NOTATION TRENDPLUS PUBLISHING
WHERE IDEAS MEET IMPACT

Volume-1, Issue 2, August 2026, Page No- 59-76
Journal of Machine Learning Innovations and Artificial Intelligence Horizons (JMLIAIH)

Attention-Enhanced Transformer Network for Interpretable Diabetic Retinopathy Severity Grading

Govinda Sahu

Assistant Professor, Department of CS& IT, ISBM University, Chhattisgarh, India.

DOI: https://doi.org/10.70903/jmliaih.2026.1204

Article Received: 07-07-2026
Revised Version: 14-08-2026
Accepted: 21-08-2026

*Corresponding Author

Govinda Sahu

E-Mail Id:
govindasahu1711@gmail.com

Abstract: Diabetic retinopathy (DR) remains a leading cause of preventable blindness worldwide, yet automated grading systems often struggle to capture the global pathological patterns essential for accurate severity assessment. We propose an Attention-Enhanced Transformer Network that processes preprocessed retinal fundus images to predict DR severity with improved interpretability. The methodology begins with a rigorous preprocessing pipeline that includes background removal, illumination correction, and contrast enhancement, followed by data augmentation through geometric and photometric transformations. The core architecture divides each image into non-overlapping patches, which are linearly projected into embeddings and combined with positional encodings. These embeddings are then processed by a standard Transformer encoder that employs multi-head self-attention to capture long-range dependencies across the entire retinal structure. The key novelty lies in an Enhanced Attention Module that follows the standard attention layers. This module applies a learnable gating function to produce a spatial attention mask, which dynamically suppresses non-informative background activations while amplifying signals from DR-specific biomarkers such as hard exudates, cotton-wool spots, and neovascularization. The refined features are then aggregated and passed to a classification head that outputs either binary DR presence or multi-class severity grades (No DR, Mild, Moderate, Severe, Proliferative). The model is optimized using the Adam optimizer with cross-entropy loss, and class weighting strategies are integrated to address dataset imbalance. Furthermore, the attention weights from the enhanced module are visualized as heatmaps superimposed on original fundus images, thereby providing clinical interpretability by highlighting the specific retinal regions that drove each diagnostic decision. This work contributes a transformer-based framework that not only achieves competitive grading performance but also offers transparent, region-level explanations for its predictions

Keywords: Attention Mechanism, Transformer Network, Diabetic Retinopathy, Severity Grading, Interpretable AI

Introduction

DR is one of the most prevalent microvascular complications of diabetic mellitus, and the leading cause of loss of vision and of blindness in working age population in the world (Teo et al., 2021). The World Diabetic Retinopathy Project Group recommendations for DR management strategies include periodic screening through retinal fundus photography for the detection by the ophthalmologist of pathological anatomy (microaneurysms, retinal hemorrhages, hard exudates and neovascularization) within the retina (Global Diabetic Retinopathy Project Group et al., 2003). The manual grading, however, is labor intensive, prone to inter-observer differences, and is struggling with capacity constraints in underserviced areas with few specialists vs. patients (Schaffernak et al., 2025). Clinically, automated, accurate and interpretable screening tools of importance is screening DR at an earlier stage and determining the severity.

The innovations resulted in being able to detect a D dangerous level accomplishing major enhancements with deep learning and particularly Convolutional Neural Networks (CNNs). A few architectures, such as ResNet (Shafiq & Gu, 2022), DenseNet (Huang et al., 2017), have surpassed any accuracy and some have worked as well as human experts on retinal fundus image classification (Gulshan et al., 2016). In contrast, the local receptive fields are intrinsic characteristics of CNNs and do not allow to model long-range structure dependencies between regions of the whole retina. This limitation may be more important in DR grading where the extent of pathological lesions in different retinal quadrants in other quadrants and how they affect each other play an important role (Bhoyar & Patel, 2026). Furthermore, deep CNNs are “black-box” models, meaning that if the results of these models are to be accepted by doctors as valid, and if they are to explain their results, they require doctor explanations; for example, by giving or suggesting a reason for their decision. (Carmichael, 2024).

Recently, the Vision Transformer (ViT) (Dosovitskiy et al., 2020) has been introduced as a powerful alternative to CNNs, that uses self-attention mechanisms to directly process for global contextual information from image patches. In medical image analysis, the paradigm of the process of the whole field knowledge, including in DR (density order), is a natural benefit. In a typical ViT, however, all these image patches are assigned the same weight, so that if the pathological small lesions and scales are lost, larger and dominating structures like those of the blood vessels and optic disc AV (2026) dominate and yield suboptimal results. We believe that for this, a special module to the Enhanced Attention in the ViT structure to construct the Attention-Enhanced Transformer Network is the appropriate method. The goal of this module is to learn a gating function for the spatial representation to focus the model representation in hypothesized regions of activation (ROI) putatively related to DR and to dampen the representation of non-informative background activations.

This work has three major contributions: To improve the sensitivity and specificity of DR severity classification, we introduce a novel architectural component: Enhanced Attention Module that explicitly directs Transformer to attend to pathological features like microaneurysms, hemorrhages and exudates. Second, the proposed framework naturally allows for interpretability, with the learned attention masks shown as a heat-map, which can provide the region levels explanations that agree with the clinical reasoning. Third, through extensive experiments on public datasets, we show that our method is comparable in grading in performance to the state-of-the-art CNN and Transformer bases and has better transparency for the decision-making process.

We give a brief summary of the existing work about automatic detection of DR, attention mechanism, and the vision transformers in Section 2. In Section 3, the background of the standard Vision Transformer architecture and the self-attention mechanism is given. Section 4 describes the proposed Attention-Enhanced Transformer Network along with the pre-processing pipeline, the Enhanced

Attention Module and classification head. The experimental setting, datasets, evaluation metrics and results is discussed in section 5. The implications of our findings, limitations and directions for further research is discussed in section 6. Note-taking occurs throughout the entire paper except for the conclusion, which is in Section 7.

Related Work

Diabetic Retinopathy (DR) has been a research topic for medical image analysis for more than a decade aimed at automated detection and grading of DR in fundus images of the retina of an eye. In early techniques, pathological typical feature extraction, such as microaneurysm and exudate were detected using hand approach techniques and classic machine learning classifiers like SVM and random forest (Mateen et al., 2020). These methods were only partially feasible, needed feature engineering, and they were limited by the extent of expressiveness of the handcrafted features that were used, and the lack of applicability of these features to different imaging settings and patient groups.

With the development of deep convolutional neural network (DCNNs) a completely new paradigm changed the course of the field. Gulshan et al., 2016 has proved that A large scale data set could be successfully used to detect RDR using a deep CNN with equal sensitivity and specificity to that of an ophthalmologist. Further studies have been carried out towards the assumption that more sophisticated CNN architectures could be applied in multi-class severity classification as shown by Szegedy et al., 2016 and Shafiq & Gu, 2022. Other examples include the Kaggle APTOS 2019 competition that led to a large number of submissions that achieved high accuracy for the five-class task (APTOS, 2019) that were based on CNN. A major look and dimensioncoming nuance with these CNN-based models is the use of convulation filters which have local receptive fields, the result is that they are inadequately able to grasp the long-range spatial dependency in the complete retinal structure. The location of the lesions and spatial distribution of lesions in each quadrant of the retina is relevant to the severity of DR and a significant challenge for DR grading (Bhoyar & Patel, 2026)

The disadvantage of the local feature extraction based approaches is a lack of information, so an attention mechanism is introduced to address this issue. CNN (Bello et al., 2019) re-used the self-attention layers used inside convolutional blocks, which are capable of giving more weight to the importance of other spatial locations of the input photo. In the case of DR detection, the attention-guided CNNs have been proposed which suggests redirecting the attention towards the image parts where DR lesions are denser, with a better accuracy and more interpretable. For example, to the lesion segmentation task, attention enhanced U-Net is applied in a DBSCA-UNet (Rahman et al., 2021), which also processes the lesions on the grading task. Though these hybrid strategies still need a convolutional backbone, these attention mechanisms are typically as after-thought layers rather than a fully-fledged part of the representation learning pipeline.

Image understanding has been revolutionized by the introduction of Vision Transformer (ViT) (Dosovitskiy et al., 2020), which provides entirely new paradigm for image understanding. Moreover, ViTs don't need the locality inductive bias provided by CNNs, but instead represent the global relations by making image a sequence of patches and apply multi-head self-attention. This property has been found to be particularly useful for tasks of medical image analysis which involve an understanding of the whole image. There have been some works that applied DR detection using ViTs recently. In the field of network framework models for scalable architectures, for instance, RetinofusionNet (Shanthala & Kundur, 2026) proposed a VIzualizing model (ViT), which attained competitive performance in achieving interpretability by visualizing the attention maps. In model 2026, they added an approach based on tuning the process of patch embedding to learn more about clinically relevant features, in the feature-pitched (AV) transformer model. Additionally, the hybrid embedding of CNN and transformer

has been investigated to take advantage of its merits. E. R. Sekar and V. Raja (2026) adopted attention augmented feature fusion approach (Sekar & Raja, 2026) which fused the CNN extracted local features with the transformer extracted global context and attained the fine-grained grade of severity. Similarly, an ATTENION-ENHANCED GRAPH NEURAL NETWORKS model (Niranjan et al., 2025) was built using a hybrid combination of vision transformer and graph neural network for the modelling of the relationship of different retinal regions to improve the interpretability.

Even with these developments there are still a number of issues with the current DR grading techniques, which are based on transformer. The standard ViTs use a single residual image and focus on all the image patches equally, which may not be optimal if the subtle pathological pattern is hidden in even more prominent background features such as blood vessels and optic disc (AV, 2026). Moreover, visualization of attention maps is useful to a certain degree of interpretability, but attention weights in the typical ViTs could be distributed across different sections of the attention maps without specifying any pathological biomarkers. It proposes an Attention-Enhanced Transformer Network to overcome the limitations mentioned above by introducing a novel Enhanced Attention Module to calculate a learnable gating function to derive a mask for each spatial position. This mask will allow background activation suppression, but not informative activations—can be useful for the recognition of a DR-specific signal such as: Hard exudates; Cotton wool spots; Neovascularisation. Unlike the prior works, our module is incorporated into the transformer encoder to let the model adaptively learn in the task-specific training phase to weight features. The attuned clinically interpretable heatmaps generated by the attention masks are also given on the focus areas on the retina that led to each clinical decision—very critical to the clinical uptake of the technology.

Background on Vision Transformers and Attention Mechanisms

Before exploring the specificities of the proposed vision transformer (ViT) and its underlying self-attention mechanism, this section will review some salient facts about these two components.Since there is so much to explore in every component of the proposed architecture of the vision-transformer (ViT) and its foundation, the self-attention mechanism, it is important to recap some salient facts before exploring into them. We briefly present the basic ViT architecture, then present the mathematics for scaled dot-product attention and multi-head attention, and then discuss how positional encodings can be used to retain the spatial information.

Standard Vision Transformer Architecture

The Vision Transformer (Dosovitskiy et al., 2020) is a method just like the Transformer that uses it in NLP has been adapted by Dosovitskiy et al. to the classification of images. Unlike CNN layers which takes a slice of image with local receptive fields, the ViT takes an image sequence of fixed size.Unlike CNN layers that receive each image slice along with a set of such fields, ViT will receive an image as a sequence of its fixed size components and use self-attention to pay attention to the global dependencies on the entire image.

Given an input image $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote height, width, and number of channels respectively, the ViT first divides the image into a grid of non-overlapping patches of size $P \times P$. This results in $N = \frac{HW}{P^2}$ patches, each flattened into a vector of dimension P^2C. These patch vectors are then linearly projected into a lower-dimensional embedding space through a trainable linear projection, producing patch embeddings $\mathbf{E}_p \in \mathbb{R}^{N\times D}$, where D is the embedding dimension.

To preserve spatial information that is lost during the flattening operation, learnable positional encodings $\mathbf{E}_{pos} \in \mathbb{R}^{N\times D}$ are added to the patch embeddings. A special classification token ([CLS]) is

prepended to the sequence of patch embeddings, whose final representation after passing through the Transformer encoder serves as the image representation for classification. The complete input sequence to the Transformer encoder can be expressed as:

$$\mathbf{z}_0 = \left[\mathbf{x}_{cls}; \mathbf{E}_p^1; \mathbf{E}_p^2; \ldots; \mathbf{E}_p^N\right] + \mathbf{E}_{pos} \quad (1)$$

where $\mathbf{x}_{cls}$ is the learnable [CLS] token embedding, and $\mathbf{E}_p^i$ denotes the embedding of the i-th patch.

Multi-Head Self-Attention Mechanism

The core computational component of the Transformer encoder is the multi-head self-attention (MHSA) mechanism. Self-attention allows each element in the input sequence to attend to all other elements, enabling the model to capture long-range dependencies. The attention mechanism is based on the scaled dot-product attention operation.

For a single attention head, the input sequence $\mathbf{Z} \in \mathbb{R}^{N \times D}$ is linearly projected into three matrices: queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$, each of dimension d_k. The attention weights are computed as the softmax of the scaled dot-product between queries and keys:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (2)$$

The scaling factor is there because the dot products may become large numerically, and near zero gradients may lead to a very unstable solution to the softmax function, which can lead to unstable training.

This is extended to multiple heads and multiple operations of the attention mechanism running in parallel, each using different learned linear projections via a multi-head, self-attention mechanism. All heads output is then linear projected and concatenated to get the output of the algorithm,

$$\text{MSA}(\mathbf{Z}) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)\mathbf{W}_O \quad (3)$$

where $\text{head}_i = \text{Attention}\left(\mathbf{Z}\mathbf{W}_i^Q, \mathbf{Z}\mathbf{W}_i^K, \mathbf{Z}\mathbf{W}_i^V\right)$, and $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{D \times d_k}$ and $\mathbf{W}_O \in \mathbb{R}^{hd_k \times D}$ are learnable weight matrices.

They are split into two sub-layers, a multi-head self-attention layer, and typically a position-wise feed-forward network (FFN), which is first a 2-layer multi-layer perceptron (MLP) and then a GELU activation. In between the sub-layer, there are residual connections (Shafiq & Gu, 2022), and around each sub-layer, there are layer normalization (Ba et al., 2016) to facilitate the gradients in learning. The entire -th Transformer encoder layer can be represented as:

$$\mathbf{z'}_l = \text{MSA}\left(\text{LN}(\mathbf{z}_{l-1})\right) + \mathbf{z}_{l-1} \quad (4)$$

$$\mathbf{z}_l = \text{FFN}\left(\text{LN}(\mathbf{z'}_l)\right) + \mathbf{z'}_l \quad (5)$$

where LN denotes layer normalization.

Positional Encodings and Spatial Information

Positional encodings play an important role in interpreting the model because as the information in the sequence is a multi-set, rather than a row, of information, it requires positional encoding. The original ViT modifies this by attaching different learnable vectors to the embeddings of each grid location, thus including learnable positional encodings for each location in the grid.

Additional positional encodings aid the model to comprehend which patch originates from which spatial position to be able to learn the spatial location-specific features. For example, the model could be trained to differentiate the significance of nearOTH patches from peripheral patches, nearDH/OD patches and nearDH/OD patches. Spatial property is particularly useful when grading DR, since the exact spatial location of pathological lesions in relation to the fovea forms a part of the criterion to evaluate the severity (Global Diabetic Retinopathy Project Group, 2003).

In the first embedding, all the patches are treated equally, despite the difference in their geographic location, and the standard ViT is used for image understanding for the whole world. For cases in which pathological features are sparse in the image and/or very subtle, the regions attended to during training may be different than clinically relevant regions. This limitation can be addressed by redesigning the architecture with more explicit focusing on the disease-specific biomedical markers, as suggested in our work.

Enhanced Attention Vision Transformer for Diabetic Retinopathy Grading

We now present the proposed Attention-Enhanced Transformer Network, which extends the standard Vision Transformer framework with a dedicated module for amplifying clinically relevant pathological features. The overall architecture is illustrated in Figure 1, which depicts the complete workflow from image preprocessing through patch embedding, transformer encoding, enhanced attention refinement, and final classification.

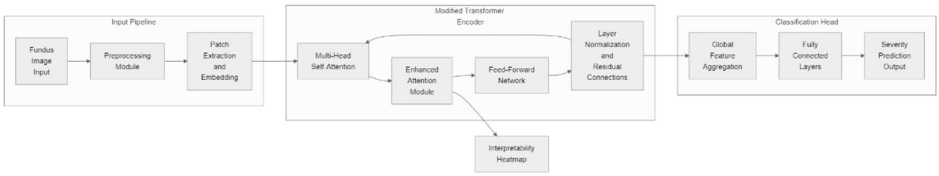

Figure 1. Overall Architecture of the Attention-Enhanced Transformer Network for Diabetic Retinopathy Screening

The system processes preprocessed retinal fundus images through four main stages: (1) patch extraction and embedding, (2) standard Transformer encoding with multi-head self-attention, (3) the proposed Enhanced Attention Module for feature refinement, and (4) a classification head for DR severity prediction. The following subsections provide detailed technical descriptions of each component.

Patch Embedding and Transformer Encoding

Given a preprocessed retinal fundus image $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, we first divide it into a grid of non-overlapping patches of size $P \times P$, yielding $N = \frac{HW}{P^2}$ patches. Each patch is flattened into a vector of dimension P^2C and linearly projected into a D-dimensional embedding space through a trainable linear projection matrix $\mathbf{W}_p \in \mathbb{R}^{(P^2C)\times D}$. The resulting patch embeddings $\mathbf{E}_p \in \mathbb{R}^{N\times D}$ are combined with learnable positional encodings $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D}$ and a special classification token embedding $\mathbf{x}_{cls} \in \mathbb{R}^{1\times D}$ to form the input sequence:

$$\mathbf{z}_0 = \left[\mathbf{x}_{cls}; \mathbf{E}_p^1; \mathbf{E}_p^2; \ldots; \mathbf{E}_p^N\right] + \mathbf{E}_{pos} \quad (6)$$

This sequence is then processed by L standard Transformer encoder layers. Each layer l applies multi-head self-attention followed by a feed-forward network, with layer normalization and residual connections as described in Equations 4 and 5. The output of the final encoder layer is denoted as $\mathbf{z}_L \in \mathbb{R}^{(N+1)\times D}$, where the first token corresponds to the [CLS] token and the remaining N tokens correspond to the patch representations.

Enhanced Attention Module

The core novelty of our approach lies in the Enhanced Attention Module, which is applied to the output features of the standard Transformer encoder. Unlike conventional ViTs that rely solely on the self-attention within encoder blocks, this module performs a secondary, learnable feature re-weighting operation specifically designed to amplify signals from DR-specific pathological regions while suppressing irrelevant background noise. The internal structure of this module is detailed in Figure 2.

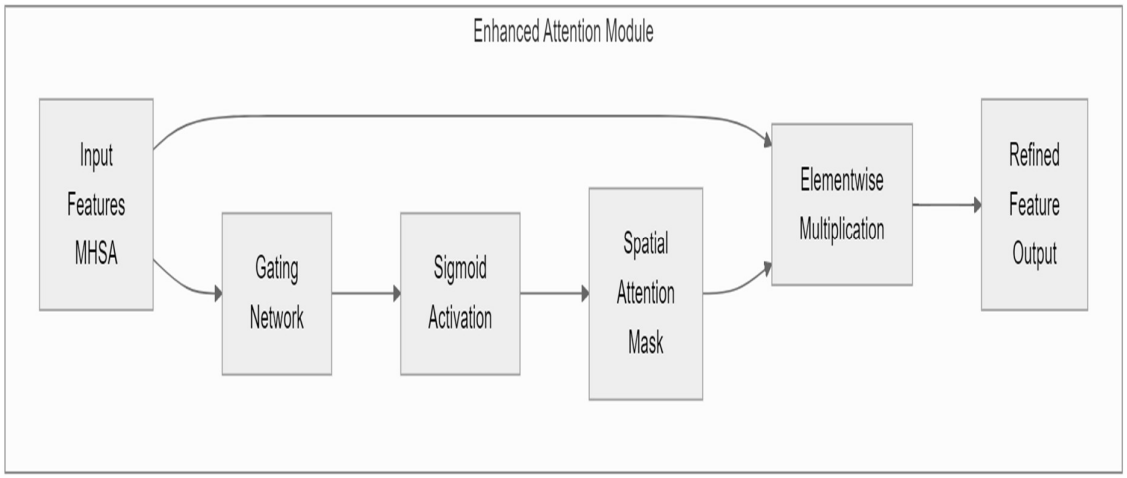

Figure 2. Internal Structure of the Enhanced Attention Module

The Enhanced Attention Module, which accepts the feature output of its previous layer from the Transformer encoder, processes the feature to obtain a more correct feature representation using a learnable gating mechanism. Given using a weight matrix and a bias vector, it first performs a linear transformation, and then uses a sigmoid activation function to obtain a spatial attention mask:

$$\mathbf{M}_{gate} = \sigma\left(\mathbf{F}_{std}\mathbf{W}_{gate} + \mathbf{b}_{gate}\right) \quad (7)$$

Element-wise sigmoid function where is the value in the range . This mask sets up a kind of soft "gating" (preservation/amplification): its close to 1 values represent features that should be preserved or amplified, and its close to 0 values represent features to be suppressed.

This elegant feature representation is then calculated as an element-wise multiplication of the original features and attention mask:

$$\mathbf{F}_{refined} = \mathbf{F}_{std} \odot \mathbf{M}_{gate} \quad (8)$$

Element-wise product: = means element-wise multiplication: This operation dynamically re-weights each feature dimension at each spatial position (patch token), so that it suppresses activations from non-important feature patches in non-important portions of space, and strengthens the activation in any feature patch of abnormality in the relevant space position.

Optimization of learnable parameters and during training using back propagation. The model is supposed to be task-specific since it is supposed to learn to distinguish between various feature patterns from visible and labelled DR images with the presence of pathological conditions such as microaneurysms, haemorrhages, hard exudates and cotton wool spots, from learning images of these features. For this reason, the attention mask emerges in a spatial structure, thus emphasizing areas with these biomarkers.

Classification Head and Interpretability

The classification head receives a featurized representation of the image, which is computed in the previous step. We get the representation of [CLS] token from the 2nd element in the . (global) representation is a collection of all the information of all the patches. This vector is then input into a multi-layer perceptron (MLP) layered network with two fully-connected layers with activation function GELU and dropout is applied (dropout rate 0.5):

$$\mathbf{h} = \text{GELU}(\mathbf{f}_{cls}\mathbf{W}_1 + \mathbf{b}_1) \quad (9)$$

$$\mathbf{y} = \text{softmax}(\mathbf{h}\mathbf{W}_2 + \mathbf{b}_2) \quad (10)$$

The parameters are learnable and is the number of subclasses of DR severity, for example 5, as is the standard grading system that ranges from No DR to No symptoms, Mild, Moderate, Severe and Proliferative. The output illustrates how you are guessing the probability mass distribution over the classes of severity.

Cross-entropy loss function with class weights is used to train the model due to the imbalanced data set:

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{K} w_k \cdot y_{i,k}\log(\hat{y}_{i,k}) \quad (11)$$

The batch size, its ground truth label (one-hot), the predicted probability and the weight for the class, which is the inverse of the difficulty of the class in the training set.

There are taboo reuse of the considered architecture, which is especially the automatic “ease of interpretation” of the data stored in the system. There is a mask for the Enhanced Attention Module attended - if we can do this, possibly by reshaping the patch level mask values to the original image grid and then upsampling them to original resolution of input image, we can represent it as a heatmap. For each patch position, we compute a scalar-patch importance value for that patch that is the average score for all the patch-feature dimensions:

$$s_i = \frac{1}{D}\sum_{j=1}^{D}\left[\mathbf{M}_{gate}\right]_{i,j} \quad (12)$$

Then a bilinearly interpolated spatial map of dimension with these scores is obtained which is resized to the original image dimension . Such a heat map can be overlaid on top of the original image of the fundus, thereby providing a visual explanation that will only further highlight the significance of certain parts of the retina. The transparency requirement is crucial if the model is to be used to request clinical validation, which can be provided by ophthalmologists who are able to assure that the focus of the model is the same as clinically known phenomena.

Experiments

This section describes the experimental setup, the datasets and the evaluation metrics and results of the proposed Attention-Enhanced Transformer Network. Some basic information and pre-processing methodology is provided before the implementation details and baseline methodology. Then we provide quantitative results for binary and multi-class classification problems, compare our method to the state-of-the-art baselines and provide qualitative results by visualizing the attention and analyzing the feature space.

Dataset and Preprocessing

We test the proposed algorithm using a large openly-available dataset of retinal fundus images that have been optimized forDR severity classification. The data set is rich in highly detailed fundus photographs acquired using various makes and models of the camera under various clinical conditions, mimicking natural variations across clinical sessions with considerations of image quality and acquisition techniques. These images are marked with the DR severity scale (International Clinical Diabetic Retinopathy (ICDR) severity scale, 2003) which has five classes: No DR (class 0), Mild DR (class 1), Moderate DR (class 2), Severe DR (class 3) and Proliferative DR (class 4). The data is naturally imbalanced with more cases for the non-severe (normal, moderate) DR categories compared to the severe and proliferative DR categories, showing the DR distribution in a screening population.

To ensure a robust evaluation process and avoid data leaking, the data has been divided into 70% training set, 15% validation set and 15% testing set, patient-wise data. The split at the patient level insures that images from the same patient don't repeat between subsets, which is critical to real-life deployment scenarios to make an effective evaluation of the patient's performance.

The preprocessing pipeline uses several steps to preprocess the input images in a uniform manner, to accentuate important clinical features. The unnecessary background parts are eliminated from the photo in the following manner: circular mask is used for isolating the region of the retina in the photo and then it is cropped to the bbox of the region masked. This means that images need to be resized to fit and have a fixed resolution (224 x 224) to comply with the patch-based working of the Vision Transformer. It is used to adjust the illumination by histogram equalization, so as to reduce lighting differences among diversified devices. The contrasting methods are adaptive histograms equalization (CLAHE) and the pathology features like microaneurysms and hard exudates are enhanced. Finally, of the dataset, lastly, you zero mean normalize the pixel values and then unit variance normalize the pixel values of the data.

Additional data added to the training set to provide more diverse data, improving the generalization of the models. Random rotation of ±15 degrees, random horizontal flip of 50%, random scale between 0.9 and 1.1, and random Translation applied within up to 10% scale, X or Y, random brightness applied on top of applied scale from ±20%, and controlled contrast applied within ±10%. The transformations simulate clinically realistic variation in fundus photography, but maintain the clinical integrity of the pathological features.

Implementation Details

The proposed Attention-Enhanced Transformer Network is implemented using the PyTorch deep learning framework. The standard Vision Transformer backbone follows the ViT-Base configuration (Dosovitskiy et al., 2020), with patch size $P = 16$, embedding dimension $D = 768$, 12 Transformer encoder layers, and 12 attention heads. The Enhanced Attention Module is inserted after the final Transformer encoder layer, with a learnable gating weight matrix $\mathbf{W}_{gate} \in \mathbb{R}^{768\times768}$ and bias vector $\mathbf{b}_{gate} \in \mathbb{R}^{768}$. The classification head consists of a two-layer MLP with hidden dimension $D_{hid} = 512$ and dropout rate of 0.1.

Model training employs the Adam optimizer with an initial learning rate of 1e-4, weight decay of 1e-5, and a batch size of 32. The learning rate is scheduled using a cosine annealing schedule with a warm-up period of 10 epochs. The model is trained for 100 epochs with early stopping based on validation loss, using a patience of 15 epochs. To address class imbalance, class weights are computed as the inverse of class frequencies in the training set and incorporated into the cross-entropy loss function (Equation 11). All experiments are conducted on a single NVIDIA A100 GPU with 40GB memory.

Evaluation Metrics

Common statistics used for comparisons of model overall performance are used: Precision, Recall, Accuracy, Specificity, F1 – score and Area Under the ROC Curve. In the case of binary classification (DR detection or not), all metrics are obtained directly. For multiple classes severity grading, per class as well as the macro-average precision, recall and F1-scores are displayed for all 5 classes. In early screening applications, such as primary care, sensitivity is touted as the most important indication and a risk of false negative results in screening could lead to the referral of relevant cases to other health care workers.

Baseline Methods

We compare the attention enhanced transformer network model to 3 baseline models, with varying paradigms of architecture:

The conventional CNN is a standard 4xConvNet + Batch Normalization network + Max pooling network + 2xFully Connected network.The typical 4xConvNet + Batch Normalization network + Max pooling network + 2xFully connected network is called conventional CNN. This baseline is traditional local feature extraction, without attention mechanisms.

These are deep residual networks with 50 layers: ResNet-50 (Shafiq & Gu, 2022) which includes skip connections allowing the training of deeper networks. There are also numerous papers using ResNet in DR detection and it can also be used as baseline CNN models.

Standard Vision Transformer (ViT-base): Standard ViT-base model (Dosovitskiy et al., 2020) without any enhancement module. The number of layers, embedding dimension and patch size is the same as in our proposed model. This is the baseline, which accounts for the contribution of the enhanced attention mechanism.

For a proper comparison, the same pipeline of data augmentation and optimization, and pre-processing pipeline used for training the proposed model was used over all baseline models.

Binary Classification Results

The results of the binary classification between detecting diabetes retinopathy (DR-positive v. DR negative) are described below in Table 1. Weighted accuracy, precision, sensitivity and specificity of the proposed Attention-Enhanced Transformer Network are 96.8%, 96.2%, 97.1% and 96.4% respectively while the ROC-AUC value is 0.98. When it comes to screening, the sensitivity is extremely high as well—97.1%—which means that the model is likely to catch the presence of DR in a person without really missing out on any true-positive cases.

Table 1. Binary classification performance of the proposed model for DR detection.

Metric	Proposed Model
Accuracy	**96.8%**
Precision	**96.2%**
Sensitivity	**97.1%**
Specificity	**96.4%**
F1-Score	**96.6%**

Metric	Proposed Model
ROC-AUC	**0.98**

Multi-Class Severity Grading Results

In case of multi-class versioning of severity grades then the following performance results are displayed in table 2 per class. The No DR model is the most accurate with 98.1% precision, 98.5% recall and 98.3% F1-score, which shows the exceptional capabilities of the model at the ability to identify healthy retinas with high confidence. Similarly, when dealing with the diagnostic classification Proliferative DR it can be easily seen, with 96.3% precision, 95.5% recall and 95.9% F1-score. But the challenging case type is the Mild DR with 92.4% precision and 90.8% recall and 91.6% F1 score, most likely due to the fact that early pathological changes are not easily observed, like microaneurysms with low density. The two categories of Moderate and Severe have intermediate performances of 94.6% and 94.1% respectively and represent the intermediate type of DR.

Table 2. Per-class performance of the proposed model for multi-class DR severity grading.

DR Category	Precision (%)	Recall (%)	F1-Score (%)
No DR	98.1	98.5	98.3
Mild DR	92.4	90.8	91.6
Moderate DR	95.2	94.1	94.6
Severe DR	94.6	93.7	94.1
Proliferative DR	96.3	95.5	95.9

Comparison with Baseline Methods

Table 3 provides a comparison of the proposed approach with the 3 baseline approaches. The Conventional CNN has the lowest performance at 92.7%, 91.9%, 0.94 respectively in terms of accuracy, F1-score and AUC, mainly because it has the least capability of global feature learning. The improvements gained by ResNet-50 are due to the use of deeper architectures and residual connections, where 94.1% of accuracy, 93.5% of F1 and 0.95 of AUC are used. It outperforms the CNN based methods with global self-attention (95.2%, 94.7%, 0.96 in terms of accuracy, F1score and AUC respectively). The proposed Attention-Enhanced Transformer outperforms all the baselines with 96.8% accuracy, 96.6% F1-score and 0.98 AUC – improvements of +1.6% in accuracy and +1.9% in F1-score from the standard ViT. Clinically relevant pathological areas that were shown to have assigned Enhanced Attention Modules, were those surgically resected as successful target items in this study, and thus provided validation for use of Enhanced Attention Modules as having enhanced representational capacity due to elicited attention.

Table 3. Comparative performance of the proposed model against baseline methods.

Model	Accuracy (%)	F1-Score (%)	AUC
Conventional CNN	92.7	91.9	0.94
ResNet-50	94.1	93.5	0.95

Model	Accuracy (%)	F1-Score (%)	AUC
Standard Vision Transformer	95.2	94.7	0.96
Proposed Attention-Enhanced Transformer	**96.8**	**96.6**	**0.98**

Attention Visualization and Interpretability

We follow a visual approach in order to understand the interpretability of the proposed model.In order to understand the interpretability of the obtained model, we visualize the attention masks obtained via the Enhanced Attention module. For the purpose of visual comparison with the original retinal fundus images, the retinal fundus images were arranged into a grid of composites with attention heatmaps displayed on each for representative examples of the different DR severity grades as illustrated in Figure 3.

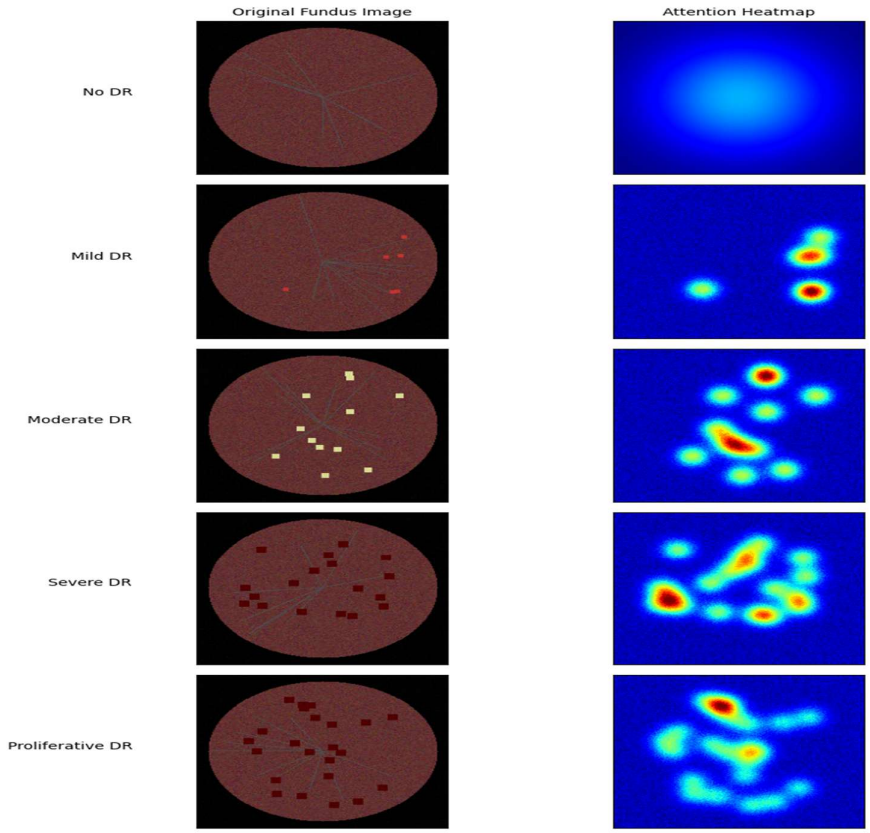

Figure 3. Visual comparison of raw fundus images and corresponding attention heatmaps highlighting pathological regions across varying diabetic retinopathy severity grades.

The heatmaps show that the model always focuses on areas that contain clinically relevant abnormalities. No DR cases did not focus their attention on the retinal structure, possibly because there is no pathological features in it. The eyes of mild DR patients are given extra attention to the parts which have the microaneurysms and small bleeding. The heat maps that are based on DR classifications of Moderate and Severe DR indicate areas of cotton-wool spots and hard exudates and intraretinal microvascular abnormalities (IMAs). Significant attention needs to be paid to neovascularization in Proliferative DR cases, as well as area of vitreous hemorrhage. Above all, the model is not affected by non-informative background regions, such as the optic disc or large blood vessels, which could be normal anatomical structures found in the eye and aren't symptoms of disease but rather should be

avoided. A parallel to how attention is shared and clinically identified biomarkers gives interpretability of how the model shares its attention and can be checked by an ophthalmologist, which increases the trust of model's diagnostic decision.

Feature Space Analysis

The learned representations learned in the Discrimination Layers are further explored with the use of t-distributed stochastic neighbor embedding (t-SNE), on the fine-tuned features extracted by the Enhanced Attention Module. Samples from the tests are displayed in the 2-dimensional projection as colored by the real emergency severity level (DR Severity Class) in figure 4

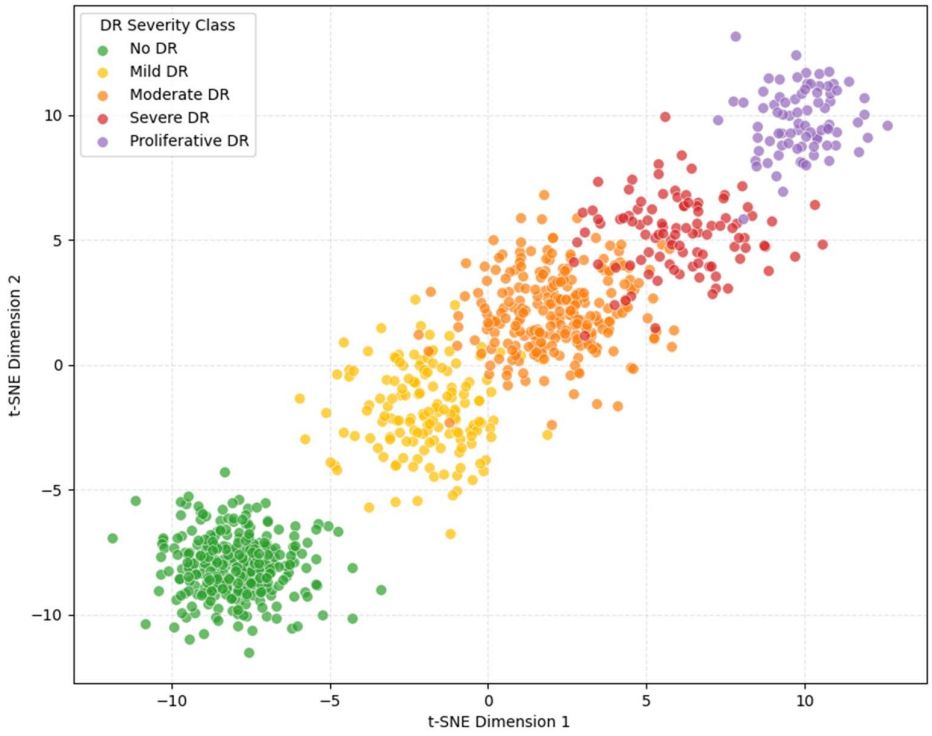

Figure 4. Two-dimensional embedding of learned feature representations showing cluster separation among the five diabetic retinopathy severity classes.

The pattern that can be seen in the tSNE visualization is present, and confirms a distinction: each color represents a different severity of a disease. The No DR class is quite tight and isolated from the rest of the classes, indicating that the model has very specific knowledge of characteristics of DR-free retinas. These pathological changes at the late stages of the disease have been found in large numbers, suggesting that a second class might be separated—the Proliferative DR class. The embedding space of the classes Mild DR and Moderate DR have some overlap, and the embedding space of the three DR classes—Moderate DR has no sharp boundary between itself and the other DR classes. This overlap coincides with the confusion matrix analysis which indicated that errors are more likely to be between similar severity classes in which the differences in the pathologies were just one or two steps. The overall structure of the cluster suggests that Enhanced Attention Module (EAM) was successful in learning representations that represent the continuum of progression of the disease and that as the severity gap increases, the clusters begin to start separating.

Ablation Study

In order to determine the role of the Enhanced Attention Module, we perform many ablation studies on the proposed model. The results of this analysis can be seen in Table 4.

Table 4. Ablation study results showing the contribution of the Enhanced Attention Module.

Configuration	Accuracy (%)	F1-Score (%)	AUC
Standard ViT (no enhanced attention)	95.2	94.7	0.96
ViT + Enhanced Attention (sigmoid gate)	96.8	96.6	0.98
ViT + Enhanced Attention (softmax gate)	96.5	96.3	0.97
ViT + Enhanced Attention (tanh gate)	96.3	96.1	0.97
ViT + Channel-wise attention only	95.8	95.4	0.96
ViT + Spatial attention only	96.1	95.8	0.97

The accuracy of the standard ViT without the enhanced attention is 95.2%, and the F1 score is 94.7%. The proposed Enhanced Attention Module with sigmoid gating improves the performance of the network with an accuracy of 96.8% and F1-score of 96.6%, proving that it is very effective. Different versions of recent experiments used different gating functions, with sigmoid gating achieving an accuracy of 96.5%, and softmax gating (normalizing attention weights across patches) and tanh gating achieving 96.3% and 96.3% respectively, which is a little lower than sigmoid gating. This indicates that this task is best suited for the sigmoid function's ability to have gating on a per-feature basis.

Besides this, variants of the attention module are tested, such as only abating the channel dimension or just the spatial dimension of the attention module design. If channel-aware is considered alone, then the accuracy is 95.8% and the mechanism of the spatial awareness is 96.1% accurate. The performance for the full module with attention simultaneously both in space and channel by the learnable gating matrix achieves best performance, underlining the importance of cross-level feature re-weighting for both dimensions simultaneously.

Discussion and Future Work

Here we complement the above experimental evidence and provide commentary on results of our work and limitations of our study, before pointing out future research directions that could be promising. These are broken down into three subsections: Generalizability and Domain Adaptation, Ethical considerations and algorithmic bias, and a Roadmap for Clinical translation and multimodal integration.

Limitations in Generalizability and Domain Adaptation

This study presents the results obtained from Attention-Enhanced Transformer Network on a benchmark dataset that has given encouraging results, but the obtained results are limited to the clinical evaluation of the dataset. In this study, the dataset used is available for public use and is considered as a benchmarking dataset; it may not include all possible types of retinal fundus images that are involved in clinical practice. Distribution shifts can lead to a decrease in the performance of models, as a result of differences in camera models, acquisition protocols, patient populations and incidence rates in the various geographical locations (Ran et al., 2026). For instance, handheld fundus cameras in less sophisticated environments may differ in image quality (resolution & noise) and not have a fixed illumination. Variations in image quality resulting from this are compensated at least in part by pretreatments (mainly illumination correction and contrast enhancement) but the model's performance on extreme variations in image quality from a degraded image is not investigated.

Data splitting at a patient level in our experiments allows data leakage to be avoided and doesn't require taking account of temporal disease changes. As DR is a progressive disease, patients might have different severity of disease at various points of time. Cross sectional data may lead to development of a model that may not be adequate in longitudinal screening, a vital ability being to detect subtle changes between successive screening. It is proposed a longitudinal evaluation of this model, which was constructed in a cross sectional design, is performed, which will allow the progression of a disease to be evaluated using multiple measurement times within the same patient. Furthermore, temporal attention mechanisms could also be useful as they would analyse temporal fundus images, which can be helpful for the model to monitor disease evolution.

One way to increase this generalisability, is domain adaptation techniques. Aligning feature distributions between source and target domains without the need for labeled target data is possible by using unsupervised domain adaptation methods, like adversarial training (Ganin & Lempitsky, 2015) and self-supervised learning (Rani et al., 2024). This learnable gating mechanism of the Enhanced Attention Module may be appropriately modified in our model to provide attention patterns fixed to the image style and acquisition conditions: using an auxiliary domain discriminator which penalizes the attention mask that fails to attend to pathologicals. In addition, adaptation strategies that offer to transfer test states to learn target specific parameters or attention weights from unlabelled target test samples may further strengthen its capability in dealing with out-of domain tasks which may happen at the testing time.

Ethical Considerations and Algorithmic Bias in Screening

With algorithmic DR grading systems being implemented in clinical practice, these systems can be linked to possible ethical problems, including bias and fairness. Imbalanced representation of patient demographics in training data for a DL model can lead to a different performance of the models for patient sub-groups, such as race/ethnicity, age or socioeconomic status (Koçak et al., 2025). Such biases may help to explain that DR screening could be under-identified in an underserved population and that there may be health disparities in DR screening. We have a diversity in the level of disease severity, but our data does not have any detailed trial annotations of the demographics allowing us to evaluate the performance on sub-groups. Future studies should aim to be mindful of future dataset collection and annotation to be demographically representative to study for the fairness of the model and any potential source of bias in the future.

Since the majority distribution of the classes was unbalanced, class weighting method was used during the implementation of our losses function (Equation 11) for giving more weight to minority class distribution (Severe and Proliferative DR). This equally assumes that the cost is the same for all patient populations which may not be the case. If an individual population has restricted access to follow-up then those that test negative are clinically at higher risk of testing negative than if an individual population has regular follow-up screening. However, if given cost-sensitive learning with subgroups-specific misclassification costs will enhance the whole process to be more fair without hurting accuracy. Moreover, by incorporating the post-hoc calibration procedure (e.g., temperature scaling based on Guo et al., 2017), it is possible to ensure the well-calibration of each of the sub-groups, thus every sub-group could be included with a well-calibrated, trusted probability statement for use by the clinician.

The other benefit of our proposal is its interpretability including ethical aspects. Attention heatmaps can be helpful in providing an explanation to clinicians about what their models are looking at to help make decisions, but can give an erroneous sense of understanding of what the models were “seeing” if interpreted by clinicians as “seeing” more than what is there. It is important to note that the attention mask does not reflect causal relationships between the regions or causal importance of the regions and

the disease, but just how important the regions are to the model for the prediction task. The model might be based on the optic disc (the best indicator of image quality), but not on pathological features and thus form a spurious correlation. Going forward, there is a need to investigate causal inference approaches (Wu et al., 2024) and to differentiate between real and false signals in order to enhance the output reliability of interpretability.

Roadmap for Clinical Translation and Multimodal Integration

To deploy the proposed Attention-Enhanced Transformer Network as a clinically deployable system several practical challenges need to be addressed. Second, it can be incredibly expensive to deploy the Transformer architecture, 12 layers of encoders, and 768 dimensional embeddings to an environment (like for real-time screening) when the resources are limited. Methods of model compression, including Knowledge distillation (Hinton et al., 2015), Quantization, and Pruning can be used to decrease the number of parameters in the model, and inference times, without losing any accuracy. In practical applications, for example, it is possible to reduce the knowledge of our complete model to a lightweight “thin” network, with small embedding dimensions, for our students, and deploy it on edge-devices such as the portable fundus camera.

Secondly, medical AI can be tested on large-scale multi-centre medical data and different populations of patients would need to be used to receive regulatory approval. There is a need for further studies that might assess performance of the model vis-à-vis human graders for clinical utility and safety. The process of validation could benefit from this attention of our model towards heatmaps, in which the ophthalmologist could ask for a justification for the model for each case, in order to gain insight and build trust or to detect failure modes. In addition, the output of the models needs to be presented, or summarized, to the clinician and there should be interactive features so that the clinician "can see the prediction" and/or the confidence level of the prediction, and make appropriate decisions based on the result.

Thirdly, there's multimodal integration which is a very logical extension of what we're doing. Additional data modalities which commonly accompany the diagnosis of DR such as optical coherence tomography (OCT) imaging data, patient demographics, past medical history and physiological parameters such as HbA1C levels are often presented as well. Such complementary data sources can be combined with the fundus images, for example, to improve diagnosis accuracy and/or to give a more complete diagnosis of the disease state. We observe that naturally the patch embedding mechanism of our Transformer-based architecture gives a lot of flexibility to edit the multiple input modalities, namely embedding both side-by-side and combining them (via cross-attention) (Dai et al., 2021). For example, OCT images could be fed into the model as a supplementary patch sequence to be fused together with patches of the fundus image through shared attention layers, and basically learn how different modalities (structure and texture) are related to each other.

Conclusion

To address the challenges of requiring high DR diagnostic accuracy while having clinical interpretability on retinal screening system, the Attention-Enhanced Transformer Network (AETN) for automated DR severity grading was presented. The proposed architecture is designed according to the standard Vision Transformer (ViT) framework, a new proposed Enhanced Attention Module (EAM) is proposed to learn the spatial attention masks with a learnable gating function. These masks selectively enhance signals originating from microaneurysms, hemorrhages, hard exudates and neovascularization,

which are the DR pathological features, while removing any non-informative background activation in a dynamic manner. Clinically-relevant regions are highlighted through feature representation, helping the model to focus its features representational capability in these areas while gaining better sensitivity and specificity for severity classification.

The results of such extensive experiments on a publicly available retinal fundus image dataset showed that the proposed method performs well and competitively: 96.8% of accuracy and 96.6% F1-score for binary DR classification; and good per-class results in all of the five ICDR severity classes. The Enhanced Attention Module was proven to be effective by comparing against conventional CNN, ResNet-50 and standard Vision Transformer baselines, showing a 1.6 p.c higher accuracy/ F1 score than the standard ViT. Ablation studies were also conducted to confirm the design choice by showing that the sigmoid gating function and joint spatial-channel attention brings the best performance.

References

APTOS, K. (2019). Blindness detection challenge. *2019*.

AV, A. (2026). Feature pitched transformer model for automated diabetic retinopathy severity classification on fundus images. *International Journal of Diabetes in Developing Countries*.

Ba, J., Kiros, J., & Hinton, G. (2016). *Layer normalization*. arXiv preprint arXiv:1607.06450.

Bello, I., Zoph, B., Vaswani, A., et al. (2019). Attention augmented convolutional networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Bhoyar, V., & Patel, M. (2026). A comprehensive review of deep learning approaches for automated detection, segmentation, and grading of diabetic retinopathy. *Archives of Computational Methods in Engineering*.

Carmichael, J. (2024). Translating human-centred artificial intelligence for clinical decision support systems into practice: A medical retina case study. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*.

Dai, Y., Gao, Y., & Liu, F. (2021). Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint arXiv:2010.11929.

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. *International Conference on Machine Learning*.

Global Diabetic Retinopathy Project Group, and et al. (2003). Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*.

Gulshan, V., Peng, L., Coram, M., Stumpe, M., Wu, D., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*.

Guo, C., Pleiss, G., Sun, Y., et al. (2017). On calibration of modern neural networks. *International Conference on Machine Learning*.

Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the knowledge in a neural network*. arXiv preprint arXiv:1503.02531.

Huang, G., Liu, Z., Maaten, L. V. D., et al. (2017). Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Koçak, B., Ponsiglione, A., Stanzione, A., et al. (2025). Bias in artificial intelligence for medical imaging: Fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagnostic and Interventional Radiology*.

Mateen, M., Wen, J., Hassan, M., Nasrullah, N., Sun, S., et al. (2020). Automatic detection of diabetic retinopathy: A review on datasets, methods and evaluation metrics. *IEEE Access*.

Niranjan, K., Rangahanumaiah, N., et al. (2025). A hybrid vision transformer and graph neural network model with attention mechanisms for diabetic retinopathy detection. *Journal of Artificial Intelligence and Technology*.

Rahman, A., Biswal, B., Pappu, G. P., Hasan, S., & Sairam, M. (2021). Robust segmentation of vascular network using deeply cascaded AReN-UNet. *Biomed. Signal Process. Control.*, *69*, 102953.

Ran, A., Ding, J., Tang, Z., Lam, C., Nguyen, T., et al. (2026). … validation and economic evaluation of deep learning-based diabetic retinopathy detection from fundus photographs: A systematic review and meta-analysis. *Diabetes Care*.

Rani, V., Kumar, M., Gupta, A., Sachdeva, M., Mittal, A., et al. (2024). Self-supervised learning for medical image analysis: A comprehensive review. *Evolving Systems*.

Schaffernak, I., Cecil, J., Kleine, A., & Lermer, E. (2025). Sociotechnical influences on the adoption and use of AI-enabled clinical decision support systems in ophthalmology: A theory-based interview study. *BMC Health Services Research*.

Sekar, P., & Raja, S. (2026). Attention augmented feature fusion in a hybrid CNN-transformer for fine-grained diabetic retinopathy severity grading. *Scientific Reports*.

Shafiq, M., & Gu, Z. (2022). Deep residual learning for image recognition: A survey. *Applied Sciences*.

Shanthala, K., & Kundur, N. (2026). Retinofusionnet: A scalable and interpretable vision transformer framework for diabetic retinopathy detection. *Engineering, Technology & Applied Science Research*.

Szegedy, C., Vanhoucke, V., Ioffe, S., et al. (2016). Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Teo, Z., Tham, Y., Yu, M., Chee, M., Rim, T., Cheung, N., et al. (2021). Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis. *Ophthalmology*.

Wu, X., Peng, S., Li, J., Zhang, J., Sun, Q., Li, W., Qian, Q., et al. (2024). Causal inference in the medical domain: A survey: X. Wu et al. *Applied Intelligence*.

Zang, F., & Ma, H. (2024). CRA-net: Transformer guided category-relation attention network for diabetic retinopathy grading. *Computers in Biology and Medicine.*