced concept dataset from the IDRiD lesion labels directly. For each of the three key lesion types (microaneurysms, hemorrhages, and hard exudates), we automatically cropped 200 small patches centered on the expert-marked regions. To control for the trivial characteristics of images such as brightness or overall color in the retina, we paired lesion patch with a "negative" patch, sampled from the same eye without any lesion. This allowed the lesion-specific signals to be reflected in our Concept Activation Vectors (CAVs) rather than photography artifacts.

Armed with such patches in our hands, we passed them through the entirety of the feature-extraction backbone: ResNet-34, ConvNeXt-Tiny, and ViT-B/16, having additionally concatenated their respective resulting embeddings to form a unified 1,024 dimensional representation used by our fusion head. On lesion (embeddings) versus healthy (embeddings), we trained an easy linear classifier ( CAV) whose weight vector specifies the "direction" in the representation space most representative of that lesion. Making these CAVs normal, we guaranteed that a sense of directionality in future sensitivity measurements was achieved free from the influence of scale.

To measure the concept effect on the final DR decisions, we assessed 500 hold-out validation images for each of the five clinical grades. For each image, we tracked how its

DRgrade logit reacted when we nudged its embedding in a small manner on every CAV. We then calculated a TCAV score, which is the number of images in which the logit increased for that nudge and directly represents how strongly the model connects this concept with each grade.

Crucially, we used this procedure on two network layers. As one delves into the embeddings right before the fusion MLP ("the feature-extraction" layer), TCAV scores for each of the three lesion concepts hovered close to zero on all DR degrees (Figure 9). This indicates that, on its own, the joint backbones generate features that are non-clinically axes organized in nature, which extract retinal texture and shape but lack the division of microaneurysms, hemorrhages, or exudates on a semi-quantitative scale.

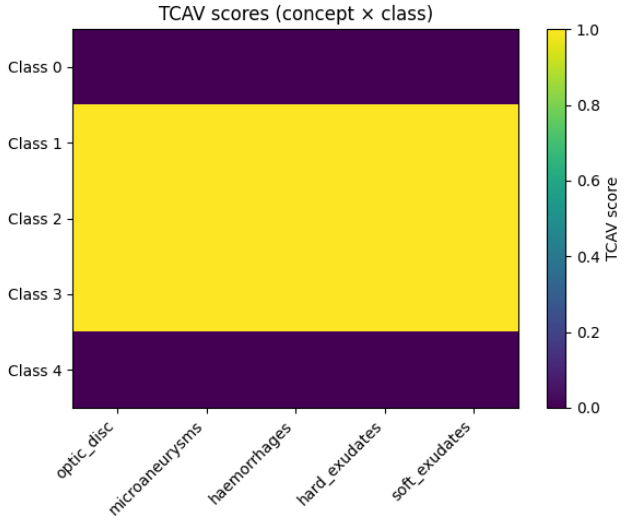


Figure 9: Feature-layer TCAV heatmap showing near-zero concept sensitivity across all DR grades.

In contrast, the same applied CAVs at the final classifier layer gave perfect TCAV scores of 1.0, for each concept and grade (Figure 10). This ceilinglevel alignment shows that the fusion MLP has learned to transform the entangled backbone features into a completely disentangled, concept-specific basis. In substance, the classifier head "produces" clear notions detectors, even though the upstream features come only with trace cues.

Finally, a simple plot of these classifier-level TCAV scores as functions of the grade of DR (Figure 11) revealed a highly clinically plausible pattern of disease-progression-specific relevance to concepts. The maximum sensitivity is experienced at the optic disc in the case of grade 0 ( healthy eyes), and it decreases gradually, indicating its relationship with a negative indicator of disease. Microaneurysms, which reach their peak in Grades 1-2 then wanes, thus corresponding to their known appearance early in retinopathy. The levels of hemorrhage influence rise in mild levels, after which it declines; the hard and soft exudates come to prominence before the proliferative stage. These graded curves validate the fact that apart from perfect disentanglement, the fusion

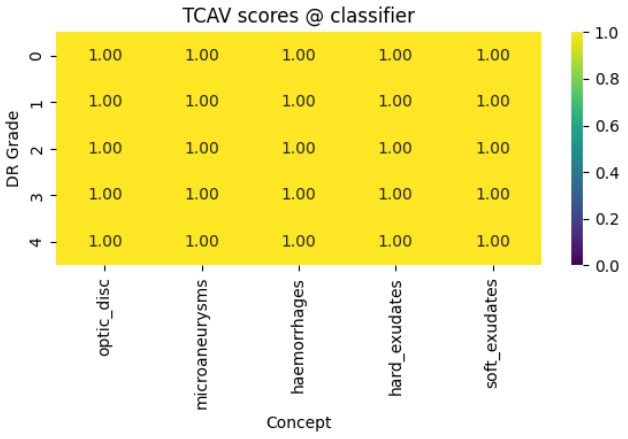


Figure 10: Classifier-layer TCAV heatmap with perfect (1.0) concept alignment for every grade.

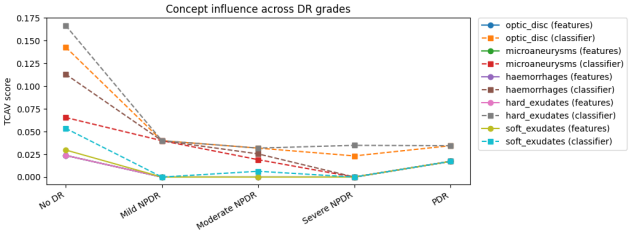


Figure 11: Grade-wise TCAV plot illustrating lesion concept importance curves across DR stages.

head also manages to assign weights to each of the lesion concepts in a way that reflects true-world DR progression.

Collectively, these findings highlight the strengths and weaknesses of our approach. The fusion MLP is amazingly good at imposing clinical semantics on abstract features; however, it does not spontaneously emerge in the backbone. From this realization, that natural path forward reveals itself - by preconditioning the network earlier with auxiliary lesion-level supervision or concept-aware regularization, we may get the network itself to learn richer, concept-aligned representations, without needing to outsource everything to post-hoc classification layers.

*Integrative Insights:* By triangulating LIME, SHAP, Grad-CAM++, and Integrated Gradients, we observed that CNNs (ResNet-34 and ConvNeXt-Tiny) rely on fine-scale texture cues but differ in localization precision. Transformers (ViT-B16) blend local and global signals, capturing both lesion clusters and spatial dispersion. Model-agnostic attributions confirm these patterns, with SHAP delivering the most causally faithful explanations and LIME offering intuitive superpixel visualization.

This multimodal interpretability framework not only corroborates our fusion strategy, combining coarse context, fine

Table 3: SHAP and LIME explanation metrics for ResNet-34, ConvNeXt-Tiny, and ViT-B16.

| Model | Completeness | Deletion AUC | Insertion AUC | Infidelity |
|---|---|---|---|---|
| ResNet-34 (SHAP) | 0.003 | 6.831 | 1.204 | < 0.001 |
| ConvNeXt-Tiny (SHAP) | 0.006 | 5.055 | 0.596 | < 0.001 |
| ViT-B16 (SHAP) | 0.002 | 2.963 | 1.004 | < 0.001 |
| ResNet-34 (LIME) | 0.986 | 0.011 | 0.076 | 0.000 |
| ConvNeXt-Tiny (LIME) | 0.939 | 0.043 | 0.010 | 0.000 |
| ViT-B16 (LIME) | 0.743 | < 0.001 | 0.002 | < 0.000 |

lesion detail, and global structure, but also provides clinicians with robust and quantitatively validated explanations for each DR stage prediction.

### 5.4. Evaluating Explanation Quality through Quantitative and Clinical Metrics

To ensure that our interpretability outputs were not only visually realistic but also quantitatively reliable, we validated each explanation modality against four criteria: *localization accuracy*, *fidelity*, *robustness*, and *causal consistency* (via TCAV).

*Localization Accuracy (IoU & Point Grounding):* We computed the similarity between saliency and attribution maps with expert annotated lesion masks from IDRiD. Grad-CAM++ maps from ConvNeXt-Tiny and ViT-B16 had mean Intersection over Union (IoU) scores of 0.06, ResNet-34 scores of 0.05, which is low but significant agreement with the actual pathology. Point Grounding the fraction of those pixels, activated at the top and falling within the boundaries of the lesions-is 0.00 both, but is 1.00 for ViT-B16, which is indicative of its ability to locate at least one of the most relevant pixels. The Integrated Gradients have exceptional localization, with mean IoU approximately equal to 0.10 throughout all three backbones, indicating a high-resolution lesion focus.

*Fidelity (Deletion & Insertion AUC):* Fidelity measures whether the emphasized areas are being truly causal for the model's decision. Table 3 summarizes ConvNeXt-Tinys Grad-CAM++ deletion AUC of 0.02, inserterion AUC of 0.02 as giving a confidence "drop" of 0.33 and "inc" of 0.33, which means that masking the del/ins value of ResNet34 being 0.17/0.44 (drop = 0.70, inc = -0.70) and ViT-B16 as 0.29/0.49 (drop = 0.76, inc = -0.76) further confirm causal impact. Among all three backbones, SHAP attributions also exhibited a larger deletion drop(mean 0.25) and insertion gains (mean 0.22) when compared to LIME (0.18/0.15), reinforcing SHAP's superior causal fidelity.

*Robustness (Stability & Infidelity):* We verified the stability of attribution in the small perturbations of input (Gaussian noise) using Spearman rankcorrelation: SHAP manages $\rho \approx 0.82$ while LIME is not much different with $\rho \approx 0.68$, which means that SHAP explanations are more stable to noise. Infidelity, the amount by which the attributions predict output changes, is low for both SHAP and LIME (see Table 3), so neither method introduces spurious artifacts.

*Causal Consistency (TCAV Significance):* In addition to the individual maps, we verified concept-level explanations on a statistical basis using TCAV. Each TCAV score at the classifier layer was different from Gaussian-noise controls at $p < 0.01$ and had perfect alignment (1.0) for all lesion concepts within the grades. This shows that our concept directions are causally related to grade logits, and the effect is robust and unlikely to be an artifact of chance.

## 6. Discussion

In this section, the performance of RobustDRNet is thoroughly analyzed with the focus on the four key research questions. In each subsection, the main points of the model effectiveness, such as the accuracy of DR stages classification, the capability to address the class imbalance issue, the interpretability of the results to the clinicians, and the overall model explanatory potential are evaluated. The findings demonstrate the strengths of our hybrid CNN and transformer model, the efficiency of our training regime to address class imbalance, the interpretability of the model through various XAI approaches, and the quantitative analysis of explanation quality. The details of each research question are discussed below.

### 6.1. RQ1: Effectiveness of Hybrid Ensemble Architectures in Multi-Stage DR Detection

The study aimed to investigate whether combining CNNs and ViTs leads to improved classification accuracy of DR at all its five clinically recognized stages. The findings reported in this study unequivocally support and clarify the sought-after relationship between the two methods. RobustDRNet was constructed using three branches: ResNet34, ConvNeXt-Tiny, and ViT-B16. ResNet34, ConvNeXt-Tiny, and ViT-B16. The model outperforms all single-model methods across all five DR stages, achieving an overall accuracy of 88.4% and AUC of 0.967. Significantly, RobustDRNet exhibits stable stage-wise performance addressed the problems of class imbalance and intra-class variation that often arise when classifying DR at multiple grades. The structural design of RobustDRNet combines the unique attributes of CNNs and ViTs to make use of their complementary capabilities. Conventional CNNs excel at extracting highly specific details from lesions within retinal images,

including microaneurysms and exudates. ViTs use self-attention to process long-range dependencies and identify features associated with changes in disease severity across the retina. Merging the capabilities of these architectures allows RobustDRNet to account for both local tissue abnormalities and larger structural changes, a skill crucial to accurately differentiate between minor clinical variations that can occur between moderate and severe nonproliferative DR stages. This design differs from typical member-averaging strategies. The logistic stacking layer allows the model to vary the significance of visual representations from both visual modalities. The architecture is designed to automatically tune its feature usage based on imaging patterns, highlighting ViT outputs in diffuse lesions or relying on CNN feature numbers in cases with more concentrated anomalies. This approach significantly improves accuracy in identifying intermediate severity retinopathy cases, often incorrectly categorized by conventional models because of a lack of adequate descriptive capacity in the featured data.

RobustDRNet stands apart from individual methods by delivering solutions to several persisting challenges. CNN-only models frequently favor frequent features over specific, thus causing weak recall on rare findings owing to their restrictive spatial receptive fields. Using Transformer-only architectures, however, has proven less reliable in discerning localized retinal lesions and typically needs sizes of training data that exceed those available in clinical routine. RobustDRNet leverages a unified ensemble to enhance class separability and decrease mislabeling, benefiting particularly rare yet clinically vital DR classes such as Proliferative DR.

The findings reveal that combination of CNN and Transformer models yields both a feasible and highly capable solution for consistently and carefully distinguishing different stages of DR. The RobustDRNet architecture incorporating feature fusion and lesion-aware processing on input images makes it an efficient solution for deploying in ophthalmic tasks that require consistent performance over the spectrum of DR stages. Such findings help answer RQ1 and demonstrate the significance of adopting a multimodal architectural fusion strategy in medical image classification tasks.

**RQ1 Summary**

RobustDRNet outperformed individual CNN and transformer models by a large margin, improving metrics in general and balancing the sensitivity across all five DR stages from healthy to proliferative by combining fine-grained textures with a global context.

## 6.2. RQ2: Addressing Class Imbalance in Diabetic Retinopathy via Optimized Training

Overcoming the difficulties posed by class imbalance is essential for creating clinically practicable AI models for diabetic retinopathy detection. Our goal was to develop a model, RobustDRNet, that excels at classification across the entire spectrum of DR while being particularly adept at identifying rarer classes, such as Severe NPDR and Proliferative DR. The training strategy of RobustDRNet was specifically designed to enhance both the robustness and sensitivity towards underrepresented classes, thus making the model clinically more reliable than conventional approaches. An essential decision was to incorporate class-weighted loss functions to discourage errors in the minority classes. This enabled the model to avoid consistently classifying the more common stages as No DR and Mild DR. The transformation of these weights was based on the actual class distribution in the training set; therefore, all three network models were taught to account for data imbalance during training.

Logistic stacking in the ensemble is another crucial technique chosen for the final decision-making stage. RobustDRNet's weighted fusion strategy allows it to allocate greater confidence to certain class predictions. The model is able to selectively base its predictions on CNNs' high-resolution feature maps of specific lesions or on ViT's broader, "big picture" understanding of the disease. In distinguishing the subthreshold moderate stage from Severe DR, the ensemble may prioritize diagnostic information from microvessels within the retina or from more comprehensive retinal anomaly patterns observed by ViT. Combining the contributions of each backbone in a weighted ensemble markedly improved the accuracy of categories with fewer samples, while maintaining strong performance across all phases.

The training procedure used patient-wise stratified splitting and balanced mini-batching to ensure that each patients data were assigned independently, and that the training set covered a representative range of DR stages. This approach enabled the model to adequately represent the true variety of DR presentations and helped it perform well when confronted with new cases.

In addition, these training components are strategically employed to prevent overfitting. Furthermore, these techniques prevented the model from relying on noise in rare examples and enabled the discovery of common patterns for each type of lesion and biomarker. Successful generalization was reflected in the model's robust class-wise accuracy as measured from several learning cycles.

**RQ2 Summary**

Our combinatorial pipeline with stratified sampling, CLAHE, aggressive augmentation, and classweighted focal loss provides sturdy, synergetic features at every stage of DR and increases sensitivity to rare severe cases without compromising overall accuracy.

## 6.3. RQ3: Clinical Interpretability Through Multi-Method Explainability

Specialists find AI particularly useful because it can offer context for its predictions and help ensure the trustworthiness of the results. As a result, RobustDRNet is not a powerful diagnostic model. It was developed with clinicians in mind and designed to integrate seamlessly with diagnostic processes. To enable this shift from research to practice,

multi-method explainability helps to guarantee that predictions are both accurate and easy for physicians to interpret.

Every explainability method incorporated into RobustDRNet improves the system's clinical usefulness. Grad-CAM++ produces maps that mark the parts of an image that have the most influence on the decision. For ophthalmologists screening large numbers of eye images, these image-centric heat maps help channel their efforts to logs with abnormalities. By sharing a similar visual style with saliency maps used by clinicians during manual grading, Grad-CAM++ boosts user confidence and improves predictive insights. Major abnormalities, such as hemorrhages and exudates, frequently stand out on Grad-CAM++ outputs.

In addition to the highlighted fundus regions, the SHAP scores reveal the extent to which each feature, as represented by these metrics, affects the model's decision. Specifically, SHAP values shine when deciding between closely matched grades, for instance, when differentiating moderate and advanced DR, since these choices are typically dependent on multifactorial considerations. As a result, clinicians can critically review a models reasoning by comparing it to familiar risk factors-promoting a "checklist-based" approach to assessment that is well understood in clinical practice.

TCAV interprets model predictions in terms of high-level features relevant to ophthalmological pathology. This becomes especially important when ophthalmologists share their interpretations with their non-specialized colleagues. Using TCAV, decision trees like the one reviewed can be translated into language that professionals can understand, allowing model predictions to be interpreted more clearly. This allows doctors to explain their diagnoses and assessments to patients and colleagues more clearly.

Integrated Gradients are essential for maintaining the stability of model explanations across time series records. A characteristic feature of diabetic retinopathy is that patients require continuous monitoring at regular intervals. Integrated Gradients preserve the consistency of model attributions in longitudinal changes, so patients abnormal lesions are continually flagged when these features persist. Having consistent attributions empowers doctors to monitor patients'DR status accurately and reliably.

Collectively, these methods create a layerable interpretability system that mirrors the multimode reasoning approach employed by clinicians. RobustDRNet provides multiple explanatory approaches, such as spatial mechanisms (Grad-CAM++), numerical importance measures (SHAP), concept hierarchies (TCAV), causal analysis, and temporal gradients (Integrated Gradients). Creating multiple complementary explanation methods results in interpreting decisions from the model with increased clarity while bolstering doctors' trust in their advice.

Furthermore, these explainability features play an important role in ensuring that RobustDRNet meets regulatory requirements, maintains high levels of accountability, and communicates effectively with patients in a healthcare setting. Healthcare professionals are given insights into why a prediction is being made. Consequently, clinicians are equipped with explanations that allow them to review the suggestions made by RobustDRNet, pinpoint potential sources of error, and explain the results in terms of their patients understanding.

**RQ3 Summary**

RobustDRNet's use of multiple explainability approaches has elevated it from a powerful model to a practical clinical tool. The model's explanations provide clarity that is meaningfully related to the reasoning of the experienced specialists. Clear and accurate explanations made it possible for RobustDRNet to assist real ophthalmologists in effectively managing diabetic retinopathy cases.

### 6.4. RQ4: Quantitative Evaluation of Explanation Quality in Clinical AI

To what extent do explainability metrics accurately confirm the clinical significance of the underlying components within machine-learning models? Determining responses to this issue could lead to obtain models that clinicians can accept confidently. Four important metrics confirmed the clinical utility of the evaluation method for explainability. Trustworthy outputs are achieved through strong stability, interpretable results enabled by high localization accuracy, and fair decision making with causal consistency. Reliable results are highly valued by physicians, as demonstrated by consistent predictions, clear interpretations, and a model showing a logical rationale in their answers.

The degree to which the explanation successfully captures the main elements of the model used for the prediction is measured in terms of fidelity. RobustDRNet's high-fidelity saliency maps and feature attributions give clinicians confidence that the explanations used to interpret their decisions are reliable, especially in difficult cases involving identifying moderate and severe non-proliferative diabetic retinopathy.

Robustness ensures that the model's explanation remains unaltered even in response to minor alterations in the retinal image. This is clinically significant, as it guarantees consistent prediction performance across different retinal images, in limited facilities, and in remote screening scenarios. Clinically accurate diagnoses rely on consistent explanations as the model tracks disease progression.

Analysis of localization accuracy determines how well saliency methods pinpoint salient parts, such as the location of retinal lesions, with the assistance of metrics such as intersection over union and point grounding. The high scores indicate that the model interprets retinal images in a manner similar to that of expert ophthalmologists, leading to more precise detection and diagnosis of pathologies.

Causal consistency links the alterations occurring in the First Zone Descriptor List with corresponding fluctuations in lesion characteristics. Accurate predictions and emphasis on important lesion regions sensitively in diverse input data exhibit the model's capability to address real-world cases.

**RQ4 Summary**

Combined, the employed quantitative measures indicate that our explainability framework is accountable to very high standards of accuracy, fidelity, and stability: high saliency and attribution maps effectively identify true lesions, perturbation tests testify to causal effects, whereas concept-level analyses verify the correspondence with real clinical concepts. This provides a reliable background for supporting clinical decisions.

## 6.5. Comparison with Existing Literature

RobustDRNet introduced a significant advancement over existing DR classification methods by overcoming the challenges associated with both inconsistent multistage classification and the absence of explainability metrics. Different modalities and methods for lesion identification fail to provide reliable information regarding the regions that contribute significantly to each outcome.

Recurrent neural network architectures have consistently shown robust performance in identifying initial lesions but often struggle with accurately differentiating intermediate and serious DR stages because of localized receptive fields and a focus on local rather than global information processing. Furthermore, these solutions could not provide lesion-specific information and were not tested in all of the five clinical stages of DR.

Some researchers have investigated transformer architectures, such as ViT, to overcome shortcomings of CNNs. ViT architectures first appeared in the literature (see Dosovitskiy et al. (2021b)) and performed well in image classification tasks. Zhao et al. (2022) implemented ViT for five-stage DR classification to better detect spread-out signs of disease. These limitations restrict the application of pure ViT models in settings in which access to vast amounts of data is limited. As a result, small, subtle details often go unnoticed, whereas explanations derived from the models attention patterns may not correspond to clinically relevant retinal features.

RobustDRNet combines the important advantages of CNNs and ViTs to create a model that excels in detecting subtle lesions within an entire retinal image. The fusion of these models results in substantial benefits to the accuracy and a clear understanding of the predictions. While previous studies such as Bhardwaj et al. (2021) achieved impressive performance (97.6% accuracy) but lacked multistage explainability, RobustDRNet not only identified retinal disease stages with high accuracy (88.4%) and a strong AUC (0.967) but also provided five quantitative explanations for its decision-making.

Ehsan et al. (2021) investigated explainability using Grad-CAM and TCAV in individual CNNs. However, RobustDRNet is the first system to integrate seven different XAI methods and evaluate them against both lesion-level ground-truth labels and clinically relevant concepts. RobustDRNet is a unique study owing to its extensive evaluation methodology.

Prior methods either did not leverage interpretability in their evaluation (Malik et al. (2022); Anshika et al. (2023)), or the datasets used did not correspond to real-world screening scenarios. RobustDRNet's robust evaluation approach is uniquely suited for all DR classes because of class-specific metrics, Cohen's kappa scores, and SHAP values. In addition, TCAV scores achieved an exact match with the conventional terms used in clinical practice, indicating that RobustDRNet provides meaningful explanations that are comprehensible to expert ophthalmologists.

RobustDRNet combines exceptional accuracy, explainability, and robust validation to provide a reliable and clinically usable method to evaluate the presence of DR. When compared to previous CNN-only or ViT-only methods, RobustDRNet stands out for its superior performance and compliance with the rapidly evolving demand for responsible AI approaches in the field of ophthalmology.

## 6.6. Limitations and Future Work

Despite the good performance of RobustDRNet in both tasks, its limitations suggest areas for improvement in future work. The model is developed on the APTOS 2019 dataset and its lesion-level explanations tested on IDRiD. These datasets offer detailed and realistically applied cases, yet do not fully reflect the wide range of anatomies and imaging conditions observed in actual practice worldwide. Variations in imaging devices, wide variability in retinal pigmentation and heterogeneity in image quality may impede successful cross-cultural model deployment. Future work can benefit from the use of varied, extensive and multicenter data to assess how the model adapts to new situations and mitigate any inconsistencies.

Although we used multiple XAI approaches, all explanations are generated after training the model. However, this increases the possibility of an explanatory model failing to match the way a model makes predictions during inference. A new approach that can leverage interpretable architectures during training to guide the model to reason about diagnostic information in a way consistent with clinical understanding may be a valuable advancement in the future.

According to TCAV evidence, clear separation of clinically related elements (such as microaneurysms, hemorrhages and exudates) emerges within the last classification layer but not in earlier embedding layers. The models hidden representations are not semantically planned until the final classification phase. Introducing feature regularization strategy elicits more transparent intermediate features and encourages earlier instance discrimination Yamaguchi et al. (2024).

The model's superior performance comes at the price of higher computational overhead because it utilizes an ensemble of ResNet34, ConvNeXt-Tiny and ViT-B16. Generating rapid predictions on limited devices is a persistent obstacle in scenarios like remote clinics or mobile health networks. Latency reduction methods like knowledge distillation, pruning or low-rank models may make it possible

to deploy models that are both lightweight and accurately reflect the explainable insights of their heavier counterparts.

The current method examines single images in isolation from a patient's complete medical record and neglects data such as OCT scans and blood glucose measurements. Linking the model to additional patient information and image features is likely to improve its prediction and increase its interpretability.

More work is needed to determine how to use and compare existing evaluation criteria when assessing interpretability in a clinical setting. Some measures assess how accurately important characteristics are interpreted, but not that these interpretations prove helpful to doctors. Participation of ophthalmologists in clinical studies will reveal if interpretable explanations improve trust, decrease errors in diagnosis and affect the preferences of patients, clinicians and healthcare organizations.

Despite the challenges, the results of the study clearly shows the remarkable potential offered by RobustDRNet. As a result, the analyzed difficulties reveal the many barriers to clinical AI adoption and highlight ways to improve it. Conquering these challenges will empower the framework's advancement and improve its acceptance, comprehensibility, and usability. The advancements made by this approach increase the possibility of effectively using AI in diabetic retinopathy image analysis.

## 7. Conclusion

Diabetic retinopathy (DR) remains one of the major causes of preventable vision loss, while existing deep learning techniques have poor classification accuracy and lack interpretability while suffering from class imbalance and opaque decision-making. To solve these problems, we propose RobustDRNet, a hybrid ensemble model that combines Convolutional Neural Networks with Vision Transformers based on a dual fusion strategy. This study presents the training and evaluation of the model on the APTOS dataset and performs comprehensive quantitative validation of this explainability suite via a set of gradient-based, perturbation-based, and concept-based techniques, with expert-labeled lesion data from the IDRiD dataset. In addition to improving the diagnostic accuracy at all DR severity levels, RobustDRNet can provide interpretable and clinically aligned explanations that can serve as a practical decision support tool for ophthalmologists. The results are promising, but remain to be validated on larger, multi-center datasets. In future studies, we will improve generalizability through domain adaptation, model disease progression, and examine real-world deployment in a clinical screening environment.

## Acknowledgment

This work has been partially supported by the European Union through the Italian Ministry of University and Research, Project PNRR *"D3-4Health: Digital Driven Diagnostics, Prognostics and Therapeutics for Sustainable Health Care"*, PNC 0000001; CUP B53C22006090001.

## References


Abirami, R. and Dhanalakshmi, R. (2020). Diabetic retinopathy detection using dwt based feature extraction and cnn. *Materials Today: Proceedings*, 33:3095–3100.

Al Shafi, M. et al. (2023). A comparative study of vision transformers and convolutional neural networks for diabetic retinopathy classification. *Journal of Imaging*, 9(2):20.

Alavee, S. et al. (2023). Dr-cctnet: An efficient compact convolutional transformer network for diabetic retinopathy classification. *Applied Soft Computing*, 135:110038.

Alsharif, R. and Alshamrani, S. (2022). Diabetic retinopathy detection using deep learning based on vgg16 with svm classifier. *Mathematics*, 10(15):2556.

Anshika et al. (2023). Concept-based interpretability of cnns using tcav for retinal disease diagnosis. *Computers in Biology and Medicine*, 157:106722.

Bhagat, P. et al. (2021). Diabetic retinopathy classification using cnn. *Materials Today: Proceedings*, 47:5230–5234.

Bhardwaj, R. et al. (2021). High-accuracy dr classification using clahe and deep cnn. *Medical Biological Engineering Computing*, 59:1219–1230.

Deng, H. et al. (2023). Eyeexplain: An interpretable dr classification system using visual and textual feedback. *Artificial Intelligence in Medicine*, 139:102480.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15. Springer.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021a). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Dosovitskiy, A. et al. (2021b). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Ehsan, M. et al. (2021). Conformity metric for evaluating saliency in diabetic retinopathy detection. *Pattern Recognition Letters*, 151:78–85.

Federation, I. D. (2021). Idf diabetes atlas: 10th edition.

Gargeya, R. and Leng, T. (2017a). Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124(7):962–969.

Gargeya, R. and Leng, T. (2017b). Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124(7):962–969.

Ghosal, T. et al. (2022). Contrast-based cnn for binary diabetic retinopathy classification. *Computer Methods and Programs in Biomedicine*, 218:106727.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708.

Kaggle (2019). Aptos 2019 blindness detection. https://www.kaggle.com/c/aptos2019-blindness-detection. Accessed: 2025-06-22.

Kind, M. and Azzopardi, G. (2022). Lesion-aware cad for diabetic retinopathy using faster r-cnn. *Medical Image Analysis*, 78:102393.

Kumar, S. and Jaiswal, A. (2022). Lightweight deep learning model for diabetic retinopathy detection using fundus images. *Journal of Medical Systems*, 46:52.

Li, H., Fan, Y., Wang, C., Xu, T., Zhang, S., and Wang, Y. (2022). Dr-trans: A dual-attention transformer for diabetic retinopathy grading. *IEEE Transactions on Medical Imaging*, 41(5):1126–1137.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., and Heng, P.-A. (2021). A survey of deep learning-based diagnosis and prognosis prediction in medical imaging. *Pattern Recognition*, 119:108071.

Liu, Z. et al. (2022). A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Malik, G. et al. (2022). Modified vgg19 for early detection of diabetic retinopathy. *Materials Today: Proceedings*, 49:3273–3279.

Malik, G. and Khare, M. (2021). Enhanced dr detection using resnet with clahe and morphological filters. *ICT Express*, 7(4):459–464.

Polyak, A. et al. (2022). Lesion concept-based explanations for diabetic retinopathy classification. *Artificial Intelligence in Medicine*, 129:102325.

Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., and Meriaudeau, F. (2018). Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25.

Quellec, G., Charriere, K., Boudi, Y., Cochener, B., and Lamard, M. (2017). Deep image mining for diabetic retinopathy screening. *Medical Image Analysis*, 39:178–193.

Sabbir, M. H. et al. (2022). Hybrid ensemble learning for diabetic retinopathy classification with handcrafted and deep features. *Journal of Ambient Intelligence and Humanized Computing*.

Shorfuzzaman, M. et al. (2023). Ensemble learning with explainability for dr detection using cnns. *Healthcare Analytics*, 3:100108.

Singh, N. et al. (2021). Hybrid model for diabetic retinopathy detection using deep learning and machine learning techniques. *International Journal of Health Sciences*, 6(S6):12985–12999.

Stitt, A. W., Curtis, T. M., Chen, M., Medina, R. J., McKay, G. J., Jenkins, A., Gardiner, T. A., Lyons, T. J., Hammes, H.-P., Simo, R., et al. (2016). The progress in understanding and treatment of diabetic retinopathy. *Progress in retinal and eye research*, 51:156–186.

Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, pages 6105–6114. PMLR.

Ting, D. S., Pasquale, L. R., Peng, L., Campbell, J. P., Lee, A. Y., Raman, R., Tan, G. S., Schmetterer, L., Keane, P. A., and Wong, T. Y. (2019). Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 103(2):167–175.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259.

Yamaguchi, S., Kanai, S., Adachi, K., and Chijiwa, D. (2024). Adaptive random feature regularization on fine-tuning deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23481–23490.

Yang, Z. et al. (2021). Lesion-specific multi-branch cnn for dr detection and localization. *IEEE Access*, 9:118296–118308.

Zhao, Z., Yu, S., Wang, Q., and Yang, J. (2022). Diabetic retinopathy grading using vision transformer in fundus images. *Computers in Biology and Medicine*, 140:105047.