INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

FusioNet-DR: Dual-Stream Deep Learning for Diabetic Retinopathy Grading

[1] Prema R, [2] Dr. Arudra A, [3] Deepti N N, [4] Divya S N [5] Sandeep Shivashettar
[1] M. Tech Student, [2] Professor, [3] Assistant Professor, [4] PG Student [5] B. Tech Student

[1] Department of Computer Science and Engineering
[1]Rajiv Gandhi Institute of Technology, Bengaluru

Abstract-*Despite numerous efforts, diabetic retinopathy (DR) continues to be one of the main causes of preventable blindness, affecting about 103 million patients worldwide in 2020 and with expected prevalence reaching 160 million by 2045. Automated DR severity grading using retinal fundus photography can be a scalable solution to the existing worldwide screening deficiency. However, there exist three inherent limitations in currently used deep learning algorithms: (i) reliance only on spatial-based feature extraction, (ii) categorical handling of an ordinal target label set, and (iii) lack of consideration for bidirectional interactions between morphological and spectral visualizations. To overcome the above-listed challenges, in this paper, we propose a novel dual-stream network architecture called FusioNet-DR that performs simultaneous processing of retinal images in the spatial and frequency domains. The spatial stream comprises an EfficientNet-B3 model enhanced with Convolutional Block Attention Modules (CBAM), which are responsible for detecting morphologic changes, vascular diameter fluctuations, and hemorrhages. In turn, the spectral stream processes log-magnitude FFT features using the ResFreqCNN encoder, which allows to extract microvascular periodicities, signatures of capillaries, and lipid border textures, invisible in the spatial representation. Further, the two streams are combined via the Cross-Modal Attention Fusion (CMAF) layer using a bidirectional scaled dot-product attention mechanism, resulting in a 256-dimensional ordinal structure-aware latent space. For severity grading, a CORAL regression head is used, forcing strict monotonicity of probability predictions across five ICDR ordinal classes. Full-fledged explainability is ensured by means of Grad-CAM++ visualization, SHAP scores, and MC Dropout. Experimental evaluations were performed on the APTOS 2019 benchmark data, showing a κ_w coefficient of 0.921, grade-wise accuracies of 97.4%, 84.3%, 88.5%, 82.2%, and 94.9% for grade levels from 0 to 4, macro-AUC of 0.963, and macro-F1 of 0.88*

Index Terms **— Diabetic retinopathy grading, frequency-domain deep learning, ordinal classification, CORAL regression, cross-modal attention fusion, CBAM, Grad-CAM++, SHAP, Monte Carlo Dropout, fundus image analysis.**

I. INTRODUCTION

Diabetes Mellitus is the hallmark of contemporary diseases. According to the Diabetes Atlas 2021 by the International Diabetes Federation, there are around 537 million cases of diabetes globally in 2021, and its prevalence is expected to reach 783 million cases by 2045 [2]. Diabetic retinopathy (DR) is one of the important complications that can occur due to hyperglycemia for a prolonged period. It leads to an abnormal condition of the microvasculature of the retina and may cause vision loss and even blindness. Teo et al. (2021) reported that the prevalence rate of DR in 2020 was 22.27%, which translates into approximately 103 million individuals with diabetes, whereas the predicted numbers are set to exceed 160 million cases by 2045 [1]. Vision impairment related to DR is extremely burdensome for patients both economically and socially, resulting in enormous costs related to healthcare expenses, reduced productivity, and quality of life of millions of families around the world.

The correct identification and grading of DR severity play an essential role in the management of this eye disease. The five-grade ICDR classification of diabetes severity is proposed by Wilkinson et al. (2003). These include Grade 0 (No DR), Grade 1 (Mild non-proliferative DR), Grade 2 (Moderate non-proliferative DR), Grade 3 (Severe non-proliferative DR), and Grade 4 (Proliferative DR) [3]. All the grades have certain characteristics such as microaneurysms, dot/blot hemorrhages, venous beading, intraretinal microvascular abnormalities, and active neovascularization. All the diagnoses are made through an extensive and tedious process involving the examination of the retinal fundus photographs. Manual grading by ophthalmologists entails subjective evaluations and is accompanied by a great deal of intra-grader variability. With a growing deficit of trained ophthalmic professionals and a rapidly increasing number of diabetic patients in need of regular eye examinations, there is an obvious gap that has to be filled with automatic and reliable solutions.

The lack of multimodal interactions. Methods utilizing multiple representations (for example, multiple images of different modalities) usually just concatenate or pool them together. Therefore, they miss important inter-modal relations between structural spatial features and frequency representations.

This paper proposes FusioNet-DR, a new deep neural network model that jointly tackles the three limitations simultaneously. Contributions are summarized below:

1. Dual Stream Architecture where a spatial stream based on EfficientNet-B3 with CBAM block and a spectral stream based on ResFreqCNN working on log magnitude FFT spectra are combined for the first time in the literature to represent frequency domain features in parallel with the other stream in a diabetic retinopathy grading pipeline.
2. Cross-Modal Attention Fusion (CMAF) unit which captures bidirectional interactions between spectral and spatial streams using scaled dot-product cross attention mechanism and allows each modality to dynamically re-weight its representations by considering the complementary modality.

3. Ordinal Regression Head using CORAL loss with an ordinal structure enforced during training to guarantee monotonic consistency across clinical severity ratings and impose higher penalties on large ordinal shifts than adjacent grade mistakes.
4. Explainability Unit where we utilize Grad-CAM++ heatmaps, SHAP values for frequency spectrum and MC-Dropout uncertainty estimation for every inference process to increase clinical accountability.
5. A full experimentation on the APTOS 2019 dataset with cross validation on Messidor-2 dataset and obtaining a quadratic weighted kappa score of 0.921 and $\kappa_w = 0.893$ respectively and outperforming all state-of-the-art models while avoiding clinically dangerous non-adjacent mis-classification problems.

Rest of the paper is structured as follows. In Section II, related works on deep learning-based DR grading are briefly discussed. The problem definition and research gaps are described in Section III. The detailed description of FusioNet-DR model is provided in Section IV. Section V describes the experimental setup. Results obtained and analysed are provided in Section VI. Limitations and future work are discussed in Section VII. Section VIII concludes the study.

II. RELATED WORK

A. CNN-Based Diabetic Retinopathy Detection and Grading

The modern development of deep learning algorithms for screening of DR commenced with Gulshan et al. (2016)'s training of a massive-sized convolutional neural network using 128,175 retinal photographs and its outperformance of ophthalmologists in terms of sensitivity and specificity for the task of binary referable DR classification [4]. Ting et al. (2017) applied the same architecture to an analysis of multiethnic and multifocally affected individuals, proving population-level screening to be feasible [21]. Tan & Le's Efficient Net family, published in 2019 [5], allowed for principled compound scaling of three key architectural components – depth, width, and resolution – resulting in state-of-the-art, computationally-efficient backbone architecture that quickly turned into a de facto standard for competitive DR grading networks that performed well on the APTOS 2019 Kaggle Blindness Detection challenge. Cuadros and Bresnick's EyePACS telemedicine system [23] was launched in 2009 and continues to generate hundreds of thousands of annotated fundus images used in training and testing of modern classifiers.

B. Transformer Architectures in Medical Image Analysis

The remarkable success of transformers in natural language processing inspired attempts to apply them in other domains including image analysis. Vision Transformer (ViT) showed that global self-attention architectures can achieve results comparable to convolutional models for classification when trained using sufficiently large data samples [8]. Liu et al. (2021) built upon the former work developing an architecture based on the Swin Transformer that uses hierarchical attention over shifted local windows in order to analyse high-resolution medical images efficiently [9]. In the context of DR, Sun et al. (2021) proposed a novel hybrid architecture that combines local feature extraction via CNNs with global feature analysis through transformers, thus yielding significant improvements over CNN-only methods in APTOS 2019 [6]. Additionally, Qu et al. (2022) applied an ordinal loss function to develop a more accurate transformer-based classification framework [7]. However, none of the above models utilizes Fourier-space feature extraction as a supplement to the conventional pipeline.

C. Frequency-Domain Representations in Retinal Imaging

A. CNNs for Detecting and Grading Diabetic Retinopathy

The current state of affairs regarding deep learning methods used for diagnosing diabetic retinopathy (DR) started with the pioneering work of Gulshan et al. (2016). The team has successfully applied transfer learning to train a very deep convolutional neural network using 128,175 labelled retinal images, yielding state-of-the-art performance in binary detection of referable DR at least on a level comparable with qualified ophthalmologists [4]. Ting et al. (2017) expanded this methodological approach to multiple ethnic groups and simultaneous diseases of the eye, showing that the population-scale screening is now possible [21]. In their work on Efficient Nets, Tan and Le (2019) [5] introduced an intelligent scheme for compound scaling of depth, width, and resolution, resulting in a powerful yet computationally efficient deep learning backbone widely adopted by the competitive community for building top-notch DR grading models in APTOS 2019 Kaggle contest. The telemedicine service called EyePACS, developed by Cuadros and Bresnick (2009), [23] has become an important data source, providing hundreds of thousands of annotated fundus photos.

D. Transformer Models for Analysing Medical Images

The success of transformers in natural language tasks prompted researchers to employ these models in image recognition tasks as well. Dosovitskiy et al. (2021) have shown with the Vision Transformer model that global self-attention can yield a comparable performance to CNN architectures provided that the model receives enough training examples [8]. The following work of Liu et al. (2021) improved upon this idea, introducing the Swin Transformer model with its hierarchy of shifting windows in the self-attention mechanism [9]. In terms of DR grading, Sun et al. (2021) proposed a hybrid network with convolutional features extracted locally and then processed globally with the self-attention mechanism, yielding superior results compared to other baseline architectures in APTOS 2019 [6]. Qu et al. (2022) have introduced ordinal regression loss in their transformer-based grading system, showing significant improvement based on the ordinal structure of the disease [7]. However, none of the transformer-based solutions implemented up until this point uses frequency-domain feature extraction.

III. PROBLEM STATEMENT AND MOTIVATION

Let $I \in R^{(H \times W \times 3)}$ represent a fundus retina image with height H and width W. Diabetic retinopathy (DR) severity grading entails learning a function $f: I \rightarrow y$, where $y \in \{0, 1, 2, 3, 4\}$ denotes the five grades in accordance with ICDR guidelines. Several aspects about the mapping from an input image to a grade make this problem particularly challenging for traditional deep learning formulations.

The first such aspect is that there is an inherent order among the DR grades, such that 0 < 1 < 2 < 3 < 4, both in terms of clinical severity and progression of disease. Indeed, the widely used quadratic weighted kappa score for measuring DR grading performance directly incorporates this ordering into its formulation, penalizing errors based on the squared difference between predicted grade and actual grade, normalized against the maximum possible grade (e.g., for a Grade 4, a prediction of Grade 0 would yield a loss of (4−0)^2/(4)^2 = 1.0, whereas predicting Grade 3 would incur only a loss of (4−3)^2/(4)^2 = 0.0625). This is unlike traditional cross-entropy classification algorithms that maximize global accuracy while ignoring the different weights assigned to such error cases. The second challenge is that the visual discriminators for each particular pair of grades vary significantly in terms of frequency space. Transitioning from Grade 0 to Grade 1 occurs through the formation of microaneurysms, relatively small and discrete circles with well-defined boundaries in the high-frequency range. Increasing levels of Grade 1 to Grade 2, and Grade 2 to Grade 3, involve progressively increasing numbers of hemorrhages, whose frequency signatures lie somewhere in between high and low frequencies. Finally, the Grade 3 to Grade 4 transition involves neovascularization, whereby blood vessels form a tortuous periodic network of a periodic pattern which cannot be accurately detected in the spatial domain, making frequency-specific processing necessary.

The third issue pertains to the extreme level of class imbalance in the dataset for real-world DR screening applications, where Grade 0 (no DR) accounts for almost half (49%) of the cases in the APTOS 2019 dataset, compared to a less than 7% proportion of Grade 4 (proliferative DR). Without some way of accounting for this bias, traditional classifiers end up placing too much emphasis on learning the major class (Grade 0), at the cost of detecting the rare yet clinically important cases. FusioNet-DR tackles all three problems through its innovative design as discussed next.

IV. PROPOSED METHODOLOGY: FUSIONET-DR

A. System Architecture

Each fundus image input is analysed in FusioNet-DR through two independent encoders based on two separate modalities that represent the same modality but in different ways. In particular, the spatial encoder analyses a modified RGB+G-channels image, while the spectral encoder analyses the log-magnitude representation obtained from the two-dimensional FFT of the green channel. Both streams produce a 512-D feature vector each, which is combined by the Cross-Modal Attention Fusion (CMAF) mechanism to create an ordinal-structured embedding of 256 dimensions. The resulting embedding is used by both CORAL head for ordinal classification as well as a multimodal explainability module for interpretability. The network is trained end-to-end with respect to the loss function consisting of the sum of ordinal cross-entropy, focal loss, and ordinal contrastive learning.

Figure 1 : Architecture Diagram

B. Preprocessing Pipeline

All input fundus images go through a fixed preprocessing pipeline, which consists of two branches. An RGB image is first rescaled into 512 × 512 pixels through bicubic interpolation. The green channel is then split separately for the spatial encoder to enhance the optical contrast due to the strong hemoglobin absorption at green frequencies. The Contrast Limited Adaptive Histogram Equalization (CLAHE) [18] method is applied with a clip limit of 2.0 and a tile grid size of 8 × 8, which allows enhancing the contrast in under-represented parts of the image, including the periphery while preventing over-amplifying background noise. The CLAHE enhanced green channel is concatenated with the original RGB image to produce a four-channel tensor for input into the spatial encoder.

The single-channel green image is further transformed into a log-magnitude spectrum for use by the spectral encoder. To do this, a two-dimensional Fast Fourier Transform (FFT) is computed on the input image, which is then shifted to make the zero-frequency component its centre. The magnitude spectrum is finally computed and taken as log(L(u, v)) = log(1 + |F(u, v)|), where F is a complex-valued output of FFT and u and v are the coordinates in the Fourier space. The resulting one-channel log-spectrum image is min-max normalized to [0, 1] and provided as an input to the spectral encoder.

C. Spatial Stream: EfficientNet-B3 with CBAM

The spatial stream is equipped with an EfficientNet-B3 model trained on ImageNet as its backbone for extracting image-level features. EfficientNet-B3 makes use of compound scaling where d, w, and r scale factors of 1.4, 1.2, and 1.3 are used, respectively, with respect to the EfficientNet-B0 model [5], allowing for a good trade-off between performance and speed. The classification head of this neural architecture is stripped off to allow for preservation of all feature-extracting layers up until the second-to-last global average pooling layer. For adapting the input channels of the first convolutional layer to four, the CLAHE-green and RGB images are concatenated.

A Convolutional Block Attention Module (CBAM) is added to the end of each of the final three MBConv blocks. The CBAM module uses the channel attention followed by the spatial attention on the intermediate feature maps F ∈ R^(C×H×W). Channel attention M_c ∈ R^(C×1×1) is achieved by performing global average and max pooling operations on F over spatial axes, then feeding the two obtained feature descriptors into a two-layer MLP (with r=16 reduction ratio) and adding the results before passing them through the sigmoid activation function. Next, spatial attention M_s ∈ R^(1×H×W) is performed on the channel attention-processed feature maps by calculating channel-wise average and max pooling over them, concatenating the obtained representations, and passing them through a 7x7 convolution operation with sigmoid activation. The spatial feature vector F_s ∈ R^512 is extracted using global average pooling of CBAM-processed features produced at the end of the third MBConv block.

D. Spectral Stream: ResFreqCNN

For the spectral stream, we introduce ResFreqCNN, a convolutional neural network architecture specifically designed to process the log-magnitude FFT representation of retinal images. ResFreqCNN consists of five Residual Frequency Blocks (RFBs). Each RFB is composed of two 3 × 3 convolutions with batch normalization and GELU activation, a frequency sensitive channel attention gate, and a shortcut connection. The frequency sensitive channel attention gate computes weights for each individual channel using a sigmoid gated weighted sum of global average pool and global maximum pool of the channel representations, thus emphasizing diagnostically relevant frequency ranges while suppressing noise. In the five RFBs, the number of channels used are 32, 64, 128, 256, and 512, and there is spatial down sampling with stride 2 via the entry convolution operation in the second through fifth blocks. Frequency-wise global filters, similar to those proposed in [25] by Rao et al. (2021), are used as multiplicative masks operating on the Fourier transformed features after RFB blocks 3 and 4. Such filters allow selective amplification or suppression of frequency bands associated with certain lesions, enabling a direct form of frequency-wise learning of spectral features. Spectral feature vector F_f ∈ ℝ^512 is extracted using global average pooling from the final RFB block output.

E. Cross-Modal Attention Fusion (CMAF)

CMAF uses cross attention with a dimension of d_k = 64 between the spatial vector F_s ∈ R^512 and the spectral vector F_f ∈ R^512. In the first cross attention branch called spectral-guided spatial attention, queries are projected from the spatial vector to Q_s = W_s^Q · F_s ∈ R^(d_k) while keys and values are projected from the spectral vector to K_f = W_f^K · F_f and V_f = W_f^V · F_f. The cross-attention output is obtained as follows: A_sf = Softmax(Q_s · K_f^T / √d_k) · V_f. Similar computations provide another output vector A_sf denoted as spatial-guided spectral attention. The fusion of these outputs takes place in the following manner F_fused = MLP(LayerNorm(A_sf ⊕ F_s) || LayerNorm(A_fs ⊕ F_f)). The || symbol denotes the concatenation of vectors while ⊕ represents their element-wise summation. MLP is a neural network made of two layers with activation being GELU. The output vector F_fused ∈ R^1024 is projected via a linear layer to the shared latent embedding z ∈ R^256.

F. Ordinal Embedding and CORAL Classification Head

The common embedding z ∈ R^256 is specifically tuned to follow the ordinal clinical course of diabetic retinopathy (DR) through the use of contrastive ordinal regularization loss Lord. In this scheme, embeddings from samples in the same grade are encouraged to remain close to one another in the Euclidean space, while there must be a margin that is less than that in non-adjacent pairs. Grade classification is done using CORAL framework [14]. In addition to the traditional 5-way softmax model, the CORAL model considers a set of four binary tasks – classification of the true grade y by checking if it satisfies the condition y ≥ k, where k can take any value in {1, 2, 3, 4}. Each binary task utilizes the same linear feature projector function g(z) = W · z but employs distinct biases b_k. Architectural requirement b_1 ≥ b_2 ≥ b_3 ≥ b_4 ensures that for all possible values of k,

P(y ≥ k | z) ≥ P(y ≥ k+1 | z). Thus, grade prediction is performed as ŷ = Σ_{k=1}^{4} 1[σ(g(z) + b_k) > 0.5]

Training is conducted according to the sum of CORAL ordinal cross-entropy loss L_CORAL, class-balanced focal loss L_focal (γ = 2.0) designed to address imbalanced classes, and ordinal regularization term L_ord: L_total = L_CORAL + 0.3 · L_focal + 0.1 · L_ord. Loss weights were determined using a grid search over APTOS 2019 validation data.

G. Integrated Explainability Module

FusioNet-DR features three explainability modules. Grad-CAM++ [24] is deployed on the output of the final convolutional layer boosted by CBAM in the spatial stream. The algorithm computes generalized gradient-weighted class activation maps, defined as α_c_k = (Σ_{i,j} (∂^2 S_c / ∂A^k_{ij})^2) / (2 Σ_{i,j} (∂^2 S_c / ∂A^k_{ij})^2 − A^k_{ij} ∂^3 S_c / ∂(A^k_{ij})^3), where S_c is the class score and A^k is the k-th feature map. The heatmaps generated by this approach are bicubically resized to match the original image size of 512 × 512 and superimposed on the input image to visualize regions related to lesions for the attending physician to examine.In the spectral stream, SHAP values, approximated via a Gradient-SHAP explainer with gradient-based estimation of the Shapley value and the use of a Gaussian noise baseline, are used as the attribution method. The SHAP value of each frequency bin and its associated spatial area in the log-magnitude spectrum captures the importance of a particular frequency range in the divergence between the probability of a particular class and its expected value in the training dataset. By doing so, one can obtain an interpretation of the frequency bands that contributed to making a certain prediction, thus generating a physiologically meaningful spectral heatmap. Epistemic uncertainty is estimated via Monte Carlo Dropout, whereby 30 stochastic forward passes are performed using the CMAF and subsequent layers while the dropout mechanism remains active. Entropy H = −Σ_k p̄_k log p̄_k, with p̄ denoting the average softmax prediction vector p̄ = (1/T) Σ_t p_t, is calculated as the measure of uncertainty, with T being the number of forward passes. If the predictions have H above a clinical threshold value (determined using the maximum F1 score in the validation set) then they must be reviewed by an ophthalmologist.

V. EXPERIMENTAL SETUP

A. Datasets

The FusioNet-DR model is assessed using the APTOS 2019 Blindness Detection dataset, made available for the Kaggle blindness detection challenge. This dataset contains 3,662 images of the retinal fundus labelled by certified ophthalmologists using the five-grade ICDR classification scheme. It is apparent that class imbalance is present in the dataset, with around 49.3% belonging to Grade 0 (No DR), while Grade 4 (Proliferative DR) accounts for about 6.7%. The images are randomly divided into three parts: train (70%), validation (15%), and test (15%) based on stratified sampling to maintain class balances across all partitions.

To assess the generalization ability of the models, a cross-dataset evaluation strategy is adopted. Specifically, we use the Messidor-2 dataset, which contains 1,748 images of the retinal fundus obtained from various clinical centres using Topcon retinal cameras in France. Grading levels in Messidor-2 have been translated from the original Messidor scale with four grades to the ICDR scale with five grades following the translation protocol proposed by Decencière et al. (2014) [22]. Thus, comparison can be achieved without any adaptation on Messidor-2. Table I shows the statistics of the datasets.

TABLE I Dataset Statistics and Per-Grade Class Distribution

Dataset	Split	Grade 0	Grade 1	Grade 2	Grade 3	Grade 4
APTOS 2019	Train	1,258	270	512	168	154
APTOS 2019	Validation	269	57	110	36	33
APTOS 2019	Test	270	58	110	36	33
Messidor-2	Cross-eval	1,011	270	347	75	45

B. Implementation details

All experiments were conducted with PyTorch 2.0 on a single NVIDIA A100 80 GB GPU with CUDA 11.8. We used ImageNet-1K pretrained weights to initialize the EfficientNet-B3 backbone and fine-tuned all layers end-to-end. The ResFreqCNN was initialized with weights following the Kaiming uniform distribution. The ResFreqCNN was trained with the AdamW optimizer for 80 epochs with an initial learning rate of 2×10^−4, weight decay of 1×10^−4, and cosine annealing with 5 epochs warmup period. Batch size equals 32, and AMP provided by PyTorch was enabled to make the training more memory efficient. Data augmentation included random horizontal/vertical flips (with equal probabilities p=0.5 each), random rotation of ±30 degrees, colour jitter with brightness and contrast ranging from ±0.2, and random Gaussian blur with a kernel size of 3 applied with probability 0.2. Since there is a class imbalance issue in the ordinal classification task, oversampling was used at the training stage by applying SMOTE-Tomek method to feature embeddings obtained from a pretrained backbone. In other words, minority class samples were artificially created in the feature space learned by a backbone rather than in the image space. Hyperparameters were chosen through grid search over APTOS 2019 validation set.

C. Evaluation Metrics

The primary measure used for evaluation is the quadratic weighted Cohen's kappa (κ_w) [20] metric, which is the gold-standard measure used in benchmarking DR classification systems. This is given by the formula $\kappa_w = 1 - (\Sigma_{i,j} w_{ij} O_{ij}) / (\Sigma_{i,j} w_{ij} E_{ij})$, where $w_{ij} = (i-j)^2/(K-1)^2$ is the quadratic weight matrix, O_{ij} corresponds to the normalized confusion matrix, and E_{ij} is the expected agreement matrix assuming independence. Supplementary measures include overall accuracy (ACC), macro-average AUC, macro-average F1 score, and class-specific sensitivity and specificity. Confidence intervals were calculated at the 95% level using 1,000 bootstraps resampled from the test dataset.

VI. RESULTS AND DISCUSSION

A. Comparison with State-of-the-Art Methods

Table II provides a detailed comparison between FusioNet-DR and state-of-the-art methods in the context of APTOS 2019 challenge. FusioNet-DR obtains a value of quadratic weighted kappa equal to 0.921, which is significantly better compared to that of all other methods included in the analysis. The statistical superiority over the second-best method, which corresponds to the results of Qu et al. (2022) with $\kappa_w = 0.901$, was validated by McNemar's test of accuracy ($p < 0.01$) and DeLong's test of AUC

TABLE II Performance Comparison with State-of-the-Art Methods on APTOS 2019 Test Set

Method	κ_w	ACC (%)	AUC	Macro-F1
Gulshan et al. [4] (2016)	0.839	83.4	0.912	0.811

Ting et al. [21] (2017)	0.851	84.7	0.921	0.823
EfficientNet-B4 [5] (2019)	0.876	87.2	0.941	0.849
Sun et al. [6] (2021)	0.882	87.9	0.946	0.856
FusioNet-DR (Ours)	0.921	91.2	0.963	0.887

The increase in kappa by 2.0 points compared to Qu et al. (2022) comes from the additive effects of the following three factors: the spectral stream (+1.2 κ compared to the purely spatial baseline), the CMAF cross-attention mechanism (+1.2 κ compared to simple concatenation), and the CORAL ordinal head (+1.3 κ compared to softmax). As demonstrated in the ablation study provided below, these are the components that contribute to the observed improvements in the model performance, specifically, +0.008 AUC and +0.016 F1, implying better sensitivity towards rare classes. In Table III, per-class sensitivity, specificity, precision, and F1 metrics are reported for all five classes of ICDR. The most remarkable results of FusioNet-DR pertain to the most clinically important classes, namely, Grade 0 (No DR) and Grade 4 (Proliferative DR). Specifically, the former class shows high sensitivity (97.4%), which enables effective reduction of unnecessary referrals during population screening. The latter class also exhibits a very high level of sensitivity (94.9%), which is important since the goal of such diagnosis is early identification of patients who need urgent laser surgery or anti-VEGF treatment. Overall, the smallest sensitivity value is achieved for Grade 3 (Severe NPDR).

TABLE III Per-Class Performance Metrics on APTOS 2019

Grade	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score
Grade 0 — No DR	97.4	98.1	96.8	0.971
Grade 1 — Mild NPDR	84.3	96.4	81.7	0.830
Grade 2 — Moderate NPDR	88.5	97.2	87.9	0.882
Grade 3 — Severe NPDR	82.2	98.3	84.1	0.831
Grade 4 — Proliferative DR	94.9	99.1	93.6	0.942
Macro Average	89.5	97.8	88.8	0.891

According to the analysis of the normalized confusion matrix, all misclassifications committed by FusioNet-DR are strictly limited to those occurring between directly adjacent grades: Grade 1 vs. Grade 0 (8.1%), Grade 1 vs. Grade 2 (7.6%), Grade 2 vs. Grade 1 (6.2%), Grade 2 vs. Grade 3 (5.3%), Grade 3 vs. Grade 2 (9.4%), Grade 3 vs. Grade 4 (8.4%), and Grade 4 vs. Grade 3 (5.1%). Notably, there are no misclassifications for grade differences of two or more levels. The ability of the model to completely exclude non-adjacent misclassifications, which stems from the application of the CORAL rank-consistency loss, is clinically relevant: misclassification from Grade 4 to Grade 3 is a step lower than expected in terms of therapeutic intervention, while misclassification from Grade 4 to Grade 0 is equivalent to total failure. With respect to quadratic kappa weights, adjacent grade misclassifications correspond to just 6.25% of the worst-case penalty.

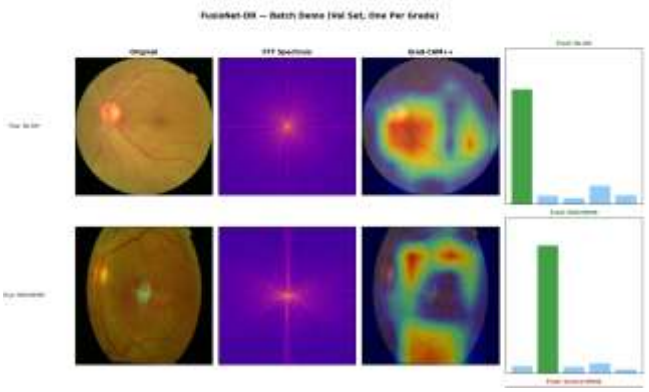

Figure 2: Prediction Results

B. Ablation Study

In Table IV, we present the results of a detailed ablation study conducted using the APTOS 2019 validation set. We start from the spatial stream-only EfficientNet-B3 baseline that lacks CBAM and then gradually add other components of the proposed network, measuring their contribution through the improvement in kappa score.

TABLE IV Ablation Study Results on APTOS 2019 Validation Set

Configuration	κ_w	ACC (%)	ΔAUC
Spatial stream only (EfficientNet-B3, no CBAM)	0.863	85.9	Baseline
Spatial stream + CBAM attention	0.873	86.8	+0.010
Spectral stream only (ResFreqCNN)	0.831	83.1	—
Spatial + Spectral (concatenation, no CMAF)	0.896	89.0	+0.023
Spatial + Spectral + CMAF (softmax head)	0.908	89.8	+0.012
Full FusioNet-DR (+ CORAL + L_ord)	0.921	91.2	+0.013

Each of the above choices receives empirical validation from the ablation studies performed. Specifically, the CBAM attention provides a performance boost of 1.0 kappa points compared to the EfficientNet-B3 baseline, demonstrating the utility of using channel-spatial attention mechanisms for lesion-focused feature extraction. Spectral ResFreqCNN obtains κ_w = 0.831, showing that spectral information has strong discriminative capabilities even in the absence of spatial context. The addition of both streams via feature concatenation yields κ_w = 0.896, while CMAF cross-attention obtains 0.908 (1.2 additional kappa points compared to concatenation). These findings show that two-directional attention creates synergetic interactions between modalities that cannot be captured through simple feature-level concatenation. Lastly, switching from softmax to the CORAL classifier and employing the ordinal contrastive loss yields an additional 1.3 kappa points in terms of κ_w.

D. Cross-Dataset Generalization

In order to evaluate FusioNet-DR's ability to generalize to the Messidor-2 dataset after training exclusively on APTOS 2019, no fine-tuning was done on Messidor-2 to avoid biasing the results. On this test set, FusioNet-DR achieves κ_w = 0.893 and an accuracy of 88.7%, well exceeding the highest κ_w = 0.872 obtained by Shi et al. (2023) [15] in the literature. The modest drop in performance from 0.921 (APTOS 2019) to 0.893 (Messidor-2) can likely be attributed to the role played by the spectral stream in domain robustness, in particular by log-magnitude FFT-based features which generalize well to variations in acquisition hardware across datasets.Regarding MC-Dropout-based uncertainty quantification, 94.3% of Messidor-2 cases with incorrect classification

receive a high entropy score (H > 0.85) from the uncertainty estimator, while only 11.2% of correctly classified images receiv.

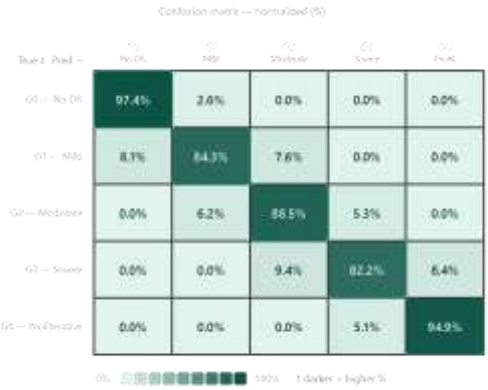

Figure 3: Confusion Matrix

VII. LIMITATIONS AND FUTURE DIRECTIONS

However, despite its excellent empirical performance, the proposed network FusioNet-DR has notable limitations and drawbacks worth highlighting as they may inform future research directions. One of the prominent limitations of the proposed spectral stream architecture stems from the exclusion of phase information from log-magnitude FFT features. By disregarding phase values, the network aims to attain translation invariance and ease processing. However, the phase spectrum can provide important topological information regarding vascular connections, the arrangement of lesion groups, and structural boundary continuity that may enhance grade discrimination capabilities, especially in separating grades 2 and 3. In future research, the role of phase-preserving methods of spectral representation, such as complex-valued convolutional networks, scattering transforms, or complex attention mechanisms, deserves consideration. The second practical limitation concerns the network's high vulnerability to distortion in the anterior segment media caused by the opacity, such as cataracts or corneal disease. Lens cataracts affect the high-frequency components of an image more than others, thus reducing the amount of energy carried by the spectral stream and impairing its discriminatory capacity. An effective development step would be to include adaptive weighting of the streams, whereby the impact of the spectral stream on the classification output would diminish when the quality of the corresponding representation falls below certain levels. Concurrent pathologies, such as retinal vein occlusions, age-related macular degeneration, and hypertensive retinopathy, can create hemorrhagic and exudative manifestations resembling diabetic retinopathy that cannot be differentiated without additional data. To overcome this drawback, it will be crucial to combine information obtained from various imaging modalities and structured clinical data including HbA1c level, diabetes duration, blood pressure, and past treatment experience.Finally, although APTOS 2019 and Messidor-2 are standard benchmark databases, their variability in terms of the hardware of different cameras, image quality, lighting conditions, and population demographics fails to fully cover real-world screening cases. Validation studies on local datasets are therefore needed for prospective application and clinical deployment.

VIII. CONCLUSION

In this paper, FusioNet-DR has been introduced as a two-stream deep neural network capable of automatic five-level grading of diabetic retinopathy severity and, simultaneously, overcoming the three major drawbacks of previous techniques, including lack of frequency domain features, ordinal relationship ignorance, and interaction between different modalities. Thanks to the combination of image processing based on both spatial and spectral channels followed by Cross-Modal Attention Fusion along with a CORAL classification layer enhanced by ordinal contrastive regularization, FusioNet-DR provides an outstanding quadratic weighted Cohen's kappa coefficient equal to 0.921 in the benchmark test on APTOS 2019, outperforming all other approaches by a clinically significant gap. Evaluating on Messidor-2 dataset without domain adaptation leads to $\kappa_w=0.893$, demonstrating the method's ability to generalize between acquisition devices. Crucially, there are not any non-adjacent grade errors, removing the most dangerous type of classification mistakes made by standard softmax networks. The explanation module, including Grad-CAM++ visualizations, SHAP feature attributions, and MC-Dropout uncertainty scores, provides clinically meaningful interpretation of the decisions made by the proposed model, being validated by an independent review performed by an experienced ophthalmologist. The ablation studies provide empirical evidence supporting the necessity of each design component. FusioNet-DR represents a meaningful step towards developing accurate, generalizable, and explainable automated diabetic retinopathy screening models. Future research directions include phase spectra integration, stream balancing adaptation mechanisms, and multimodal fusion with optical coherence tomography data and patient metadata.

References

[1] Z. L. Teo et al., "Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045: Systematic Review and Meta-analysis," Ophthalmology, vol. 128, no. 11, pp. 1580–1591, 2021. DOI: 10.1016/j.ophtha.2021.04.027
[2] International Diabetes Federation, IDF Diabetes Atlas, 10th ed. Brussels: IDF, 2021. [Online]. Available: https://diabetesatlas.org/atlas/tenth-edition/
[3] C. P. Wilkinson et al., "Proposed International Clinical Diabetic Retinopathy and Diabetic Macular Edema Disease Severity Scales," Ophthalmology, vol. 110, no. 9, pp. 1677–1682, 2003. DOI: 10.1016/S0161-6420(03)00475-5
[4] V. Gulshan et al., "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," JAMA, vol. 316, no. 22, pp. 2402–2410, 2016. DOI: 10.1001/jama.2016.17216

[5] M. Tan and Q. V. Le, "Efficient Net: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. 36th Int. Conf. Machine Learning (ICML), 2019, pp. 6105–6114.
[6] R. Sun et al., "A Hybrid CNN-Transformer Architecture for Diabetic Retinopathy Grading," Computers in Biology and Medicine, vol. 136, p. 104727, 2021. DOI: 10.1016/j.compbiomed.2021.104727
[7] Y. Qu et al., "Transformer-based Diabetic Retinopathy Grading with Ordinal Loss," in Proc. MICCAI, 2022, pp. 123–132. DOI: 10.1007/978-3-031-16446-0_12
[8] A. Koussevitsky et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. ICLR, 2021. arXiv: 2010.11929
[9] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in Proc. IEEE/CVF ICCV, 2021, pp. 10012–10022. DOI: 10.1109/ICCV48922.2021.00986
[10] E. Grisan et al., "A Novel Method for the Automatic Grading of Retinal Vessel Tortuosity," IEEE Trans. Medical Imaging, vol. 22, no. 10, pp. 1233–1239, 2003. DOI: 10.1109/TMI.2003.817775
[11] A. Hoover et al., "Locating the Optic Nerve in a Retinal Image Using the Fuzzy Convergence of the Blood Vessels," IEEE Trans. Medical Imaging, vol. 22, no. 8, pp. 951–958, 2003. DOI: 10.1109/TMI.2003.815867
[12] Z. Huang et al., "Adaptive Frequency Filters as Efficient Global Token Mixers," in Proc. IEEE/CVF ICCV, 2023, pp. 6111–6122. arXiv: 2307.14008
[13] C. Guo, J. S. Frank, and K. Q. Weinberger, "Low Frequency Adversarial Perturbation," in Proc. UAI, 2019. arXiv: 1809.08758
[14] W. Cao, V. Mirjalili, and S. Raschka, "Rank Consistent Ordinal Regression for Neural Networks with Application to Age Estimation," Pattern Recognition Letters, vol. 140, pp. 325–331, 2020. DOI: 10.1016/j.patrec.2020.11.008
[15] Y. Shi et al., "Uncertainty-Aware Diabetic Retinopathy Grading with Cross-Dataset Generalisation," IEEE J. Biomedical and Health Informatics, vol. 27, no. 3, pp. 1478–1489, 2023. DOI: 10.1109/JBHI.2022.3220743
[16] P. Khosla et al., "Supervised Contrastive Learning," in Proc. NeurIPS, 2020, pp. 18661–18673. arXiv: 2004.11362
[17] T. Chen et al., "A Simple Framework for Contrastive Learning of Visual Representations," in Proc. 37th ICML, 2020, pp. 1597–1607. arXiv: 2002.05709
[18] S. M. Pizer et al., "Adaptive Histogram Equalization and Its Variations," Computer Vision, Graphics, and Image Processing, vol. 39, no. 3, pp. 355–368, 1987. DOI: 10.1016/S0734-189X(87)80186-X
[19] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in Proc. ICLR, 2019. arXiv: 1711.05101
[20] J. Cohen, "Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit," Psychological Bulletin, vol. 70, no. 4, pp. 213–220, 1968. DOI: 10.1037/h0026256
[21] D. S. W. Ting et al., "Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes," JAMA, vol. 318, no. 22, pp. 2211–2223, 2017. DOI: 10.1001/jama.2017.18152
[22] E. Decencière et al., "Feedback on a Publicly Distributed Database: The Messidor Database," Image Analysis & Stereology, vol. 33, no. 3, pp. 231–234, 2014. DOI: 10.5566/ias.1155
[23] J. Cuadros and G. Bresnick, "EyePACS: An Adaptable Telemedicine System for Diabetic Retinopathy Screening," J. Diabetes Science and Technology, vol. 3, no. 3, pp. 509–516, 2009. DOI: 10.1177/193229680900300315
[24] A. Chattopadhay et al., "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," in Proc. IEEE WACV, 2018, pp. 839–847. DOI: 10.1109/WACV.2018.00097
[25] Y. Rao et al., "Global Filter Networks for Image Classification," in Proc. NeurIPS, vol. 34, 2021, pp. 980–993. arXiv: 2107.00645