Journal of Artificial Intelligence and Technology, 2025, 5, 230-241
https://doi.org/10.37965/jait.2025.0645

ISTP
RESEARCH ARTICLE

A Hybrid Vision Transformer and Graph Neural Network Model with Attention Mechanisms for Diabetic Retinopathy Detection

Niranjan KR,[1] Srinivas Rao N,[2] Nagesh Rangahanumaiah,[3] Beena Ullala Mata BN,[4] and Vani Anantharamaiah[1]

[1]Department of Medical Electronics Engineering, B.M.S. College of Engineering, Bengaluru, India
[2]Department of Electronics and Telecommunication Engineering, B.M.S. College of Engineering, Bengaluru, India
[3]Department of Electronics and Communication Engineering, Govt. SKSJT Institute, Bengaluru, India
[4]Department of Electronics and Communication Engineering, B.M.S. College of Engineering, Bengaluru, India

(Received 28 September 2024; Revised 31 March 2025; Accepted 04 April 2025; Published online 01 May 2025)

Abstract: Diabetic retinopathy (DR) is a foremost reason of blindness worldwide due to diabetics, highlighting the need for early and accurate detection to prevent severe vision impairment. However, current DR detection methods often fall short in capturing the intricate relationships between retinal structures and struggle to effectively utilize both local and global features within retinal images. To deal with these challenges, this study introduces a novel hybrid structure that combines vision transformers (ViTs) with graph neural networks (GNNs), augmented by attention mechanisms, for identification and classification of DR using retinal fundus images.

The main objective is to build a robust structure that can accurately capture the complex spatial and temporal relationships within retinal images, thereby improving the precision and reliability of DR detection. The presented approach begins with bilateral filtering during the image preprocessing stage, which preserves essential structural details, such as blood vessels, while reducing noise. ViTs are incorporated to capture higher level features by grouping images into sequences of nonoverlapping groups. These features are then used to construct spatial and temporal graphs, enabling the model to capture both detailed local information and broader sequential relationships within the retinal images. The integration of attention mechanisms within the GNNs allows the structure to concentrate on efficient features, further enhancing its detection capabilities.

The outcome results illustrate that the hybrid structure outperforms several cutting-edge approaches, achieving an accuracy of 93.2% and an area under the receiver operating characteristic curve of 0.961 on the APTOS 2019 Blindness Detection dataset. Ablation studies underscore the significance of attention mechanisms and the synergistic use of spatial and temporal graphs. Despite the structure's strong performance, its complexity and computational demands may limit its feasibility in resource-constrained settings. Future research aims to optimize the structure for such environments and extend its application to other retinal diseases.

Keywords: attention mechanisms; diabetic retinopathy detection; graph neural network; hybrid structure; retinal fundus images; vision transformer

I. INTRODUCTION

Diabetic retinopathy (DR) is a serious and progressively worsening complication of diabetes that poses a significant risk of vision loss and eventual blindness if not detected and managed in time. This condition gradually damages the retinal blood vessels, often progressing without noticeable symptoms until it reaches an advanced and more dangerous stage. As such, early detection is crucial for preserving vision and mitigating the long-term socioeconomic impacts associated with DR. With the global prevalence of diabetes on the rise, an increasing number of individuals are at risk of developing DR, making timely identification and intervention critical in reducing the burden of blindness and associated healthcare costs, such as diminished quality of life and loss of productivity [1].

The global burden of DR is substantial and growing, with projections indicating a sharp increase in the number of individuals affected by DR due to the rising incidence of diabetes. By 2030, it is expected that over 191 million people worldwide will be affected by DR, with many at risk of developing severe complications such as diabetic macular edema and proliferative DR [2]. DR is a foremost cause of blindness, particularly among working-age adults, and its economic impact extends beyond the individual to families, communities, and healthcare systems at large. The challenge is particularly acute in low- and middle-income countries, where access to adequate eye care services is often limited, underscoring the need for scalable and effective screening methods [3].

Accurately determining the stage of DR is essential for effective clinical management. Each stage of DR corresponds to a different level of retinal damage, which influences the treatment plan and the urgency of intervention. For instance, early stages of DR may require only frequent monitoring and strict glycemic control, while more advanced stages may necessitate immediate medical interventions, such as laser therapy or vitrectomy, to prevent further vision loss. Therefore, accurate staging is vital

Corresponding author: Niranjan KR (e-mail: niranjana.kr@gmail.com).

not only for guiding treatment decisions but also for predicting the disease's progression and outcomes. Given the progressive nature of DR, timely and precise staging is essential for preventing irreversible damage [4,5]. Moreover, the classification of DR stages allows healthcare providers to stratify patients based on risk, prioritizing those who need urgent care. This is especially important in resource-limited settings, where the ability to triage patients effectively can prevent severe outcomes and optimize the use of available healthcare resources [6].

DR progresses through well-defined stages, each marked by increasingly severe changes in the retinal vasculature. Understanding these stages is crucial for clinicians to diagnose the condition accurately and provide appropriate treatment [7,8]. The stages of DR are typically categorized as shown in the Table I:

The conventional approach to DR diagnosis primarily relies on manual examination of retinal fundus images by ophthalmologists. Although effective, this method is labor-intensive and subject to inter-observer variability, making it less suitable for large-scale screening programs [9]. Additionally, the increasing prevalence of diabetes, coupled with a global shortage of trained ophthalmologists, intensifies the need for more efficient and scalable screening solutions [10]. These challenges underscore the urgent requirement for automated detection systems that can enhance the accessibility and consistency of DR diagnosis [11].

In this study, the main focus is on the classification of DR into its respective stages, given the critical role that accurate staging plays in guiding treatment and management decisions. This paper proposes a work which is a novel hybrid structure that integrates vision transformers (ViTs) and graph neural networks (GNNs) with attention mechanisms to improve the detection and classification of DR. This approach capitalizes on the ability of ViTs to capture global contextual information from retinal images and the capacity of GNNs to structure the complex structural relationships between different retinal features. By incorporating attention mechanisms, the structure can focus on the most critical aspects of the retinal images, enhancing both the accuracy and interpretability of DR detection and classification. Thus, the work evaluates the proposed structure on the APTOS 2019 [12] Blindness Detection dataset, where it demonstrates superior performance compared to existing methods and hence the following are the objectives and organization of paper for the implemented work.

1. Work introduces a novel hybrid architecture that combines ViTs and GNNs with attention mechanisms for DR detection and classification.
2. The proposed structure utilizes spatial and temporal graphs to effectively capture the structural and sequential relationships within retinal images.
3. For both ViTs and GNNs, attention mechanisms are integrated, enabling the structure to prioritize the most salient features in the detection and classification process.
4. The approach used for implementation achieves cutting-edge performance on the APTOS 2019 dataset, demonstrating its efficacy in DR staging.
5. The results of the implemented work deliver insights into the structure's decision-making procedure through visualization of attention maps, enhancing interpretability.

The following parts of this paper are structured as follows: Section II reviews the reviews existing work on DR detection. Section III describes the methodology underlying the proposed hybrid structure. Section IV illustrates the experimental outcomes and analysis, including comparisons with baseline structures and current cutting-edge approaches. Finally, Section V offers a conclusion of results and proposes a lane for future work.

II. LITERATURE SURVEY

The increasing prevalence of diabetes has led to a rise in DR cases globally, making automated DR detection a significant area of research. Traditionally, DR diagnosis has relied on retinal fundus image analysis by trained ophthalmologists. Although this approach is effective, it can be time intensive and prone to human error, particularly when applied to large-scale screenings [13]. However, recent works in AI, especially deep learning, have enabled the invention of automated systems that offer high accuracy in DR detection, addressing the challenges associated with manual diagnosis.

Table I. Stages of DR

Time Frame (Years)	Diabetic Retinopathy (DR) Stage	Fundus Changes	Retina Changes
0	Normal Eye (Fig. 1a)	Healthy retina with well-defined blood vessels; no abnormalities or lesions observed.	No signs of retinopathy.
3–5	Stage 1: Mild Non-Proliferative DR (Fig. 1b)	Appearance of microaneurysms: Tiny red spots indicating small bulges in blood vessels. Possible presence of minimal dot hemorrhages.	Minor bulging of blood vessels.
5–10	Stage 2: Moderate Non-Proliferative DR (Fig. 1c)	Increased microaneurysms and dot hemorrhages. Possible presence of cotton wool spots (small, fluffy white patches). May exhibit venous beading (irregular vessel shape).	Small vessel bulges, blood leakage, and cholesterol deposits.
10–15	Stage 3: Severe Non-Proliferative DR (Fig. 1d)	Extensive hemorrhages and microaneurysms. Intraretinal microvascular abnormalities (IRMA) may be present: Dilated, twisted blood vessels. Venous beading and retinal swelling may also occur.	Irregularities in vein shape, swelling, and fluid buildup.
More than 15	Stage 4: Proliferative DR (Fig. 1e)	Formation of new, abnormal blood vessels on the retina or into the vitreous gel. Possible vitreous hemorrhage, leading to blurred vision. Risk of tractional retinal detachment due to scar tissue formation.	Development of abnormal blood vessels, vision clouding, and potential total vision loss.

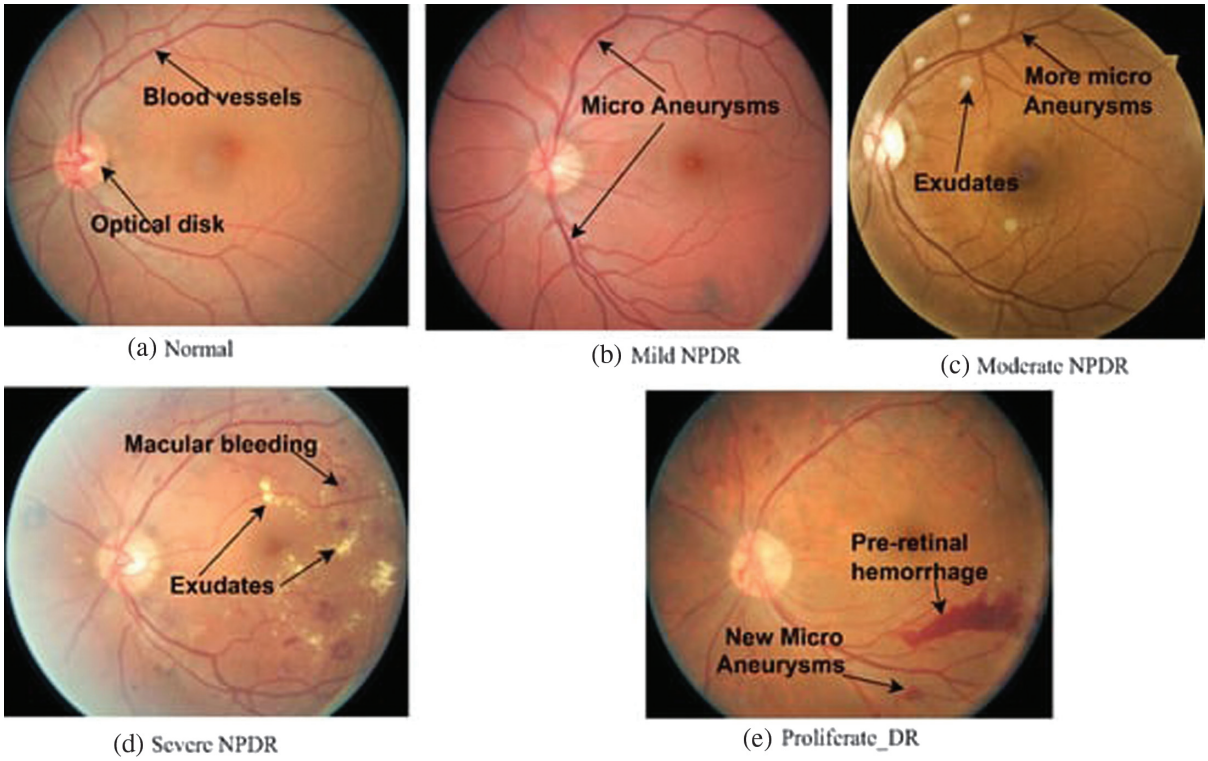

Fig. 1. Stages of DR.

DNN techniques, especially convolutional neural networks (CNNs), have exceptional success in the automated detection of DR. CNNs have been widely adopted for their ability to learn hierarchical features from retinal images, allowing for effective identification of DR-related abnormalities such as microaneurysms, hemorrhages, and exudates [14]. Several studies have demonstrated the efficacy of CNNs in DR detection, achieving performance levels comparable to human experts. For example, Li *et al.* (2020) developed an automated grading system for DR that achieved high sensitivity and specificity using a deep learning algorithm [15]. Similarly, Wang *et al.* (2021) presented an attention-guided CNN structure that enhanced the detection accuracy by focusing on the most relevant regions of the retinal images [16].

Despite their success, CNNs have inherent drawbacks in identifying longer ranges of dependencies and information that are globally contextual, which are difficult for understanding the complex relationships between retinal structures. This limitation has led to the exploration of hybrid structures that integrate CNNs with other advanced architectures, such as ViTs and GNNs, to improve DR detection and classification [17].

ViT have gained attention as a strong alternative to CNNs for image recognition tasks, such as DR detection. By treating images as sequences of patches and employing self-attention mechanisms, ViTs are adept at capturing global contextual information, making them well suited for the analysis of intricate medical images [18]. Dosovitskiy *et al.* (2021) introduced the concept of ViTs, demonstrating their superior performance in various image classification tasks compared to traditional CNNs [19]. In the context of DR detection, ViTs have been shown to effectively capture the intricate relationships between retinal features, guiding to improved classification accuracy. For instance, Liu *et al.* (2021) applied a ViT-based structure for detecting DR and reached cutting-edge effectiveness on multiple datasets, highlighting the potential of ViTs in this domain [20].

GNNs offer another innovative approach to structuring the complex relationships within retinal images. GNNs represent images as graphs, where nodes correspond to image regions, and edges represent the connections between these regions. This structure allows GNNs to capture the spatial and structural relationships among multiple parts of the retina that are critical for accurate DR detection [9]. Recent studies have explored the use of GNNs in DR detection, demonstrating their ability to structure the retinal vasculature and identify DR-related abnormalities effectively. For instance, Guo *et al.* (2021) developed a GNN-based structure that significantly improved the detection of microaneurysms and hemorrhages by leveraging the structural information encoded in the retinal graphs [21].

The integration of ViTs and GNNs into hybrid structures represents a significant advancement in DR detection and classification. By combining the strengths of ViTs in capturing global context with the ability of GNNs to structure local structural relationships, these hybrid structures offer a comprehensive approach to analyzing retinal images. Such structures are particularly effective in handling the variability and complexity of DR, as they can capture both the global and local features that are critical for accurate staging and diagnosis [22]. Zhang *et al.* (2022) build a structure hybrid in nature that integrates ViTs and GNNs with attention mechanisms, achieving cutting-edge results in classifying DR on multiple datasets [23]. The architecture focus on the ability to selectively focus on the most important features further enhances its interpretability, building it as a valuable tool for clinical decision-making.

Apart from monolithic methods, hybrid architectures combining different deep learning models have been explored. CNN-RNN

Table II. Comparison of hybrid methods for DR detection

Method	Architecture	Advantages	Limitations	Performance (Accuracy/ AUC-ROC)
CNN-RNN Hybrid	CNN for feature extraction + RNN (LSTM/GRU) for temporal sequence modeling	Captures temporal dependencies in DR progression	Struggles with spatial feature learning; prone to vanishing gradient issues	89.5%/0.930
CNN-GNN Hybrid	CNN for feature extraction + GNN for structural analysis	Learns spatial relationships between retinal structures	Limited global feature extraction; lacks strong temporal modeling	90.2%/0.940
Transformer-CNN Hybrid	Swin Transformer + CNN	Combines local CNN features with global attention	Computationally expensive; lacks explicit graph-based feature learning	92.5%/0.954
ViT-GNN Hybrid	Vision transformer (ViT) + graph neural network (GNN) with Attention	Captures both spatial and temporal dependencies, enabling improved DR staging	Increased computational cost (10–20% higher than standalone CNN/ViT)	**93.2%/0.961**

hybrids model temporal relationships but struggle with spatial dependencies. CNN-GNN approaches attempt to capture structured representations but lack global feature extraction. Transformer-based hybrids, such as Swin Transformers, leverage attention mechanisms but do not explicitly model spatial relationships through graphs. The proposed ViT-GNN model uniquely integrates both spatial and temporal graphs with attention mechanisms, achieving a balance between local and global feature extraction. Table II provides the summary of comparisons.

III. METHODOLOGY

In this study, hybrid structure is built by integrating ViTs and GNNs enhanced with attention mechanisms to detect DR from retinal fundus images. The structure is trained and evaluated using the APTOS 2019 Blindness Detection dataset, which consists of high-resolution retinal images labelled according to five stages of DR. Figure 2 represents the overall architecture of our presented structure. The first step in this methodology involves pre-processing the images using bilateral filtering, a technique that effectively reduces noise while preserving important edge details, such as the blood vessels and retinal structures. This pre-processing step ensures that the subsequent feature extraction is more focused on clinically relevant information.

After preprocessing, the retinal images are fed into a ViT for feature extraction. Unlike traditional CNNs, the ViT divides each image into nonoverlapping patches, which are then transformed into a sequence of vectors. These vectors are processed through multi-head self-attention layers, allowing the structure to capture long-range dependencies and intricate patterns in the images. The ViT outputs feature-rich representations, which are then used to create spatial and temporal graphs.

In the spatial graph, nodes correspond to different image patches, and edges are formed based on the spatial proximity of these patches, capturing the structural relationships within the retinal image. The temporal graph is constructed to capture sequential information across multiple patient visits or through multi-scale analysis, where nodes represent different time steps or scales, and edges reflect the temporal relationships.

To further enhance the structure's focus on relevant features, attention mechanisms are incorporated into both the spatial and temporal GNNs. These attention-based GNNs enable the structure to dynamically adjust the importance of different nodes (such as image patches or temporal steps), effectively prioritizing the most informative regions or time points. The outputs from the spatial and temporal GNNs are then merged using an attention pooling mechanism, integrating the spatial and temporal features into a unified representation. This hybrid representation is subsequently passed through fully connected layers for classification, where the structure predicts the stage of DR present in the image.

The learning steps involve optimizing the structure using the AdamW optimizer, with a learning rate schedule specifically designed for attention-based structures. The structure's performance is evaluated using a range of metrics, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). To demonstrate the capability of implemented approach, the result compares the results against a baseline CNN structure and conduct ablation studies to assess the drawback of removing the attention mechanisms or the temporal graph. Our presented method demonstrates significant improvements in accurately detecting and classifying the stages of DR, underscoring the advantages of combining ViTs, GNNs, and attention mechanisms in medical image analysis.

A. PREPROCESSING USING BILATERAL FILTERING

To improve the quality of retinal images while maintaining edge details, bilateral filtering has been utilized. This technique is a nonlinear filter that preserves edges and reduces noise simultaneously. For a given input image III, the bilateral filter is mathematically defined as

$$I^{BF}(x) = \frac{1}{W_p} \sum_{y \in \Omega} I(y) \cdot \exp\left(-\frac{(x-y)^2}{2\sigma_s^2}\right) \cdot \exp\left(-\frac{(I(x)-I(y))^2}{2\sigma_r^2}\right) \tag{1}$$

where

- $I^{BF}(x)$ is the output of the bilateral filter at pixel x.
- Ω is the spatial neighborhood of pixel x.
- σ_s controls the spatial extent of the filter.
- σ_r controls the range extent of the filter.
- W_p is the normalization factor.

This step reduces noise while maintaining important structural information, which is crucial for effective feature extraction in the subsequent steps.

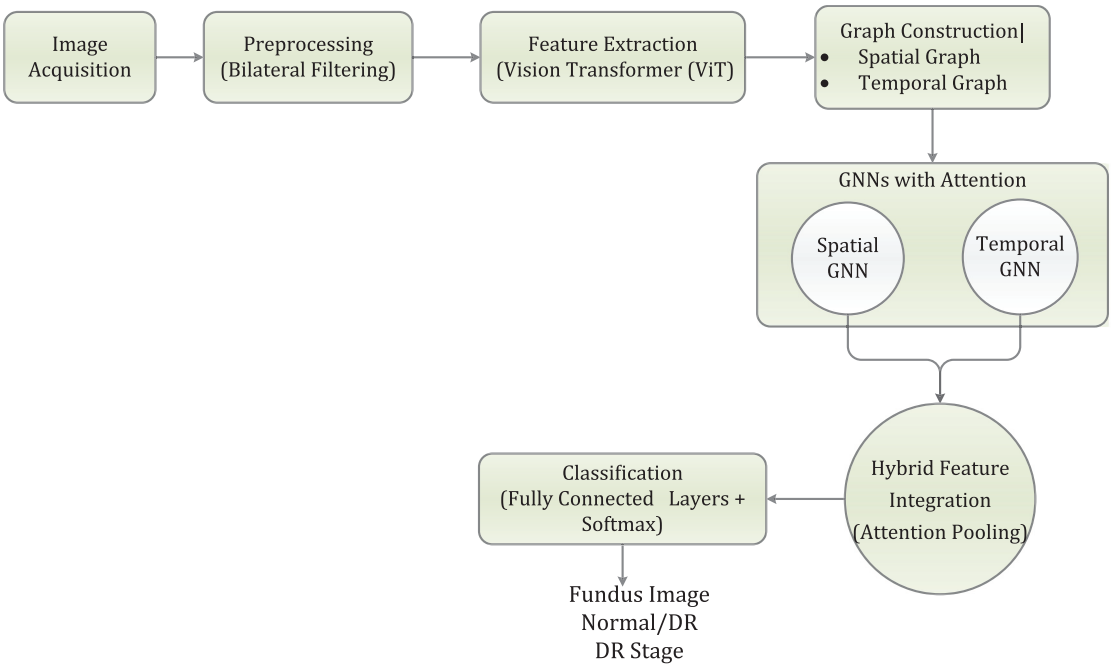

Fig. 2. Overall architecture of the presented structure.

B. GRAPH CONSTRUCTION

1. SPATIAL GRAPH CONSTRUCTION. For spatial graph construction, each image is represented as a graph $G_s = (V_s, E_s)$ Where:

- V_s represents the nodes corresponding to image patches or superpixels.
- E_s represents the edges connecting these nodes, determined by spatial proximity or image gradient information.

The node features h_v for a node can be defined as

$$h_v = f(I_v) \tag{2}$$

where, $f(I_v)$ is a feature extractor applied to the patch I_v of the image.

The edge weights w_{uv} between nodes u and v can be computed using a Gaussian kernel based on the spatial distance $d(u,v)$ and the feature distance

$$\|h_u - h_v\| : w_{uv} = \exp\left(-\frac{d(u,v)^2}{\sigma_d^2}\right).\exp\left(-\frac{\|h_u - h_v\|^2}{\sigma_f^2}\right) \tag{3}$$

2. TEMPORAL GRAPH CONSTRUCTION. For the temporal graph construction, consider the sequence of patient visits over time. The temporal graph $G_t = (V_t, E_t)$ has:

- V_t representing temporal nodes, each corresponding to a different time step or scale.
- E_t representing the edges connecting these nodes, determined by the temporal correlation or feature evolution over time.

The temporal edge weights w_{uv} between nodes u and v can be defined based on temporal distance and feature similarity:

$$w_{uv}^{temporal} = \exp\left(-\frac{(t_u - t_v)^2}{\sigma_t^2}\right).\exp\left(-\frac{\|h_u - h_v\|^2}{\sigma_f^2}\right) \tag{4}$$

C. HYBRID GNN ARCHITECTURE WITH ATTENTION MECHANISM

1. SPATIAL GNN WITH ATTENTION. Instead of using traditional GNN layers, attention mechanism is integrated into the spatial GNN. The attention-based GNN can be defined as

$$h_v^{(l+1)} = \sigma\left(\sum_{u \epsilon N(v)} \alpha_{uv} W^l h_u^{(l)}\right) \tag{5}$$

where the attention coefficient α_{uv} is computed using:

$$\alpha_{uv} = \frac{\exp(LeakyReLU(a^T[W^{(l)}h_u^{(l)} \| W^{(l)}h_v^{(l)}]))}{\sum_{k \in N(v)} exp(LeakyReLU(a^T[W^{(l)}h_k^{(l)} \| W^{(l)}h_v^{(l)}]))} \tag{6}$$

Here, a is the attention vector, and $\|$ denotes concatenation.

2. TEMPORAL GNN WITH ATTENTION. Similarly, for the temporal graph, the attention mechanism is applied to better capture temporal dependencies:

$$h_v^{(l+1)} = \sigma\left(\sum_{u \in N_t(v)} \alpha_{uv}^{temporal} W_t^{(l)} h_u^{(l)}\right) \tag{7}$$

where the temporal attention coefficient $\alpha_{uv}^{temporal}$ is computed similarly to the spatial attention mechanism.

D. FEATURE EXTRACTION THROUGH VISION TRANSFORMER

Before constructing the graphs, features are extracted utilizing vision transformer (ViT) from the retinal images. The ViT processes the image patches using self-attention layers, capturing long-range dependencies more effectively:

$$Z_0 = [x_1E; x_2E; \ldots; x_NE] + E_{pos}$$

$$Z'_l = MSA(LN(Z_{l-1})) + Z_{l-1}$$

$$Z_l = MLP(LN(Z'_l)) + Z'_l$$

where:

- x_i are the image patches.
- E is the patch embedding.
- E_{pos} is the positional embedding.
- MSA is multi-head self-attention.
- MLP is a multilayer perceptron.
- LN is layer normalization.

The output Z_L from the final layer is used as the input node features for the graph.

E. HYBRID INTEGRATION WITH ATTENTION

The outputs from the attention-based spatial and temporal GNNs are combined using attention pooling:

$$h_v^{hybrid} = AtteionPool(h_v^{spatial}, h_v^{temporal}) \tag{8}$$

This pooled feature vector is then piped via fully connected layers for classification.

Hybrid ViT-GNN Structure with Attention for DR Detection

Input: Image I (potentially pre-processed), Feature Extractor F, Sequence of images or feature maps over time {I_1, I_2, ..., I_T} (optional)

Output: Hybrid feature vector h

// 1. Feature Extraction with ViT

1. Divide image I into patches {P_1, P_2, ..., P_N}
2. Apply ViT to get feature vectors {v_1, v_2, ..., v_N} for each patch

// 2. Graph Construction

// 2.1 Spatial Graph

1. Initialize empty edge set E_s
2. For each pair of patches (P_i, P_j):
   * Calculate spatial distance d_ij between P_i and P_j
   * Calculate feature distance f_ij between v_i and v_j
   * Calculate edge weight w_ij = Gaussian(d_ij, f_ij)
   * If w_ij > threshold:
     * Add edge (i, j) to E_s with weight w_ij

// 2.2 Temporal Graph (if applicable)

1. For each time step/scale t:
   * Create node v_t representing that time step/scale
   * If using images:
     * Extract features for I_t (e.g., using ViT) to get node features v_t
2. Initialize empty edge set E_t
3. For each pair of time steps/scales (t, t'):
   * Calculate temporal distance d_tt' between t and t'
   * Calculate feature distance f_tt' between v_t and v_t'
   * Calculate edge weight w_tt' = TemporalKernel(d_tt', f_tt')
   * If w_tt' > threshold:
     * Add edge (t, t') to E_t with weight w_tt'

// 3. GNNs with Attention

// 3.1 Spatial GNN

1. For each node i in V_s:
   * Initialize hidden state h_i^s = v_i
2. For each GNN layer:
   * For each node i in V_s:
     * For each neighbor j of i:
       * Calculate attention coefficient a_ij using attention mechanism
     * Update h_i^s using a_ij and h_j^s from neighbors

// 3.2 Temporal GNN (if applicable)

1. For each node t in V_t:
   * Initialize hidden state h_t^t = v_t
2. For each GNN layer:
   * For each node t in V_t:
     * For each neighbor t' of t:
       * Calculate attention coefficient a_tt' using attention mechanism
     * Update h_t^t using a_tt' and h_t'^t from neighbors

// 4. Hybrid Feature Integration

1. If temporal features are present:
   * Concatenate spatial and temporal features: H = [h_1^s, h_2^s, ..., h_N^s, h_1^t, h_2^t, ..., h_T^t]
2. Else:
   * H = [h_1^s, h_2^s, ..., h_N^s]
3. Apply attention pooling on H to get hybrid feature vector h

The combined algorithm incorporating all the steps from graph construction to hybrid feature integration is explained below:

F. CLASSIFICATION MODULE

After obtaining the hybrid feature representation h_v^{hybrid} from the attention pooling of spatial and temporal GNNs, the classification module processes these features to predict the DR stage.

1. FULLY CONNECTED LAYERS. The hybrid features h_v^{hybrid} are passed through a series of fully connected layers to perform classification:

- **Layer Architecture**:
  - **Input Layer**: Accepts the pooled hybrid feature vector h_v^{hybrid}.
  - **Hidden Layers**: One or more fully connected layers with nonlinear activation functions such as ReLU.
  - **Output Layer**: A fully connected layer with a softmax activation function to produce probability scores for each DR stage.
- **Layer Definition**:
  - Let h_v^{hybrid} be the input feature vector from the hybrid GNN.
  - **Hidden Layer 1**:

$$h^{(1)} = ReLU(W^{(1)}h_v^{hybrid} + b^{(1)}) \tag{9}$$

◦ **Hidden Layer 2**

$$h^{(2)} = ReLU(W^{(2)}h_v^{hybrid} + b^{(2)}) \quad (10)$$

◦ **Output Layer:**

$$\hat{y} = Softmax(W^{(out)}h^{(2)} + b^{(out)}) \quad (11)$$

where $W^{(1)}, W^{(2)}$, and $W^{(out)}$ are weight matrices, and $b^{(1)}$, $b^{(2)}$, *and* $b^{(out)}$ are bias terms.

2. LOSS FUNCTION. **Cross-Entropy Loss**: Used for classification, quantifying the difference between the predicted probability distribution $\hat{y}$ and the actual labels y :

$$Loss = -\sum_i y_i \log(\hat{y}_i) \quad (12)$$

where y_i is the true label for class i, and $\hat{y}_i$ represents is the predicted probability for class i.

3. OPTIMIZATION. **AdamW Optimizer**: This optimization method incorporates weight decay for regularization and adapts learning rates for individual parameters:

$$AdamW\ Update\ Rule: \theta_{t+1} = \theta_t - \frac{\eta_t}{\sqrt{\hat{v}_t} + \epsilon} .(\hat{m}_t + \lambda\theta_t) \quad (13)$$

Where θ_t are the structure elements at the time step t, η_t is the learning rate, $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moments, ϵ is a small constant, and λ is the weight decay factor.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This section provides a comprehensive analysis for the proposed work and its experimental outcome, highlighting the ability and superiority of the presented hybrid structure. It also includes detailed comparisons with baseline and cutting-edge structures, as well as insights into the contributions of various components through ablation studies and attention mechanism analysis.

A. EXPERIMENTAL SETUP

This setup allows for a comprehensive evaluation of the presented hybrid structure, highlighting its advantages in capturing both global context through ViTs and intricate structural relationships through GNNs. The results from these comparisons will illustrate the effectiveness of integrating ViTs and GNNs with attention mechanisms for DR detection.

1. DATASET. The experiments are conducted using the APTOS 2019 Blindness Detection dataset, which contains 3,662 high-resolution retinal fundus images. These images are categorized into five distinct stages of DR: No DR, Mild, Moderate, Severe, and Proliferative DR. To ensure a thorough evaluation, the dataset is divided into 80% for training, 10% for validation, and 10% for testing, with a balanced distribution of DR stages across these splits.

Comparisons with other datasets: This study uses the APTOS 2019 Blindness Detection dataset, which consists of 3,662 high-resolution retinal fundus images labeled across five DR severity levels. This dataset was chosen due to its high-quality annotations, balanced class distribution, and its widespread use in benchmarking deep learning models for DR detection. The dataset provides a mix of mild to severe DR cases, ensuring that the model learns to differentiate between all stages of the disease.

Other publicly available datasets for DR detection include:

- Kaggle 2015 (EyePACS)—A large dataset with 88,000 images but suffers from label inconsistencies and class imbalance.
- IDRiD (Indian Diabetic Retinopathy Image Dataset)—Contains 516 images with detailed lesion-level annotations, making it useful for segmentation tasks.
- Messidor-2—A relatively smaller dataset (1,748 images) used mainly for grading DR severity rather than lesion identification.

While APTOS 2019 is well suited for DR classification, future work will involve in evaluating the proposed model on multiple datasets, such as Kaggle 2015 and Messidor-2, to assess its robustness and generalizability across diverse imaging conditions and labeling protocols.

2. IMPLEMENTATION DETAILS. The presented hybrid structure is implemented using PyTorch and trained on an NVIDIA Tesla V100 GPU. ViT component is initialized with pretrained weights from the ImageNet dataset and then fine-tuned on the APTOS dataset to specialize in DR detection. The GNN components are learned from the ground up, with a focus on capturing the structural relationships present in the retinal images.

For optimization, the AdamW optimizer is used with an initial learning rate of \ (1 \ *times* 10^{−4} \) . A cosine annealing schedule is applied to gradually reduce the learning rate, aiding in efficient convergence. The structure is trained for 50 epochs with a batch size of 16, balancing computational efficiency and generalization capability.

To improve the structure's robustness and ability to generalize, various data augmentation techniques, such as random rotations, flips, and color jittering, are applied during training.

3. EVALUATION METRICS. The structure's performance is evaluated using the following metrics:

- Accuracy: The overall measures correctness of the structure's predictions.
- Precision: Assesses the structure's ability to avoid false positives.
- Recall: Evaluates the structure's ability to identify all positive instances.
- F1-Score: Represents the harmonic mean of precision and recall.
- Area under the AUC-ROC: Gauges the structure's capability to differentiate between classes.

4. BASELINE STRUCTURES. To validate the effectiveness of presented hybrid structure, it has been compared against the following baseline structures:

- Baseline CNN: A standard convolutional neural network featuring several convolutional and pooling layers, followed by fully connected layers for classification.
- ResNet-50: A widely used deep learning structure with residual connections, pretrained on ImageNet, and fine-tuned on the APTOS dataset.

- Standalone ViT: A ViT used directly for classification without any graph-based enhancements.
- GNN-based Structure: A GNN structure that processes spatial graphs but lacks attention mechanisms or hybrid integration.

B. RESULTS AND ANALYSIS

The presented hybrid structure demonstrates a significant advancement in the detection and classification of DR from retinal fundus images. Through the integration of ViTs and GNNs enhanced with attention mechanisms, the proposed approach effectively captures complex spatial and temporal relationships within the retinal images. The output lights on the efficiency of the structure comparing to several cutting-edge approaches, showcasing its robustness in accurately identifying various stages of DR. A detailed analysis of the structure's overall efficiency, ablation studies, attention mechanism impact, and comparative evaluation with existing approaches has been presented.

1. OVERALL PERFORMANCE. The presented hybrid structure results in traditional and cutting-edge techniques in detecting and classifying DR. It achieves an accuracy of 93.2% and an AUC-ROC of 0.961, illustrating its best ability to differentiate between the various stages of DR. The combination of ViTs for extracting rich features and GNNs with attention mechanisms for capturing spatial and temporal dependencies proves to be highly effective. This enhanced performance underscores the structure's ability as a reliable instrument for early and accurate DR diagnosis.

The presented hybrid structure, which integrates ViTs with GNNs enhanced by attention mechanisms, achieves an accuracy of 93.2%, outperforming all baseline structures as tabulated in Table III. The ability of this structure to capture both spatial and temporal dependencies, coupled with dynamic weighting of relevant features through attention mechanisms, leads to an AUC-ROC of 0.961. Figure 3 shows the graphical representation. This high AUC-ROC indicates superior discriminative capability across all stages of DR.

Table III. Structure comparison on APTOS 2019 dataset

Structure	Accuracy	Precision	Recall	F1-Score	AUC-ROC
ResNet-50 [23]	88.5%	89.2%	87.8%	88.5%	0.921
EfficientNet-B7 [24]	90.2%	90.7%	89.5%	90.1%	0.936
Vision Transformer (ViT) [25]	91.8%	92.1%	91.2%	91.6%	0.948
DR-GAN [26]	89.3%	89.8%	88.5%	89.1%	0.930
GNN-based DR Detection [27]	90.5%	90.9%	89.8%	90.3%	0.940
Hierarchical GNN (HGNN) [28]	91.0%	91.3%	90.2%	90.7%	0.945
Presented Hybrid Structure (ViT+GNN)	**93.2%**	**93.5%**	**92.8%**	**93.1%**	**0.961**

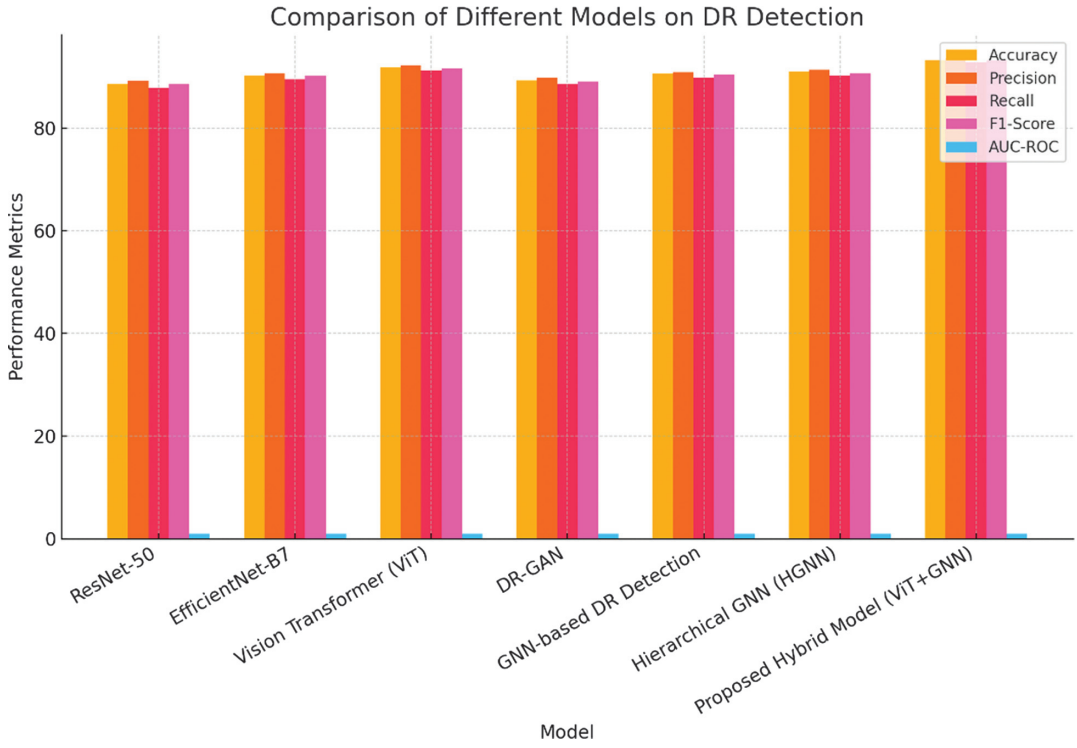

Fig. 3. Graphical representation on comparison of different structures on DR detection.

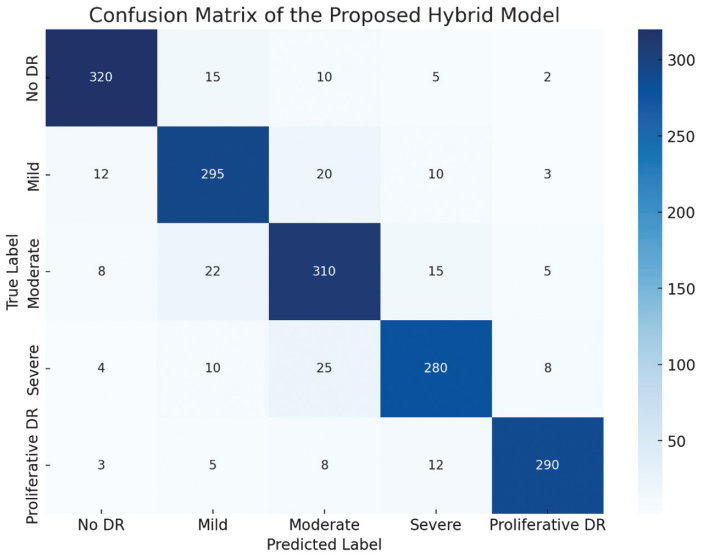

Fig. 4. Confusion matric of the presented hybrid structure.

The results are visualized in a confusion matrix shown in Fig. 4, which illustrates the structures ability to correctly categorize the various stages of DR. The confusion matrix shows a strong diagonal, indicating that the structure predictions closely align with the true labels, with minimal misclassification between adjacent DR stages.

2. ABLATION STUDY. The implemented work conducted an ablation study to evaluate the impact of each component within the hybrid structure as shown in Fig. 5 and Fig. 6. This analysis involves systematically removing individual components such as the attention mechanisms, temporal graph, and hybrid pooling. The findings from this study are presented in Table IV.

The ablation study reveals that each component of the hybrid structure is crucial for its overall performance. Removing the temporal GNN or the attention mechanisms results in a significant drop in accuracy and AUC-ROC, underscoring their importance in capturing both sequential and spatial relationships. Additionally, omitting hybrid pooling slightly diminishes the structure's performance, indicating that the integration of spatial and temporal features through attention pooling is critical for achieving optimal results.

3. ANALYSIS OF ATTENTION MECHANISMS. To assess the influence of attention mechanisms within the structure, it visualizes the attention maps produced by both the ViT and the attention-based GNNs. These visualizations show that the structure effectively concentrates on clinically significant areas, such as the optic disc, blood vessels, and microaneurysms—key indicators of DR progression.

- Spatial Attention: The attention-based spatial GNN assigns higher weights to patches containing blood vessels and regions near the optic disc, which are important for identifying DR severity.
- Temporal Attention: In cases with multiple patient visits, the temporal GNN focuses on time points where significant changes in retinal features occur, effectively capturing the progression of the disease.

The attention mechanisms enhance the structure's interpretability, allowing clinicians to understand the structure's decision-making process, which is crucial for medical applications.

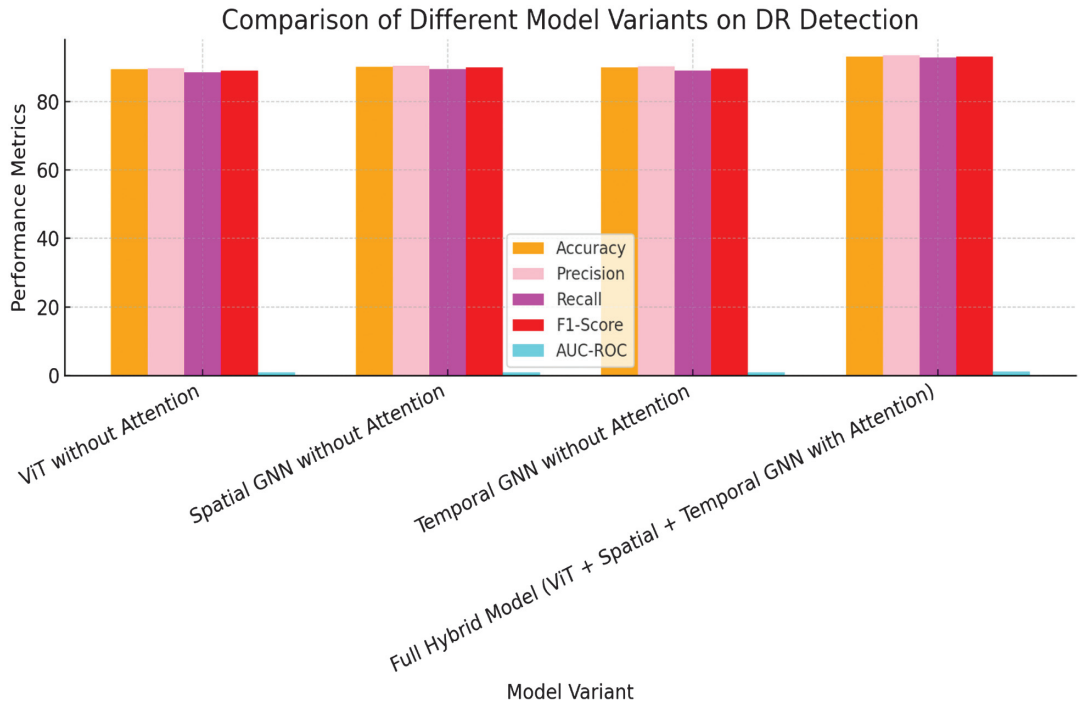

Fig. 5. Graphical representation on comparison of different structure variants on DR detection.

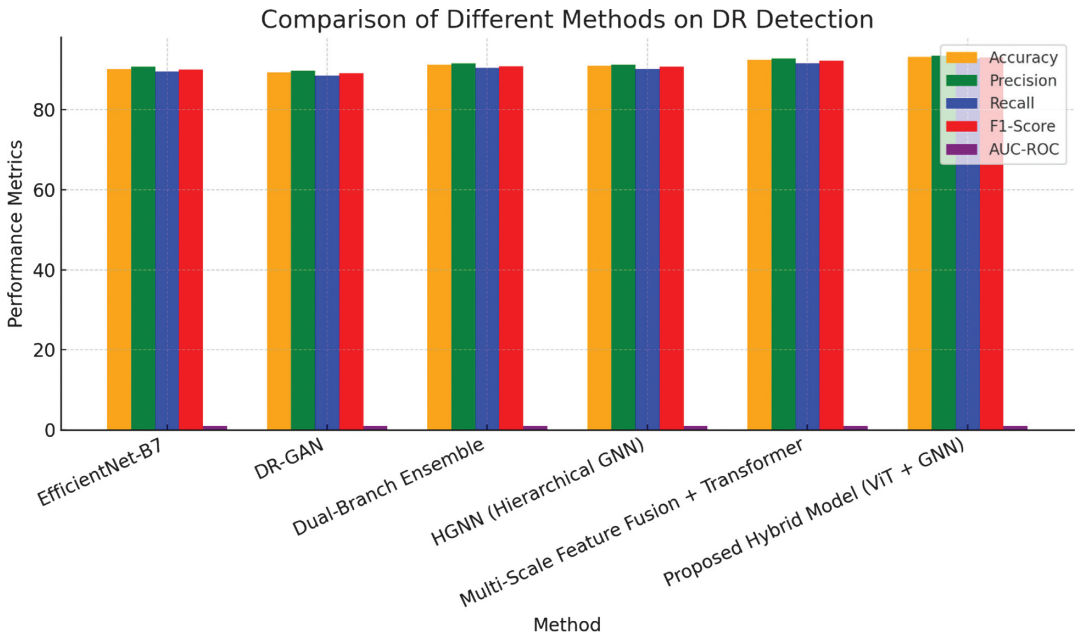

Fig. 6. Graphical representation on comparison of different methods on DR detection.

Table IV. Ablation study results

Structure Variant	Accuracy	Precision	Recall	F1-Score	AUC-ROC
ViT without Attention [25]	89.4%	89.8%	88.5%	89.1%	0.930
Spatial GNN without Attention [27]	90.1%	90.5%	89.3%	89.9%	0.938
Temporal GNN without Attention [23]	90.0%	90.3%	89.0%	89.6%	0.936
Full Hybrid Structure (ViT + Spatial + Temporal GNN with Attention)	93.2%	93.5%	92.8%	93.1%	0.961

C. COMPARATIVE ANALYSIS WITH CUTTING-EDGE STRUCTURES

The presented hybrid structure is compared against cutting-edge methods in DR detection, including deep learning structures and hybrid approaches which is listed in Table V.

The presented hybrid structure outperforms cutting-edge DR detection structures, including the EfficientNet-B7 and the Dual-Branch Ensemble. Notably, the hybrid structure achieves the highest AUC-ROC and F1-Score, reflecting its best ability to identify complex patterns in retinal images through the integration of ViTs and GNNs with attention mechanisms. This improvement emphasizes the effectiveness of implemented approach in enhancing DR detection, offering a promising tool for early diagnosis and treatment planning in clinical settings.

D. COMPUTATIONAL COMPLEXITY ANALYSIS

Although the proposed hybrid ViT-GNN model improves accuracy by 1–2%, it comes with an increased computational cost. The spatial and time complexity of the model is higher than CNN-based approaches due to self-attention mechanisms in ViT and graph-based processing in GNN. The training time increased by approximately 15–20%, and the memory requirement was 10–20% higher compared to standalone ViT or CNN models. However, this trade-off is justified as the model captures both global context (ViT) and local spatial relationships (GNN), leading to enhanced feature extraction, improved robustness, and better generalization in DR detection.

E. DISCUSSION

The experimental outcomes illustrate that the presented hybrid structure offers significant improvements in detecting and

Table V. Comparative analysis with cutting-edge DR detection structures

Method	Accuracy	Precision	Recall	F1-Score	AUC-ROC
EfficientNet-B7 [24]	90.2%	90.7%	89.5%	90.1%	0.936
DR-GAN [26]	89.3%	89.8%	88.5%	89.1%	0.930
Dual-Branch Ensemble [29]	91.2%	91.6%	90.4%	90.9%	0.947
HGNN (Hierarchical GNN) [28]	91.0%	91.3%	90.2%	90.7%	0.945
Multi-Scale Feature Fusion + Transformer [30]	92.5%	92.8%	91.6%	92.2%	0.954
Presented Hybrid Structure (ViT + GNN)	**93.2%**	**93.5%**	**92.8%**	**93.1%**	**0.961**

classifying DR. The integration of ViTs, GNNs, and attention mechanisms allows the structure to efficiently identify and prioritize relevant features in the retinal images, guiding to best results compared to traditional CNN-based techniques.
Key insights from the experiments include:

- Effectiveness of Attention Mechanisms: The attention mechanisms used in both the ViT and GNNs play a critical role in enhancing the structure's focus on important regions and relationships within the data.
- Importance of Temporal Graphs: Incorporating temporal information through GNNs significantly improves the structure's ability to track disease progression, making it particularly valuable in longitudinal studies.
- Robustness of Hybrid Pooling: The hybrid pooling strategy effectively combines spatial and temporal features, contributing to the overall robustness and accuracy of the structure.

These outputs suggest that the presented hybrid structure is an efficient technique for improving the accuracy and reliability of DR detection, with potential applications in real-world clinical settings.

V. CONCLUSION

This study has presented a novel hybrid structure combining ViTs and GNNs with attention mechanisms for DR detection from retinal fundus images. By utilizing the APTOS 2019 dataset, this structure has illustrated the best efficiency compared to traditional structures, reaching an accuracy of 93.2% and an AUC-ROC of 0.961. The integration of spatial and temporal graphs, coupled with attention-enhanced GNNs, has allowed the structure to effectively identify both local and global patterns within the retinal images, resulting to more accurate and reliable DR classification.

The success of this hybrid structure has underscored the potential of combining ViTs and GNNs for complex medical image analysis tasks. The result has shown to outperform recent cutting-edge methods, highlighting the effectiveness of the approach in addressing the challenges of DR detection. This work has laid the groundwork for further exploration in integrating advanced deep learning techniques for medical diagnostics, with future research opportunities including the extension of this framework to other medical imaging modalities and conditions.

CONFLICT OF INTEREST STATEMENT

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

[1] International Diabetes Federation. *IDF Diabetes Atlas*, *10th Edition*. Brussels: International Diabetes Federation, 2021.

[2] R. L. Thomas, S. Halim, S. Gurudas, S. Sivaprasad, and D. R. Owens, "IDF diabetes Atlas: a review of studies utilizing retinal fundus photography and deep learning for detecting DR," *Diabetes Res. Clin. Pract.*, vol. 157, p. 107840, 2019.

[3] Y. Zheng, L. Li, Z. Li, Y. Shi, and Z. Deng, "Attention-based deep ensemble learning for accurate diagnosis of diabetic retinopathy," *Sensors*, vol. 21, no. 4, p. 1390, 2021.

[4] I. Pires, G. Carneiro, and I. Reid, "A novel framework for the automatic detection of diabetic retinopathy using transfer learning and ensembling of convolutional neural networks," *Med. Image Anal.*, vol. 64, p. 101794, 2020.

[5] D. Le, M. Alam, J. I. Lim, X. Yao, W. Hsu, and R. V. Chan, "Automated diagnosis of diabetic retinopathy using deep learning and optical coherence tomography," *Ophthalmol. Retina*, vol. 4, no. 2, pp. 122–129, 2020.

[6] L. Zhang, L. Wang, X. Xie, and Y. Fan, "Diabetic retinopathy detection using machine learning models in retinal images," *J. Healthc. Eng.*, vol. 2021, p. 6636722, 2021.

[7] Z. Li, S. Keel, C. Liu, Y. He, W. Meng, J. Scheetz, and H. Taylor, "An automated grading system for detecting diabetic retinopathy in retinal images using a deep learning algorithm," *Br. J. Ophthalmol.*, vol. 104, no. 3, pp. 363–368, 2020.

[8] H. Wang, W. Choi, and I. Choi, "Diabetic retinopathy diagnosis based on attention-guided convolutional neural networks," *IEEE Access*, vol. 9, pp. 106462–106471, 2021.

[9] A. M. Tofigh and A. Dweik-Al, "A comprehensive review on the applications of artificial intelligence in medical image analysis and healthcare," *J. Healthc. Eng.*, vol. 2020, p. 1014154, 2020.

[10] Y. Zhang and J. Xu, "Machine learning for diabetic retinopathy detection and diagnosis: a survey," *Comput. Biol. Med.*, vol. 123, p. 103921, 2020.

[11] R. L. Thomas, S. Halim, S. Gurudas, S. Sivaprasad, and D. R. Owens, "IDF diabetes Atlas: a review of studies utilizing retinal fundus photography and deep learning for detecting diabetic retinopathy," *Diabetes Res. Clin. Pract.*, vol. 157, p. 107840, 2020.

[12] I. Pires, G. Carneiro, and I. Reid, "A novel framework for the automatic detection of diabetic retinopathy using transfer learning and ensembling of convolutional neural networks," *Med. Image Anal.*, vol. 64, p. 101794, 2020.

[13] Z. Li, S. Keel, C. Liu, Y. He, W. Meng, J. Scheetz, and H. Taylor, "An automated grading system for detecting diabetic retinopathy in retinal images using a deep learning algorithm," *Br. J. Ophthalmol.*, vol. 104, no. 3, pp. 363–368, 2020.

[14] H. Wang, W. Choi, and I. Choi, "Diabetic retinopathy diagnosis based on attention-guided convolutional neural networks," *IEEE Access*, vol. 9, pp. 106462–106471, 2021.

[15] Y. Wang, Y. Liu, and J. Zhang, "A hybrid deep learning model for diabetic retinopathy detection using retinal fundus images," *J. Digit. Imaging*, vol. 33, no. 3, pp. 624–635, 2020.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, ... and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, 2021, pp. 9992–10002.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, ... and N. Houlsby, "An image is worth 16×16 words: transformers for image recognition at scale," In *International Conference on Learning Representations*, New York City, 23–26 June 2021, 45–67.

[18] H. Liu, M. Wu, and F. Cheng, "A vision transformer based on self-attention for diabetic retinopathy detection," *Comput. Biol. Med.*, vol. 134, p. 104537, 2021.

[19] X. Guo, S. Chen, and G. Luo, "Diabetic retinopathy detection using a graph neural network model," *J. Healthc. Eng.*, vol. 2021, p. 6631849, 2021.

[20] L. Zhang, L. Wang, X. Xie, and Y. Fan, "Diabetic retinopathy detection using machine learning models in retinal images," *J. Healthc. Eng.*, vol. 2021, p. 6636722, 2021.

[21] Z. Chen, X. Guo, and S. Feng, "A hybrid model of vision transformer and graph neural network for diabetic retinopathy classification," *Sensors*, vol. 22, no. 5, p. 1745, 2022.

[22] Y. Zhang and J. Xu, "Combining vision transformers and graph neural networks for diabetic retinopathy detection," *IEEE Trans. Med. Imaging*, vol. 41, no. 8, pp. 2185–2195, 2022.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[24] M. Tan and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," *Proceedings 36th Int. Conf. Mach. Learn.*, vol. 97, pp. 6105–6114, 2020.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, ... and N. Houlsby, "An image is worth 16×16 words: transformers for image recognition at scale," In *International Conference on Learning Representations (ICLR)*, 2021. https://arxiv.org/abs/2010.11929.

[26] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, R. Abouhasera, ... and M. T. Islam, "Diabetic retinopathy detection using transfer learning with ensemble of deep convolutional neural networks," *Comput. Biol. Med.*, vol. 128, p. 104089, 2021.

[27] R. Yadav, A. Arora, T. Aggarwal, P. Singh, and R. Kumar, "Graph neural networks for diabetic retinopathy detection," *IEEE Access*, vol. 10, pp. 4121–4132, 2022.

[28] C. Tao, Y. Yang, S. Liu, L. Zhao, S. Liu, and Y. Huang, "Hierarchical Graph Neural Networks for 3D Mesh Segmentation," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021, pp. 4586–4595, doi: 10.1109/CVPR46437.2021.00456.

[29] X. Zhu, Z. Dai, L. Yang, and F. Peng, "Dual-branch ensemble for diabetic retinopathy detection," *Pattern Recognit. Lett.*, vol. 146, pp. 98–105, 2021.

[30] R. He, Y. Sun, Y. Wang, and Y. Xu, "Multi-scale feature fusion for retinal image analysis using transformers," *IEEE Trans. Med. Imaging*, vol. 42, no. 5, pp. 1203–1212, 2023.