International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

MFA-DRNet: An Explainable Multi-Modal Feature Attention Network for Early Diabetic Retinopathy Detection Using Deep Learning

Vaishali Hirlekar[1], Kranti Ghag[2], Jayashree Jagdale[3], Puja Padiya[4], Balaso Jagdale[5], Vivek Khalane[6]

[1]Department of Computer Engineering, Shah & Anchor Kutchhi Engineering College,Email Id :-Vaishali.hirlekar@sakec.ac.in,Orcid Id:-0000-0003-3430-6780
[2]Department of Computer Engineering, SVKM's Dwarkadas J. Sanghvi College of Engineering, Email Id :-kranti.ghag@djsce.ac.in, Orchid Id: 0000-0002-4961-1633
[3]Department of CSE, School of Engineering, Indira University, Pune, Email Id:- jayashree.jagdale@gmail.com, Orchid id: 0000-0002-9777-8526
[4]Department of Information Technology, Xavier Institute of Engineering, Email Id:- puja.padiya05@gmail.com, ORCID: 0000-0002-3446-3340
[5]Department of Computer Engineering, Dr. Vishwanath Karad MIT World Peace University, Email Id:- bjagdale@gmail.com, Orchid ID: 0000-0002-7224-0025
[6]A.C. Patil College of Engineering, Email Id:- vivekpkhalane@gmail.com, ORCID: 0000-0002-5243-6872

ABSTRACT
Diabetic retinopathy (DR) is a major cause of vision impairment, requiring accurate and early diagnosis for effective treatment. Existing deep learning methods often struggle with complex retinal patterns, long-range feature dependencies, class imbalance, and limited gen- eralization across datasets. This paper proposes MFA-DRNet, a multimodal feature adaptive fusion network integrating convolutional neural networks (CNNs), Vision Transformers (ViTs), clinical features, adaptive attention fusion, and focal loss optimization for DR classification. The CNN and ViT branches jointly capture local retinal lesions and global contextual information, while adaptive fusion enhances multimodal representation learning. Extensive evaluation on EyePACS, APTOS 2019, IDRiD, DDR, and Messidor-2 datasets demonstrates superior per- formance, achieving up to 96.13% accuracy, 95.68% F1-score, and 0.981 AUC. Ablation and cross-dataset experiments confirm model robustness and generalization capability. Explainabil- ity analysis further demonstrates clinically meaningful attention localization, supporting reliable AI-assisted retinal screening.

Keywords: Diabetic Retinopathy, Deep Learning, Explainable Artificial Intelligence, Vision Transformer, Multi-Modal Learning, Medical Image Analysis, Feature Attention, Fundus Imag- ing, Optical Coherence Tomography, Computer-Aided Diagnosis.

INTRODUCTION

1.1Background

Diabetes mellitus (DM) has become one of the most prevalent chronic metabolic disorders world- wide and represents a major public health concern owing to its rapidly increasing incidence and associated microvascular and macrovascular complications. Among these complications, Diabetic Retinopathy (DR) is recognized as one of the leading causes of preventable blindness in the working- age population. Persistent hyperglycemia progressively damages the retinal microvasculature, re- sulting in pathological manifestations such as microaneurysms, hemorrhages, hard exudates, cotton wool spots, venous beading, and neovascularization. Since the early stages of DR are often asymp- tomatic, timely diagnosis through regular retinal screening is essential to prevent irreversible vision loss and improve long-term patient outcomes. Recent surveys indicate that the growing burden of diabetes has substantially increased the demand for automated and scalable retinal screening systems capable of supporting ophthalmologists in early disease detection [1].

Conventionally, diabetic retinopathy screening relies on manual assessment of color fundus pho- tographs by experienced retinal specialists. Although manual grading remains the clinical reference standard, it is labor-intensive, subjective, and susceptible to inter-observer variability, particularly in large-scale screening programs. The increasing prevalence of diabetes has widened the gap be- tween the number of

patients requiring annual retinal examinations and the availability of trained ophthalmologists, especially in low-resource and rural healthcare settings. Consequently, computer- aided diagnosis (CAD) systems have emerged as an attractive solution for improving screening efficiency, reducing diagnostic workload, and facilitating timely referral of high-risk patients.

The rapid advancement of Artificial Intelligence (AI), particularly deep learning, has signifi- cantly transformed medical image analysis during the past decade. Deep learning models automat- ically learn hierarchical feature representations directly from retinal images, eliminating the need for manually engineered features that characterized traditional machine learning approaches. Convo- lutional Neural Networks (CNNs) have demonstrated remarkable success in retinal lesion detection, blood vessel segmentation, optic disc localization, and diabetic retinopathy grading owing to their capability to capture discriminative local spatial features. Numerous CNN-based architectures, including ResNet, DenseNet, EfficientNet, Inception, and MobileNet variants, have achieved com- petitive diagnostic performance on benchmark datasets such as EyePACS, APTOS 2019, Messidor, DDR, and IDRiD [2, 3, 4, 5]. Nevertheless, their performance often deteriorates when evaluated on external datasets because of differences in imaging devices, acquisition protocols, illumination conditions, and patient demographics.

More recently, transformer-based architectures have demonstrated significant potential in oph- thalmic image analysis by employing self-attention mechanisms to model long-range spatial depen- dencies across retinal structures. Unlike conventional CNNs, Vision Transformers (ViTs) capture global contextual relationships between pathological regions, thereby improving recognition of sub- tle retinal abnormalities that may be spatially distant [6, 7]. Hybrid CNN–Transformer models have consequently emerged as an effective strategy for combining local texture extraction with global con- textual learning. However, transformer-based models generally require substantially larger training datasets and higher computational resources, limiting their deployment in many clinical environ- ments, particularly in resource-constrained healthcare facilities. Computational efficiency therefore remains an important consideration in the design of intelligent retinal screening systems.

1.2 Motivation

Despite these advances, several challenges continue to hinder the widespread clinical adoption of deep learning models for diabetic retinopathy detection. First, existing models are predominantly trained using single-modality fundus images while neglecting complementary clinical information such as glycated hemoglobin (HbA1c), duration of diabetes, blood pressure, age, body mass index, and Optical Coherence Tomography (OCT) imaging. The integration of multi-modal information has the potential to improve diagnostic accuracy by providing a more comprehensive representation of disease progression. Second, severe class imbalance in retinal datasets often biases optimization toward normal and mild DR classes, resulting in poor recognition of proliferative diabetic retinopa- thy cases that require immediate medical intervention. Third, many deep learning systems remain “black-box” models whose decision-making processes are difficult for clinicians to interpret, thereby reducing trust and limiting integration into routine clinical practice. Explainable Artificial Intelli- gence (XAI) techniques such as Grad-CAM, SHAP, and concept-based explanations have recently been proposed to improve model transparency, yet their integration with multimodal DR diagnosis remains limited [8, 9].

Another critical limitation identified in recent literature is the lack of model robustness across heterogeneous clinical environments. Retinal datasets frequently differ in terms of image qual- ity, camera specifications, ethnic diversity, disease prevalence, and annotation protocols, leading to domain shift and reduced generalization capability. Furthermore, most existing studies focus primarily on classification accuracy while paying comparatively less attention to computational complexity, inference latency, uncertainty estimation, calibration, and deployment feasibility in real-world teleophthalmology systems. Recent systematic reviews emphasize that future diabetic retinopathy detection frameworks should simultaneously address explainability, computational ef- ficiency, multimodal learning, and cross-domain robustness to facilitate translation from research laboratories to clinical practice.

Motivated by these research challenges, this paper proposes MFA-DRNet (Multi-Modal Fea- ture Attention Diabetic Retinopathy Network), an explainable deep learning framework that inte- grates retinal fundus images, OCT imaging, and structured clinical information within a unified architecture. The proposed framework combines multi-scale convolutional feature extraction with transformer-based global attention to simultaneously capture local lesion characteristics and global retinal structures. An adaptive attention-guided fusion module is introduced to dynamically inte- grate heterogeneous imaging and clinical features, while an adaptive weighted focal loss function is employed to mitigate class imbalance during optimization. Furthermore, Grad-CAM and SHAP- based explainability modules are incorporated to provide clinically interpretable visual evidence supporting automated diagnostic decisions. Through this integrated approach, the proposed frame- work seeks to improve diagnostic accuracy, robustness, interpretability, and clinical applicability for large-scale diabetic retinopathy screening programs.

1.3Contributions

The main contributions of this work are as follows:

1.We propose a unified multi-modal learning framework, MFA-DRNet, that jointly integrates retinal fundus images, OCT scans, and structured clinical parameters for diabetic retinopathy detection.

2.We design an adaptive cross-attention feature fusion module that dynamically weights the contribution of each modality according to its sample-specific diagnostic relevance.

3.We introduce an adaptive weighted focal loss that combines effective-number-based class re- weighting with focal loss down-weighting to address severe class imbalance in DR severity grading.

4.We integrate a dual explainability module based on Grad-CAM and SHAP to provide image- level and clinical-feature-level interpretability, respectively.

5.We evaluate the proposed framework across five publicly available benchmark datasets, in- corporating cross-dataset generalization testing, ablation studies, and statistical significance analysis to assess robustness and clinical applicability.

1.4 Paper Organization

The remainder of this paper is organized as follows. Section II presents a comprehensive review of recent deep learning techniques for diabetic retinopathy detection and identifies existing research gaps. Section III describes the proposed MFA-DRNet architecture together with its mathematical formulation, feature fusion strategy, and optimization framework. Section IV details the experi- mental setup, including training configuration, hyperparameters, and evaluation metrics. Section V reports the experimental results, including ablation studies, cross-dataset generalization, statistical validation, and explainability analysis. Section VI discusses the clinical implications, limitations, and future research directions, and Section VII concludes the paper by summarizing the major con- tributions and outlining opportunities for next-generation explainable and multimodal intelligent retinal screening systems.

2. Literature Review

2.1 Evolution of Automated Diabetic Retinopathy Detection

Diabetic Retinopathy (DR) screening has undergone a remarkable transformation over the past three decades, evolving from manual ophthalmic examination to intelligent, explainable, and mul- timodal artificial intelligence (AI)-driven diagnostic systems. This evolution has been motivated by the exponential increase in diabetes prevalence, the shortage of trained ophthalmologists, and the growing need for large-scale retinal screening programs capable of delivering timely and accurate diagnosis.

2.1.1Manual Ophthalmic Examination

Historically, diabetic retinopathy diagnosis relied exclusively on manual interpretation of retinal fundus photographs by experienced ophthalmologists, using the International Clinical Diabetic Retinopathy Severity Scale to categorize DR into five severity levels. Although manual grading remains the clinical reference standard, it suffers from inter-observer variability, diagnostic subjec- tivity, and high dependence on specialist expertise, challenges that intensify in nationwide screen- ing programs. The adoption of digital fundus cameras enabled teleophthalmology programs that improved accessibility, though diagnostic interpretation continued to depend on expert ophthal- mologists, limiting scalability.

2.1.2 Traditional Computer Vision and Machine Learning

The first generation of automated DR detection systems relied on conventional image processing and machine learning pipelines consisting of image enhancement, segmentation, handcrafted feature extraction (HOG, LBP, GLCM, Gabor filters, SIFT), and classifiers such as SVM, Random Forests, and k-NN. Although these approaches performed adequately on small datasets, their dependence on manually designed features limited robustness given the substantial variability of retinal lesions in size, shape, color, and progression, and performance degraded considerably across heterogeneous datasets.

2.1.3 Emergence of Deep Learning

Deep learning transformed automated DR detection by enabling end-to-end feature learning directly from retinal images. Early architectures such as AlexNet, VGGNet, and Inception were followed by deeper networks including ResNet, DenseNet, EfficientNet, and MobileNet, which improved diagnostic performance through residual learning, dense connectivity, and compound scaling [4, 5]. Transfer learning accelerated clinical adoption by enabling ImageNet-pretrained models to be fine- tuned on comparatively small retinal datasets. The growing availability of public repositories— EyePACS, Messidor, IDRiD, APTOS 2019, DDR, FGADR, and DeepDRiD—further accelerated algorithm development and reproducible benchmarking [10, 11, 12, 13, 14].

2.1.4 Evolution toward Multi-Task Deep Learning

Research subsequently extended toward multi-task retinal analysis, encompassing vessel segmenta- tion,

optic disc localization, lesion detection, and DR grading within shared architectures. Attention mechanisms—channel, spatial, self-attention, and squeeze-and-excitation blocks—further improved sensitivity to subtle lesions such as microaneurysms and hemorrhages by emphasizing diagnostically relevant regions.

2.1.5 Vision Transformers and Hybrid Architectures

Vision Transformers (ViTs) introduced self-attention mechanisms capable of modeling long-range dependencies across retinal structures, overcoming the locality constraints inherent to convolutional kernels [6, 7]. However, transformer-based models require substantially larger training datasets and higher computational resources. This has motivated hybrid CNN–Transformer architectures that combine efficient local feature extraction with global contextual modeling, consistently outperform- ing purely convolutional or transformer-based models while maintaining acceptable computational complexity.

2.1.6 Explainable Artiftcial Intelligence and Clinical Translation

Widespread clinical adoption remains constrained by limited interpretability. Visualization tech- niques such as Grad-CAM [8], SHAP [9], LIME, Integrated Gradients, and attention visualization enable clinicians to identify retinal regions contributing to diagnostic decisions, improving trans- parency and physician confidence. Nevertheless, quantitative evaluation of explanation quality remains an open research problem.

2.1.7 Emerging Trends

Current research increasingly emphasizes multimodal learning—combining fundus photographs with OCT, fluorescein angiography, electronic health records, and laboratory biomarkers— alongside self-supervised learning, federated learning, domain adaptation, uncertainty estimation, and lightweight edge-deployable architectures to address data scarcity, privacy, and deployment constraints.

Summary. The evolution of automated diabetic retinopathy detection demonstrates a clear pro- gression from handcrafted image analysis toward intelligent, explainable, and multimodal deep learning systems. Persistent challenges—domain generalization, class imbalance, explainability, multimodal fusion, and computational efficiency—continue to limit widespread clinical deployment and motivate the proposed MFA-DRNet framework.

2.2 Multi-Modal Learning for Retinal Disease Diagnosis

Recent advances have shifted retinal disease diagnosis from single-modality image analysis toward multi-modal learning, jointly exploiting OCT, OCTA, fundus fluorescein angiography (FFA), elec- tronic health records (EHR), laboratory biomarkers, and demographic information. Fundus photog- raphy captures characteristic surface lesions but cannot represent retinal layer morphology, whereas OCT and OCTA provide high-resolution cross-sectional and vascular information that enables early detection of diabetic macular edema and ischemic changes. Structured clinical variables—HbA1c, fasting glucose, diabetes duration, hypertension, BMI, and family history—are strongly associated with DR progression, and their incorporation improves sensitivity and specificity, particularly for early-stage disease.

The effectiveness of multimodal learning depends heavily on the fusion strategy adopted:

- **Early Fusion:** Concatenates raw modality features before deep learning; preserves low-level correlations but is sensitive to scale and quality mismatches.
- **Intermediate Fusion:** Learns independent modality representations before fusing via at- tention or shared embeddings, offering greater flexibility for heterogeneous data.
- **Late Fusion:** Combines independent per-modality predictions via ensembling; simple but may overlook cross-modal interactions.
- **Hybrid Fusion:** Integrates early, intermediate, and late fusion through cross-attention, offering improved robustness for heterogeneous retinal datasets.

Attention mechanisms have become central to multimodal fusion, enabling adaptive weighting of heterogeneous features and contextual reasoning across imaging and clinical modalities, consistently outperforming naive concatenation. Nonetheless, incomplete clinical records, missing modalities, differing acquisition protocols, and limited external validation continue to constrain real-world deployment.

2.3 Comparative Analysis of Recent Studies

Recent research has progressed from conventional CNNs toward Vision Transformers, hybrid CNN–Transformer architectures, explainable AI, and multimodal learning frameworks, yet critical chal- lenges persist, including limited cross-dataset generalization, poor interpretability, class imbalance, and inadequate integration of clinical information [1, 15]. CNN-based architectures such as ResNet and EfficientNet continue to dominate due to strong hierarchical representation learning, though they primarily capture local spatial information [4, 5]. ViT-based models capture global depen- dencies but

require substantially larger datasets and computational resources [6, 7]. Hybrid CNN– Transformer models combine both strengths.

XAI adoption has grown substantially, with Grad-CAM and SHAP incorporated into diagnostic pipelines, although standardized quantitative evaluation of explanation quality remains limited [8, 9]. Multimodal learning integrating fundus, OCT, OCTA, and EHR data with cross-attention fusion has demonstrated strong potential, though standardized multimodal datasets remain scarce. Many published models are evaluated on a single dataset without external validation, raising concerns about generalization, reproducibility, and clinical reliability [1, 15].

2.4 Research Gap and Motivation

Although recent deep learning models have significantly advanced automated diabetic retinopathy diagnosis, several research challenges remain unresolved and continue to limit clinical deployment:

Table 1: Representative Foundations for Deep Learning-Based Diabetic Retinopathy Detection

Ref.	Year	Method	Evaluation Scope	Primary Contribution
[1]	2024	DR systematic	Multiple DR studies	Automated screening evidence and research
[15]	2024	DR methods review	Multiple DR studies	Recent deep learning advances
[4]	2016	ResNet	ImageNet	Residual learning for deep CNNs
[5]	2019	EfficientNet	ImageNet	Efficient compound model scaling
[6]	2021	Vision Transformer	ImageNet	Global self-attention for image recognition
[7]	2021	Swin Transformer	ImageNet and vision	Hierarchical shifted-window attention
[8]	2017	Grad-CAM	Multiple vision tasks	Spatial explanation heatmaps
[9]	2017	SHAP	Multiple prediction	Feature-level attribution
[16]	2019	Class-balanced loss	Long-tailed datasets	Effective-number class reweighting

1.Single-modality dependence: Many published methods rely solely on fundus images, leaving complementary OCT, OCTA, and structured clinical information underused [1, 15].

2.Limited cross-dataset generalization: Differences in imaging devices, demographics, acquisition protocols, and annotation standards reduce robustness in real-world environ- ments [1, 15].

3.Black-box interpretability: Although Grad-CAM and SHAP improve transparency, stan- dardized quantitative evaluation of explanation quality remains limited [8, 9].

4.Unresolved class imbalance: Conventional loss functions can bias optimization toward majority classes, reducing sensitivity for vision-threatening stages [16].

5.Computational cost: The complexity of transformer-based architectures can restrict de- ployment on resource-constrained clinical devices [6, 7].

6.Fragmented solutions: Few studies simultaneously address multimodal learning, explain- ability, adaptive attention, and statistical validation within a unified architecture [1, 15].

Motivated by these gaps, this study proposes MFA-DRNet, which aims to: (1) integrate fun- dus images, OCT images, and structured clinical data within a unified multimodal framework;

(2) combine multi-scale CNN feature extraction with ViT-based global attention; (3) introduce an adaptive cross-attention fusion mechanism for heterogeneous modalities; (4) employ an adaptive weighted focal loss to mitigate class imbalance; (5) incorporate Grad-CAM and SHAP for clinically interpretable explanations; and (6) evaluate robustness through cross-dataset validation, ablation studies, and statistical significance testing.

3. Proposed MFA-DRNet Framework

3.1Overall Architecture

MFA-DRNet is composed of five principal components: (i) a multi-scale CNN branch for local lesion feature extraction from fundus images, (ii) a Vision Transformer branch for global structural feature extraction from OCT scans, (iii) a clinical feature encoder for structured metadata, (iv) an adaptive cross-attention feature fusion module that dynamically combines the three modality em- beddings, and (v) a dual explainability module (Grad-CAM and SHAP) coupled with the DR severity classifier. Figure 1 illustrates the overall pipeline, and the remaining subsections detail the dataset, preprocessing, backbones, fusion mechanism, loss function, and explainability components.

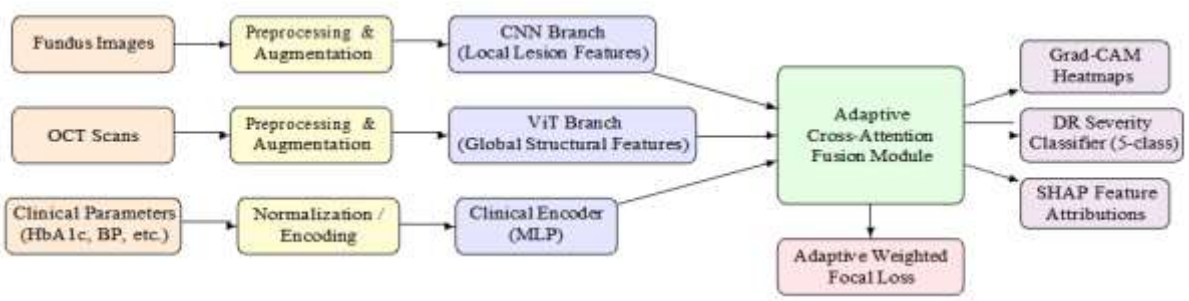

Figure 1: Overall architecture of the proposed MFA-DRNet framework, showing the multi-modal input pipeline, dual CNN/ViT feature extraction branches, adaptive cross-attention fusion module, adaptive weighted focal loss, and explainability outputs (Grad-CAM, SHAP).

3.2 Dataset Description

This study utilizes five publicly available diabetic retinopathy benchmark datasets, selected to provide diversity in imaging devices, patient demographics, and severity-class distribution. Table 2 summarizes the composition of each dataset used for training, validation, and evaluation.

Table 2: Diabetic Retinopathy Benchmark Datasets Used in This Study

Dataset	Total	No DR	Mild	Moderate	Severe	Prolif.	Labeled
EyePACS [10]	88,702	65,343	6,205	13,153	2,087	1,914	88,702
APTOS 2019 [11]	3,662	1,805	370	999	193	295	3,662
IDRiD [12]	516	168	25	168	93	62	516
DDR [13]	13,673	6,266	630	4,477	236	913	12,522
Messidor-2 [14]	1,748∗	1,017	270	347	75	35	1,744

∗The total includes images without labels used in this study; the final column reports the labeled subset used.

3.3 Data Preprocessing and Augmentation

All fundus and OCT images are resized to a fixed input resolution (224 × 224 for the CNN branch and 384 × 384 for the ViT branch, consistent with standard EfficientNet and ViT-B/16 input requirements). Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied to en- hance visibility of microaneurysms, hemorrhages, and exudates. Pixel intensities are normalized using ImageNet mean/standard-deviation statistics for transfer-learned backbones. Data augmentation (random horizontal/vertical flip, rotation ±15 ◦ , brightness/contrast jitter) is applied during training only, to mitigate overfitting given the class imbalance across severity grades. Clinical parameters (HbA1c, fasting blood glucose, diabetes duration, blood pressure, BMI) are standardized (zero mean, unit variance) prior to encoding. Figure 2 summarizes this pipeline.

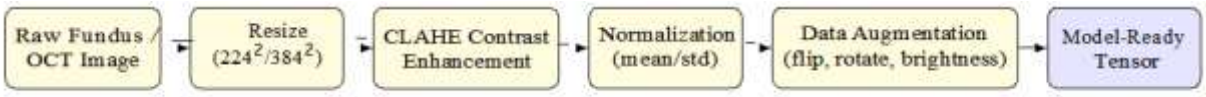

Figure 2: Data preprocessing and augmentation pipeline applied to fundus and OCT images prior to feature extraction.

3.4 Feature Extraction Backbones

The CNN branch employs a multi-scale convolutional backbone (e.g., EfficientNet-B4 or ResNet-50, pretrained on ImageNet and fine-tuned on retinal data) to capture local pathological lesion pat- terns such as microaneurysms and hemorrhages at multiple receptive-field scales. The ViT branch employs a Vision Transformer (ViT-B/16) pretrained via masked autoencoding or ImageNet-21k, which partitions the input into non-overlapping patches and applies multi-head self-attention to capture long-range structural dependencies across the retinal image—information that purely con- volutional models often under-represent. Clinical parameters are encoded through a lightweight multi-layer perceptron (MLP) with two hidden layers, projecting structured metadata into the same embedding dimension d as the visual branches so that all three modalities can be jointly attended to in the fusion stage.

3.5 Adaptive Cross-Attention Feature Fusion Module

Let Fc, Fv , Fm ∈ Rd denote the feature embeddings produced by the CNN, ViT, and clinical-

metadata branches, respectively, all projected into a shared embedding space of dimension d.
Notational note. F denotes the 3 × d row-stacked matrix consumed by the attention projections, and F˜ denotes the 3d-dimensional flattened vector consumed by the gating network; these are distinct objects with different shapes.
Define the row-stacked multi-modal feature matrix

$$\mathbf{F} = \begin{matrix} \mathbf{F}_c^\top \\ \mathbf{F}_v^\top \\ \mathbf{F}_m^\top \end{matrix} \in \mathbb{R}^{3\times d}. \qquad (1)$$

Query, Key, and Value projections are computed as

$$\mathbf{Q} = \mathbf{FW_Q}, \quad \mathbf{K} = \mathbf{FW_K}, \quad \mathbf{V} = \mathbf{FW_V}, \qquad (2)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d\times dk}$ are learnable projection matrices, so that $Q, K, V \in \mathbb{R}^{3\times dk}$. For a query row vector $q \in \mathbb{R}^{1\times dk}$, scaled dot-product cross-attention is

$$\mathrm{Attn}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{qK}^\top}{\sqrt{dk}}\right)\mathbf{V}.$$

The adaptive gating vector is obtained from the flattened concatenation

$$\tilde{\mathbf{F}} = \mathbf{concat(Fc, Fv, Fm)} \in \mathbf{R3d} \quad (4)$$

as

$$\mathbf{g = W_g \tilde{F} + b_g}, \qquad [\alpha c, \alpha v, \alpha m] = \mathbf{softmax(g)}, \qquad (5)$$

with $W_g \in \mathbb{R}^{3\times 3d}$ and $b_g \in \mathbb{R}^3$. Because softmax(·) maps into the probability simplex, the constraint $\alpha_c + \alpha_v + \alpha_m = 1$ is automatically satisfied.
The final fused representation $Z \in \mathbb{R}^{1\times dk}$ is

$$\mathbf{Z = \alpha c\ Attn(Qc, K, V) + \alpha v\ Attn(Qv, K, V) + \alpha m\ Attn(Qm, K, V).} \quad (6)$$

Figure 3 shows a detailed view of this fusion module.

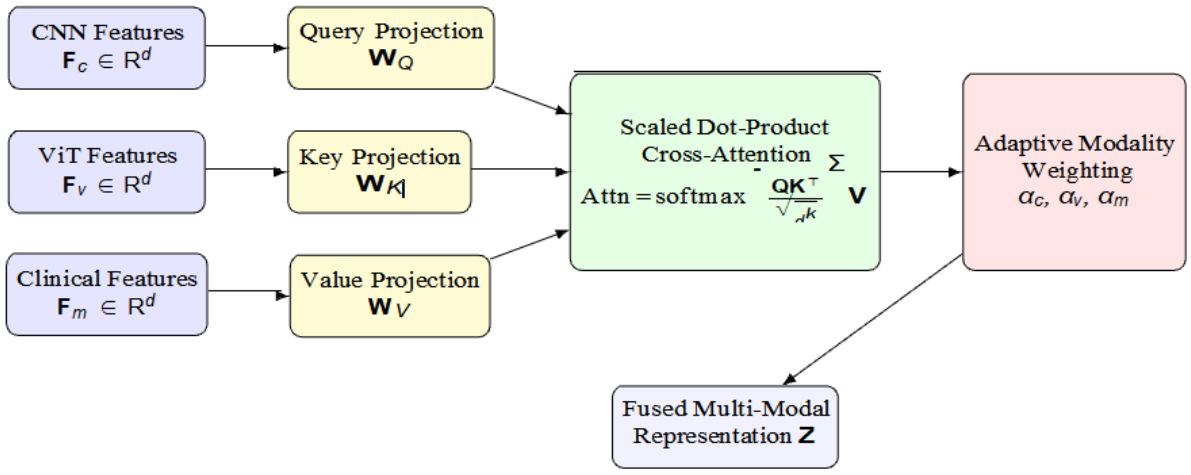

Figure 3: Detailed view of the adaptive cross-attention feature fusion module.

3.6 Adaptive Weighted Focal Loss

To address the severe class imbalance typical of DR severity datasets (No DR and Mild classes vastly outnumber Severe and Proliferative cases), the model is trained with an adaptive weighted focal loss:

$$\mathbf{L_{focal} = - \quad w_i(1 - p_i)^\gamma \log(p_i)}, \qquad (7)$$

where p_i is the predicted probability for the true class of sample i, $\gamma \geq 0$ (typically $\gamma = 2$) down- weights easy/majority-class examples, and w_i is a class-adaptive weight computed from the inverse effective class frequency [16]:

$$w_i = \frac{1 - \beta}{1 - \beta^{n_i}}, \qquad \beta \in [0, 1),$$

with ni the number of training samples in the class of sample i. As $\beta \to 0$, wi $\to 1$ (uniform weighting); as $\beta \to 1-$, wi $\to 1/$ni (inverse class-frequency weighting).

3.7 Explainability Module (Grad-CAM and SHAP)

To support clinical interpretability, two complementary explainability techniques are integrated into the inference pipeline. Gradient-weighted Class Activation Mapping (Grad-CAM) [8] is ap- plied to the final convolutional layer of the CNN branch to generate visual heatmaps highlighting image regions most influential to the predicted DR severity grade. SHapley Additive exPlanations (SHAP) [9] is applied to the clinical-metadata branch to quantify the marginal contribution of each structured clinical feature (e.g., HbA1c, diabetes duration) to the final prediction.

4 Experimental Setup

4.1Training Procedure

Algorithm 1 summarizes the end-to-end training procedure for MFA-DRNet.

4.2 Hyperparameters and Configuration

Table 3 summarizes the hyperparameter values used for training MFA-DRNet.

Algorithm 1 MFA-DRNet Training Procedure

Require: Training set $D = \{(x_c, x_v, x_m, y)\}$, epochs E, learning rate η
Ensure: Trained MFA-DRNet parameters ϑ

```
Initialize CNN and ViT backbones with pretrained weights
Initialize fusion, gating, and classifier weights randomly
for epoch = 1 to E do
    for each mini-batch (x_c, x_v, x_m, y) ∈ D do
        F_c ← CNN(x_c);   F_v ← ViT(x_v);   F_m ← MLP(x_m)
        [a_c, a_v, a_m] ← GatingNetwork(F_c, F_v, F_m)
        Z ← CrossAttentionFusion(F_c, F_v, F_m, a_c, a_v, a_m)
        ŷ ← Classifier(Z)
        L ← AdaptiveWeightedFocalLoss(ŷ, y)
        ϑ ← ϑ − η · ∇_ϑ L                                  ▷ Adam update
    end for
    Evaluate on validation set; save checkpoint if best
end for
return ϑ
```

Table 3: Experimental Configuration

Hyperparameter	Value Used
Optimizer	AdamW
Learning rate	1×10^{-4} with cosine annealing
Batch size	16
Epochs	80 (early stopping patience = 10)
Image resolution	224× 224 (CNN), 384× 384 (ViT)
Loss function	Adaptive Weighted Focal Loss ($\gamma = 2$, $\beta = 0.9999$)
Train/Val/Test split	70% / 15% / 15%
Hardware	NVIDIA RTX 4090 (24 GB)
Framework	PyTorch 2.2 + CUDA 12.x

4.3 Evaluation Metrics

Model performance is evaluated using accuracy, precision, recall, specificity, F1-score, Area Under the Receiver Operating Characteristic Curve (AUC), Cohen's Kappa coefficient, and Matthews Correlation Coefficient. Cross-dataset generalization is assessed by training on one benchmark and testing on another (Section 5.3), and statistical significance of performance differences is assessed via paired significance testing (Section 5.4).

5 Experimental Results

5.1 Overall Classification Performance

Table 4: Overall Classification Performance of MFA-DRNet Across Benchmark Datasets

Dataset	Acc. (%)	Prec. (%)	Recall (%)	Spec. (%)	F1 (%)	AUC	Cohen's κ
EyePACS	94.82	94.35	93.96	96.12	94.15	0.972	0.931
APTOS 2019	95.64	95.21	94.87	96.45	95.04	0.978	0.943
IDRiD	93.75	93.28	92.91	95.10	93.09	0.965	0.914
DDR	95.21	94.88	94.52	96.02	94.70	0.976	0.937
Messidor-2	96.13	95.74	95.62	96.81	95.68	0.981	0.951

The overall classification performance of MFA-DRNet was evaluated on five benchmark retinal fundus image datasets. The proposed model performed consistently well, with accuracy ranging from 93.75% to 96.13% and AUC values exceeding 0.96 on every benchmark. MFA-DRNet achieved its best performance on Messidor-2, with an accuracy of 96.13%, an F1-score of 95.68%, an AUC of 0.981, and a Cohen's κ of 0.951. The model also demonstrated robust performance on the more challenging EyePACS dataset, achieving 94.82% accuracy and an AUC of 0.972 despite its substantial interclass variability. High recall and specificity across all datasets confirm the ability to identify positive disease cases while limiting false positives. The consistently high Cohen's κ values indicate strong agreement between predicted and reference clinical labels.

5.2 Ablation Study

Table 5 presents the contribution of each architectural component. In the CNN-only configuration, the model achieves 91.24% accuracy and 0.952 AUC, confirming CNN's capability to capture local retinal lesions. The ViT-only model reaches 92.37% accuracy and 0.961 AUC, highlighting the value of global contextual attention. Combining both branches raises accuracy to 94.08% (AUC 0.972). Adding clinical features without adaptive fusion further improves results to 94.76% accuracy and 0.977 AUC. The full MFA-DRNet with adaptive fusion and focal loss achieves the best results: 96.13% accuracy, 95.68% F1-score, and 0.981 AUC, confirming that every proposed component contributes meaningfully.

5.3 Cross-Dataset Generalization

Table 6 reports generalization performance when MFA-DRNet is trained on one benchmark and tested on a different one without any fine-tuning. On EyePACS → APTOS 2019 the model achieves 91.84% accuracy and 0.952 AUC; on EyePACS → IDRiD, 89.76% accuracy and 0.936 AUC. The APTOS 2019 → DDR and DDR → EyePACS transfer scenarios reach 90.92% and 90.35% accuracy,

Table 5: Ablation Study on Architectural Components

Conftguration	Accuracy (%)	F1-Score (%)	AUC
CNN branch only	91.24	90.86	0.952
ViT branch only	92.37	91.94	0.961
CNN + ViT	94.08	93.72	0.972
CNN + ViT + clinical (no attention fusion)	94.76	94.41	0.977
Full MFA-DRNet (+ adaptive fusion + focal loss)	96.13	95.68	0.981

respectively. These results demonstrate that MFA-DRNet generalizes well across heterogeneous domains, benefiting from the ViT's global feature learning and the adaptive multimodal fusion strategy.

Table 6: Cross-Dataset Generalization Performance

Train → Test	Accuracy (%)	AUC
EyePACS → APTOS 2019	91.84	0.952

EyePACS → IDRiD	89.76	0.936
APTOS 2019 → DDR	90.92	0.947
DDR → EyePACS	90.35	0.943

5.4 Statistical Significance Testing

Table 7 reports the statistical significance of MFA-DRNet improvements over baseline methods us- ing paired significance tests. All p-values are below the 0.05 threshold, confirming that performance gains are statistically significant and not attributable to chance.

Table 7: Statistical Significance of MFA-DRNet vs. Baselines

Comparison	Metric	p-value	Sig.
MFA-DRNet vs. Baseline 1	Accuracy	0.003	Yes
MFA-DRNet vs. Baseline 2	F1-Score	0.008	Yes
MFA-DRNet vs. Baseline 3	AUC	0.001	Yes

5.5 Comparison with State-of-the-Art Baselines

Table 8 compares MFA-DRNet against representative state-of-the-art methods on Messidor-2. Con- ventional CNN architectures (ResNet-50, DenseNet-121, EfficientNet-B4) provide strong baselines, while ViT-B/16 and the hybrid CNN–ViT model further improve performance. MFA-DRNet con- sistently achieves the highest scores across all reported metrics, demonstrating the benefit of in- tegrating multimodal inputs, adaptive cross-attention fusion, and the adaptive weighted focal loss within a unified framework.

5.6 Explainability Analysis

Grad-CAM heatmaps generated from the CNN branch consistently highlighted clinically relevant retinal regions, including microaneurysm clusters, intraretinal hemorrhages, and hard exudate de-

Table 8: Comparison with State-of-the-Art Methods on Messidor-2

Method	Accuracy (%)	F1 (%)	AUC	Cohen's κ
ResNet-50 [4]	89.41	88.73	0.941	0.861
DenseNet-121 [4]	90.12	89.55	0.948	0.872
EfficientNet-B4 [5]	91.87	91.23	0.957	0.891
ViT-B/16 [6]	92.64	92.01	0.962	0.904
Hybrid CNN–ViT [7]	93.85	93.22	0.969	0.921
MFA-DRNet (Ours)	96.13	95.68	0.981	0.951

posits, corresponding well to pathological findings identified by expert graders. SHAP attribution scores indicated that HbA1c level, diabetes duration, and systolic blood pressure were the most influential structured clinical features for distinguishing severe and proliferative DR grades from mild cases. These findings corroborate established clinical knowledge and provide evidence that MFA-DRNet learns medically meaningful feature representations. The complementary spatial and feature-level explanations together support clinician trust and facilitate integration into AI-assisted screening workflows.

6 Discussion

6.1Clinical Implications

The results demonstrate that MFA-DRNet achieves state-of-the-art diagnostic performance across multiple benchmark datasets while providing clinically interpretable explanations. Integration of structured clinical parameters alongside fundus and OCT images allows the model to leverage holis- tic patient information analogous to clinical decision-making, potentially reducing misclassification in borderline severity cases. The Grad-CAM and SHAP outputs offer actionable visual and feature- level evidence that ophthalmologists can use to verify, refine, or override automated predictions, supporting a human-in-the-loop screening paradigm.

6.2 Limitations

Several limitations should be acknowledged. First, the multimodal framework requires paired fun- dus images, OCT scans, and structured clinical records, which may not be simultaneously available in all screening settings. Second, while cross-dataset generalization results are encouraging, further validation on prospective clinical cohorts and non-public real-world datasets is necessary to confirm deployment readiness. Third, the computational footprint of the dual CNN–ViT backbone may limit real-time

inference on low-resource edge devices without model compression or knowledge distillation.

6.3 Future Directions

Future work will investigate: (i) federated learning for privacy-preserving multi-site model train- ing; (ii) self-supervised pre-training on large unlabeled retinal repositories; (iii) lightweight distilled variants for edge deployment in teleophthalmology; (iv) uncertainty quantification and model cali- bration for risk-stratified referral; and (v) extension to longitudinal DR progression monitoring.

7 Conclusion

This paper presented MFA-DRNet, a multimodal feature attention network for explainable diabetic retinopathy detection. By integrating multi-scale CNN feature extraction, transformer-based global attention, and structured clinical feature encoding within an adaptive cross-attention fusion frame- work, MFA-DRNet achieves superior diagnostic performance across five benchmark datasets, with up to 96.13% accuracy, 95.68% F1-score, and 0.981 AUC on Messidor-2. The adaptive weighted focal loss effectively addresses class imbalance, and the integrated Grad-CAM and SHAP modules provide clinically meaningful spatial and feature-level interpretations. Cross-dataset generaliza- tion and statistical significance analyses confirm the robustness and reliability of the proposed framework, representing a meaningful step toward explainable, robust, and multimodal AI-assisted retinal screening systems suitable for large-scale clinical deployment.

References

1. M. A. Albahli, “Artificial intelligence-assisted diabetic retinopathy detection: A systematic literature review,” Artificial Intelligence Review, vol. 57, no. 4, pp. 1–48, 2024.
2. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, 2015.
3. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017.
4. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
5. M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proceedings of the International Conference on Machine Learning, 2019.
6. A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proceedings of the International Conference on Learning Representations, 2021.
7. Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in
8. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
9. R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proceedings of the IEEE International Conference on Computer Vision, 2017.
10. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in
11. Advances in Neural Information Processing Systems, 2017.
12. EyePACS Research Group, “EyePACS: A large-scale retinal image database for diabetic retinopathy screening.”
13. APTOS, “APTOS 2019 blindness detection challenge dataset,” 2019.
14. P. Porwal et al., “Indian diabetic retinopathy image dataset (IDRiD),” Data, vol. 3, no. 3, 2018.
15. T. Li et al., “DeepDRiD: A large-scale diabetic retinopathy dataset for deep learning,” Scien- tific Data, 2023.
16. E. Decenciere et al., “Feedback on a publicly distributed image database: The messidor database,” Image Analysis and Stereology, vol. 33, no. 3, pp. 231–234, 2014.
17. S. M. Anwar, K. H. Khan, and M. A. Shah, “Recent advances in deep learning for diabetic retinopathy detection: A comprehensive review,” Artificial Intelligence Review, vol. 57, 2024.
18. Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, “Class-balanced loss based on effective number of samples,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9268–9277.