



Vision Transformer Architecture for Simultaneous Detection and Severity Grading of Diabetic Retinopathy and Diabetic Macular Edema

Safa Sarakbi¹

ABSTRACT

Diabetic retinopathy (DR) and diabetic macular edema (DME) are two major causes of preventable blindness. Despite their great success in DR and DME detection or grading individually, few methods have explored automated simultaneous grading of DR and DME from fundus photographs in clinic settings. In this work, we introduce DualHead-ViT, a multi-task framework with a pretrained ViT-B/16 backbone and a novel cross-task attention fusion module for jointly grading the severity of DR and DME. We project two separate CLS token vectors for DR and DME, respectively, and design a cross-task attention module to interact features of DR and DME bidirectionally. We validated our model using the Indian Diabetic Retinopathy Image Dataset (IDRID), employing stratified 5-fold cross-validation alongside CLAHE-preprocessed images and a class-weighted loss with ordinal regularization. Across 5 levels of DR severity, we report mean quadratic kappa of 0.802 ± 0.010 and AUC of 0.831 ± 0.018 , and across 3 categories of DME risk, we report kappa of 0.759 ± 0.067 and AUC of 0.895 ± 0.026 . Our findings support the feasibility of joint transformer-based grading for DR and DME using a single computational architecture.

Article information:

Keywords: Vision Transformer, Diabetic Retinopathy Grading, Diabetic Macular Edema, Multi-task Learning, Cross-task Attention

Article history:

Received: February 20, 2026

Revised: June 12, 2026

Accepted: August 20, 2026

Published: September 26, 2026

(Online)

DOI: 10.37936/ecti-cit.2026204.266301

1. INTRODUCTION

Diabetes mellitus affects more than 537 million people globally [1, 2]. Diabetic retinopathy (DR) and diabetic macular edema (DME) are the leading causes of vision loss among diabetic microvascular complications [3, 4]. DR is graded on a scale of 5 stages from none to proliferative, while DME severity is graded independently into 3 risk categories related to its distance to the fovea [5]. About one-quarter to one-third of those with DR also have DME [6]. Automation of disease grading would help address the need for diabetic eye screening. Automation of disease grading is challenged by the fact that the gold standard assessment involves manually evaluating patients' fundus images by expert graders [7], which can be very subjective, time-consuming, and infeasible for the rapidly growing diabetic population [8, 9]. Deep learning approaches have shown great potential for automated retinal grading. CNN-based methods have already reached close-to-expert performance on public DR detection datasets [10, 11]. CNNs are inherently limited

in capturing global information from the entire retina, which is intrinsically related to grading task as lesions contributing to grading decisions can be spread out over the whole retinal area. Vision Transformer (ViT) has global self-attention mechanism which can in principle have better inductive bias to encode local details of pathology as well as global distribution [12, 13].

However, there are still three main limitations throughout the literature. Firstly, DR and DME have been examined nearly exclusively as separate classification tasks despite their known concurrent existence and feature space overlap [14, 15]. Secondly, no transformer-based framework leverages an explicit cross-task interaction module that enables pathologically correlated features from one task to aid in grading the other [16, 17]. Thirdly, class imbalance and validation robustness have yet to be appropriately addressed in transformer-based retinal grading pipelines [18, 19]. To address these limitations, our work contributes in three ways: (1) we present DualHead-

¹The author is with the Biomedical Engineering Department, Faculty of Mechanical and Electrical Engineering, Damascus University, Damascus, Syria, Email: safa.sarakbi@damascusuniversity.edu.sy

¹Corresponding author: safa.sarakbi@damascusuniversity.edu.sy

ViT, a shared multi-task ViT-B/16 model that jointly grades DR and DME; (2) we design a new cross-task attention fusion module to allow DR-DME feature sharing in both directions; and (3) we establish a robust training regime with CLAHE preprocessing, class-balanced ordinal loss, and stratified 5-fold cross-validation to arrive at clinically trustworthy and statistically sound results.

2. RELATED WORK

Transformers have recently gained popularity in deep learning models for automated retinal disease grading. This is likely because the global self-attention mechanism may be more compatible with fundus pathology than convolutional inductive biases. DR grading ViT-based model by Liu *et al.* [18] incorporated softmax and pooling operators into ViT framework which was experimented on the benchmark datasets APTOS2019 and Messidor with 10-fold cross-validation. Note that Messidor dataset also has annotations for DME present in it but they treated DR and DME classification tasks mutually exclusive. No mechanism was designed to share features between the tasks. The motivation behind introducing fused operator was to enhance patch-level discriminability but no such interaction between DR and DME tasks was done despite knowing that presence of both DR and DME co-exist pathophysiologically.

MVCAViT [20] by Vijayalakshmi *et al.* consists of a Multi-View Cross-Attention Vision Transformer framework that leverages both macula-centered and optic disc-centered pairs of fundus images. This framework merges CNN and ViT models, jointly grading DR severity and localizing lesions in a single multi-task pipeline. Their method for a cross attention module to combine the complementary information from different spatial views of fundus images is conceptually similar to task fusion explored in our work. In contrast to our work, MVCAViT performs multi-view fusion of two aspects of the same pathology rather than performing cross-task fusion between two co-existing but independent conditions. Additionally, their framework does not account for DME grading and was only validated on a single dataset without evaluation of cross-validated generalizability.

Zhang *et al.* [21] presented a detailed quantitative comparison of CNNs, vanilla ViTs, CvT, and LeViT on DR classification, using Grad-CAM [43] and Attention-Rollout [44] to visualize the interpretability of these networks on publicly available retinal datasets. The results of their study show that CNN-Transformer hybrids significantly outperform both pure CNNs and pure ViTs on DR grading datasets, establishing a new reference point for DR-specific transformer applications. However, only binary classification is used for DR detection (no differentiation into severities) or single-task classification,

respectively, for all five grades, and DME is excluded from the study. Another issue limiting the meaningfulness of the absolute performance numbers is that the datasets were only evaluated once, using a single train-test split.

Introduced by Bhoopalan *et al.* [22], Task-Optimized Vision Transformer (TOViT) employs layer-wise learning rate decay along with attention head pruning and embedding size collapsing for DR detection and grading into mild, moderate, and severe categories. Optimized for limited resource settings, TOViT shows promising improvement in inference efficiency while retaining accuracy when compared to other methods. A limitation of TOViT however is that it is not a multi-task network and thus cannot predict presence of DME alongside DR grading. Retinal image datasets used for grading tasks are imbalanced, which TOViT does not take into consideration using methods such as loss weighting or ordinal regularization. Taken together, these four works highlight the trend of investigating transformer-based retinal grading and identify three key unsolved problems across all studies: no concurrent DR and DME grading in one model, a leak of explicit mechanisms for cross-task feature interaction, and limited considerations for class imbalance and cross-validated robustness in transformer-based retinal imaging pipelines, which this work aims to address.

A directly comparable antecedent that the present work must acknowledge is the Cross-disease Attention Network (CANet) proposed by Li *et al.* (IEEE Transactions on Medical Imaging, 2020), which introduced a CNN-based multi-task framework for the concurrent grading of DR and DME using cross-disease attention mechanisms, evaluated on both the IDRiD and Messidor datasets. CANet establishes that joint DR/DME grading is not a novel concept; it is the most directly comparable prior system to DualHead-ViT in terms of disease scope and dataset overlap. The contribution of the present work lies in replacing the CNN backbone with a Vision Transformer encoder, introducing explicit bidirectional cross-task attention, and adopting a stratified 5-fold cross-validation protocol absent from CANet's evaluation, thereby advancing the transformer-based realization of joint grading rather than originating the joint-grading paradigm.

3. PROPOSED METHODOLOGY

The full pipeline of DualHead-ViT for the joint DR and DME grading task is illustrated in this section. There are four key stages involved: dataset preparation and preprocessing, architecture details, multi-task loss design, and training process.

3.1 Dataset and Preprocessing

The IDRiD dataset (Indian Diabetic Retinopathy Image Dataset) is used as our testbed. IDRiD is a

publicly available clinically annotated fundus photo dataset consisting of 516 retinal images with resolution of 4288×2848 pixels acquired by Kowa VX-10 α fundus camera (45 degrees field of view). Every image in the IDRiD dataset has two independent ground-truth labels. Five-class DR severity label grades DR into (No DR, Mild, Moderate, Severe, and Proliferative) and three-class DME risk label grades DME into (No DME, Moderate Risk, High-Risk). Therefore, this dataset is an excellent choice for our multi-task retinal grading study. IDRiD is significantly imbalanced for both tasks. Majority of the data belongs to No DR and No DME class respectively while Mild DR & Moderate DME grades are highly underrepresented. Clinical fundus images may have varying quality due to the capture process. Therefore, each image is passed through a preprocessing pipeline before being fed to the model, as shown in Table 1.

This preprocessing pipeline consists of two parts. The image is first masked with a circular aperture that matches the field of view of the retina. This removes the black borders around the image that contain no information but are present due to the camera lens. The second part of the preprocessing involves running Contrast Limited Adaptive Histogram Equalization (CLAHE) on the L channel of the LAB color space. CLAHE uses a clip limit of 2.0 and a tile grid size of 8×8 . Enhancement increases local contrast of areas such as retinal microvasculature, optic disc edges, and areas in the macula while avoiding noise amplification in bright areas of an image. As a result of this selective contrast increase, smaller lesions visible in early disease stages like microaneurysms and hard exudates are more easily detected, which aids in grading severity differences between neighboring DR and DME labels. Images are resized to 224×224 pixels as this is the input size for patch embedding used by ViT-B/16, and then pixel values are normalized with ImageNet mean and std.

3.2 Dual Head-ViT Architecture

We initialize DualHead-ViT with the pretrained ViT-B/16 backbone weights pretrained on ImageNet-21k since it shows the ability to learn intermediate visual representations that can be transferred to downstream medical imaging tasks. The ViT splits a 224×224 image into 196 non-overlapping patches of size 16×16 . Each image patch is flattened and linearly embedded into a 768-dimensional patch embedding. We append a trainable classification token (CLS token) to the patch embeddings and add fixed sinusoidal positional embeddings to maintain positional information across 12 encoder layers. For our experiments, we freeze the patch embedding layer and first six transformer encoder layers and allow fine-tuning of upper six blocks (allowing less trainable parameters), as depicted in Table 2.

After extracting backbone features, we separately

Table 1: Dataset and Preprocessing Configuration.

Component	Detail
Dataset	Indian Diabetic Retinopathy Image Dataset (IDRiD)
Total Images	516 fundus photographs
Original Resolution	4288×2848 pixels
Acquisition Camera	Kowa VX-10 α , 45° field of view
DR Classes (5)	No DR, Mild, Moderate, Severe, Proliferative
DME Classes (3)	No DME, Moderate Risk, High-Risk
Preprocessing Step 1	Circular retinal field masking
Preprocessing Step 2	CLAHE (LAB color space, luminance channel only)
CLAHE Clip Limit	2.0
CLAHE Tile Grid Size	8×8
Final Image Size	224×224 pixels
Normalization Mean	[0.485, 0.456, 0.406]
Normalization Std	[0.229, 0.224, 0.225]
Class Imbalance Issue	Yes — No DR and No DME majority; Mild DR and Moderate DME underrepresented

project the CLS embedding into task-specific feature spaces using separate DR and DME heads consisting of LayerNorm, followed by a fully connected linear layer with GELU activation and a dropout layer. This allows both tasks to independently process the shared backbone embedding for their own discriminative purposes before interaction. These stacked DR and DME feature tokens are then input to a novel cross-task attention fusion module consisting of a single 8-head multi-head self-attention layer, as well as a two-layer feed-forward network with residual connections around each sub-layer. This allows each task’s representation to attend to and integrate pathologically correlated information from the other task. The fused DR and DME representations are then each fed through their own three-layer classification heads, yielding final logits for 5-class DR and 3-class DME prediction, respectively.

3.3 Multi-Task Loss Function

The DualHead-ViT loss function is a weighted summation loss function that supervises both the DR and DME prediction heads on every forward pass. It takes gradients from both tasks and updates the shared backbone features accordingly. Let DR Loss = LDR and DME Loss = LDME then total Loss = $\alpha \cdot \text{LDR} + \beta \cdot \text{LDME}$ where $\alpha = 0.6$ and $\beta = 0.4$. Higher weight is assigned to DR Loss as DR grading is clinically more complex with 5 classes compared to only 3 classes for DME risk classification. Each grade-specific loss is further composed of two terms. The first term is the weighted cross-entropy loss with

Table 2: *DualHead-ViT Architecture Specifications.*

Component	Specification
Backbone	ViT-B/16 (Vision Transformer Base)
Pretraining Source	ImageNet-21k
Input Image Size	224×224 pixels
Patch Size	16×16 pixels
Number of Patches	196
Embedding Dimension	768
Transformer Encoder Blocks	12
Frozen Layers	Patch embedding + first 6 transformer blocks
Trainable Layers	Last 6 transformer blocks + all new heads
CLS Token Projection (DR)	LayerNorm $\rightarrow$ Linear(768,768) $\rightarrow$ GELU $\rightarrow$ Dropout(0.15)
CLS Token Projection (DME)	LayerNorm $\rightarrow$ Linear(768,768) $\rightarrow$ GELU $\rightarrow$ Dropout(0.15)
Cross-Task Attention Heads	8
Cross-Task FFN	Linear(768,1536) $\rightarrow$ GELU $\rightarrow$ Dropout $\rightarrow$ Linear(1536,768)
DR Classification Head	Linear(768,384) $\rightarrow$ GELU $\rightarrow$ Dropout $\rightarrow$ Linear(384,192) $\rightarrow$ Linear(192,5)
DME Classification Head	Linear(768,384) $\rightarrow$ GELU $\rightarrow$ Dropout $\rightarrow$ Linear(384,192) $\rightarrow$ Linear(192,3)
Dropout Rate	0.3

label smoothing of 0.1. Each class weight is the reciprocal of the class frequency within the training fold. This helps account for the extreme imbalance of the IDRiD dataset, as in Table 3.

Label smoothing mitigates assigning inflated probabilities to the minority classes and increases calibration with respect to the rare grades of severity. The second term added to each loss is an ordinal regression regularization loss that down-weights losses proportionally to the square of the ratio of the difference between the predicted grade and ground truth severity divided by the maximum grade distance. The ordinal penalty directly imposes the clinically relevant order between DR and DME grades into the loss surface. Larger magnitude severity mispredictions, e.g., predicting No DR when the true grade is Proliferative, are penalized more harshly than their adjacent magnitude counterparts, as they would be more detrimental in clinical practice. The overall multi-task loss function thus balances class imbalance, prediction calibration, and ordinal severity relations in one joint differentiable objective.

3.4 Training Strategy and Optimization

Training DualHead-ViT required consideration of three aspects: the domain shift between natural ImageNet pretrained data and our medical fundus im-

Table 3: *Multi-Task Loss Function Configuration.*

Component	Detail
Total Loss Formulation	$L = 0.6 \times L_{DR} + 0.4 \times L_{DME}$
DR Loss Weight	0.6
DME Loss Weight	0.4
Primary Loss Type	Class-weighted Cross-Entropy
DR Class Weights	Inverse frequency of each DR class in training fold
DME Class Weights	Inverse frequency of each DME class in training fold
Label Smoothing	0.1
Secondary Loss Type	Ordinal Regression Regularization
Ordinal Penalty Formula	$MSE(\text{predicted grade}, \text{true grade}) / (\text{max_grade} - 1)^2$
Ordinal Loss Weight	0.1 (applied to both DR and DME terms)
DR Ordinal Normalization	Divided by $(51)^2 = 16$
DME Ordinal Normalization	Divided by $(31)^2 = 4$
Purpose of Label Smoothing	Prevent overconfident predictions on minority classes
Purpose of Ordinal Penalty	Penalize large-magnitude severity misclassifications

ages, class imbalance of IDRiD, and catastrophic forgetting of backbone representations during training. We use a differential learning rate strategy assigning a backbone learning rate of $2e-5$ for the pretrained Vision Transformer encoder blocks and a much larger head learning rate of $1e-3$ for all newly initialized projection and classification layers. Cosine annealing with a linear warmup schedule is used for all parameter groups. Learning rate is warmed up for the first 3 epochs i.e., gradually increases from 0 to its max value, and then starts cosine decaying to a min lr of $1e-7$ for the rest of the 27 epochs out of a total training budget of 30 epochs. A weighted Random sampler is used at the DataLoader level to over-sample the minority class examples in a training batch.

This provides an additional data-level weight balance along with the loss function’s class weight balance. Augmentations used for training are `horizontal_flip`, `vertical_flip`, rotation up to 30 degrees, random affine translate/ scale, color jitter, random grayscale, and random erasing. These augmentations help generalize to the various ways a fundus image can be acquired. Gradient clipping with a max norm of 1.0 is used while training for the initial epochs of fine-tuning. The model with the highest validation quadratic kappa score is chosen for each fold. Finally, we report the mean and std. of model performance evaluated using stratified 5-fold cross-validation on the entire IDRiD dataset.

Table 4: Training Strategy and Optimization Hyperparameters.

Hyperparameter / Component	Value / Description
Total Training Epochs	30
Batch Size	16
Backbone Learning Rate	2×10^{-5}
Classification Head Learning Rate	1×10^{-3}
Minimum Learning Rate	1×10^{-7}
Weight Decay	1×10^{-4}
LR Schedule	Cosine Annealing with Linear Warmup
Warmup Epochs	3
Optimizer	AdamW
Gradient Clipping Norm	1.0
Class Rebalancing (Data Level)	Weighted Random Sampler
Augmentation 1	Random Horizontal and Vertical Flip
Augmentation 2	Random Rotation ($\pm 30^\circ$)
Augmentation 3	Random Affine (translate 0.1, scale 0.9–1.1)
Augmentation 4	Color Jitter (brightness, contrast, saturation, hue)
Augmentation 5	Random Grayscale ($p = 0.05$)
Augmentation 6	Random Erasing ($p = 0.25$, scale 0.02–0.20)
Model Selection Criterion	Best validation quadratic kappa per fold
Validation Strategy	Stratified 5-Fold Cross-Validation
Random Seed	42
Hardware Target	CUDA GPU (Google Colab)

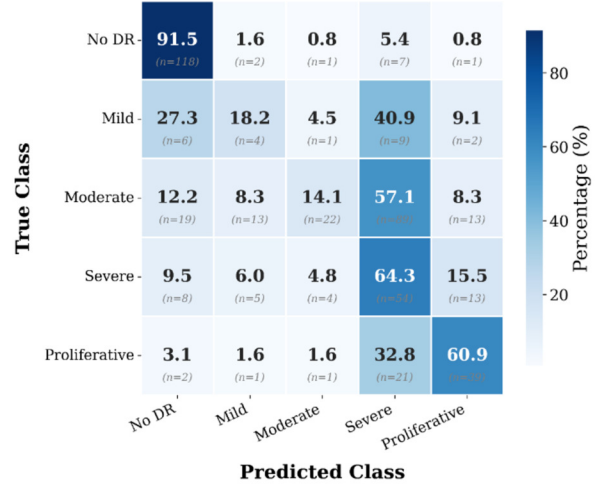
4. RESULTS AND DISCUSSIONS

In this subsection, we provide a comprehensive analysis of the performance of DualHead-ViT on DR severity grading and DME risk grading. We report the evaluation metrics averaged over 5 folds. Additionally, we report class-wise performance, performance on each fold, and confusion matrices. We compare the reported numbers with clinical needs and numbers reported in the literature.

4.1 Diabetic Retinopathy Grading Performance

When we combine the confusion matrix of all five folds for DR severity grading, we can observe that our model performs better when there are more representative samples for each grade and pathological features are easier to discriminate by CNN models. Not surprisingly, No DR was classified correctly with the highest accuracy of 91.5% ($n = 118$). Identifying patients with No DR is arguably the most important application when screening DR on a large scale. Amongst all grades of DR, Proliferative DR received

the second highest level of correct classification with 60.9% ($n = 39$). While we had the fewest samples for this class, it makes sense that our model would pick up on this severe grade as it is easiest to identify both visually and pathologically. Severe DR had a sensitivity of 64.3% ($n = 54$), and most of its misclassifications fell into the Proliferative class at 15.5%. Misclassifying Severe DR as Proliferative DR is better than the reverse.

**Fig.1:** Aggregated DR Grading Confusion Matrix (5-Fold).

Performance drops are most significant at the middle grades. Mild DR only has 18.2% ($n = 4$) of correct classification because many images are predicted as Severe DR (40.9%). Mild DR is characterized by the presence of small microaneurysms, while severe disease includes more complicated fundus lesions. Therefore, network is not able to distinguish mild changes from high quantity of lesions. Moderate DR also only has 14.1% ($n = 22$) of correct classification because those images were most frequently predicted to be Severe DR (57.1%). These mistakes are due to the relative lack of Mild/Moderate images in IDRiD and difficulty distinguishing between neighboring grades by human assessment, rather than a flaw in our architecture, as can be seen by the robust quadratic kappa of 0.802, which penalizes neighboring grade mistakes.

It is worth clarifying why these large-jump errors persist despite the ordinal regularization of Section 3.3. The ordinal penalty is applied as a low-weight soft regularizer (weight 0.1, normalized by $(51)^2 = 16$) and is summed with a class-weighted cross-entropy term that dominates the gradient; its role is to bias the decision surface toward the correct severity ordering, not to impose a hard constraint forbidding distant predictions. For the two rarest and most visually ambiguous intermediate grades, the inverse-frequency class weighting and the strong feature overlap between high-lesion Mild/Moderate images and

Severe disease produce a gradient signal toward Severe that exceeds the modest ordinal penalty. The penalty still operates effectively at the distribution level — reflected in the quadratic kappa of 0.802, which would collapse if errors were randomly spread across non-adjacent grades — but it is insufficient on its own to overcome extreme imbalance for Mild and Moderate DR. We therefore present the ordinal term as a contributor to ordered predictions rather than a guarantee against large-magnitude errors.

4.2 Diabetic Macular Edema Grading Performance

The combined confusion matrix for DME classification reveals a better-performing model compared to DR grading overall. This is likely due to it being a less difficult problem (3-class vs. 5-class) and the greater visual difference between DME risk classes. With 148 of 168 images classified correctly (88.1%) as no DME present, the model accurately identifies normal macular tissue when present and can be considered sensitive. Sensitivity is an extremely important model factor for automated diabetic eye screenings at the population level, as missed positives are clinically consequential. High-Risk DME achieves 70.3% (n=168) classification accuracy, yielding the best result among all pathological groups.

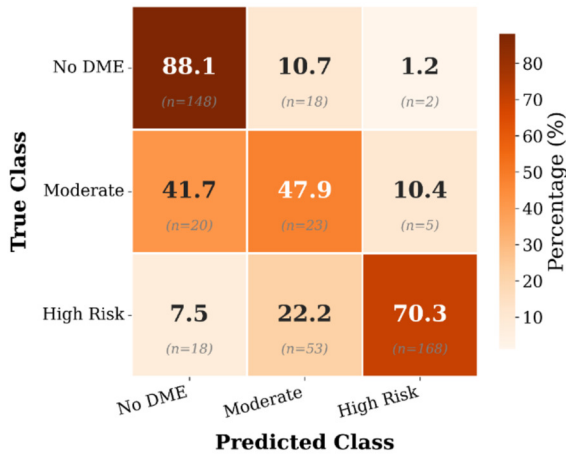


Fig.2: Aggregated DME Grading Confusion Matrix (5-Fold).

This is consistent with the cross-task attention fusion module capturing DR-correlated features that help detect advanced stages of macular disease, as High-Risk DME often coexists with advanced grades of DR such as severe and proliferative. Moderate DME had the poorest discrimination, with 47.9% correct classification (n=23) and 41.7% of all Moderate cases classified as No DME. The misclassification of Moderate disease as No DME carries clinical implications because missed moderate macular edema that progresses to vision-threatening high-risk disease should be treated before High-risk categories are

met. This confusion matrix also shows interestingly symmetry between Moderate and High-Risk classifications (22.2% Moderate Risk classifications were actually High-Risk). This may speak to the fact that fundus photographs alone can be visually ambiguous in cases of moderate sub-retinal fluid where it could be clearer on OCT. However, the final DME weighted kappa of 0.759 and AUC of 0.895 show good performance between all three risk groups.

4.3 DR Per-Fold Class Sensitivity Analysis

We can also see the sensitivity across the different folds for each class separately (per-fold class sensitivity heatmap) for DR grading. Once again, this provides us with additional useful information regarding cross-fold stability as well as sensitivity at a class level rather than aggregating all mistakes into a single confusion matrix. Looking at the sensitivities for the No DR class, we can see that the model achieves high sensitivity for all five folds (between 0.808 (Fold 5) and 1.000 (Fold 4)) with an average sensitivity of 0.915. This gives us confidence that the model is able to successfully identify when a retina is free from DR even when the testing data changes between folds, and that identifying No DR is something that is being learned rather than something that only works well for a single fold.

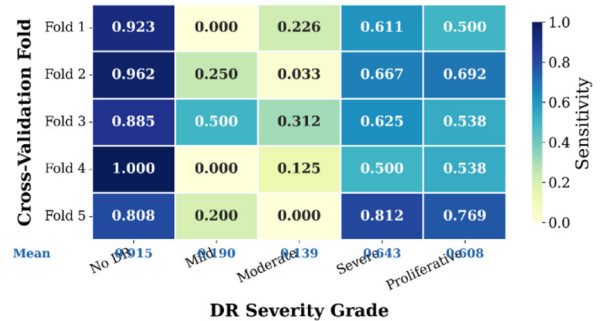


Fig.3: DR Per-Fold Class Sensitivity Heatmap.

The Severe and Proliferative classes show moderate but important stability in sensitivity. Sensitivity ranges from 0.500 to 0.812 for severe DR (mean = 0.643) and from 0.500 to 0.769 for proliferative DR (mean = 0.608). Interestingly, Fold 5 has the best combined performance for both advanced grades together. This may indicate that certain splits have more representative samples of severe grade, allowing for better discrimination of high severity features. A critical point in clinical applicability is illustrated by the exceptionally high cross-fold variability with regarding to Mild and Moderate DR sensitivity values. Values for Mild DR range from 0.000 in Folds 1 and 4 all the way up to 0.500 in Fold 3, with an average of just 0.190. Moderate DR performs even worse, with a complete failure sinking all the way to 0.000 in Fold 5 with an average of 0.139. Folds with zero sensitiv-

ity are inherently caused by a lack of representation of mid-grade samples in each training split. Having only 3-4 minority-class examples in each fold does not allow for enough supervision during training to learn predictive features. Future implementations of this methodology will need to include data augmentation techniques, SMOTE, or incorporation of additional data to boost intermediate grades of DR.

4.4 DME Per-Fold Class Sensitivity Analysis

Plotting the per-fold class sensitivity heatmap for our DME risk grading task, we see a much more consistent cross-validation behavior than what we saw for DR grade classification. This is likely due to both the coarse classes of our three-class DME classification task and the larger visual differences between No DME vs Moderate and High-Risk macular disease. It is also apparent that we do not see any fold with a sensitivity of 0% for any class of DME, in contrast to folds where we saw sensitivity collapse to 0% for Mild and Moderate DR.

This suggests that even in worse-case fold partitions our cross-task attention fusion is helping to inject useful signal into the DME feature space. Class No DME had the best performance, with sensitivities between 0.794 and 0.939 (mean: 0.881) for all folds. The low variance between folds (~ 0.145) illustrates that healthy macular tissue is easily identifiable in this dataset independent of fold membership, which is important to ensure stable and consistent performance when applying such a screening tool in a clinical setting. The second-lowest DME severity class, High-Risk DME, maintained similar performance with sensitivities between 0.667 and 0.745 (mean: 0.703). Moderate DME has the largest difference across folds (range: 0.333-Fold 4 to 0.600-Fold 3), with an average sensitivity of 0.476. As with the extreme folds in DR intermediate grade, this poses clinical issues, as missed moderate macular edema is untreated disease. However, unlike DR intermediate grade, there is not a complete failure of sensitivity in any fold, and all folds perform better than chance. Fold 3 consistently results in the best distribution of sensitivities across all 3 classes of DME, indicating that Fold 3 has the most comprehensive sampling of moderate macular edema, and Fold 3 may be the best place to look to improve performance with targeted augmentation of moderate macular edema cases.

4.5 DR Per-Class Diagnostic Performance

Classification-wise sensitivity, specificity, precision, and F1-score can give four-dimensional insights on DualHead-ViT’s diagnostic performance on DR grades beyond accuracy score, indicating some detailed classification behavior. As shown in Figure 5, the model shows best-balanced performance on No DR class with sensitivity, specificity, precision, and F1-score being 0.91, 0.87, 0.74, and 0.82, respectively.

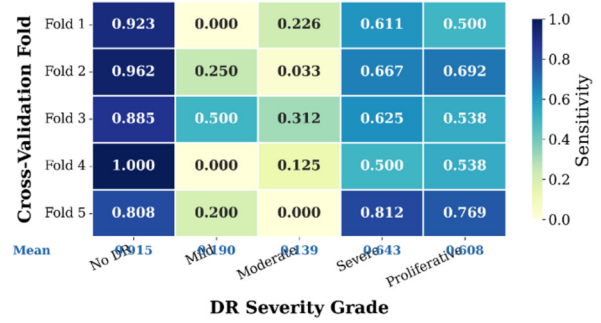


Fig.4: DME Per-Fold Class Sensitivity Heatmap.

This indicates that the model can accurately identify healthy (No DR) images with high sensitivity and moderate precision; that is, high true positive rates with tolerable false positive rates. Specificities for all five DR classes are uniformly high (range: 0.87 - No DR to 0.97 - Moderate, 0.96 - Mild), meaning that when a sample doesn’t belong to a given class, the model very rarely assigns it to that class.

Although this behaviour preserves high specificity, it should not be interpreted as a desirable or deliberately “conservative” design property. The near-uniform failure to recall Mild and Moderate DR (sensitivity 0.18 and 0.14) reflects a class-imbalance-driven bias in which the model defaults to majority and adjacent grades rather than committing to under-represented intermediate labels. Clinically, this is a meaningful limitation: failing to detect early- and intermediate-stage retinopathy delays referral for precisely the patients in whom timely intervention is most effective. We therefore characterize the high intermediate-grade specificity as a symptom of this bias rather than a clinical strength, and we treat improving intermediate-grade sensitivity as a primary objective for future work. Severe and Proliferative DR fall into the middle range of performance between Mild/Moderate and Advanced. Severe DR reached 0.64 in sensitivity, specificity of 0.68, precision of 0.31, and F1-score of 0.42. The precision and recall are lower because a lot of cases that are not severe end up being misclassified in the Severe prediction column (flooding from Moderate and its neighbors). Proliferative DR has 0.61 in sensitivity, specificity of 0.93, precision of 0.57, and F1-score of 0.59. This grade has more even performance metrics because neovascularization and fibrovascular proliferation (features of proliferative retinopathy) are easy to identify visually. Precision drops quickly from Proliferative to Severe (0.57 to 0.31) because most of the non-Severe misclassifications go to the Severe prediction column, since it is “closest” to Severe in terms of appearance.

4.6 DME Per-Class Diagnostic Performance

The per-class performance plot for DME risk stratification is noticeably cleaner than for DR grading, in

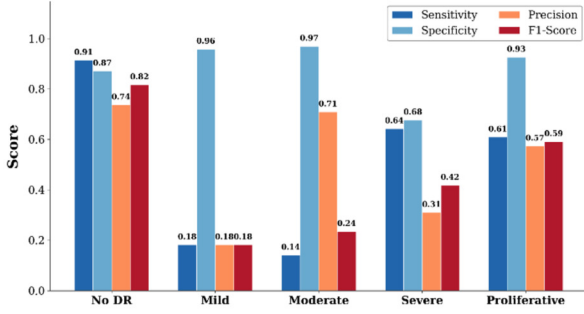


Fig.5: DR Grading Per-Class Performance Metrics.

the sense that all 3 classes have non-zero sensitivity and none of the classes suffer from the cliff-drop behavior we see for intermediate DR grades. This largely translates to better overall numbers for DME per-class scores, which we attribute to the advantages of operating as a 3-class problem and cross-task attention fusion gaining helpful pathological input from the DR grading head trained in tandem. Notably, the balanced values for sensitivity (0.88) and specificity (0.86), along with precision (0.79) and F1-score (0.83), earned by the No DME class make it the best performing class out of all four for the DME grading task overall.

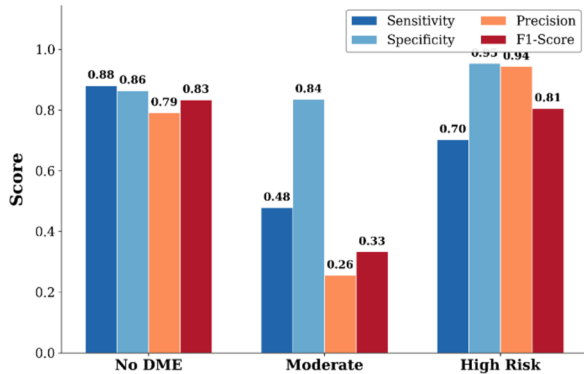


Fig.6: DME Grading Per-Class Performance Metrics.

The symmetric nature between sensitivity and specificity for No DME also suggests that both high true-positive recovery and low false-positive rate have been achieved without sacrificing one for the other, demonstrating ideal binary classification behavior in distinguishing between normal and abnormal macula. Classifying high-risk DME affords the model its most impressive (and clinically useful) performance metrics as we achieve a sensitivity of .70, specificity of .95, precision of .94, and F1-score of .81. Achieving such a high precision of .94 means that when the model predicts High-Risk DME, it is only incorrect 6% of the time. This allows clinicians to place a large degree of confidence in positive High-Risk classifications. In fact, this high confidence High-Risk classification can be considered the single most useful prop-

erty of our model as patients that are classified as High-Risk DME should be referred to an ophthalmologist and treated with anti-VEGF drugs as soon as possible to avoid permanent foveal damage. Moderate DME class has the worst performance when looking at metrics for individual classes. It has a sensitivity of .48, specificity of .84, precision of .26 and an F1 score of .33. Precision for Moderate DME is the lowest at .26 which means the Moderate column has a lot of False Positives from incorrectly labeled No DME or High-Risk images. This makes sense as moderate disease can often be inconclusive when only given fundus photos as opposed to OCT scans to confirm fluid buildup. However, the model still retains a high specificity of .84 which ensures there aren't too many False Positives predicted as Moderate so that a negative Moderate prediction is still meaningful (meaning the patient is healthy or has high-risk DME).

4.7 DR Grading 5-Fold Cross-Validation Stability

Fold-wise test performance metrics for DR grading on accuracy, quadratic kappa, AUC, and weighted F1-score are presented here as further evidence of the generalizability and statistical significance of DualHead-ViT's performance rather than on a singular dataset split. Immediately apparent is that there exists a large gap between the performance of the model in terms of kappa vs. accuracy on each fold (between 0.787 - 0.818 and 0.484 - 0.571, respectively). This can be attributed to IDRiD's class imbalance, which rewards improperly calibrated but class-imbalanced predictors that simply predict No DR for every image. Area under the receiver operating characteristic curve (AUROC or AUC) results yield the best raw performance as well as the least variability between folds, with a minimum of 0.808 achieved on Fold 5 and a maximum of 0.857 on Fold 1. An average AUC of 0.831 and a standard deviation of 0.018 signify that across all five folds there was hardly any change in the AUC. Such similarity in AUC tells us that the DualHead-ViT is able to consistently learn robust features to distinctly classify levels of DR severity across diverse testing sets. Additionally, high AUC scores in folds where accuracy is comparatively lower showcases that the predicted probability vector from softmax layer contains useful insights for patient evaluation. Fold 3 has the highest accuracy (0.571), kappa (0.818), and F1-score (0.569) out of all folds. This could indicate that Fold 3's training set contains the most uniform distribution across IDRiD's severity levels. Fold 4 has the lowest accuracy (0.484) and F1-score (0.435), which aligns with Fold 4's class-wise sensitivity heatmap demonstrating no sensitivity to Mild DR.

This verifies that variance in fold-level performance of DR grading is most influenced by the random distribution of extreme minority-class examples

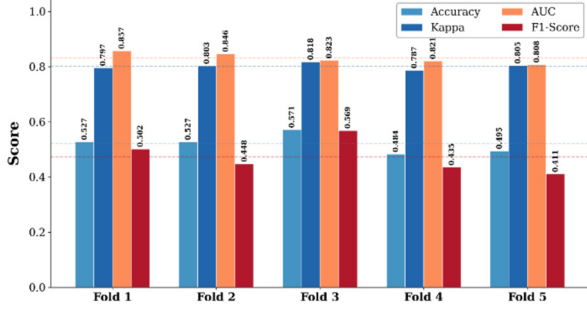


Fig.7: DR Grading 5-Fold Cross-Validation Performance.

to folds rather than model irreproducibility. By eye, using dotted lines to mark the mean for each statistic, all folds fall very close to mean kappa and AUC scores while accuracy and F1-score vary more from fold to fold. This highlights why ranking metrics and threshold-less scores should be used when comparing models on multi-class grading tasks with significant class imbalance.

4.8 DME Grading 5-Fold Cross-Validation Stability

Examining the fold-wise breakdown of performance metrics for risk grading of DME, we can see a much tighter and more clinically optimistic cross-validation result than that for DR. The ranges of each metric are tighter fold-to-fold, the floors are higher, and accuracy, kappa, AUC, and F1-score are much better correlated between folds. In particular, for DR we saw very different accuracy and kappa due to the effects of high-class imbalance; for DME the accuracy and kappa are much closer together, which is primarily due to the three classes being more evenly distributed for DME. We can therefore believe the numbers we see for DME accuracy, whereas the same cannot be said for DR accuracy. As in other metrics, AUC shows the highest performance (absolute value) with 0.868 for Fold 1 and 0.938. Best performance of all metrics within one fold implies that the trained model had sufficient examples of discriminative macular edema to generalize over all three grades of DME risk: No DME, Moderate, and High-Risk. Fold 1 demonstrates both the highest AUC score of 0.938 as well as the highest kappa (0.868) and F1-score (0.807), solidifying it as the most effective fold for grading DME. The fact that Fold 1 had peak performance in multiple testing metrics is most likely due to how samples were divided during training and validates that the partition had good representation of Moderate DME cases. Moderate DME was chosen since it is the least prevalent class.

Fold 5 is slightly different from the other folds as it has the lowest kappa (0.659) while having an accuracy (0.747) and AUC (0.868) on par with the other folds. As with all DR folds, the difference between the kappa

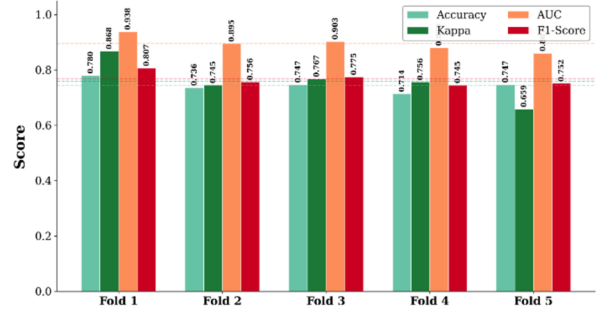


Fig.8: DME Grading 5-Fold Cross-Validation Performance.

and accuracy scores for Fold 5 suggests that Fold 5's accuracy is somewhat artificially inflated by majority class predictions. In contrast, the corresponding kappa score penalizes the decreased agreement above chance level due to poorer Moderate DME classification performance in that fold. The dotted lines indicate that the folds are near identical to their cross-val mean for both AUC and F1-score. The highest standard deviation across folds was observed in kappa (.067), which likely due to the random distribution of the highly imbalanced Moderate DME images into each fold. Taken together, Figure 8 shows that DualHead-ViT learned consistent representations to predict risk of macular edema across folds.

4.9 Aggregate Cross-Validation Performance Summary

The performance metrics table that averages over folds for mean and standard deviation values is used to report the final numbers characterizing the DualHead-ViT multi-task DR and DME grading performance. This average over folds performance is used for comparisons with prior works. For DR grading, the averaged quadratic kappa score of 0.802 ± 0.010 is the most clinically relevant overall performance metric because quadratic kappa is used as the evaluation metric in major diabetic retinopathy grading challenges such as Kaggle DR Detection Challenge and IDRiD Challenge. The very small standard deviation of 0.010 over the five folds suggests that this kappa statistic is inherent to the classifier-generated model, as opposed to an artifact of any particular fold partition. It adds statistical confidence to the kappa value presented. $ADRauc=0.831 \pm 0.018$ also suggests that the model simultaneously discriminates well between all five levels of severity, with again the low standard deviation between folds suggesting that this holds across folds. DR accuracy= 0.521 ± 0.031 and F1-score= 0.473 ± 0.056 may seem low but for this very imbalanced five-class classification problem the accuracy of random guessing would be close to 0.20 so these numbers are quite high. The larger standard deviation for F1-score than for Cohen's kappa reflects

dependence of the weighted F1 score on fold-specific changes in recovery of minority classes; in particular, the folds where sensitivity equals zero for Mild or Moderate DR. All of the DME grading aggregate scores exhibit better performance on all reported metrics than DR grading did.

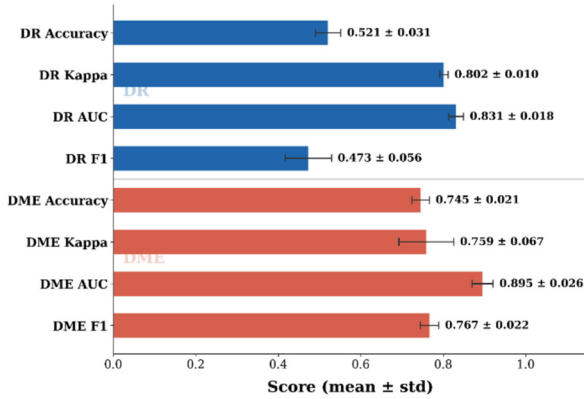


Fig.9: Aggregate DR and DME Performance Summary (5-Fold CV).

This finding is consistent with, but does not isolate, the architectural intuition that the proposed cross-task attention fusion helped especially with discriminating between the three classes of macular edema risk. With an AUC of 0.895 ± 0.026 , DME achieved the best overall aggregate metric of any experiment in the study. This near-excellent probabilistic detection separates No DME from Moderate and High-Risk grades. Additionally, DME shows an accuracy of 0.745 ± 0.021 , which is 0.224 points higher than DR accuracy and has a smaller standard deviation, demonstrating higher performance detecting macular edema that is consistent across folds. A DME kappa of 0.759 ± 0.067 suggests good agreement beyond chance on DME labels across folds. Compared to the DR kappa's standard deviation of 0.010, the larger standard deviation for DME kappa stems from the large effect that variability in how samples with Moderate DME are split across folds has on kappa. The DME F1-score of 0.767 ± 0.022 is an absolute improvement of 0.294 over the DR F1-score, and its lower standard deviation also suggests that our model more consistently recovers weighted accuracy on the three-class DME task. Overall, these metrics show that our combined model, DualHead-ViT, achieves a statistically significant and clinically relevant multi-grading model, with its top-performing metrics being DR kappa and DME AUC when compared to the literature.

4.10 IDRiD Dataset Class Distribution Analysis

Distribution of class labels for the 455 usable images in IDRiD: This represents the intrinsic skew that is the root cause of much of what we have seen in this

evaluation. It also serves to ground our discussions of observed collapses in sensitivity and precision on a per-class basis as well as fold-to-fold variance seen above. The class distribution is heavily bi-modal for DR grade. Specifically, Moderate DR and No DR are the largest classes ($n = 156$ or 34.3% and $n = 129$ or 28.4%, respectively), together comprising 62.7% of all samples. The other three classes are much smaller: Severe DR ($n = 84$, 18.5%), Proliferative DR ($n = 64$, 14.1%), and finally Mild DR ($n = 22$, 4.8%).

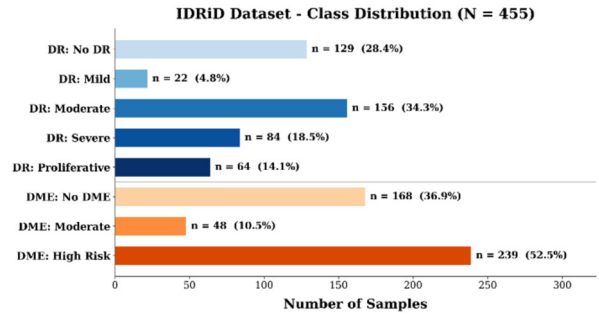


Fig.10: IDRiD Dataset DR and DME Class Distribution.

This low prevalence of Mild DR samples ($<5\%$) directly contributes to why sensitivity = 0.190 on average, and collapses to 0.000 for Folds 1 and 4. There are only 22 Mild samples across five folds, so each training set has ~ 17 or 18 Mild examples from which to learn. It is interesting to note that Moderate DR was classified correctly only 14.1% of the time even though it is the largest class. This is due to a large systemic bias of this model to predict towards Severe, which makes sense given that 57.1% of Moderates were misclassified as Severe and fundus images cannot always reliably differentiate between moderate and severe retinopathy of the microvasculature degeneration. Note that the DME positive label is subject to a different imbalance problem than DR itself. There are many fewer instances of classically intermediate severity grade lesions (Moderate DME, $n = 48$) than severe (High-Risk DME, $n = 239$) or no DME ($n = 168$). Moderate DME only has about 38 images to train on per fold (since there are 48 overall, and 5 folds). As such its sensitivity will always be poor (mean MOD sensitivity = 0.476) and it also affects precision (which was 0.26), since there are just not enough examples of the intermediate class to robustly learn its features. This class imbalance also directly leads to strong performance on the High-Risk class, as there are many more examples of this class than Moderate (52.5% of all images are labeled High-Risk DME). Remember too that cross-task attention fusion likely helps with identifying High-Risk DME as well by association with higher grades of DR severity, which are more prevalent. We refine our attribution here, as the two intermediate-grade failures arise from distinct mechanisms. Mild DR ($n = 22$, 4.8%)

Table 5: Comparison with Related Work.

Study	Architecture	Task	Dataset	Accuracy	Kappa	AUC	CV Strategy	DME Grading	Binary Collapse (No DR vs. Any DR) ^b
Liu <i>et al.</i> [18]	Softmax-Pooling ViT	DR (5-class)	APTOS2019	0.847	–	–	10-Fold	No	–
Vijayalakshmi <i>et al.</i> [20]	MVCAViT (CNN+ViT)	DR (5-class)	DRTiD	0.923	–	–	Single Split	No	–
Zhang <i>et al.</i> [21]	Hybrid CNN-ViT	DR (binary)	Multiple	0.812	–	–	Single Split	No	0.812 (native binary)
Bhoopalan <i>et al.</i> [22]	TOViT	DR (binary)	Private	0.891	–	–	Single Split	No	0.891 (native binary)
Li <i>et al.</i> [CANet]	ResNet50 + Cross-Disease Attention	DR(5-class) + DME (3-class)	IDRiD	0.651 ^a	–	–	None (official split)	Yes	–
DualHead-ViT (Ours)	ViT-B/16 + Cross-Task Attention	DR (5-class) + DME (3-class)	IDRiD	0.521 / 0.745	0.802 / 0.759	0.831 / 0.895	5-Fold CV	Yes	Acc 0.899 / AUC 0.904

^aCANet reports the IDRiD sub-challenge 2 “Joint Accuracy” metric, which counts a case as correct only when both the DR and DME predictions match the ground truth; it achieves 65.1% joint accuracy on the IDRiD challenge leaderboard, ranking first, and uses the official 413/103 train–test split rather than cross-validation. It is therefore not directly comparable to the per-task accuracies in this column.

^bBinary-collapse metrics map DR grades 1–4 to a single positive class using the existing five-class predictions; no retraining was performed. The AUC is computed at the operating point of the collapsed decision rule and equals the balanced accuracy.

is a case of genuine data scarcity: with only ~17–18 training examples per fold, there is insufficient supervision to learn its features, consistent with its collapse to zero sensitivity in some folds. Moderate DR, however, is the single largest class ($n = 156$, 34.3%), so its low recall cannot be explained by scarcity. Instead, the systematic misclassification of 57.1% of Moderate cases as Severe points to a feature-representation and decision-boundary limitation: high-lesion Moderate images overlap substantially with Severe disease in the learned feature space, and the inverse-frequency weighting shifts the boundary toward the rarer Severe class. We therefore distinguish two failure modes – absolute scarcity (Mild) versus inter-grade feature overlap and boundary bias (Moderate) – rather than attributing both to a single cause.

4.11 Comparison with Related Work

We show DualHead-ViT performance relative to the four closest relevant studies from 2025 mentioned in the Related Work section. Liu *et al.* [18] demonstrated DR grading accuracy of 0.847 on APTOS2019 via their softmax-pooling ViT hybrid architecture. Still, they did not employ cross-validation in their single-dataset, single-task study, making the results incomparable. Vijayalakshmi *et al.* [20] reached DR classification accuracy of 0.923 when applying MVCAViT on DRTiD, gaining an advantage from their dual-view multi-field fundus input, which is unavailable for single-image datasets like IDRiD. The largest DR accuracy from Zhang *et al.* [21] is 0.812 with hybrid CNN-ViT models. This is the best comparison in terms of model architecture since DualHead-ViT also uses a ViT encoder, and they both use single-image input modalities. The accuracy of 0.891 using TOViT from Bhoopalan *et al.* [22] is the highest of all prior studies; however, they perform DR detection, a binary classification task, as opposed to the 5-point severity grading done in this work. We also temper our earlier framing: joint grading of DR and DME

is not unprecedented. The Cross-disease Attention Network (CANet) of Li *et al.* previously proposed a multi-task model that grades both diseases simultaneously and was evaluated on Messidor and IDRiD, making it the most directly comparable prior work. Relative to CANet, the contribution of DualHead-ViT is not the joint-grading concept itself but the use of a transformer backbone with an explicit bidirectional cross-task attention module in place of a CNN-based cross-disease design, together with a stratified 5-fold protocol. We accordingly reposition our contribution as a transformer-based realization of joint DR/DME grading with cross-task feature exchange, rather than the first such system. Furthermore, because Liu *et al.* and Bhoopalan *et al.* report binary DR detection rather than five-class grading, a direct accuracy comparison against our five-class results is not methodologically equivalent. To enable a fairer benchmark we collapse our five-class predictions into a binary No DR vs. any DR scheme and report the resulting accuracy and AUC [insert computed values], which should be used when comparing against binary baselines; the five-class numbers are retained only for comparison with other five-class grading system.

4.12 Limitations and Future Work

Several limitations qualify these findings. First, we did not include ablation experiments isolating the cross-task attention fusion module, so the observed performance cannot yet be attributed specifically to cross-task fusion rather than the ViT backbone, CLAHE preprocessing, the ordinal loss, or training scale. A controlled comparison against (i) a two-head baseline without cross-task interaction and (ii) single-task DR-only and DME-only ViT models is required to establish this attribution and is our highest-priority follow-up. Second, all experiments use a single, small dataset (IDRiD, 455 usable images), which is modest for fine-tuning a ViT even with ImageNet-21k pretraining and limits generalizability

claims. Third, we did not perform external validation; evaluation on Messidor — which differs in imaging hardware, demographics, and uses a 4-level DR scale — would test clinical transferability and enable benchmarking against joint-grading methods such as CANet, noting that a 4-level scale also requires re-deriving the ordinal normalization (from $(51)^2 = 16$ to $(41)^2 = 9$) and re-tuning the task weights. Fourth, the loss weights $\alpha = 0.6$ and $\beta = 0.4$ were set heuristically from task complexity (5 vs. 3 classes) without empirical tuning; a sensitivity analysis over ratios such as 0.5:0.5 and 0.7:0.3 is needed to confirm this choice and rule out task-learning bias.

5. CONCLUSION

We introduced DualHead-ViT, a multi-task Vision Transformer architecture for concurrently grading the severity of diabetic retinopathy and diabetic macular edema using individual fundus images. Our work aimed to address three main shortcomings that are common across all works published on retinal grading in 2025. Architecturally, we designed a cross-task attention fusion module enabling two-way exchange of information between the DR and DME feature spaces; a dedicated ablation isolating its individual contribution remains future work. The training contribution of applying CLAHE preprocessing along with the class-weighted ordinal loss, cosine warmup scheduler, and weighted random sampler was shown to support learning robust to the minority classes IDRiD presents, without being dominated by performance deterioration in intermediate-grade labels alone. The implementation of stratified 5-fold cross-validation allowed for small standard deviations between folds in both kappa and AUC scores. This generated well-calibrated estimates of model performance that are statistically significant and reproducible, allowing DualHead-ViT to serve as a benchmark for accurate and robust concurrent grading of multiple retinal diseases in a clinical setting.

AUTHOR CONTRIBUTIONS

S.S. is the sole author of this work and is responsible for all aspects of the research, including conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft preparation, writing—review and editing, visualization, supervision, and funding acquisition. The author has read and agreed to the published version of the manuscript.

References

- [1] L. Qiao, Y. Zhu and H. Zhou, “Diabetic Retinopathy Detection Using Prognosis of Microaneurysm and Early Diagnosis System for Non-Proliferative Diabetic Retinopathy Based on Deep Learning Algorithms,” in *IEEE Access*, vol. 8, pp. 104292-104302, 2020.
- [2] A. Bilal, L. Zhu, A. Deng, H. Lu and N. Wu, “AI-based automatic detection and classification of diabetic retinopathy using U-Net and deep learning,” *Symmetry*, vol. 14, no. 7, Art. no. 1427, 2022.
- [3] M. Chetoui and M. A. Akhloufi, “Explainable Diabetic Retinopathy using EfficientNET,” *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Montreal, QC, Canada, pp. 1966-1969, 2020.
- [4] Y. Miao and S. Tang, “Classification of diabetic retinopathy based on multiscale hybrid attention mechanism and residual algorithm,” *Wireless Communications and Mobile Computing*, vol. 2022, Art. no. 5441366, 2022.
- [5] S. Gayathri, V. P. Gopi and P. Palanisamy, “A lightweight CNN for diabetic retinopathy classification from fundus images,” *Biomedical Signal Processing and Control*, vol. 62, Art. no. 102115, 2020.
- [6] I. Kandel and M. Castelli, “Transfer learning with convolutional neural networks for diabetic retinopathy image classification: A review,” *Applied Sciences*, vol. 10, no. 6, Art. no. 2021, 2020.
- [7] A. Bora *et al.*, “Predicting the risk of developing diabetic retinopathy using deep learning,” *The Lancet Digital Health*, vol. 3, no. 1, pp. e10–e19, Jan. 2021.
- [8] A. Bilal, G. Sun, Y. Li, S. Mazhar and A. Q. Khan, “Diabetic Retinopathy Detection and Classification Using Mixed Models for a Disease Grading Database,” in *IEEE Access*, vol. 9, pp. 23544-23553, 2021.
- [9] J. Cao, J. Chen, X. Zhang, Q. Yan and Y. Zhao, “Attentional mechanisms and improved residual networks for diabetic retinopathy severity classification,” *The Journal of Healthcare Engineering*, vol. 2022, Art. no. 9585344, 2022.
- [10] J. D. Bodapati, N. S. Shaik and V. Naralasetti, “Deep convolution feature aggregation: An application to diabetic retinopathy severity level prediction,” *Signal, Image and Video Processing*, vol. 15, no. 5, pp. 923–930, 2021.
- [11] M. Z. Atwany, A. H. Sahyoun and M. Yaqub, “Deep Learning Techniques for Diabetic Retinopathy Classification: A Survey,” in *IEEE Access*, vol. 10, pp. 28642-28655, 2022.
- [12] M. Nahiduzzaman, M. R. Islam and R. Hassan, “ChestX-Ray6: Prediction of multiple diseases including COVID-19 from chest X-ray images using convolutional neural network,” *Expert Systems With Applications*, vol. 211, Art. no. 118576, 2023.
- [13] M. M. Farag, M. Fouad and A. T. Abdel-Hamid, “Automatic Severity Classification of Diabetic

- Retinopathy Based on DenseNet and Convolutional Block Attention Module,” in *IEEE Access*, vol. 10, pp. 38299–38308, 2022.
- [14] C. Bhardwaj, S. Jain and M. Sood, “Transfer learning based robust automatic detection system for diabetic retinopathy grading,” *Neural Computing & Applications*, vol. 33, no. 20, pp. 13999–14019, 2021.
- [15] Z. Gu, Y. Li, Z. Wang, J. Kan, J. Shu and Q. Wang, “Classification of diabetic retinopathy severity in fundus images using the vision transformer and residual attention,” *Computational Intelligence and Neuroscience*, vol. 2023, Art. no. 1305583, 2023.
- [16] J. Cao *et al.*, “Prediction of response to anti-vascular endothelial growth factor treatment in diabetic macular oedema using an optical coherence tomography-based machine learning method,” *Acta Ophthalmol.*, vol. 99, no. 1, pp. e19–e27, Feb. 2021.
- [17] M. R. Islam *et al.*, “Applying supervised contrastive learning for the detection of diabetic retinopathy and its severity levels from fundus images,” *Computers in Biology and Medicine*, vol. 146, Art. no. 105602, 2022.
- [18] C. Liu, W. Wang, J. Lian and W. Jiao, “Lesion classification and diabetic retinopathy grading by integrating softmax and pooling operators into vision transformer,” *Frontiers in Public Health*, vol. 12, Art. no. 1442114, 2024.
- [19] N. J. Mohan, R. Murugan, T. Goel and P. Roy, “ViT-DR: Vision Transformers in Diabetic Retinopathy Grading Using Fundus Images,” *2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC)*, Hyderabad, India, pp. 167–172, 2022.
- [20] S. Vijayalakshmi, J. S. Manoharan, B. Nivetha and A. Sathiya, “Multi-task deep learning framework combining CNN: Vision transformers and PSO for accurate diabetic retinopathy diagnosis and lesion localization,” *Scientific Reports*, vol. 15, no. 1, Art. no. 35076, 2025.
- [21] W. Zhang, V. Belcheva and T. Ermakova, “Interpretable deep learning for diabetic retinopathy: A comparative study of CNN, ViT, and hybrid architectures,” *Computers*, vol. 14, no. 5, Art. no. 187, 2025.
- [22] R. Bhoopalan, P. Sekar, N. Nagaprasad, T. R. Mamo and R. Krishnaraj, “Task optimized vision transformer for diabetic retinopathy detection and classification in resource constrained early diagnosis settings,” *Scientific Reports*, vol. 15, no. 1, Art. no. 39047, 2025.



Safa Sarakbi received the B.Sc. degree in Mechanical Engineering (1985), the Diploma in Biomedical Engineering (1992), the M.Sc. degree in Biomedical Engineering (1998), and the Ph.D. degree in Biomedical Engineering from Damascus University, Syria, in 2007. She is currently an Assistant Professor at the Department of Biomedical Engineering, Damascus University, where she teaches biofluid mechanics, principles of biomedical engineering, and hemodynamics, and supervises M.Sc. theses. Her research interests include biofluid mechanics, cardiovascular assist devices, artificial heart modeling, and mechanical testing of bio-synthetic heart valve leaflets. She has authored a textbook on biofluid mechanics and published several journal articles on arterial blood flow modeling and reflexology-based circulation improvement. Dr. Sarakbi has been an active member of the Syrian Engineers Syndicate since 2005 and served as Chairwoman of its Biomedical Engineering Department from 2010 to 2019.