

---

# VR-FUSENET: A FUSION OF HETEROGENEOUS FUNDUS DATA AND EXPLAINABLE DEEP NETWORK FOR DIABETIC RETINOPATHY CLASSIFICATION

---

**Shamim Rahim Refat**

Software Engineer (AI/ML)  
Technohaven Company Ltd.  
Dhaka, Bangladesh  
n.a.refat2000@gmail.com

**Ziyan Shirin Raha**

Department of Computer Science and Engineering  
Ahsanullah University of Science and Technology  
Dhaka, Bangladesh  
ziyanraha@gmail.com

**Shuvashis Sarker**

Department of Computer Science and Engineering  
Ahsanullah University of Science and Technology  
Dhaka, Bangladesh  
shuvashisofficial@gmail.com

**Faika Fairuj Preotee**

Department of Computer Science and Engineering  
Ahsanullah University of Science and Technology  
Dhaka, Bangladesh  
faikafairuj2001@gmail.com

**MD. Musfikur Rahman**

Department of Computer Science and Engineering  
Ahsanullah University of Science and Technology  
Dhaka, Bangladesh  
mush9031@gmail.com

**Tashreef Muhammad**

Department of Computer Science and Engineering  
Southeast University  
Dhaka, Bangladesh  
tashreef.muhammad@seu.edu.bd

**Mohammad Shafiu1 Alam**

Department of Computer Science and Engineering  
Ahsanullah University of Science and Technology  
Dhaka, Bangladesh  
shafiu1.cse@aust.edu

June 24, 2025

**ABSTRACT**

Diabetic retinopathy is a severe eye condition caused by diabetes where the retinal blood vessels get damaged and can lead to vision loss and blindness if not treated. Early and accurate detection is key to intervention and stopping the disease progressing. For addressing this disease properly, this paper presents a comprehensive approach for automated diabetic retinopathy (DR) detection by proposing a new hybrid deep learning model called VR-FuseNet. Diabetic retinopathy is a major eye disease and leading cause of blindness, especially among diabetic patients, so accurate and efficient automated detection methods are required. To address the limitations of existing methods including dataset imbalance, diversity and generalization issues this paper presents a hybrid dataset created from five publicly available diabetic retinopathy datasets: APTOS 2019, DDR, IDRiD, Messidor 2, and Retino. Essential preprocessing techniques such as Synthetic Minority Over-sampling Technique (SMOTE) for class balancing and Contrast Limited Adaptive Histogram Equalization (CLAHE) for image enhancement are applied systematically to the dataset to improve the robustness and generalizability of the dataset. The proposed VR-FuseNet model combines the strengths of two state-of-the-art convolutional neural networks—VGG19 which captures fine-grained spatial features and ResNet50V2 which is known for its deep hierarchical feature extraction. This fusion improves

the diagnostic performance and achieves an accuracy of 91.824%, precision of 92.612%, recall of 92.233%, F1-score of 92.392% and AUC of 98.749%. The model outperforms individual architectures on all performance metrics demonstrating the effectiveness of hybrid feature extraction in Diabetic Retinopathy (DR) classification tasks. To make the proposed model more clinically useful and interpretable, this paper incorporates multiple Explainable Artificial Intelligence (XAI) techniques including Gradient-weighted Class Activation Mapping (Grad-CAM), Grad-CAM++, Score-CAM, Faster Score-CAM and Layer-CAM. These techniques generate visual explanations that clearly indicate the retinal features affecting the model's prediction such as microaneurysms, hemorrhages and exudates so that clinicians can interpret and validate.

**Keywords** Diabetic Retinopathy · Hybrid Dataset · Synthetic Minority Over-Sampling Technique (SMOTE) · Contrast Limited Adaptive Histogram Equalization (CLAHE) · Hybrid Model · Feature Fusion · VR-FuseNet · VGG19 · ResNet50V2 · Convolutional Neural Network (CNN) · Explainable AI (XAI)

## 1 Introduction

The human eye is an amazing sensory organ that is the gateway to visual perception. It detects light and sends signals to the brain through the optic nerve to form images, colors and depth perception [1]. This complex process allows us to navigate, recognize faces and do many activities that require visual input. But many medical conditions can impair the eye and cause significant visual loss. Retinopathy is a condition that affects the retina and is often a result of systemic diseases like hypertension, chronic kidney disease, heart disease and most importantly diabetes. Diabetes is a chronic metabolic disorder characterized by high blood sugar, it is a global health problem affecting millions of people worldwide. It occurs when the body either doesn't produce insulin (Type 1 diabetes) or can't use insulin properly (Type 2 diabetes) and results in consistently high blood sugar levels. Over time, uncontrolled diabetes results in complications affecting multiple organ systems including cardiovascular, renal and nervous systems. Among these complications ocular diseases like Diabetic Retinopathy (DR) is a major threat to vision and can be irreversible if left untreated [2, 3]. The prevalence of diabetes is alarming, 537 million people are affected globally as of 2021, projected to rise to 578 million by 2030 and 700 million by 2045 according to the report of International Diabetic Federation [4]. This burden is disproportionately affecting low and middle income countries where 75% of diabetes cases are reported. In these regions lack of healthcare access, limited awareness and inadequate screening programs contribute to the high prevalence of diabetes related complications including DR [5, 6].

Among the many complications of diabetes, Diabetic Retinopathy (DR) is the leading cause of vision loss and blindness in people with diabetes. As of 2020, 103 million adults worldwide had DR, and by 2045 that will increase to 161 million. In 2020 alone, 1.07 million (95% UI: 0.76–1.51 million) were blind from DR and 3.28 million (95% UI: 2.41–4.34 million) had moderate to severe visual impairment (MSVI) [7]. In the United States of America (USA), 9.6 million (26.43%) people with diabetes had DR in 2021, and 1.84 million (5.06%) had Vision-Threatening Diabetic Retinopathy (VTDR) [8]. Regionally, the highest burden of DR blindness and MSVI was in Latin America & the Caribbean (6.95%, 95% UI: 5.08–9.51). North America and the Caribbean had high prevalence rates (33.30%), while South and Central America had the lowest rates (13.37%). Elsewhere, Africa had the highest DR prevalence at 35.90%, followed by North Africa & the Middle East (2.12%, 95% UI: 1.55–2.79) [9]. Ethnic groups also varied in their susceptibility, with Hispanics (OR: 2.92; 95% CI: 1.22–6.98) and Middle Easterners (OR: 2.44; 95% CI: 1.51–3.94) more likely to develop DR than Asians [10].

Diabetic retinopathy happens when high blood sugar over time damages the small blood vessels in the retina causing microvascular problems like microaneurysms, haemorrhages, hard exudates and neovascularization. Advanced stages are called Proliferative Diabetic Retinopathy (PDR) where new abnormal blood vessels grow and the risk of retinal detachment and total blindness increases. People with diabetes are also prone to other eye problems like glaucoma and cataracts. As shown in Figure 1, the pathological changes in a diabetic retina compared to a normal one are microaneurysms, cotton wool spots, hard exudates and neovascularization which leads to progressive visual impairment. Symptoms can range from blurry vision and dark spots to total blindness as the disease progresses. Since DR is the leading cause of blindness among working age adults, early detection through regular eye check ups, prompt medical attention and good diabetes management is key to prevent further progression of the disease.

Machine learning (ML) has changed Diabetic Retinopathy (DR) diagnosis by automated and accurate detection, reducing manual assessments and healthcare burden [12, 13]. Traditional diagnosis using retinal fundus imaging is time consuming, expert dependent and variable, making early detection challenging; hence researchers use ML and Deep Learning (DL) for stage classification, lesion detection and severity grading [14]. Hybrid models combining ML, metaheuristic optimization and computational methods have improved performance and generalization [13, 15]. Large public datasets have enabled successful binary and multi-class classifications, but multi-class approaches face challenges

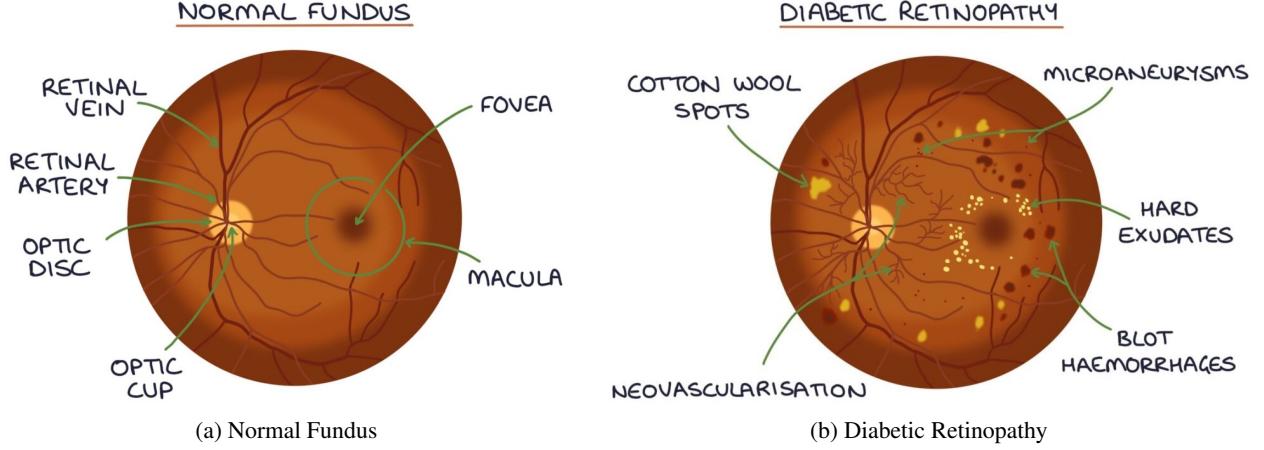


Figure 1: Differences between Normal Fundus Image and Diabetic Retinopathy Fundus Image. [11]

like data imbalance, imaging quality issues and suboptimal feature extraction [16, 17, 18, 19, 20, 21, 22, 23]. To address these issues, preprocessing methods like contrast enhancement, generative modeling, attention based extraction have improved diagnostic accuracy [24]. Ensemble methods, transfer learning and hybrid classifications have advanced automated detection of critical DR indicators like microaneurysms, hemorrhages and lesion segmentation. Optimized classifiers and advanced architectures have further improved accuracy and performance in multi-view joint learning using public datasets [25]. Despite these improvements, issues like dataset imbalance, noise in fundus images and overfitting still persist, affecting model generalization while computational demands pose challenge for real world clinical integration. But ML driven DR diagnosis has revolutionized ophthalmic diagnostics, making early screening more accessible, efficient disease monitoring and overall patient care outcomes [26].

Despite the progress in Machine learning based Diabetic Retinopathy (DR) detection, we need to address the current limitations. Existing models although work well in many controlled settings, fail to generalise across different datasets and real world clinical environments. This is mainly due to issues like data imbalance, limited diversity in training datasets and inability to capture temporal and contextual features which are important for detecting subtle disease progression. So we focus on improving model generalisation across different populations, incorporating sequential dependencies in retinal features to better represent disease evolution over time and developing consistent, robust and clinically interpretable explainable AI (XAI) methods to support clinical decision making. We also aim to reduce the dependency on large annotated datasets through efficient learning strategies and ensure our models are computationally efficient and scalable for deployment in resource constrained environments. By using advanced computational techniques, state of the art learning frameworks and robust validation processes, we hope to set a new benchmark in DR detection. To achieve this we propose the following new contributions:

- i **Hybrid Dataset:** Combines five public datasets to generalize, reduce biases and increase data diversity for robust performance across different imaging conditions.
- ii **Model Evaluation Across Individual Datasets:** Conducts an extensive evaluation of individual deep learning models, including VGG16, VGG19, ResNet50V2, MobileNetV2, and Xception, on each dataset independently before applying the combined hybrid dataset, to ensure accurate performance assessment and robust generalization.
- iii **VR-FuseNet Model:** Proposes and develops VR-FuseNet, a hybrid deep learning model inspired by the VGG19 and ResNet50V2 architectures, effectively leveraging their complementary strengths for improved diabetic retinopathy classification performance.
- iv **Explainable AI (XAI) Methods:** Uses Grad-CAM, Grad-CAM++, Layer-CAM, Score-CAM, Faster Score-CAM to evaluate and find the best approach to highlight DR specific features.
- v **Clinical Interpretability Heatmaps:** Shows multiple XAI heatmaps comparison, highlighting microaneurysms, exudates and hemorrhages and provides a clinician friendly interface for model validation.

By addressing these challenges, our approach gives better generalization, better interpretability and better accuracy, hence a more robust and scalable DR detection framework for clinical use. This paper is structured as follows: Section 2 reviews the existing research and methods for DR classification. Section 3 outlines the approaches to research, including the techniques and strategies utilized throughout the study. Section 4 presents the evaluation metrics and results of the

proposed model. Section 5 shows various Explainable AI (XAI) techniques for DR diagnosis. Section 6 summarizes the findings, limitations and future works. Section 7 concludes the paper.

## 2 Review of the Literature

The automation of DR detection has emerged as a critical area of research due to the rising global prevalence of diabetes and the urgent need for early diagnosis to prevent vision loss. In response, numerous studies have explored a wide range of computational techniques—including traditional machine learning, deep learning, and advanced image analysis—to enhance diagnostic accuracy and clinical applicability. These efforts have involved diverse strategies such as innovative data preprocessing, sophisticated model architectures, and the integration of interpretability frameworks.

Saxena et al. [27] used InceptionV3 and InceptionResNetV2 on EyePACS and Messidor datasets and got AUC values ranging from 0.92 to 0.958. Deep convolutional networks are powerful for DR presence detection. But their model was only for binary classification and not for severity levels of DR—a limitation same as Yi et al. [28] who used transfer learning CNNs and improved performance by dataset cleaning and got accuracy from 71.18% to 85.13%. But their model was also only for presence/absence classification and not for clinical granularity. Mukherjee et al. [29] targeted this gap by integrating ROI extraction, Z-score normalization and EfficientNet B1 to do severity grading. Their system got impressive results with AUC values of 0.98 for referable DR screening and DR severity grading. But their interpretability framework was limited as they only used Grad-CAM without benchmarking it against other explanation methods. Xu et al. [25] proposed RT2Net with robust preprocessing techniques like CLAHE and background removal and got 85.4%–88.2% accuracy on EyePACS and APTOS. But their fusion of multi-view features increased computational complexity and made class imbalance more challenging. Tariq et al. [30] highlighted the importance of choosing the right classifier by comparing SVM and DenseNet121 and got a huge jump in accuracy from 52.57% to 93%. But absence of oversampling details limited replication and reliability.

To address class imbalance, Alwakid et al. [24] and Alwakid et al. [31] conducted experiments on APTOS and DDR datasets using DenseNet121 and advanced image enhancement techniques like ESRGAN, histogram equalization and CLAHE. They reported accuracy of 98.7% which shows the potential of high resolution image synthesis in improving the classification accuracy. But they didn't compare with other deep models so the merit of their approach is unclear. Same issue found in Mohanty et al. [32] who developed a hybrid model with VGG16, XGBoost and DenseNet121 and got 97.3% accuracy but didn't compare with more recent architectures. Wong et al. [33] tested how the performance changes across DR grading scales using ShuffleNet and ResNet18 and found significant drop in accuracy with finer class granularity—highlighting the need for more complex multi-class training pipelines. They used only one optimizer and froze the model layers which may limit the generalization. Mutawa et al. [34] broadened the architectural comparison by testing VGG16, MobileNetV2 and DenseNet121 but didn't use image enhancement or denoising which may compromise the robustness in real world clinical settings. Nahiduzzaman et al. [35] used PCNN-ELM and got decent accuracy on EyePACS and APTOS but didn't benchmark with CNNs which dominate the domain. Murugappan et al. [36] introduced DRNet and got 99.7% in binary and 98.1% in multi-class classification but their reliance on limited datasets and Grad-CAM alone makes it difficult to evaluate the generalizability and model transparency.

More recent studies have explored interpretability, hybrid modeling and preprocessing customization for DR detection. Abdel Maksoud et al. [37] introduced E-DenseNet with HEPDS and median filtering and got 91.2% average accuracy across multiple datasets but underperformed for normal DR cases. Bilal et al. [38] used U-Net and CNN-SVD hybrids with green channel transformations and got nearly 98% on EyePACS but didn't define oversampling protocol to handle imbalance. Beham et al. [39] integrated InceptionV3, custom CNN and PBIL and scored high sensitivity and specificity (88.13% and 96.56% respectively) but didn't provide cross comparison with standard models. Butt et al. [40] combined GoogleNet and ResNet18 with SVM and got 97.8% binary classification but reduced DR classes from 5 to 3 which may undermine nuanced diagnosis. Abbood et al. [41] used Gaussian blur, foreground identification and DRRNet and scored up to 93.6% but struggled to detect mild DR due to weak supervision and illumination inconsistencies. Similarly Ali et al. [42] applied IR-CNN with histogram equalization and normalization and got 96.85% but evaluated only 2 deep models (ResNet50 and InceptionV3). Raiaan et al. [43] used RetNet-10 with comprehensive set of preprocessing methods (CLAHE, morphological opening, ROI extraction) and got 98.65% but limited availability of annotated DR grade samples reduced reliability. Menauer et al. [44] implemented hybrid CNN+VGG16+VGG19 architecture and got 90.6% but didn't provide baseline comparison and mentioned need for longer training epochs. Mercaldo et al. [45] compared CNNs like VGG16, MobileNet and AlexNet for both binary and severity classification and got up to 97% but struggled with class imbalance and didn't specify oversampling details. Vasireddi et al. [46] used DFNN with LOA and got 98.04% but oversimplified DR progression and used only LIME for explainability. Alghamdi et al. [47] tested VGG16, DenseNet121 and ResNet18 and got limited performance in multiclass (48.43%) and used only Grad-CAM which restricted interpretability depth. Many works adopted only a single explainable AI technique. In this context,

Table 1 provides the findings and methodologies from the 24 research papers that were systematically reviewed as part of this study.

Table 1: Comparative Analysis of Previous Studies

Author	Dataset	Preprocessing Techniques	Methodology	Result Analysis	Explainable AI	Limitation
Saxena et al. [27]	EyePACS, Messidor 1, Messidor 2	Cropping and Centering Data Augmentation	InceptionV3, InceptionResNetV2	AUC of 0.927 on EyePACS, AUC of 0.958 on Messidor 1, AUC of 0.92 on Messidor 2	Model name wasn't mentioned	Only checked if DR were present or not, didn't separate different levels of DR severity
Tariq et al. [30]	EyePACS, APTOS 2019	Image Resizing	ResNet50, DenseNet121, SVM	EyePACS: SVM 52.57% APTOS 2019: DenseNet121 93%	X	Class imbalance, exact amount of data oversampling details were missing
Mukherjee et al. [29]	EyePACS, Hybrid (APTOS 2019 + Messidor 1+ Messidor 2)	ROI Extraction, Ben Graham Approach, Z-score Normalization, Data Augmentation	Transfer Learning CNNs	EfficientNet B1 recorded an accuracy of 0.9511 in DR severity grading, 0.9934 in DR-Screening and 0.9798 in referable DR	Grad-CAM	Exact amount of data oversampling details were missing, only Grad-CAM was used for explanation, without comparing other methods
Yi et al. [28]	APTOS 2019, Messidor 2	Didn't mention any preprocessing technique	Transfer Learning CNNs	Binary Classification for the APTOS 2019 Dataset: average accuracy after cleaning increased from 71.18% to 85.13%	X	Classified only DR as only present or absent, didn't distinguish between different severity levels
Xu et al. [25]	APTOS 2019, Eye-PACS	Image Resizing, Removing Background, Image Gray-scaling, CLAHE	RT2Net	APTOS 2019: 85.4%, EyePACS: 88.2%	X	Class imbalance, computational complexity due to multi-view feature fusion

Table continues to next page

Author	Dataset	Preprocessing Techniques	Methodology	Result Analysis	Explainable AI	Limitation
Alwakid et al. [24]	APPOS 2019, DDR	CLAHE, ESRGAN, Data Augmentation	DenseNet121	APPOS 2019: 98.7%, DDR: 79.67%	X	Small sample size and dataset biases, only DenseNet121 was used, without testing other models
Mohanty et al. [32]	APPOS 2019	Image Resizing, Gaussian Blur, Ben Graham Approach	Hybrid Model (VGG16 + XGBoost Classifier), DenseNet121	Hybrid Model: 79.50%, DenseNet 121: 97.30%	X	Only VGG16 and DenseNet 121 were tested, no comparison with other deep learning models
Moustari et al. [26]	APPOS 2019	Laplacian of Gaussian (LoG), CLAHE	DenseNet121	98.42%	Grad-CAM	Imbalance Dataset, only DenseNet121 was used, limiting model comparison, only Grad-CAM was used without comparing other interpretation methods
Alwakid et al. [31]	APPOS 2019	Histogram, CLAHE, ESRGAN, Data Augmentation	DenseNet121	98.36%	X	Only DenseNet121 was used, without testing other models, ESRGAN-based image enhancement requires high computation, extensive hyperparameter tuning was needed for optimization
Wong et al. [33]	APPOS 2019, Hybrid (Eye-PACS+ Messidor 2)	CLAHE	ShuffleNet, ResNet18	APPOS 2019 2 Class Grading: 96%, APPOS 2019 5 Class Grading: 82%, Hybrid Dataset 3 Class Grading: 75%	X	ADE was the only optimizer tested, without comparing others, frozen layers were used from only two transfer learning models
Nahiduzzaman et al. [35]	EyePACS, APPOS 2019	CLAHE	PCNN-ELM	EyePACS: 91.78%, APPOS 2019: 97.27%	X	Imbalanced dataset, used ML techniques for evaluation metrics, no comparison with deep learning models

Table continues to next page

Author	Dataset	Preprocessing Techniques	Methodology	Result Analysis	Explainable AI	Limitation
Mutawa et al. [34]	APPOS 2019, Eye-PACS, ODIR	Data Augmentation	VGG16, InceptionV3, MobileNetV2, DenseNet 121	By using hybrid dataset, DenseNet 121: 98.97%, VGG16: 98.79%, MobileNetV2: 98.51%, InceptionV3: 97.21%	X	Didn't used any image enhancement, noise removal preprocessing technique
Murugappan et al. [36]	APPOS 2019	Image Resizing	DRNet	Binary Classification: 99.73%, Multiclass Classification: 98.18%	Grad-CAM	No benchmarked datasets for proper few-shot classification evaluation, 100-episode limit may exclude some training samples, only Grad-CAM was used, without comparing other methods
Abdel Maksoud et al. [37]	APPOS 2019, IDRiD, Eye-PACS, Messidor	Histogram Equalization for Brightness Preservation based on a Dynamic Stretching (HEBPDS) and Median Filter, Data Augmentation	Hybrid E-DenseNet	Average 91.2%	X	Low AUC for the normal class (35% in most cases) except higher AUC (67%) only in the MESSIDOR dataset
Bilal et al. [38]	Messidor 2, Eye-PACS 1, DI-ARETDB0	Green Channel Image, Top-Bottom Hat Transformation, Data Augmentation	U-Net models, Hybrid CNN-SVD	InceptionV3, EyePACS 1: 97.92%, Messidor 2: 94.59%, DI-ARETDB0: 93.52%	X	Class imbalance was present, lack of defined oversampling counts
Beham et al. [39]	EyePACS	Didn't mention any preprocessing technique	InceptionV3, Custom CNN, PBIL	PBIL Sensitivity: 88.13%, PBIL Specificity: 96.56%	X	Only specific CNN architectures were used without comparing other models

Table continues to next page

Author	Dataset	Preprocessing Techniques	Methodology	Result Analysis	Explainable AI	Limitation
Butt et al. [40]	APPOS 2019	Normalization, Image Resizing	(GoogleNet and ResNet18) +SVM	Binary Classification: 97.80%, Multiclass Classification: 89.29%	X	Classes were reduced from five to three to manage data imbalance
Abbood et al. [41]	EyePACS, Messidor	Foreground Identification and Circle Crop, Gaussian Blur	DRRNet	EyePACS: 92% , Messidor: 93.6%	X	Limited image level supervision affects DR feature detection, difficulty in distinguishing Mild DR due to the similarity with normal grades, illumination issues remain unresolved
Ali et al. [42]	EyePACS	Histogram Equalization, Intensity Normalization	IR-CNN	96.85%	X	Class imbalance, Only ResNet50 and InceptionV3 were used, without testing other architectures
Raiyan et al. [43]	APPOS 2019, Messidor 2, IDRiD	OTSU Thresholding, Contour Detection and Sorting, ROI Extraction, Morphological Opening, NLMD, CLAHE, Data Augmentation	RetNet-10	98.65%	X	Few images available for DR grades three and four after combining datasets, needed better preprocessing to handle noisy images
Menaouer et al. [44]	APPOS 2019	Data Augmentation	Hybrid Model (CNN+ VGG16+ VGG19)	90.6%	X	Only hybrid deep learning features were used, without comparing other techniques limited dataset size, more training epochs are needed to improve accuracy
Alghamdi et al. [47]	APPOS 2019	Image Resizing	VGG16, DenseNet121, ResNet18	VGG16 for, Binary Classification: 73.04%, Multiclass Classification: 48.43%	Grad-CAM	Only Grad-CAM was used, without comparing other methods

Table continues to next page

Author	Dataset	Preprocessing Techniques	Methodology	Result Analysis	Explainable AI	Limitation
Mercaldo et al. [45]	Hybrid (Eye-PACS+ APTOS 2019)	Gaussian Filter, Cropping, Data Augmentation	StandardCNN, VGG16, AlexNet, MobileNet	Binary Classification for, Presence of DR: VGG16 97%, Severity of DR: VGG16 and MobileNet 91%	Grad-CAM, Score-CAM	Class imbalance was present, lack of defined oversampling counts
Vasireddi et al. [46]	Messidor	AHE, Gamma Correction, CLAHE	DFNN with LOA	98.04%	LIME	Assumption of constant learned features ignoring DR progression, low amount of data, only LIME was used, without comparing other interpretability methods

### 3 Approaches to Research

Our proposed methodology uses a systematic approach to detect diabetic retinopathy using advanced deep learning techniques. Figure 2 shows the workflow of our method. We selected five public datasets to get a big and representative dataset. After that we did comprehensive preprocessing including Random Undersampling and SMOTE for class imbalance, CLAHE for contrast enhancement, normalization for same data scale and resizing images to a standard size for model input. Each preprocessed dataset was evaluated individually using popular transfer learning models like VGG16, VGG19, ResNet50V2, MobileNetV2, and Xception to get baseline metrics. Then we combined all the datasets into one hybrid dataset to test the transfer learning and our proposed models robustness and generalizability across different retinal images. To make it more transparent and interpretable we applied five gradient based Explainable AI (XAI) methods using the final convolutional layers of our trained models. We did comprehensive performance evaluation and statistical analysis to quantify the improvements and validate our approach.

#### 3.1 Dataset

To ensure a robust and generalized model for diabetic retinopathy (DR) detection, we utilize five publicly available datasets: APTOS 2019, DDR, IDRiD, Messidor 2 and Retino. The integration of multiple datasets enhances diversity in imaging conditions, camera settings, and population demographics, reducing biases and improving model performance across different clinical scenarios. Table 2 and Figure 3 presents the distribution of various Diabetic Retinopathy stages across different datasets, with Figure 3 employing a logarithmic scale to enhance the visibility of class imbalances. Figure 4 displays the sample images of various datasets.

##### 3.1.1 Description of used dataset

**APTOS 2019:** The APTOS 2019 Blindness Detection Dataset, a publicly available Kaggle dataset [48], consists of 3,662 high-resolution retinal fundus images, categorized into five severity levels of diabetic retinopathy. The dataset includes images labeled as no DR, mild DR, moderate DR, severe DR, and proliferative DR, with 1805, 370, 999, 193, and 295 images, respectively. The images, with a resolution of  $3216 \times 2136$  pixels, were collected from various clinics using different imaging devices, introducing natural variations in image quality and lighting conditions, which help create a more diverse and challenging dataset.

**DDR:** The DDR dataset [49] has six stages; however, for uniformity with other datasets, we employ just five stages of diabetic retinopathy (DR), from stage 0 (no DR) to stage 4 (proliferative DR), omitting the sixth stage. This dataset,

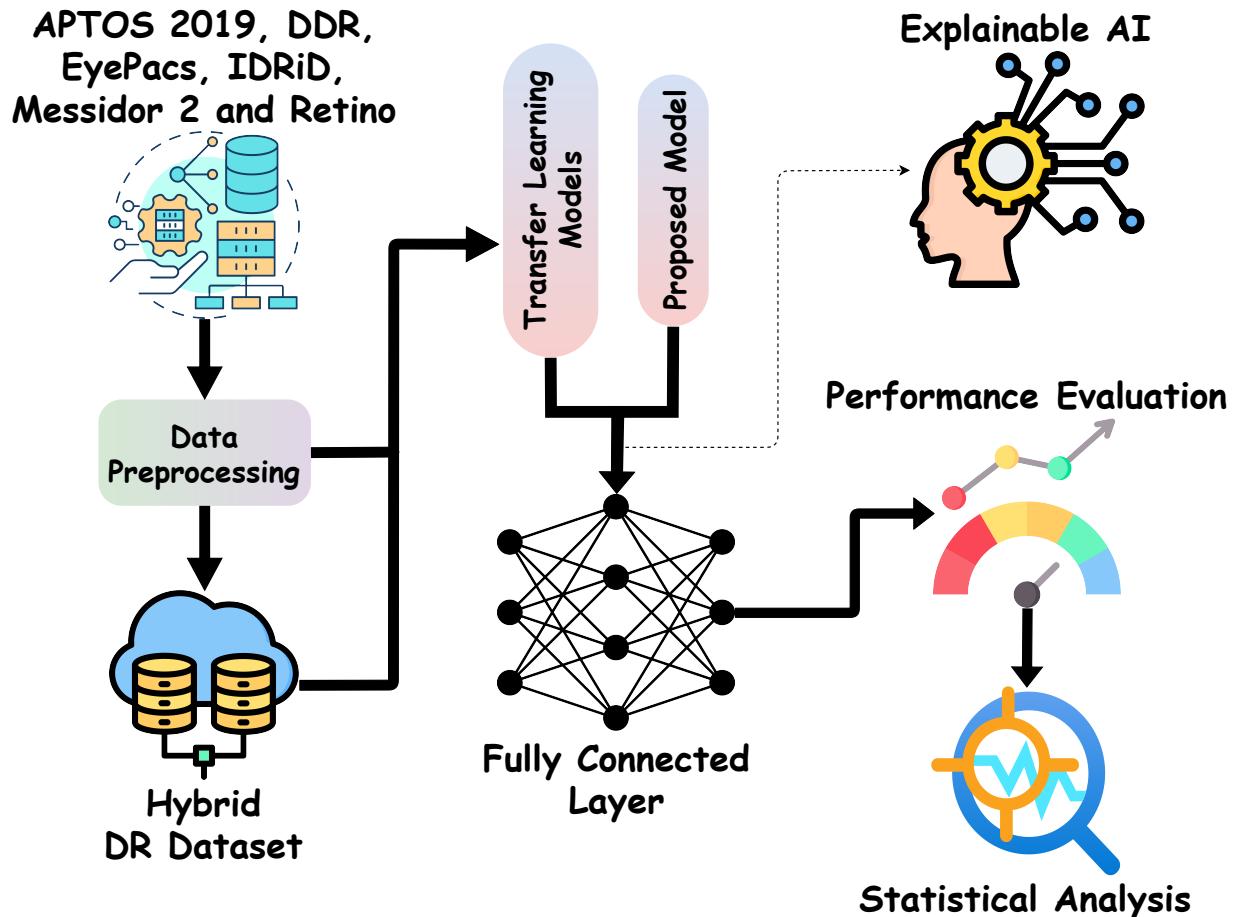


Figure 2: Proposed Methodology

Table 2: Distribution of Diabetic Retinopathy Stages Across Different Datasets

Dataset Name	Mild	Moderate	No DR	Proliferative DR	Severe
APTOS 2019	370	999	1805	295	193
DDR	630	4477	6266	913	236
IDRiD	25	168	168	62	93
Messidor 2	270	347	1017	35	75
Retino	30	480	112	265	505

accessible on Github [50], comprises 12,522 high-resolution retinal images gathered from 147 universities across 23 provinces in China. Furthermore, all photos have been subjected to preprocessing to eliminate black backgrounds, thereby providing cleaner and more dependable inputs for model training.

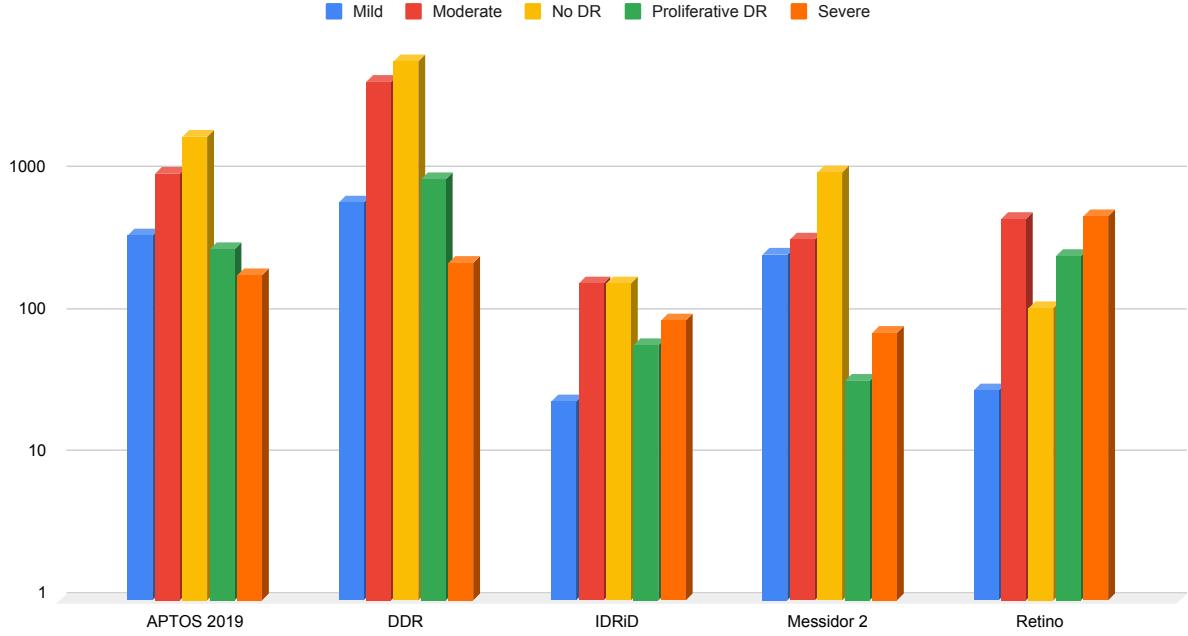


Figure 3: Class Distribution Across Different Datasets Visualized on a Logarithmic Scale

**IDRiD:** The Indian Diabetic Retinopathy Image Dataset (IDRiD) [51], available on IEEE DataPort [52], consists of 516 fundus images, including both disease grading and lesion segmentation annotations. The images were collected from Shiva Netralaya, Uttar Pradesh, India. IDRiD mainly focuses on moderate to severe DR cases, making it particularly valuable for studying disease progression and lesion localization. This dataset provides detailed ground truth information for developing deep learning models that not only classify DR severity but also detect important retinal abnormalities.

**Messidor 2:** The Messidor 2 dataset is an extension of the original Messidor dataset [53], consisting of 1,748 high-resolution macula-centered retinal fundus images from 874 patient examinations, sourced from a website [54]. The dataset is divided into two parts: Messidor-Original, which contains 1,058 images, and Messidor-Extension, which contains 690 images. The images were captured using a Topcon TRC NW6 non-mydriatic fundus camera at Brest University Hospital in France between 2009 and 2010. While the original dataset lacks official DR severity labels, we utilize an annotated version obtained from a third-party Kaggle dataset [55], proposed by Krause et al. [56], providing detailed DR grading for classification tasks. Out of the total 1,748 images available for classification, 4 images lacked labels and were therefore excluded from analysis. Consequently, the classification experiments were conducted utilizing the remaining 1,744 labeled images. Messidor-2 is widely used in medical imaging research due to its high-quality fundus images, making it suitable for developing and testing deep learning models for DR detection.

**RETINO:** The RETINO dataset [57], available on figshare [58], is for automated DR grading using retinal fundus images. It has 1,392 images, each for one patient, from Shanghai Tenth People’s Hospital using Canon non-mydriatic fundus cameras. Images are labeled with the 5 DR levels, from no DR to proliferative DR. The average image size is  $2736 \times 1824$ , and all images were re-checked by an expert ophthalmologist with over 10 years of experience to ensure the labels are correct. This dataset is very useful for evaluating automated screening methods for DR, reducing the dependence on manual diagnosis and early detection.

### 3.2 Hybrid Dataset

The hybrid dataset is a combination of 5 publicly available diabetic retinopathy datasets: APTOS 2019, DDR, IDRiD, Messidor 2 and Retino. Each of these datasets is publicly available and curated independently by different sources, clinical settings and research institutions. They together provide a wide range of diabetic retinopathy conditions and

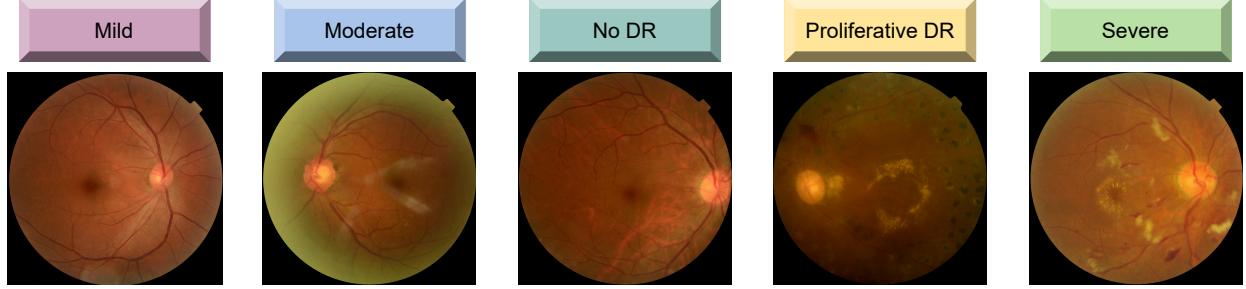


Figure 4: Sample Images of the Dataset

severities. The integration process was to merge these datasets into one. This resulted in a very diverse dataset with huge variability in terms of image quality, resolution, clinical severity and diagnostic characteristics. Figure 5 shows the distribution of Sample Images in the Hybrid Dataset.

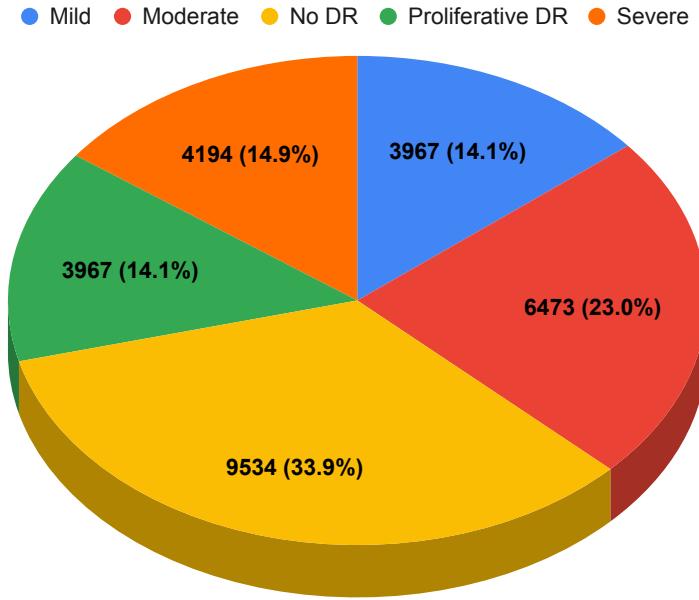


Figure 5: Distribution of Sample Images in the Hybrid Dataset

The hybrid dataset used for training deep learning models in Diabetic Retinopathy (DR) classification has images in multiple severity levels. It has 3,967 images as Mild DR, 4,194 images as Moderate DR, 9,534 images as No DR, 3,967 images as Proliferative DR, and 6,473 images as Severe DR. By combining multiple publicly available datasets, this hybrid dataset covers the clinical and imaging variability that is seen in real world. The images are from different sources so the dataset takes care of the image quality, lighting conditions and retinal structures from different patient population. This diversity is very important for generalizability of deep learning models so they can adapt to different clinical settings and imaging conditions. So this hybrid dataset is a good and reliable base to train, validate and test deep learning models for accurate and robust DR diagnosis across all severity levels.

### 3.3 Data Preprocessing

Image preprocessing plays a vital role in medical image analysis, as the effectiveness of classification models heavily depend on the quality of preprocessing applied to the images. Figure 6 shows the workflow of the pre-processing techniques. In this case, initially, class imbalance is handled through SMOTE oversampling techniques. CLAHE is applied to enhance the contrast of the retinal images. Following contrast enhancement, this preprocessing steps involve implementing image normalization and resizing techniques. Normalization initially adjusts pixel intensity values, generally within the range of 0 to 1, to enhance consistency in visual representation. Images are subsequently

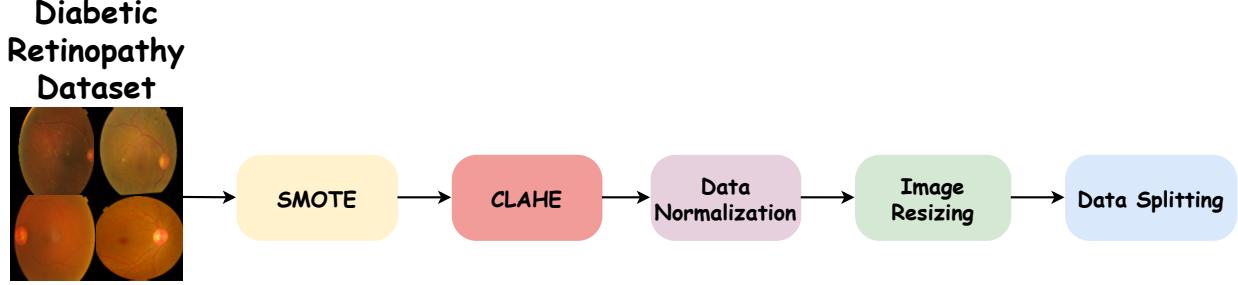


Figure 6: Workflow of the pre-processing techniques

downsized to a specified dimension of  $128 \times 128$  pixels, maintaining uniformity and keeping essential visual qualities. These preprocessing approaches are crucial for diminishing computational complexity and improving performance in subsequent picture analysis or model training activities. Finally, the dataset is partitioned into training, validation and test sets in an 80:10:10 ratio, designating 80% for training, 10% for validation and 10% for testing [59].

### 3.3.1 Synthetic Minority Over-Sampling Technique (SMOTE)

Synthetic Minority Over-sampling Technique also known as SMOTE is a resampling method that addresses class imbalance by generating synthetic data points rather than duplicating existing samples. This was initially inspired by handwritten character recognition where rotation and skewing was applied to create more training data [60]. Unlike those domain specific augmentations, SMOTE generates new synthetic samples in the feature space by interpolating between real data points to create a balanced dataset and expand the decision boundary of the minority class. This technique helps expand the feature space representation of the minority class, preventing overfitting and improving the generalization capability of the classifier. The synthetic instance  $x_{\text{new}}$  is generated using the following interpolation formula:

$$x_{\text{new}} = x_i + \delta \cdot (x_{zi} - x_i) \quad [60] \quad (1)$$

where  $x_i$  is a minority class sample,  $x_{zi}$  is one of its  $k$ -nearest neighbors, and  $\delta \in [0, 1]$  is a random number. To generalize this process when generating  $N$  synthetic samples per minority class instance, the interpolation can be repeated across different neighbors and values of  $\delta$ :

$$x_{\text{new}}^{(n)} = x_i + \delta_n \cdot (x_{zi}^{(n)} - x_i), \quad n = 1, 2, \dots, N \quad [60] \quad (2)$$

SMOTE can also be applied to image datasets particularly in medical imaging tasks. Since images are high dimensional data, applying SMOTE directly on raw pixel values can distort meaningful medical patterns. Instead, the process involves first converting images to feature vectors using a deep learning based embedding model. These feature vectors represent the underlying structure and characteristics of the images so we can apply SMOTE in a more meaningful way. After extracting these features, SMOTE identifies  $k$ -nearest neighbors of each minority class image and generates synthetic feature vectors through interpolation. Then reshape the synthetic feature vectors back to image to use for training. By ensuring the synthetic samples have clinically relevant features, this technique helps improve classification accuracy while minimizing the effect of class imbalance in the dataset [61, 62].

This Algorithm 1 is applied to image datasets by first converting images into feature vectors that represent the patterns and feature distributions of the images. Since direct pixel level interpolation can distort images, SMOTE operates in feature space where each image is represented as a high dimensional numerical vector. After finding  $k$ -nearest neighbors for each minority class image, a random neighbor is selected and a new synthetic feature vector is generated by interpolating between the original vector and its neighbor with a random gap value  $\lambda$ . The synthetic feature vector is then reshaped back into image to make sure the newly generated images preserve the meaningful medical features like microaneurysms, hemorrhages and exudates. Finally, the synthetic images are added to the original dataset to improve class balance, reduce bias and enhance the model's ability to classify diabetic retinopathy cases [63]. Following this process, SMOTE has been applied across the entire five datasets. In every dataset, the oversampling process is conducted by computing the mean value among the five stages, ensuring representative distribution of DR stages while mitigating the risks of class bias. Table 3 and Figure 7 illustrate the distribution of different stages across various datasets following the application of SMOTE, with a logarithmic scale applied to improve the clarity of class imbalance visualization.

**Algorithm 1** SMOTE-Based Image Oversampling

---

```

1: Input: Minority class samples  $T$ , oversampling percentage  $N$ , nearest neighbors  $k$ 
2: if  $N < 100\%$  then
3:   Randomly select a subset of  $T$  samples
4: end if
5: Compute number of synthetic samples:
6:  $num\_synthetic = (N/100) * len(T)$ 
7: for each sample  $x_i$  do
8:   Compute Euclidean distance to k-nearest neighbors:
9:    $d(x_i, x_{NN}) = \sqrt{\sum(x_i - x_{NN})^2}$ 
10: end for
11: for each sample  $x_i$  do
12:   Select a random neighbor  $x_{NN}$ 
13:   Compute difference:
14:    $diff = x_{NN} - x_i$ 
15:   Generate a random gap  $\lambda$  between 0 and 1:
16:    $lambda = random(0, 1)$ 
17:   Compute synthetic sample:
18:    $x_{syn} = x_i + lambda * diff$ 
19:   Store  $x_{syn}$  in dataset
20: end for
21: Repeat Until Balance is Achieved: Continue generating synthetic samples until  $N\%$  oversampling is met.
22: Return New Balanced Dataset: Merge synthetic samples with the original dataset.

```

---

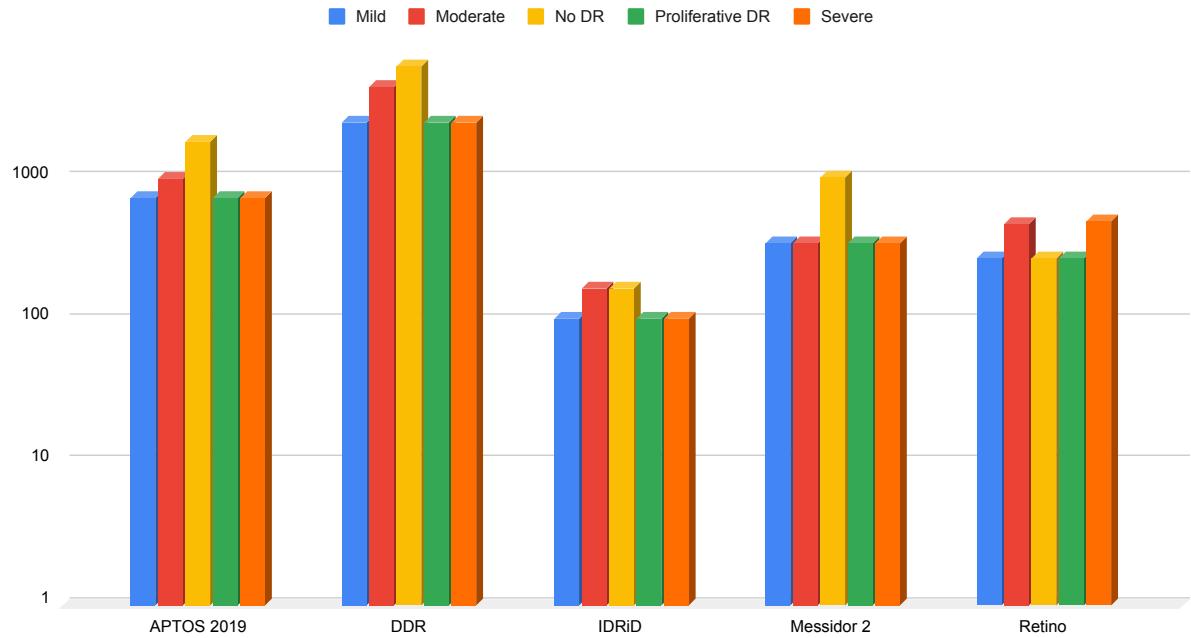


Figure 7: Comparison of Different Datasets by Categories after SMOTE using Log Scale Plot

Table 3: Distribution of Diabetic Retinopathy Stages Across Different Datasets after SMOTE

Dataset Name	Mild	Moderate	No DR	Proliferative DR	Severe
APOTOS 2019	733	999	1805	733	733
DDR	2504	4477	6266	2504	2504
EyePACS	3025	5292	5810	3025	3025
IDRiD	103	168	168	103	103
Messidor 2	349	347	1017	349	349
Retino	278	480	278	278	505

### 3.3.2 Contrast Limited Adaptive Histogram Equalization (CLAHE)

Contrast Limited Adaptive Histogram Equalization (CLAHE) [64] is an improved version of Adaptive Histogram Equalization (AHE) [65] which enhances contrast without amplifying noise in uniform areas. Unlike global histogram equalization which applies uniform contrast to the whole image, AHE works on local regions, so it's more suitable for images with non uniform lighting. But AHE tends to over enhance noise in uniform areas and produces artifacts. CLAHE overcomes this by introducing a clip limit to prevent excessive contrast changes. The CLAHE process starts by dividing the image into non overlapping tiles (sub-regions), usually  $8 \times 8$  for a  $512 \times 512$  image, to ensure local contrast enhancement. Each tile does histogram computation, where pixel intensity distribution is computed independently. The contrast clip limit is then applied, which is a multiple of the average histogram contents, so no histogram bin exceeds a specified threshold. If there are excess pixels beyond the clip limit, they are redistributed across the whole histogram, keeping the total pixel count balanced. Then a Cumulative Distribution Function (CDF) is computed to get intensity mapping. Since equalizing each tile separately produces artificial boundaries, CLAHE applies bilinear interpolation to blend adjacent tiles smoothly, so there's no gap between regions.

To formalize the mapping function, for each region, the grayscale mapping is obtained by estimating the histogram density function. The mapping is calculated as:

$$f_{i,j}(n) = \frac{(N-1)}{M} \sum_{k=0}^n h_{i,j}(k) \quad (3)$$

where  $M$  is the total number of pixels in the region,  $N$  is the total number of grayscale levels, and  $h_{i,j}(n)$  is the histogram count for intensity level  $n$  [66]. To limit the contrast, the clip factor  $\beta$  is introduced:

$$\beta = \frac{M}{N} \left( 1 + \frac{\alpha}{100} (s_{\max} - 1) \right) \quad (4)$$

where  $\alpha$  is the clip factor percentage, and  $s_{\max}$  controls the maximum allowable histogram slope [66].

After CLAHE modifies the histogram it redistributes excess histogram counts across lower bins to balance the transformation. The new grayscale value for a pixel is determined by the 4 nearest neighbours using the following equation:

$$p_{\text{new}} = \frac{s}{r+s} \left( \frac{y}{x+y} f_{i-1,j-1}(p_{\text{old}}) + \frac{x}{x+y} f_{i,j-1}(p_{\text{old}}) \right) + \frac{r}{r+s} \left( \frac{y}{x+y} f_{i-1,j}(p_{\text{old}}) + \frac{x}{x+y} f_{i,j}(p_{\text{old}}) \right) \quad (5)$$

where  $x, y, r, s$  represent the relative distances between the pixel and the centers of the neighboring regions [66].

The redistribution algorithm ensures no pixel value overshoots so no artificial contrast spikes. This is repeated until all excess counts are evenly spread across the histogram bins. The final step is to merge the enhanced tiles back into a full image so the contrast enhancement is globally consistent.

CLAHE is used in medical imaging where fine details such as microaneurysms, hemorrhages, exudates in diabetic retinopathy detection need to be preserved while making them visible. By applying CLAHE enhances critical medical features without blowing up bright regions or amplifying noise making it a useful tool in computer aided diagnosis (CAD) and medical image analysis. Here, this approach converts the image to the LAB color space, processing only the

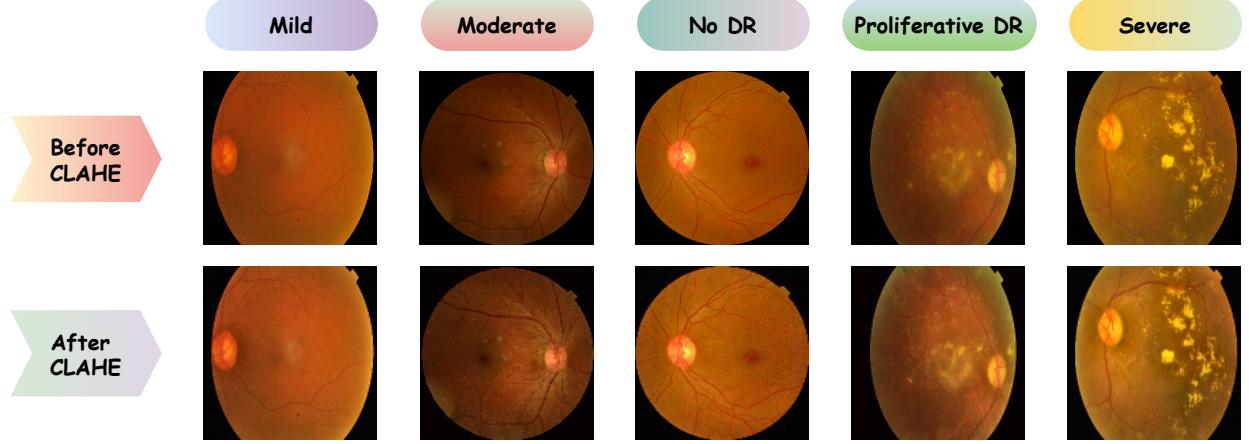


Figure 8: Comparison of Retinal Images Before and After CLAHE Enhancement.

L (luminance) channel to maintain color integrity. The CLAHE algorithm is utilized on the L channel with a clipping limit of 0.5 and a tile grid size of (8,8), enabling local contrast improvement while preventing excessive brightness fluctuations. Following enhancement, the modified L channel is re-integrated with the A and B channels, and the image is reverted to the BGR color space. Figure 8 presents a visual comparison of retinal images before and after the application of CLAHE.

### 3.4 Transfer Learning

Several popular convolutional neural network architectures were used in this study because of their good performance in image classification tasks. These are VGG16, VGG19, ResNet50V2, MobileNetV2, Xception [67, 68]. VGG16 and VGG19 mainly use  $3 \times 3$  convolutional layers and are commonly used for feature extraction, VGG19 has additional convolutional layers to detect more complex features. ResNet50V2 has residual connections to mitigate the vanishing gradient problem and to improve training stability and speed. Xception uses depthwise separable convolutions to further improve computational efficiency. Finally, MobileNetV2 is optimized for environments with limited computational resources, uses inverted residual structures and bottleneck layers to balance model performance and computational cost.

The convolutional neural networks (CNNs) architectures VGG16, VGG19, ResNet50V2, MobileNetV2, Xception are used with their initial weights from ImageNet pre-trained [69]. These models are fine-tuned with the Diabetic Retinopathy dataset, so the pre-learned features are adapted to this specific task. The modified model structure has a flattening layer, dense layers of 1024 and 512 neurons, dropout layers to prevent overfitting and a softmax output layer for classification. Fine-tuning these pre-trained models on the DR disease training dataset improves model performance and reduces the overall training time by leveraging the learned representations from ImageNet initialization. The final dense layer with softmax activation function classifies the input data into 5 classes.

### 3.5 Feature Fusion Method

Feature Fusion is a key technique in pattern recognition applications as it expands the feature space and improves classification accuracy and object detection [70]. The combination of multiple feature maps results in a denser and more representative feature space, which improves the discriminative power of the model. However, concatenating feature vectors directly can introduce redundant features and increase computational complexity [71]. To address this issue a novel feature concatenation approach is used in our method. Figure 9 shows the model architecture of our proposed VR-FuseNet model. In our study two pre-trained deep convolutional neural networks (CNNs) VGG19 and ResNet50V2 are used for feature extraction. Let the extracted deep CNN feature vectors from VGG19 and ResNet50V2 be  $m_1$  and  $m_2$  respectively. The dimensions of these feature vectors are  $q \times r$  and  $q \times s$  where  $q$  is the number of images and  $r$  and  $s$  are the respective lengths of the feature attributes. In our case, both  $r$  and  $s$  are 512 and 2048, so the feature dimensions are  $q \times 512$  and  $q \times 2048$  for both networks. To pad feature space dimensions before concatenation mean is computed and added with it. The transformation of feature vectors into time series can be expressed as follows :

$$z_1 = 1^T m_1 \quad (6)$$

$$z_2 = 2^T m_2 \quad (7)$$

The maximum covariance of  $m_1$  and  $m_2$  is defined as:

$$\hat{f} = \text{Cov}[z_1, z_2] \quad (8)$$

$$\hat{f} = \text{Cov}(1^T m_1, 1^T m_1) \quad (9)$$

$$\hat{f} = \frac{1}{n-1} \text{Cov}(1^T m_1, 1^T m_1) \quad (10)$$

The covariance of the final concatenated feature space is given by:

$$\hat{f} = \frac{1}{(F_{1,2})^2} 2 \quad (11)$$

where:

$$F_{1,2} = \frac{1}{n-1} (m_1 m_2^T) \quad (12)$$

Here  $F_{1,2}$  is the covariance between the two feature maps with features indexed by  $i$  and  $j$ . The final concatenated feature space using parallel maximum covariance strategy gives a feature vector of dimension  $q \times 2560$ . This dense representation helps in feature selection and avoid feature redundancy. By using this feature fusion approach our model benefits from the strengths of both VGG19 and ResNet50V2 feature extractors and gets better performance and efficiency.

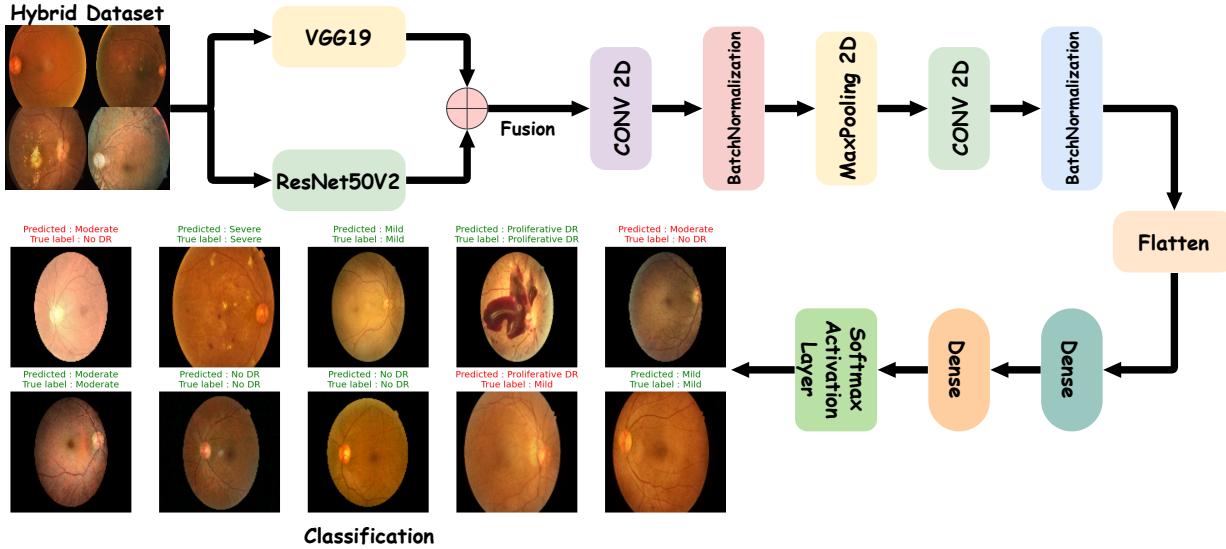


Figure 9: Model Architecture of VR-FuseNet Model

After analysis we select VGG19 and ResNet50V2 for feature fusion as they have complementary strengths and perform better than other models. VGG19 with its deep yet simple architecture of stacked convolutional layers with small 3x3 filters is good at capturing fine grained local features. ResNet50V2 with its residual connections allows training of deeper networks by mitigating the vanishing gradient problem and is good at capturing abstract high level features. By fusing features from both models we leverage the local detail extraction capability of VGG19 and the global contextual understanding of ResNet50V2 and create a more robust and comprehensive feature representation. This fusion outperforms individual models or other combinations as it integrates both depth and residual learning and helps the model to generalize across diverse retinal images and improve classification accuracy.

The VR-FuseNet Model is a hybrid deep learning model which combines two convolutional neural networks (CNNs) VGG19 and ResNet50V2 to leverage their complementary feature extraction capabilities. ResNet50V2 with its deep residual learning captures high level abstract features and VGG19 captures fine grained spatial features in retinal images. The input hybrid dataset images are passed through both the networks separately, each network extracts unique and

meaningful features. The model then uses fusion mechanism to combine the features from both the networks so that the best of both the architectures is used for better classification.

After fusion, the fused feature maps undergo further refinement through additional Conv2D layers, batch normalization, and max-pooling operations to enhance spatial hierarchies. Then the model passes the combined features through fully connected (dense) layers which refines the learned representations. Specifically it has a flatten layer followed by 2 dense layers of 256 and 64 neurons with dropout layers to prevent overfitting so the model can generalize to different datasets. The last dense layer of the model uses softmax activation function which assigns probabilities to different DR severity levels and classifies the input data into 5 classes. This way the model can distinguish different stages of DR, mild, moderate, severe, proliferative DR and normal cases with no DR. By combining two powerful CNN architectures, VR-FuseNet improves feature extraction and hence better generalization and higher accuracy. The model can learn from diverse set of retinal images and is a promising solution for automated DR diagnosis which can help ophthalmologists in early detection and treatment planning.

## 4 Result Analysis

### 4.1 Evaluation Metrics

To evaluate the proposed VR-FuseNet model for diabetic retinopathy classification, we used a set of well-known performance metrics. These metrics give insights into different aspects of the model's predictability, such as overall correctness, true cases detection and balance between precision and recall. Below are the mathematical definitions and explanations of each metric used in this study.

- **Accuracy:** The proportion of true results (both true positives and true negatives) in the population, computed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$

- **Precision:** The ratio of true positive to total predicted positives, calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (14)$$

- **Recall:** Also known as true positive rate, the ratio of true positive to all actual positives, expressed as:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (15)$$

- **F1-Score:** The harmonic mean of precision and recall, formulated as:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (16)$$

- **Receiver Operating Characteristic Curve:** ROC curve is a fundamental tool to evaluate the performance of classification models, especially in binary classification, such as presence or absence of a disease. It provides a graphical representation by plotting the True Positive Rate (TPR or Recall) against the False Positive Rate (FPR) at different threshold settings [72]. The FPR is calculated as:

$$\text{FPR} = \frac{FP}{FP + TN} \quad (17)$$

Originally developed for military radar signal detection in 1940s, ROC curve has become a standard metric in machine learning and medical diagnostics. It helps to balance sensitivity (TPR) and specificity (1 - FPR), giving insights into the model's discriminative ability. The Area Under the ROC Curve (AUC) is a quantitative measure of this performance. AUC values range from 0.5 (random guessing) to 1.0 (perfect classification), with higher values indicating better model reliability and accuracy. In this study, AUC-ROC is used to evaluate the proposed model's ability to detect diabetic retinopathy. ROC curves closer to the upper left corner of the plot space indicates better classification performance, the model is able to correctly identify both positive and negative cases at different thresholds [73].

## 4.2 Comparative Analysis Across Datasets

Results of several deep learning models on 5 publicly available diabetic retinopathy (DR) datasets—APTOS 2019, DDR, IDRiD, Messidor 2, and Retino—are shown in Table 4. Models are evaluated in terms of accuracy, precision, recall and F1-score. Models are trained with a learning rate of 1e-5 to ensure convergence during optimization. Different datasets require different batch sizes; 16, 32 and 128 are used based on computational constraints and dataset complexity. 50 to 110 epochs are used to avoid overfitting and learn features.

Table 4: Performance Metrics for Different Models

Dataset	Model	Batch Size	Epoch	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
APTOS 2019	VGG16	32	70	95.400	95.628	94.511	95.019
	<b>VGG19</b>		80	<b>96.000</b>	<b>96.135</b>	<b>95.164</b>	<b>95.620</b>
	ResNet50V2		50	92.400	92.173	91.348	91.678
	MobileNetV2		70	92.200	91.562	90.489	90.870
	Xception			92.000	91.605	90.720	91.076
DDR	<b>VGG16</b>	128	70	<b>90.630</b>	<b>93.023</b>	<b>93.227</b>	<b>93.122</b>
	VGG19			90.027	92.630	93.015	92.727
	ResNet50V2			87.726	90.567	90.365	90.354
	MobileNetV2			90.301	92.742	92.051	92.264
	Xception			89.151	91.567	91.387	91.441
IDRiD	VGG16	16	70	79.688	80.770	79.692	79.844
	<b>VGG19</b>			<b>81.250</b>	<b>84.902</b>	<b>82.082</b>	<b>82.924</b>
	ResNet50V2			76.562	78.452	77.042	77.600
	MobileNetV2		100	78.125	82.094	78.581	79.340
	Xception		50	73.438	72.873	72.686	72.655
Messidor 2	VGG16	16	60	90.871	87.272	87.561	87.383
	<b>VGG19</b>			<b>91.286</b>	<b>88.038</b>	<b>88.190</b>	<b>88.010</b>
	ResNet50V2			88.797	84.869	84.737	84.561
	MobileNetV2		90	87.552	82.655	83.295	82.298
	Xception		60	87.344	82.847	83.128	82.664
Retino	VGG16	16	70	92.265	94.903	94.545	94.607
	<b>VGG19</b>			<b>93.923</b>	<b>95.698</b>	<b>95.157</b>	<b>95.357</b>
	ResNet50V2		110	87.845	90.928	90.802	90.834
	MobileNetV2			87.845	91.078	91.048	91.025
	Xception		70	88.398	91.481	91.682	91.554

The APTOS 2019 dataset shows strong performance across all models, with VGG19 achieving the highest accuracy at 96.000%. VGG16 follows closely at 95.400%. The ResNet50V2 model achieved 92.400%, while MobileNetV2 and Xception models showed slightly lower accuracies at 92.200% and 92.000% respectively. This dataset appears to be well-suited for the VGG architecture, with both VGG variants outperforming the other models by a notable margin of approximately 3-4%. For the DDR dataset, VGG16 leads with an accuracy of 90.630%, followed by VGG19 at 90.027%. MobileNetV2 performs well at 90.301%, with Xception slightly behind at 89.151%. ResNet50V2 shows the lowest accuracy at 87.726%. The performance gap between models is less pronounced on this dataset compared to APTOS 2019, suggesting that the DDR dataset presents challenges that are handled somewhat similarly by the different architectures.

The IDRiD dataset shows generally lower accuracy scores across all models compared to other datasets. VGG19 achieves the highest accuracy at 81.250%, followed by VGG16 at 79.688%. MobileNetV2 reaches 78.125%, while ResNet50V2 and Xception trail at 76.562% and 73.438% respectively. The consistent drop in performance across all models suggests that this dataset presents more challenging classification problems compared to the others. On the

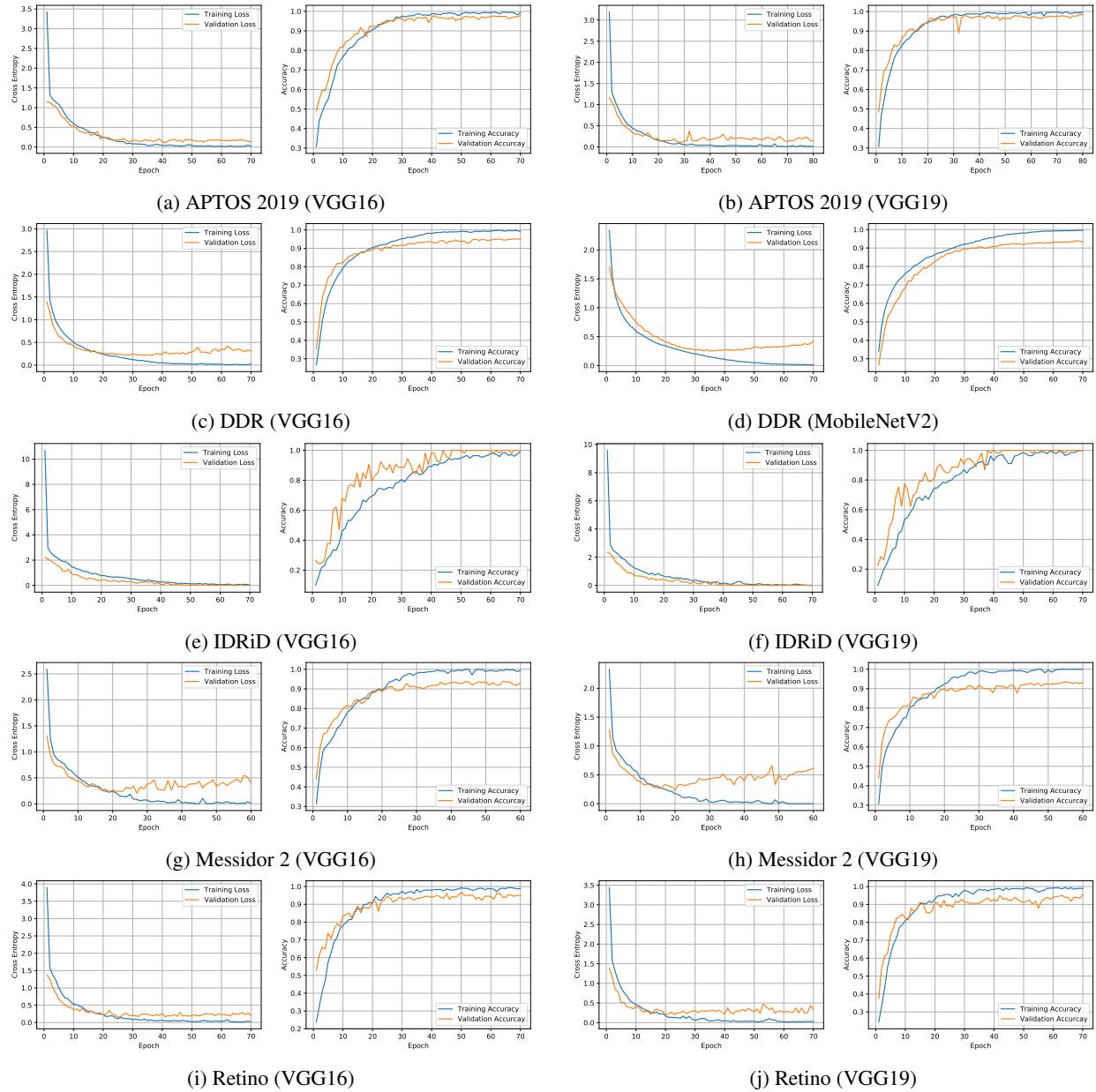


Figure 10: Individual Dataset Loss and Accuracy Curve

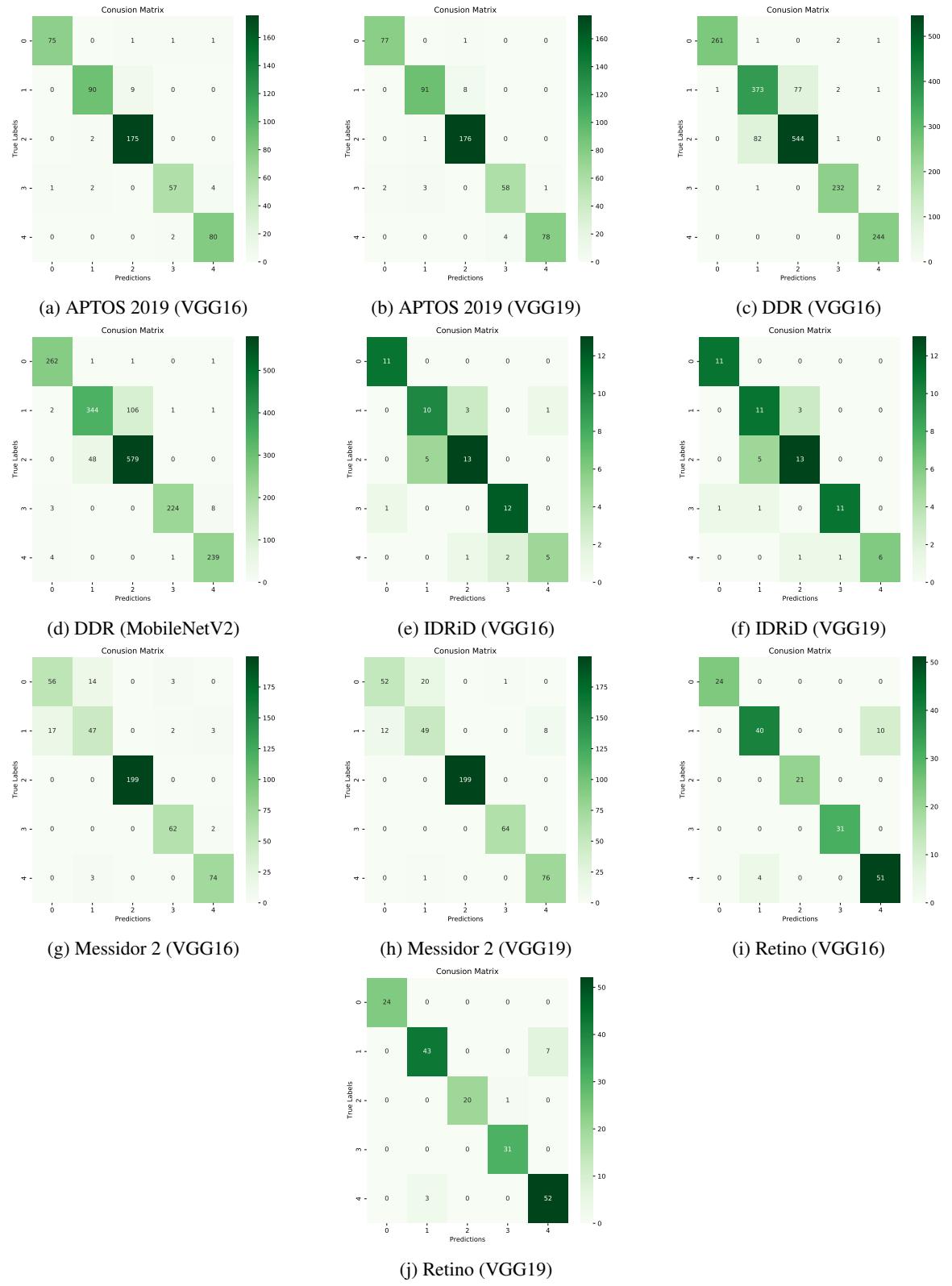


Figure 11: Individual Dataset Confusion Matrix

Messidor 2 dataset, VGG19 leads with 91.286% accuracy, followed by VGG16 at 90.871%. The remaining models show a pronounced gap, with ResNet50V2 at 88.797%, MobileNetV2 at 87.552%, and Xception at 87.344%. This dataset again demonstrates the strength of the VGG architecture, particularly VGG19, which outperforms other architectures by approximately 2-4%. Retino dataset shows the best overall performance, VGG19 is the highest with 93.923%, followed by VGG16 with 92.265%. Xception is 88.398%, and both ResNet50V2 and MobileNetV2 are 87.845%. VGG models are way ahead of the others by 4-6%.

Across all datasets, VGG19 is the best, followed by VGG16. The newer architectures like ResNet50V2, MobileNetV2, Xception are not as good as expected despite their advantages in other domains. Figure 10 and Figure 11 are loss and accuracy curves and confusion matrices for training and validation data across multiple datasets used for diabetic retinopathy classification. Each row is for a specific dataset, APTOS 2019, DDR, IDRiD, Messidor 2, Retino, with best two models for that dataset. Blue curve is training, orange curve is validation. These curves show how well the models converge during training, difference in learning stability, overfitting and generalization across datasets. At the same time, each confusion matrix shows the number of correct and incorrect instances across different diabetic retinopathy severity levels. This helps to compare the performance of different models on different datasets and understand their strengths and weaknesses in detecting diabetic retinopathy.

#### 4.3 Impact of Hybrid dataset on Model Performance

Analyzing the performance results presented in Table 5 for the hybrid dataset which integrates all the individual datasets reveals valuable insights into the generalizability and robustness of deep learning models in diagnosing diabetic retinopathy using a fixed batch size of 128, a learning rate of 1e-5 and 60 training epochs across diverse imaging conditions. The hybrid dataset combines different imaging standards, resolutions and qualities from multiple sources to test model adaptability and overcome dataset specific biases and limitations. Training models on a comprehensive hybrid dataset encourages the development of generalized features and the resulting models are highly applicable to real world clinical scenarios. When looking at individual models for the hybrid dataset, VGG19 is performing very strong and consistent, with the highest accuracy of 90.935%. VGG16 is close second with 90.473% reflecting the strength of VGG architectures in learning features across different datasets. ResNet50V2 is slightly lower at 90.153% indicating it has strong generalization but slightly less than VGG models, there is room to tune feature extraction. MobileNetV2 although valued for its efficiency and lightweight is moderate at 90.082% indicating challenges in capturing diverse and complex features in the large hybrid dataset. Xception are lower at 88.482% showing their limitations in generalizing across the dataset's heterogeneous imaging conditions.

In contrast, the proposed VR-FuseNet model, a combination of VGG19 and ResNet50V2, performs better on the hybrid dataset with an accuracy of 91.824%, precision of 92.612%, recall of 92.233%, F1-score of 92.392% and AUC of 98.749%. VR-FuseNet combines the deep hierarchical representation of ResNet50V2 with the detailed and fine-grained spatial features of VGG19. This combination enables VR-FuseNet to handle the diversity in the hybrid dataset and outperforms individual models in all metrics. The combination in VR-FuseNet addresses the complexity and heterogeneity of the hybrid dataset and leverages the strengths of each architecture to demonstrate great generalization. Overall the hybrid dataset and VR-FuseNet model is an approach to highly accurate, reliable and generalized diabetic retinopathy detection, suitable and applicable in different imaging conditions and patient population.

Table 5: Performance Metrics for Hybrid Dataset Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
VGG16	90.473	90.912	91.632	91.189	98.425
VGG19	90.935	91.439	91.960	91.677	98.714
ResNet50V2	90.153	91.215	90.840	90.993	98.354
MobileNetV2	90.082	90.472	90.798	90.605	98.602
Xception	88.482	89.463	89.201	89.253	97.700
<b>VR-FuseNet (Proposed)</b>	<b>91.824</b>	<b>92.612</b>	<b>92.233</b>	<b>92.392</b>	<b>98.749</b>

Figure 12 and Figure 13 Displays the training and validation accuracy and loss curves and confusion matrices for various deep learning models. Each subplot shows two graphs: one for loss (indicating how well the model is minimizing errors) and one for accuracy (indicating model performance over training epochs). The blue curves represent training data, while the orange curves represent validation data. The comparison of models reveals differences in learning rate, consistency, and ability to generalize effectively. Some models show fast convergence, while others require more

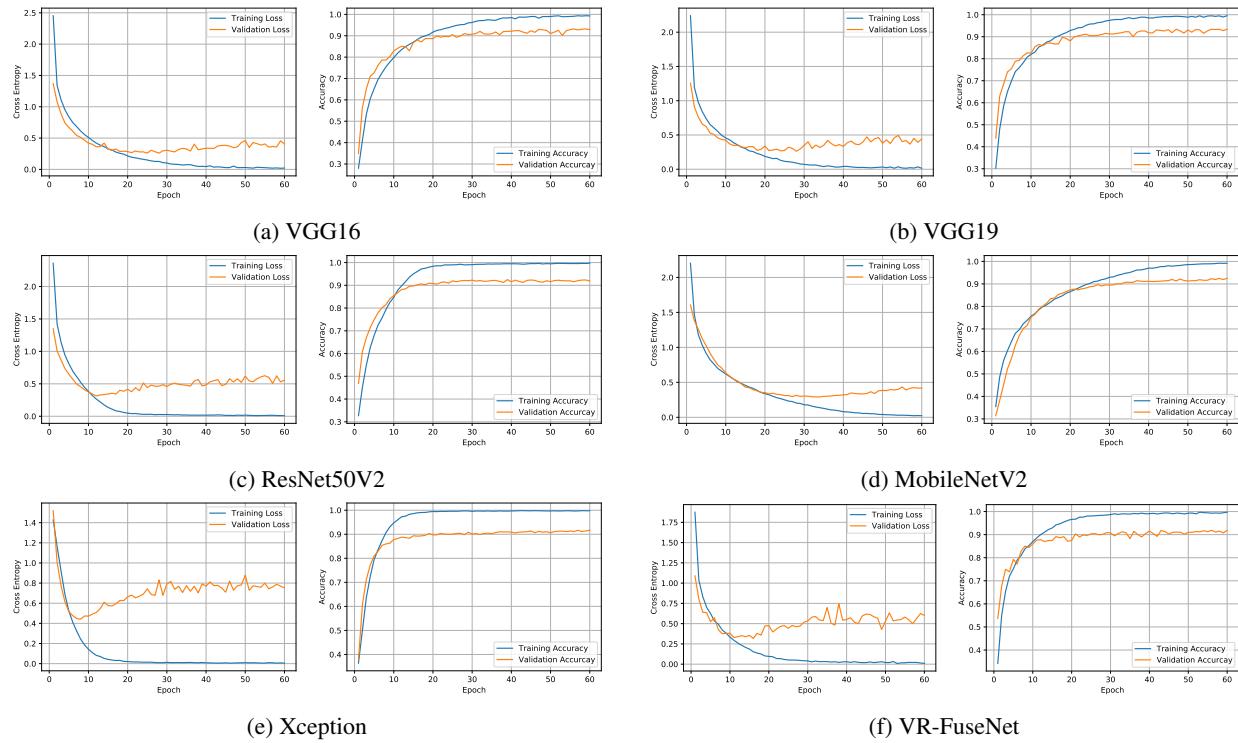


Figure 12: Hybrid Dataset Loss vs Accuracy Graph

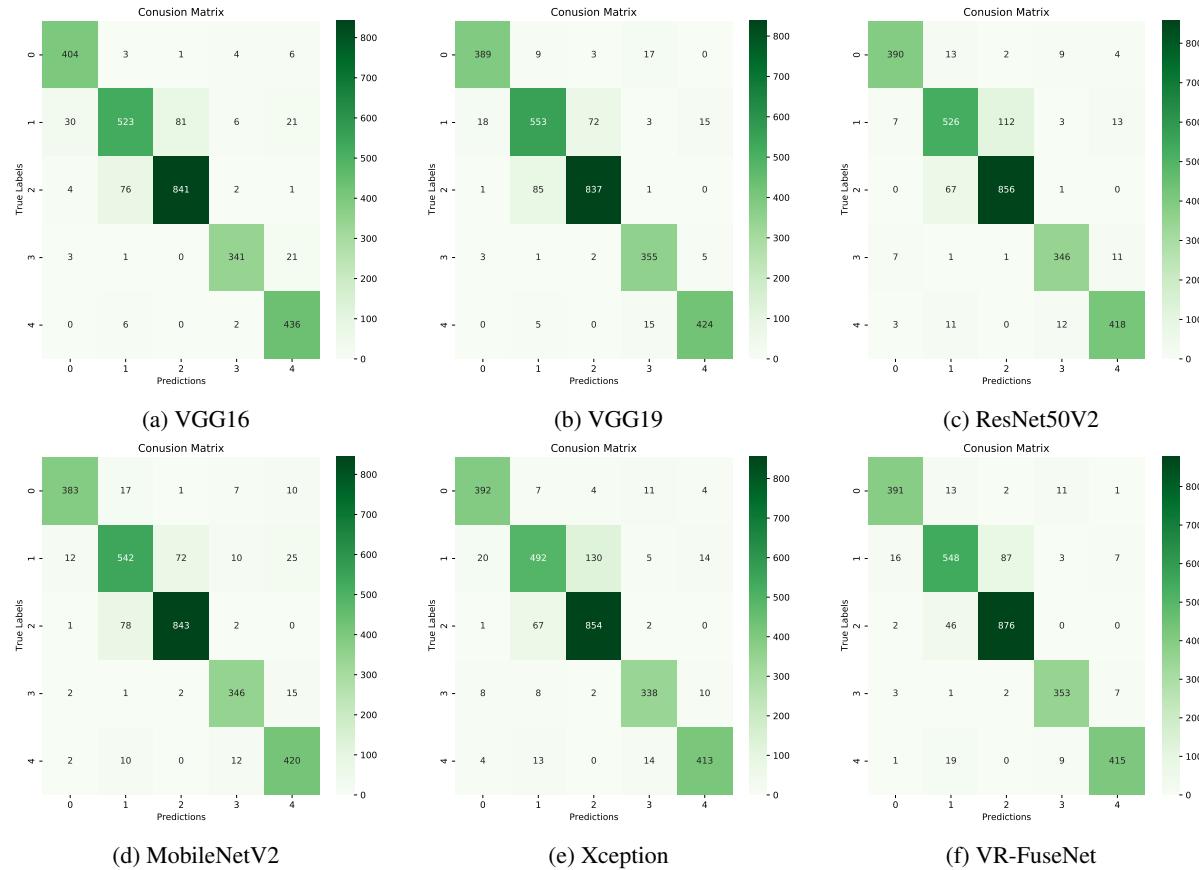


Figure 13: Hybrid Dataset Confusion Matrix

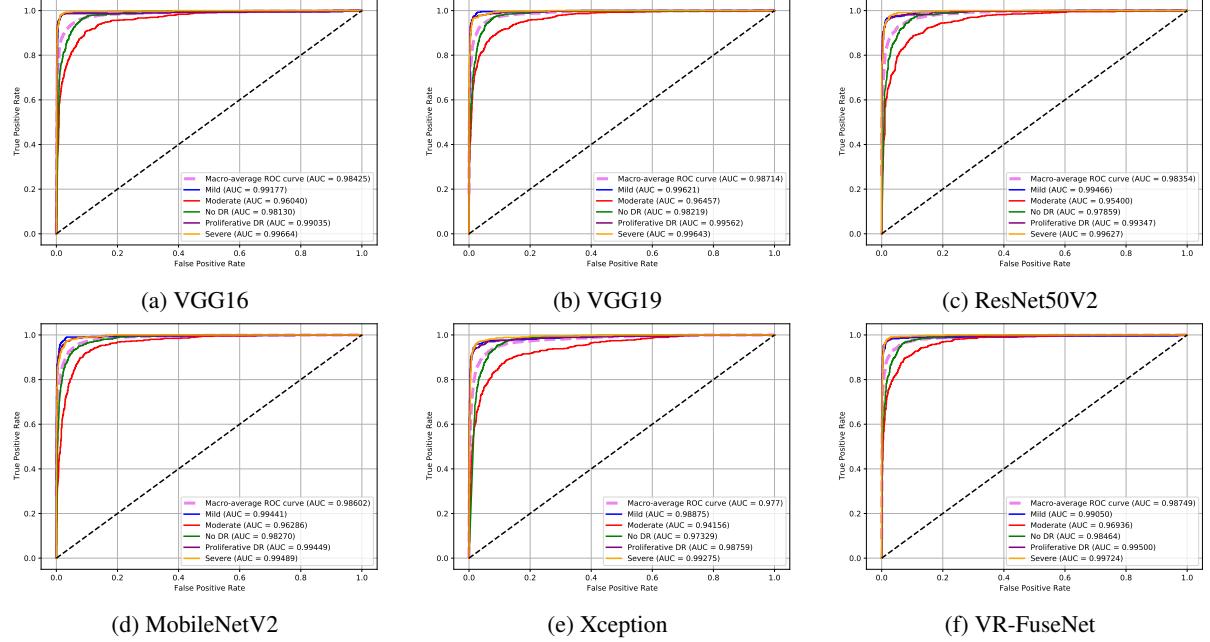


Figure 14: Hybrid Dataset ROC Curve

epochs to stabilize. Certain models exhibit overfitting, where the training accuracy is significantly higher than validation accuracy, accompanied by fluctuations in validation loss. On the other hand, each confusion matrix visualizes the model’s performance by displaying the number of correct and misclassified instances. At the bottom, the proposed VR-FuseNet model is evaluated using the hybrid dataset. This comparison provides insights into how different models classify various severity levels and highlights their advantages and limitations in diabetic retinopathy detection.

Also Figure 14 shows Receiver Operating Characteristic (ROC) curves for different deep learning models trained on the hybrid dataset. Each ROC curve plots true positive rate (sensitivity) vs false positive rate (1-specificity) to evaluate the performance of different models. The dashed diagonal line is random classification, the closer to top left corner is better. The AUC (Area Under the Curve) values displayed in each figure is the performance, higher is better. The proposed VR-FuseNet aims to improve the accuracy. By comparing different models we can see how well they can differentiate between different diabetic retinopathy severity levels and what are their strengths and weaknesses.

## 5 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) is a research area that makes the decision making process of artificial intelligence (AI) systems, especially deep learning models, more interpretable and transparent to users. Traditional deep neural networks (DNNs) are **black boxes**, they make good predictions but provide no insight into how those decisions are made, which is a problem in high stakes applications like healthcare, finance and autonomous systems [74, 75]. XAI techniques like Gradient-weighted Class Activation Mapping (Grad-CAM) [76] help to address this by generating visual explanations that highlight the most relevant input features that contribute to the model’s output. In our work we have used five gradient based XAI methods to increase model explainability and get a deeper understanding of the decision making process. By providing human understandable insights XAI makes the model more transparent, helps in debugging and ensures fairness so stakeholders can trust and refine AI driven systems.

### 5.1 Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM is a popular explainability technique to give visual explanations for the decision making process of deep convolutional neural networks (CNNs). Grad-CAM achieves this by using the gradient information flowing into the last convolutional layer to generate a class-discriminative localization map, highlighting the regions in the input image that contribute most to the model’s prediction. Unlike earlier visualization methods like Class Activation Mapping (CAM) which required architectural changes and retraining, Grad-CAM can be applied to all CNN architectures without any changes [76]. The technique computes the gradients of the target class score  $y^c$  (before the softmax) w.r.t the feature

maps  $A^k$  of the last convolutional layer to capture spatial importance. The gradient based importance weights  $\alpha_k^c$  for each feature map  $A^k$  is computed using global average pooling of the gradients as follows:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (18)$$

where  $Z$  is the total number of pixels in the feature map. These weights represent the importance of each feature map to the model's decision for class  $c$ . The final Grad-CAM heatmap  $L_{\text{Grad-CAM}}^c$  is obtained by combining the activation maps weighted by these scores and then applying ReLU to retain only positive contributions:

$$L_{\text{Grad-CAM}}^c = \text{ReLU} \left( \sum_k \alpha_k^c A^k \right) \quad (19)$$

By applying ReLU, the method focuses on features that have a positive impact on the class of interest and ignore the negative ones. The resulting heatmap shows the most important regions in the input image for the given prediction. Grad-CAM is very useful in medical imaging such as diabetic retinopathy detection as it allows clinicians to interpret the model predictions by visualizing the regions of interest in retinal fundus images, to identify the pathological features such as microaneurysms, hemorrhages and exudates. To make it even more interpretable, Visual Question Answering (VQA) is integrated in Explainable AI (XAI) with Grad-CAM so humans can interact with AI models like they do with each other. VQA adds an extra layer of explainability by allowing users to ask natural language questions about model predictions and get answers with visual explanations. In medical imaging, VQA models use a mix of image features from CNNs and text features from transformers to provide meaningful answers. Given an image  $I$  and a text question  $T$ , the VQA model extracts image features  $V$  and text embeddings  $Q$  as:

$$V = f_{\text{img}}(I), \quad Q = f_{\text{text}}(T) \quad (20)$$

where  $f_{\text{img}}$  is a vision model (e.g., ResNet, Vision Transformers) and  $f_{\text{text}}$  is a text model (e.g., BERT, LLaMA). The answer  $A$  is predicted as:

$$A = \arg \max p(A|V, Q) \quad (21)$$

To integrate VQA with Grad-CAM, the model produces a heatmap-based explanation along with the text answer so the answer is backed by a highlighted region of interest in the image. For example, if a doctor asks, *What features indicate diabetic retinopathy?*, the system not only gives a text answer but also overlays a Grad-CAM heatmap over the retinal image to show lesions, hemorrhages or exudates. By combining VQA with Grad-CAM, this makes deep learning models more transparent, interactive and interpretable in applications like medical diagnostics, autonomous decision making and security.

## 5.2 Gradient-weighted Class Activation Mapping++ (Grad-CAM++)

Grad-CAM++ is an enhanced version of Grad-CAM that improves the localisation accuracy and the detection of multiple instances of a class in an image. While Grad-CAM provides a class-discriminative heatmap by computing the gradients of the target class score with respect to the last convolutional layer feature maps, it has limitations in localizing object boundaries and identifying multiple objects of the same class [77]. Grad-CAM++ addresses these issues by reweighting the activation maps based on higher order derivatives so that the generated heatmaps capture finer spatial details. The formulation of Grad-CAM++ builds upon Grad-CAM but with a more refined weighting mechanism for the activation maps. Instead of using a global average pooling of the gradients, Grad-CAM++ uses the second and third order gradients to compute the importance of each feature map  $A^k$  so that better object localisation is achieved. The class specific importance weight  $\alpha_k^c$  is given by:

$$\alpha_k^c = \sum_i \sum_j \alpha_{ij}^c \cdot \text{ReLU} \left( \frac{\partial y^c}{\partial A_{ij}^k} \right) \quad (22)$$

where  $\alpha_{ij}^c$  is derived using the second-order and third-order derivatives:

$$\alpha_{ij}^c = \frac{\partial^2 y^c / \partial A_{ij}^{k2}}{2(\partial^2 y^c / \partial A_{ij}^{k2}) + \sum_a \sum_b A_{ab}^k (\partial^3 y^c / \partial A_{ab}^{k3})} \quad (23)$$

This ensures regions of the image that contribute most to the class prediction get more attention and irrelevant activations get suppressed. The final Grad-CAM++ heatmap  $L_{\text{Grad-CAM}++}^c$  is calculated as:

$$L_{\text{Grad-CAM}++}^c = \text{ReLU} \left( \sum_k \alpha_k^c A^k \right) \quad (24)$$

The main advantage of Grad-CAM++ over Grad-CAM is that it assigns different importance to different pixels in the same activation map, so it's better at localizing multiple objects of the same class and capturing more details in the image. This is very useful for medical imaging tasks where you need to identify multiple lesions. By generating more precise and interpretable heatmaps, Grad-CAM++ increases model transparency and trust in deep learning based medical diagnosis systems.

### 5.3 Fine-Grained Class Activation Mapping (Layer-CAM)

Layer-CAM is an advanced class activation mapping (CAM) method that deepens the understanding of convolutional neural networks (CNNs) by generating fine-grained object localization maps across multiple layers. Unlike traditional methods like Grad-CAM which only rely on the final convolutional layer and produce coarse localization maps, Layer-CAM computes activation maps from both deep and shallow layers so it balances coarse spatial localization and fine-grained object details [78]. This is very useful for image classification tasks like retinal fundus image analysis where precise localization of critical regions is important. Layer-CAM works by computing activation maps from different layers of the CNN, considering the impact of each spatial location on the class score. Formally, the weight of each spatial location  $(i,j)$  in the  $k$ -th feature map  $A_k$  is defined as:

$$w_{ij}^{ck} = \text{ReLU} \left( \frac{\partial y^c}{\partial A_{ij}^k} \right) \quad (25)$$

where  $y^c$  is the predicted score of the target class  $c$ , and  $\frac{\partial y^c}{\partial A_{ij}^k}$  denotes the gradient of the target class score with respect to the feature map  $A_k$ . The class activation map  $M^c$  is then computed as:

$$M^c = \text{ReLU} \left( \sum_k w_{ij}^{ck} \cdot A_{ij}^k \right) \quad (26)$$

And a more general representation of Layer-CAM is:

$$S_c(x) = \text{ReLU} \sum_{c=1}^C (W_c \times \text{Score}(c, F(x))) \quad (27)$$

where  $F(x)$  represents the output of the final convolutional layer,  $W_c$  denotes the weight for each channel, and  $\text{Score}(c, F(x))$  computes the contribution of each channel to the class score. By utilizing location-specific gradient weights rather than global channel-wise weights, Layer-CAM effectively highlights the most relevant image regions while filtering out irrelevant background information. Its ability to generate class activation maps from different CNN layers provides complementary insights, resulting in more precise and integral class-specific object regions. Furthermore, its ease of integration with off-the-shelf CNN-based image classifiers, without requiring architectural modifications or changes in back-propagation, makes it a versatile tool for various image analysis tasks, including weakly-supervised object localization, semantic segmentation, and medical image classification.

### 5.4 Score-weighted Class Activation Mapping (Score-CAM)

Score-CAM is a way to show the parts of an input image that matter most for a neural network's classification output. It builds upon Class Activation Mapping (CAM) and Grad-CAM and adds class specific information while not using gradients to compute importance of activations. Unlike Grad-CAM which uses the gradients of the target class score with respect to feature maps, Score-CAM derives the importance of activation maps based on their contribution to the model's output score through forward pass. This gives more stable and noise free visual explanations especially in complex neural network architectures and improved class discriminative localization [79]. In the Score-CAM method the input image goes through the network and produces final convolutional feature maps. These activation maps are extracted and upsampled to the size of the input image and used as spatial masks. Each activation map is used to generate modified inputs that highlight different regions and then forward passed through the model. The importance of each feature map is determined by how much it contributed to the model's confidence score for a specific class. The class-specific importance score  $\alpha_k^c$  for each feature map is defined as:

$$\alpha_k^c = f_c(X \circ s(\text{Up}(A^k))) - f_c(X_b) \quad (28)$$

where  $f_c(\cdot)$  is the model's confidence score for class  $c$ ,  $X$  is the original image,  $X_b$  is the baseline image (e.g., black or zero-input),  $Up(A^k)$  denotes the upsampled activation map, and  $s(\cdot)$  is a normalization function mapping values to  $[0, 1]$ . The class scores are computed by linearly combining CAM values with weights learned from the fully connected layer. Unlike traditional gradient-based methods, Score-CAM doesn't backpropagate gradients, so no gradient saturation and vanishing. After computing the contribution scores, a softmax is applied to rescale the scores so they sum to one. The final Score-CAM heatmap  $L_{\text{Score-CAM}}^c$  is computed by combining the activation maps with their importance scores and then applying a ReLU:

$$L_{\text{Score-CAM}}^c = \text{ReLU} \left( \sum_k \alpha_k^c A^k \right) \quad (29)$$

Using forward-pass confidence scores instead of gradients, Score-CAM is better than Grad-CAM as it avoids gradient related issues and provides class discriminative and less noisy saliency maps. This is very important in medical imaging applications where precise localization of pathological features is crucial. Its ability to produce fine grain explanations and be robust across different CNN architectures makes it a must have tool for Explainable AI in medical diagnostics.

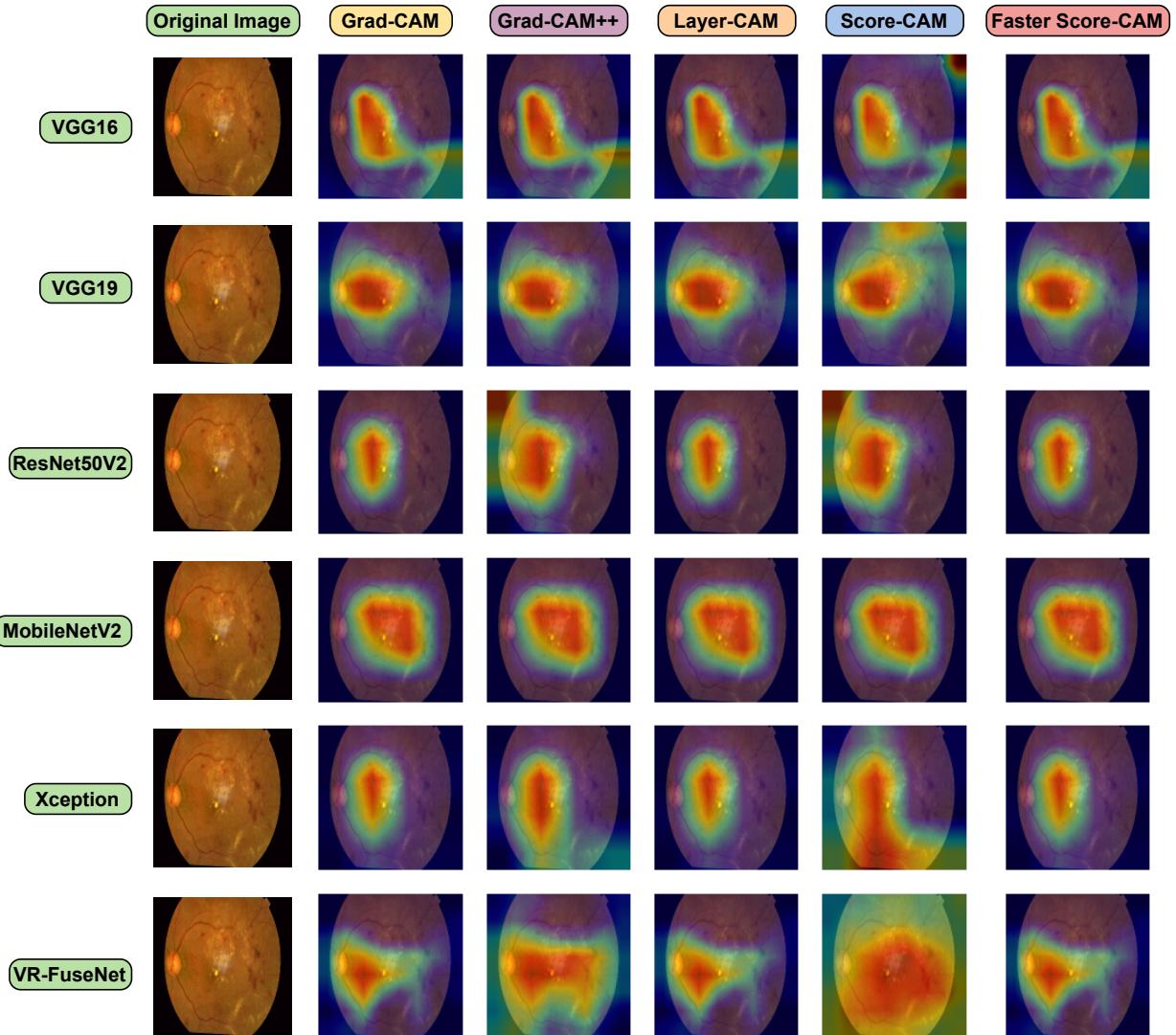


Figure 15: Comparative Analysis of Explainable AI Methods

## 5.5 Optimized Score-weighted Class Activation Mapping (Faster Score-CAM)

Faster Score-CAM is an optimization technique to improve the efficiency of convolutional neural networks (CNNs) in retinal fundus image classification. Traditional Score-CAM computes weights for all channels in a given feature map which can be computationally expensive [80]. To address this, Faster Score-CAM selects only the most informative channels by focusing on the ones with high variance, assuming they have more discriminative information. The process involves calculating the mean and variance for each channel, selecting the top  $N$  channels with the highest variance and then use those channels to compute weights and generate a heatmap. This reduces the computational overhead while preserving the essential discriminative features making it more practical for large dataset analysis in medical imaging tasks like diabetic retinopathy and glaucoma detection. Mathematically the weight assigned to channel  $k$  is given by:

$$\alpha_k = \frac{\text{Var}(A_k)}{\sum_{j=1}^N \text{Var}(A_j)} \quad (30)$$

where  $\text{Var}(A_k)$  is the variance of activations in channel  $k$  and  $N$  is the number of selected channels in the feature map. By reducing the number of channels Faster Score-CAM achieves a balance between computational efficiency and interpretability, the heatmap will highlight the relevant regions without requiring a lot of computational resources. However selecting the optimal value of  $N$  is crucial, smaller  $N$  may exclude useful features and larger  $N$  may reintroduce redundancy. This tradeoff requires experimentation to maximize both speed and quality of the visual explanation especially in applications where model interpretability is critical like retinal disease classification

### 5.5.1 Insights Derived from XAI Analysis

Figure 15 shows the activation patterns of different Transfer Learning (TL) models for Diabetic Retinopathy using various XAI techniques. All models show consistent activation around the critical regions of the retina, especially around the lesions, microaneurysms or hemorrhages. Grad-CAM shows big and widespread activation across the retina, highlighting multiple affected areas but sometimes going beyond the most relevant regions. Grad-CAM++ generates more focused and precise heatmaps, concentrating on specific regions with higher pathological relevance. Layer-CAM shows broader activation but still focuses on the significant areas, so even peripheral signs of the disease are considered. Score-CAM spreads attention more evenly across the retinal surface, capturing general structural patterns while keeping a reasonable focus on disease-prone areas. Faster Score-CAM shows sharper and more confident activation in highly affected regions, pinpointing potential lesions with higher accuracy. Across the models, Grad-CAM and Grad-CAM++ highlight the key pathological regions, Layer-CAM has a wider but still targeted focus. Score-CAM and Faster Score-CAM are similar but have different focal points, Faster Score-CAM prioritizes the areas of higher confidence. The results show that despite the variations, the TL models recognize the same regions of interest, so they can consistently detect the signs of Diabetic Retinopathy. The variations in the heatmap distribution emphasize the importance of choosing the right XAI method to get the most reliable and interpretable insights from deep learning models in medical diagnosis.

## 6 Limitations and Future Works

The Kaggle EyePACS dataset is widely used in medical image classification tasks but has some challenges as mentioned in previous studies. These include large volume, noisy data and incorrect labeling which can affect the model's accuracy and reliability [30]. Due to these limitations and the difficulty in annotating such a large dataset with accurate labels we didn't use the EyePACS dataset in our research.

Despite the effectiveness of DR detection the model has some limitations to be addressed in future research. Computational cost is one of the main issues as the transformer model increases the number of parameters which leads to longer training and inference times. Due to these computational constraints we couldn't implement Vision Transformers (ViTs) which have shown to be better in capturing long range dependencies and contextual information in medical imaging tasks. Future work will focus on integrating ViTs in the model architecture to leverage their ability to process global image features which could lead to better accuracy and interpretability of the model. However implementing ViTs will require optimized training strategies and computational resources which will be key moving forward.

Another limitation is imbalanced dataset which can affect the model's ability to generalize well across different DR severity levels. Although the hybrid dataset combines multiple publicly available sources to increase diversity, some severity classes are underrepresented which can bias the model's predictions. To address this future research will explore the use of Generative Adversarial Networks (GANs) to generate synthetic retinal images to balance the dataset and improve the model's robustness. GANs are widely used in medical imaging to generate high quality synthetic data while maintaining clinical relevance. Their application in this case can reduce bias in the model's predictions and improve its ability to generalize across different patient populations and imaging conditions.

Another area to improve is generalization to real world clinical data. The dataset used in this research is mainly composed of publicly available images which may not fully represent the variability seen in real world patient populations, different imaging devices or clinical settings. Domain adaptation techniques such as unsupervised domain adaptation will be explored in future work to make the model more adaptable to new environments. Adding more real world clinical images to the dataset and fine tuning the model to handle variations in lighting, resolution and noise levels will make the model more robust and reliable across different healthcare institutions. Finally, although the model is image based, combining multi-modal data such as patient demographics, clinical history and genetic factors could improve the accuracy. Many diseases including diabetic retinopathy are not just image based and incorporating additional clinical parameters could lead to more accurate diagnoses. Future work will explore methods to integrate image based deep learning models with structured clinical data to have a more complete AI driven diagnostic tool.

## 7 Conclusion

The VR-FuseNet model in this study does well in DR classification tasks by combining the strengths of two advanced convolutional neural network architectures, VGG19 and ResNet50V2. It uses VGG19's ability to capture fine-grained spatial features and ResNet50V2's deep hierarchical feature extraction. As a result, it scores high in accuracy, precision, recall and F1-score and outperforms individual models on the hybrid dataset of retinal images from various public sources. The careful fusion of these architectures helps the model to generalise across different clinical imaging conditions, hence more applicable in real-world healthcare settings. Moreover, the use of Explainable Artificial Intelligence (XAI) techniques, particularly gradient-based methods such as Grad-CAM, Grad-CAM++, Layer-CAM, Score-CAM and Faster Score-CAM adds value by providing interpretable visual explanations. These techniques allow medical professionals to see and understand which features in the retinal images contribute to the classification, improving model transparency, clinical trust and decision making. The model not only focuses on high accuracy but also ensures the predictions are interpretable and clinically meaningful, hence better validation and adoption by healthcare providers. However, there are still challenges to be addressed to fully utilise AI-driven DR detection systems. Specifically, computational complexity and dataset imbalance are big issues. Addressing these will be key in deploying these models in clinical settings. Future work should explore integrating Vision Transformers (ViTs) which have shown to capture long-range dependencies and global context in images. Also, using advanced data augmentation techniques including GANs can help balance the dataset and mitigate class imbalance issue, hence improve the model's performance and robustness across different patient demographics and imaging modalities.

## References

- [1] Pratima Yadav and Nagendra Pratap Singh. Classification of normal and abnormal retinal images by using feature-based machine learning approach. In *Recent Trends in Communication, Computing, and Electronics: Select Proceedings of IC3E 2018*, pages 387–396. Springer, 2019.
- [2] Pouya Saeedi, Inga Petersohn, Paraskevi Salpea, Belma Malanda, Suvi Karuranga, Nigel Unwin, Stephen Colagiuri, Leonor Guariguata, Ayesha A Motala, Katherine Ogurtsova, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas. *Diabetes research and clinical practice*, 157:107843, 2019.
- [3] Pablo Aschner, Suvi Karuranga, Steven James, David Simmons, Abdul Basit, Jonathan E Shaw, Sarah H Wild, Katherine Ogurtsova, and Pouya Saeedi. The international diabetes federation’s guide for diabetes epidemiological studies. *Diabetes research and clinical practice*, 172, 2021.
- [4] International Diabetes Federation. Idf diabetes atlas, 10th edition. [https://diabetesatlas.org/atlas/te nth-edition/](https://diabetesatlas.org/atlas/tenth-edition/), 2021. Accessed on 15 January 2025.
- [5] Tien Yin Wong and Charumathi Sabanayagam. Strategies to tackle the global burden of diabetic retinopathy: from epidemiology to artificial intelligence. *Ophthalmologica*, 243(1):9–20, 2020.
- [6] Priscilla Peixi Choo, Norshamsiah Md Din, Nooranah Azmi, and Mae-Lynn Catherine Bastion. Review of the management of sight-threatening diabetic retinopathy during pregnancy. *World Journal of Diabetes*, 12(9):1386, 2021.
- [7] Katie Curran, Tunde Peto, Jost B. Jonas, David Friedman, Judy E. Kim, Janet Leasher, Ian Tapply, Arthur G. Fernandes, Maria Vittoria Cincinelli, Alessandro Arrigo, Nicolas Leveziel, Serge Resnikoff, Hugh R. Taylor, Tabassom Sedighi, Seth Flaxman, Mukharram M. Bikbov, Tasanee Braithwaite, Alain Bron, Ching-Yu Cheng, Monte A. Del Monte, Joshua R. Ehrlich, João M. Furtado, Gus Gazzard, M. Elizabeth Hartnett, Rim Kahloun, John H. Kempen, Moncef Khairallah, Rohit C. Khanna, Van Charles Lanssingh, Kovin S. Naidoo, Vinay Nangia, Michal Nowak, Konrad Pesudovs, Pradeep Ramulu, Fotis Topouzis, Mitiadis Tsilimbaris, Ya Xing Wang, Ningli

- Wang, Rupert R. A. Bourne, Bright Opoku Ahinkorah, Hamid Ahmadieh, Ayman Ahmed, Ahmad Samir Alfaar, Louay Almidani, Hubert Amu, Sofia Androudi, Jalal Arabloo, Aleksandr Y. Aravkin, Mulu Tiruneh Asemu, Ahmed Y. Azzam, Nayereh Baghcheghi, Freddie Bailey, Mehmet Firat Baran, Mainak Bardhan, Till Winfried Bärnighausen, Amadou Barrow, Pankaj Bhardwaj, Paul Svitil Briant, Katrin Burkart, Luis Alberto Cámera, Kaleb Coberly, Omid Dadras, Xiaochen Dai, Amin Dehghan, Berecha Hundessa Demessa, Mengistie Diress, Thanh Chi Do, Thao Huynh Phuong Do, Klara Georgieva Dokova, Bruce B. Duncan, Michael Ekholueneiale, Muhammed Elhadi, Mohammad Hassan Emamian, Mehdi Emamverdi, Hossein Farrokhpour, Ali Fatehizadeh, Lorenzo Ferro Desideri, Mesfin Gebrehiwot, Fariba Ghassemi, Mesay Dechasa Gudeta, Sapna Gupta, Veer Bala Gupta, Vivek Kumar Gupta, Billy Randall Hammond, Mehdi Harorani, Hamidreza Hasani, Golnaz Heidari, Mehdi Hosseinzadeh, John J. Huang, Sheikh Mohammed Shariful Islam, Nilofer Javadi, Aida Jimenez-Corona, Mohammad Jokar, Charity Ehimwenma Joshua, Vidya Kadashetti, Himal Kandel, Hengameh Kasraei, Rimple Jeet Kaur, Sudarshan Khanal, Zahra Khorrami, Hamid Reza Kohestani, Kewal Krishan, Stephen S. Lim, Mohammed Magdy Abd El Razek, Vahid Mansouri, Andrea Maugeri, Tomislav Mestrovic, Awoke Misganaw, Ali H. Mokdad, Hamed Momeni-Moghaddam, Sara Momtazmanesh, Christopher J. L. Murray, Hadush Negash, Uchechukwu Levi Osuagwu, Shahina Pardhan, Jay Patel, Shrikant Pawar, Ionela-Roxana Petcu, Hoang Tran Pham, Mohsen Pourazizi, Ibrahim Qattea, Mosiur Rahman, Umar Saeed, Amirhossein Sahebkar, Mohammad Amin Salehi, Maryam Shayan, Aminu Shittu, Jamie D. Steinmetz, Yao Tan, Aristidis Tsatsakis, Muhammad Umair, Theo Vos, Hong Xiao, Yuyi You, Mikhail Sergeevich Zastrozhin, Zhi-Jiang Zhang, Peng Zheng, Vision Loss Expert Group of the Global Burden of Disease Study, the GBD 2019 Blindness, and Vision Impairment Collaborators. Global estimates on the number of people blind or visually impaired by diabetic retinopathy: a meta-analysis from 2000 to 2020. *Eye*, 38(11):2047–2057, 2024.
- [8] Elizabeth A Lundeen, Zeb Burke-Conte, David B Rein, John S Wittenborn, Jinan Saaddine, Aaron Y Lee, and Abraham D Flaxman. Prevalence of diabetic retinopathy in the us in 2021. *JAMA ophthalmology*, 141(8):747–754, 2023.
  - [9] Zhen Ling Teo, Yih-Chung Tham, Marco Yu, Miao Li Chee, Tyler Hyuntaek Rim, Ning Cheung, Mukharram M Bikbov, Ya Xing Wang, Yating Tang, Yi Lu, et al. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. *Ophthalmology*, 128(11):1580–1591, 2021.
  - [10] Paraskevi Salpea, Yadi Huang, Irini Genitsaridi, Dianna J Magliano, and Edward J Boyko. Call for data contribution to the 11th edition of idf's diabetes atlas (2024). *Diabetes Research and Clinical Practice*, 198, 2023.
  - [11] ZERO TO FINALS. Diabetic retinopathy. <https://zerotofinals.com/medicine/ophthalmology/diabeticretinopathy/>, October 2023. Accessed on 16 December 2024.
  - [12] Monika Lakhwani, Omkar Dastane, Nurhizam Safie Mohd Satar, and Zainudin Johari. The impact of technology adoption on organizational productivity. *The Journal of Industrial Distribution & Business*, 11(4):7–18, 2020.
  - [13] Sachin Bhandari, Sunil Pathak, and Sonal Amit Jain. A literature review of early-stage diabetic retinopathy detection using deep learning and evolutionary computing techniques. *Archives of Computational Methods in Engineering*, 30(2):799–810, 2023.
  - [14] Nurul Mirza Afiqah Tajudin, Kuryati Kipli, Muhammad Hamdi Mahmood, Lik Thai Lim, Dayang Azra Awang Mat, Rohana Sapawi, Siti Kudnie Sahari, Kasumawati Lias, Suriati Khartini Jali, and Mohammed Enamul Hoque. Deep learning in the grading of diabetic retinopathy: A review. *IET Computer Vision*, 16(8):667–682, 2022.
  - [15] A Mary Dayana and WR Sam Emmanuel. A comprehensive review of diabetic retinopathy detection and grading based on deep learning and metaheuristic optimization techniques. *Archives of Computational Methods in Engineering*, 30(7):4565–4599, 2023.
  - [16] Hossein Shakibania, Sina Raoufi, Behnam Pourafkham, Hassan Khotanlou, and Muharram Mansoorizadeh. Dual branch deep learning network for detection and stage grading of diabetic retinopathy. *Biomedical Signal Processing and Control*, 93:106168, 2024.
  - [17] Dharmalingam Muthusamy and Parimala Palani. Deep learning model using classification for diabetic retinopathy detection: an overview. *Artificial Intelligence Review*, 57(7):185, 2024.
  - [18] Samia Akhtar, Shabib Aftab, Munir Ahmad, and Asma Akhtar. Diabetic retinopathy severity grading using transfer learning techniques. *International Journal of Engineering and Manufacturing*, 14:41–53, 2024.
  - [19] Lakshay Arora, Sunil K Singh, Sudhakar Kumar, Hardik Gupta, Wadee Alhalabi, Varsha Arya, Shavi Bansal, Kwok Tai Chui, and Brij B Gupta. Ensemble deep learning and efficientnet for accurate diagnosis of diabetic retinopathy. *Scientific Reports*, 14(1):30554, 2024.
  - [20] Mustafa Youldash, Atta Rahman, Manar Alsayed, Abrar Sebiany, Joury Alzayat, Noor Aljishi, Ghaida Alshammari, and Mona Alqahtani. Early detection and classification of diabetic retinopathy: A deep learning approach. *AI*, 5(4), 2024.

- [21] Sunday O Olatunji, Nawal Alsheikh, Lujain Alnajrani, Alhatoon Alanazy, Meshael Almusairii, Salam Alshammasi, Aisha Alansari, Rim Zaghdoud, Alaa Alahmadi, Mohammed Imran Basheer Ahmed, et al. Comprehensible machine-learning-based models for the pre-emptive diagnosis of multiple sclerosis using clinical data: a retrospective study in the eastern province of saudi arabia. *International Journal of Environmental Research and Public Health*, 20(5):4261, 2023.
- [22] Jie Yao, Joshua Lim, Gilbert Yong San Lim, Jasmine Chiat Ling Ong, Yuhe Ke, Ting Fang Tan, Tien-En Tan, Stela Vujosevic, and Daniel Shu Wei Ting. Novel artificial intelligence algorithms for diabetic retinopathy and diabetic macular edema. *Eye and Vision*, 11(1):23, 2024.
- [23] Manoj Saligrama Harisha, Arya Arun Bhosale, and M Narender. Deep learning-powered segmentation and classification of diabetic retinopathy for enhanced diagnostic precision. *Artificial Intelligence in Health*, 1(4):30–42, 2024.
- [24] Ghadah Alwakid, Walaa Gouda, Mamoonah Humayun, and Noor Zaman Jhanjhi. Deep learning-enhanced diabetic retinopathy image classification. *Digital Health*, 9:20552076231194942, 2023.
- [25] Xuebin Xu, Dehua Liu, Guohua Huang, Muyu Wang, Meng Lei, and Yang Jia. Computer aided diagnosis of diabetic retinopathy based on multi-view joint learning. *Computers in Biology and Medicine*, 174:108428, 2024.
- [26] Abderaouf M Moustari, Youcef Brik, Bilal Moustari, and Rafik Bouaouina. Two-stage deep learning classification for diabetic retinopathy using gradient weighted class activation mapping. *Automatika: časopis za automatiku, mjerjenje, elektroniku, računarstvo i komunikacije*, 65(3):1284–1299, 2024.
- [27] Gaurav Saxena, Dhirendra Kumar Verma, Amit Paraye, Alpana Rajan, and Anil Rawat. Improved and robust deep learning agent for preliminary detection of diabetic retinopathy using public datasets. *Intelligence-Based Medicine*, 3:100022, 2020.
- [28] Sanli Yi and Ziyan Chen. Midc: Medical image dataset cleaning framework based on deep learning. *Heliyon*, 10(19), 2024.
- [29] Nilarun Mukherjee and Souvik Sengupta. Comparing deep feature extraction strategies for diabetic retinopathy stage classification from fundus images. *Arabian Journal for Science and Engineering*, 48(8):10335–10354, 2023.
- [30] Maria Tariq, Vasile Palade, and YingLiang Ma. Transfer learning based classification of diabetic retinopathy on the kaggle eyepacs dataset. In *International Conference on Medical Imaging and Computer-Aided Diagnosis*, pages 89–99. Springer, 2022.
- [31] Ghadah Alwakid, Walaa Gouda, and Mamoonah Humayun. Enhancement of diabetic retinopathy prognostication using deep learning, clahe, and esrgan. *Diagnostics*, 13(14):2375, 2023.
- [32] Cheena Mohanty, Sakuntala Mahapatra, Biswaranjan Acharya, Fotis Kokkoras, Vassilis C Gerogiannis, Ioannis Karamitsos, and Andreas Kanavos. Using deep learning architectures for detection and classification of diabetic retinopathy. *Sensors*, 23(12):5726, 2023.
- [33] Wei Kitt Wong, Filbert H Juwono, and Catur Apriono. Diabetic retinopathy detection and grading: A transfer learning approach using simultaneous parameter optimization and feature-weighted ecoc ensemble. *IEEE Access*, 11:83004–83016, 2023.
- [34] AM Mutawa, Shahad Alnajdi, and Sai Sruthi. Transfer learning for diabetic retinopathy detection: A study of dataset combination and model performance. *Applied Sciences*, 13(9):5685, 2023.
- [35] Md Nahiduzzaman, Md Robiul Islam, Md Omaer Faruq Goni, Md Shamim Anower, Mominul Ahsan, Julfikar Haider, and Marcin Kowalski. Diabetic retinopathy identification using parallel convolutional neural network based feature extractor and elm classifier. *Expert Systems with Applications*, 217:119557, 2023.
- [36] M Murugappan, NB Prakash, R Jeya, A Mohanarathinam, GR Hemalakshmi, and Mufti Mahmud. A novel few-shot classification framework for diabetic retinopathy detection and grading. *Measurement*, 200:111485, 2022.
- [37] Eman AbdelMaksoud, Sherif Barakat, and Mohammed Elmogy. A computer-aided diagnosis system for detecting various diabetic retinopathy grades based on a hybrid deep learning technique. *Medical & Biological Engineering & Computing*, 60(7):2015–2038, 2022.
- [38] Anas Bilal, Liucun Zhu, Anan Deng, Huihui Lu, and Ning Wu. Ai-based automatic detection and classification of diabetic retinopathy using u-net and deep learning. *Symmetry*, 14(7):1427, 2022.
- [39] A Rafega Beham and V Thanikaiselvan. An optimized deep-learning algorithm for the automated detection of diabetic retinopathy. *Soft Computing*, pages 1–11, 2023.

- [40] Muhammad Mohsin Butt, DNF Awang Iskandar, Sherif E Abdelhamid, Ghazanfar Latif, and Runna Alghazo. Diabetic retinopathy detection from fundus images of the eye using hybrid deep learning features. *Diagnostics*, 12(7):1607, 2022.
- [41] Saif Hameed Abbood, Haza Nuzly Abdull Hamed, Mohd Shafry Mohd Rahim, Amjad Rehman, Tanzila Saba, and Saeed Ali Bahaj. Hybrid retinal image enhancement algorithm for diabetic retinopathy diagnostic using deep learning model. *IEEE Access*, 10:73079–73086, 2022.
- [42] Ghulam Ali, Aqsa Dastgir, Muhammad Waseem Iqbal, Muhammad Anwar, and Muhammad Faheem. A hybrid convolutional neural network model for automatic diabetic retinopathy classification from fundus images. *IEEE Journal of Translational Engineering in Health and Medicine*, 11:341–350, 2023.
- [43] Mohaimenul Azam Khan Raiaan, Kaniz Fatema, Inam Ullah Khan, Sami Azam, Md Rafi Ur Rashid, Md Sad-dam Hossain Mukta, Mirjam Jonkman, and Friso De Boer. A lightweight robust deep learning model gained high accuracy in classifying a wide range of diabetic retinopathy images. *IEEE Access*, 11:42361–42388, 2023.
- [44] Brahami Menaouer, Zoulikha Dermane, Nour El Houda Kebir, and Nada Matta. Diabetic retinopathy classification using hybrid deep learning approach. *SN Computer Science*, 3(5):357, 2022.
- [45] Francesco Mercaldo, Marcello Di Giammarco, Arianna Apicella, Giacomo Di Iadarola, Mario Cesarelli, Fabio Martinelli, and Antonella Santone. Diabetic retinopathy detection and diagnosis by means of robust and explainable convolutional neural networks. *Neural Computing and Applications*, 35(23):17429–17441, 2023.
- [46] Hemanth Kumar Vasireddi, K Suganya Devi, and GNV Raja Reddy. Dr-xai: Explainable deep learning model for accurate diabetic retinopathy severity assessment. *Arabian Journal for Science and Engineering*, 49(9):12899–12917, 2024.
- [47] Hanan Saleh Alghamdi. Towards explainable deep neural networks for the automatic detection of diabetic retinopathy. *Applied Sciences*, 12(19):9435, 2022.
- [48] Karthik, Maggie, and Sohier Dane. Aptos 2019 blindness detection. <https://www.kaggle.com/competitions/aptos2019-blindness-detection>, September 2019. Accessed on 06 April 2024.
- [49] Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501:511–522, 2019.
- [50] Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang. Ddr dataset. <https://github.com/nkicsl/DDR-dataset>, July 2019. Accessed on 27 December 2024.
- [51] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.
- [52] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid). <https://dx.doi.org/10.21227/H25W98>, April 2018. Accessed on 19 September 2024.
- [53] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, John-Richard Ordóñez-Varela, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, pages 231–234, 2014.
- [54] ADCIS Image Processing Algorithm Evaluation and Data Management. Messidor 2. <https://www.adcis.net/en/third-party/messidor2/>. Accessed on 15 May 2024.
- [55] Google Brain and Dale Webster. Messidor-2 dr grades. <https://www.kaggle.com/datasets/google-brain/messidor2-dr-grades>, July 2018. Accessed on 15 May 2024.
- [56] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S Corrado, Lily Peng, and Dale R Webster. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, 125(8):1264–1272, 2018.
- [57] Dequan Wang, Xiosong Wang, Lilong Wang, Mengzhang Li, Qian Da, Xiaoqiang Liu, Xiangyu Gao, Jun Shen, Junjun He, Tian Shen, et al. A real-world dataset and benchmark for foundation model adaptation in medical image classification. *Scientific Data*, 10(1):574, 2023.
- [58] Dequan Wang, Xiosong Wang, Lilong Wang, Mengzhang Li, Qian Da, Xiaoqiang Liu, Xiangyu Gao, Jun Shen, Junjun He, Tian Shen, Qi Duan, Jie Zhao, Kang Li, Yu Qiao, and Shaoting Zhang. Retino: Diabetic Retinopathy Grading in Retina Images. [https://springernature.figshare.com/articles/dataset/Retino\\_Diabetic\\_Retinopathy\\_Grading\\_in\\_Retina\\_Images/22302562](https://springernature.figshare.com/articles/dataset/Retino_Diabetic_Retinopathy_Grading_in_Retina_Images/22302562), August 2023. Accessed on 20 November 2024.

- [59] Shuvashis Sarker Shuvo, Shamim Rahim Refat, Faika Fairuj Preotee, and Tashreef Muhammad. Advanced cnn and explainable ai based architecture for interpretable brain mri analysis. In *Proceedings of the 3rd International Conference on Computing Advancements*, pages 319–326, 2024.
- [60] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
- [61] Javad Hassannataj Joloudari, Abdolreza Marefat, Mohammad Ali Nematollahi, Solomon Sunday Oyelere, and Sadiq Hussain. Effective class-imbalance learning based on smote and convolutional neural networks. *Applied Sciences*, 13(6):4006, 2023.
- [62] Damien Dablain, Bartosz Krawczyk, and Nitesh V Chawla. Deepsmote: Fusing deep learning and smote for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):6390–6404, 2022.
- [63] Fatimazahra Belharar. Enhancing classification accuracy for imbalanced image data using smote. <https://medium.com/@fatimazahra.belharar/enhancing-classification-accuracy-for-imbalance-d-image-data-using-smote-41737783a720>, June 2024. Accessed on 07 February 2025.
- [64] Karel J Zuiderveld et al. Contrast limited adaptive histogram equalization. *Graphics gems*, 4(1):474–485, 1994.
- [65] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.
- [66] Ali M Reza. Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38:35–44, 2004.
- [67] Shuvashis Sarker, Shamim Rahim Refat, Faika Fairuj Preotee, Tanvir Rouf Shawon, and Raihan Tanvir. Comprehensive lung disease detection using deep learning models and hybrid chest x-ray data with explainable ai. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pages 2279–2284. IEEE, 2024.
- [68] Md Rakibul Hasan, Nusrat Jahan, Shifat Islam, and Tashreef Muhammad. Deep learning-based multi-class brain abnormality diagnosis on bangladeshi mri images. In *2024 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON)*, pages 252–257. IEEE, 2024.
- [69] Shuvashis Sarker, Shamim Rahim Refat, Faika Fairuj Preotee, Shifat Islam, Tashreef Muhammad, and Mohammad Ashraful Hoque. An exploratory approach towards investigating and explaining vision transformer and transfer learning for brain disease detection. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pages 3224–3229. IEEE, 2024.
- [70] Muhammad Attique Khan, Majed Alhaisoni, Usman Tariq, Nazar Hussain, Abdul Majid, Robertas Damaševičius, and Rytis Maskeliūnas. Covid-19 case recognition from chest ct images by deep learning, entropy-controlled firefly optimization, and parallel feature fusion. *Sensors*, 21(21):7286, 2021.
- [71] Farah Mohammad and Muna Al-Razgan. Deep feature fusion and optimization-based approach for stomach disease classification. *Sensors*, 22(7):2801, 2022.
- [72] Jocelyn D’Souza. Let’s learn about auc roc curve! <https://medium.com/greyatom/lets-learn-about-auc-roc-curve-4a94b4d88152>, March 2018. Accessed on 07 February 2025.
- [73] Zhe Hui Hoo, Jane Candlish, and Dawn Teare. What is an roc curve?, 2017.
- [74] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.
- [75] Fatema Tuj Johora Faria, Mukaffi Bin Moin, Pronay Debnath, Asif Iftekher Fahim, and Faisal Muhammad Shah. Explainable convolutional neural networks for retinal fundus classification and cutting-edge segmentation models for retinal blood vessels from fundus images. *arXiv preprint arXiv:2405.07338*, 2024.
- [76] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [77] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.
- [78] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.
- [79] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Scorecam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

- [80] Jing Li, Dongbo Zhang, Bumin Meng, Yongxing Li, and Lufeng Luo. Fimf score-cam: fast score-cam based on local multi-feature integration for visual interpretation of cnns. *IET Image Processing*, 17(3):761–772, 2023.