scientific reports

Check for updates

OPEN

Attention dual transformer with adaptive temporal convolutional for diabetic retinopathy detection

Mishmala Sushith[1]✉, Ajanthaa Lakkshmanan[2], M. Saravanan[3] & S. Castro[4]

An Attention Dual Transformer with Adaptive Temporal Convolutional (ADT-ATC) model is proposed in this research work for enhanced detection of Diabetic Retinopathy (DR) from retinal fundus images. Unlike traditional methods which evolved so far in DR analysis, the proposed model specifically processes the multi-scale spatial features through dual spatial transformer network and captures the temporal dependencies through adaptive temporal convolutional unit. The fine patterns like microaneurysms, and larger anatomical regions, including hemorrhages are focused on dual spatial transformer block which provides comprehensive and detailed analysis of spatial features. Additionally, a hierarchical cross attention module is included to fuse the spatial and temporal features which is essential to identify the DR. Experimentation of the proposed model using DRIVE and Diabetic Retinopathy datasets demonstrates the better performance of proposed ADTATC model with an accuracy of 98.2% on DRIVE and 97.7% on Diabetic Retinopathy datasets compared to conventional deep learning models.

Keywords Transformer network, Temporal convolutional Network, Diabetic retinopathy detection, Deep learning, Spatial and temporal features

Diabetic retinopathy (DR) represents a severe complication of diabetes and remains one of the leading causes of blindness among adults worldwide. Early detection of DR is essential to prevent vision loss, and timely interventions mitigate progression and reduce severity. As the number of diabetes patients increases globally it parallelly increases the number of individuals affected by DR[1]. This indicates the necessity of developing better screening tools for early-stage diagnosis and monitoring of disease progression. Due to the advancements in digital imaging techniques in recent years the ability to capture detailed visual information from the retina is improved. This provides a strong foundation for automated DR detection systems and the development of Artificial intelligence methodologies are additionally supports and increases the chances to develop DR detection modules with improved diagnostic accuracy, and increased accessibility. This reduces the dependence on specialized ophthalmologic expertise and provides enhanced patient outcomes in early diagnosis particularly in areas with limited healthcare resources.

Though technological progress and numerous methods for DR detection it faces several limitations[2]. Since the traditional diagnostic techniques mainly involve manual evaluation of retinal images by trained specialists which is time-consuming, expensive, and susceptible to subjective interpretation. The dependence on expert results limits scalability and makes the screening procedure challenging in regions with limited access to specialized care. Additionally, retinal imaging equipment, such as optical coherence tomography (OCT) is effective but costly and introduce discomfort for patients. Thus, it reduces the feasibility of regular screenings. The variability in retinal image quality due to differences in imaging equipment, lighting conditions, and patient cooperation further complicates the diagnosis. Due to this, the demand for automated, accessible, and efficient DR detection solutions evolved by highlighting the need for advanced methodologies which are capable of processing large volumes of retinal images with high accuracy and consistency.

Machine learning and deep learning methods have demonstrated considerable performances in various fields and particularly in medical imaging analysis[3,4]. DL models are efficient in extracting spatial features from retinal images which are then used to identify the pathological structures like microaneurysms, hemorrhages,

[1]Department of Information Technology, Adithya Institute of Technology, Kurumbapalayam, Coimbatore, Tamil Nadu 641107, India. [2]Department of Computing Technologies, School of Computing, SRM Institute of Science and Technology, Kattankaluthur, Chengalpattu, Tamil Nadu 603203, India. [3]Department of Electronics and Communication Engineering, Faculty of Engineering and Technology, Annamalai University, Chidambaram, Tamil Nadu 608002, India. [4]Department of Information Technology, Karpagam College of Engineering, Coimbatore, Tamil Nadu 641032, India. ✉email: mishmalasushith1926@gmail.com

and exudates[5]. Recent research works which combine different learning algorithms as hybrid models have explored the chances of analyzing temporal information in addition to sequential data to capture disease progression. Advanced architectures, including DenseNet, VGG16, InceptionV3[6] have also been adapted for DR detection and provide better detection performance on benchmark datasets[7,8]. In a few cases, ensemble learning approaches are developed by combining multiple models to improve the detection performance. Techniques like transfer learning in addition have further enhanced detection accuracy making these models a better tool for automated DR analysis. Though the developed models have potential merits it faces several challenges which provide variability in detection performance across DR stages[9].

ML and DL models for DR detection exhibit limitations which reduces their adaptability in clinical settings[10]. Though the existing models are accurate they are constrained by their inability to consistently capture both fine-grained spatial features and long-term temporal dependencies in retinal images. This results in suboptimal sensitivity particularly in detecting early-stage DR. Additionally, DL model ability to adapt to the hierarchical and multi-scale nature of retinal images is different for each model which produces different results while analyzing the critical information at both micro and macro scales[11]. The hybrid models in DR detection are beneficial for temporal analysis but face computational challenges with sequence modeling, specially processing an extended temporal data sequence. Also, the black-box nature of deep learning models reduces the interpretability and makes it difficult for clinicians to understand and validate model decisions[12]. Addressing these limitations requires models that can dynamically integrate spatial and temporal data and offer better performance across varying DR stages.

Addressing the limitations observed in current DR detection methodologies is the motivation of this research. Existing methodologies lacks sensitivity in early-stage disease diagnosis due to their limited ability in processing spatial–temporal dependencies. Also, the existing approaches struggle to generalize the model across diverse datasets. Based on this the research objective is defined as to develop a novel model for improved DR detection accuracy and enhance the adaptability across different stages of disease. The proposed work captures the complex spatial and temporal patterns in retinal images associated with DR progression. The proposed model Attention Dual Transformer with Adaptive Temporal Convolutional (ADTATC) introduces several novel elements to overcome the challenges in traditional DL models. The proposed ADTATC model combines dual spatial transformers with an adaptive temporal convolutional memory structure to analyze complex spatial dependencies and long-term temporal changes relevant to DR. Also, the proposed model includes a dual spatial transformer which is optimized for capturing multi-scale retinal features and provides more precise detection of early-stage DR. The adaptive temporal convolutional units provide enhanced memory retention across extended image sequences which is helpful in diagnosing individual disease progression rates without computational constraints. The hierarchical cross-attention mechanism further fuses spatial and temporal features so that key factors on DR are processed to enhance the interpretability and diagnostic performance. The primary contributions of this research are presented as follows.

- Presented a novel dual spatial transformer model to process both small and large-scale patterns in retinal images to attain enhanced spatial sensitivity to DR analysis. Also, adaptive temporal convolutional memory units are proposed for effective temporal analysis of DR progression.
- Presented a hierarchical cross-attention module to fuse spatial and temporal data at multiple stages which highlights the critical DR features and minimizing background noise in DR analysis. Also, an attention mechanism is included within spatial and temporal processing modules to provide a clear insight into model decisions.
- Presented a detailed experimental analysis using benchmark DRIVE and Diabetic Retinopathy datasets demonstrates the better performance of proposed ADTATC model in comparison with existing deep learning models like CNN, RNN, VGG19, Inception V3, Long Short-Term Memory (LSTM) networks and temporal aware hybrid deep learning model (TAHDL).

The remaining discussions in the article are presented as follows. Section "Related works" presents the related works, section "Proposed work" presents the proposed work mathematical model, Section "Results and discussion" presents the results and discussion, and conclusion is presented in section "Conclusion".

Related works

This section presents a brief literature review of existing research on diabetic retinopathy detection. Machine learning and deep learning models for DR detection are considered for analysis and the observations are presented. Various ML models are used over time for DR detection. ML learning classifiers like logistic regression, K-nearest neighbors (KNN), support vector machine (SVM), bagged tree, and boosted tree are comparatively analyzed in[13] for predicting DR from health records. The comparative analysis finds that it boosted tree classifier superior performance over other ML algorithms. The hybrid ML based DR detection model presented in[14] includes multi-stage methodology such as pre-processing, segmentation, feature extraction, and classification to process and colored retinal fundus images. During pre-processing, images are normalized for brightness and contrast enhancement. Segmentation is performed through encoding–decoding layers which isolate blood vessels and retinal regions which are critical for accurate DR detection. Feature extraction utilizes multiple instances learning which capture disease-relevant details from segmented images to enhance classification. Experimental evaluations using benchmark dataset highlights the presented model better performance over existing ML models.

The deep learning-based DR detection model presented in[15] initially performs pre-processing steps such as scaling and adaptive contrast enhancement to enhance image quality. Then the preprocessed images are fed into CNN to extract complex retinal features and classify to detect different stage of DR. Experimental

analysis exhibits that the model achieves high performance in terms of accuracy, sensitivity, and specificity when compared with traditional machine learning classifiers like KNN and SVM. The DR detection model reported in[16] incorporates pretrained VGG16 to classify lesions such as soft exudates, microaneurysms, hard exudates, and hemorrhages. By utilizing image preprocessing and feature extraction techniques the presented model enhances accuracy in DR stage classification. Experimental analysis demonstrated that the model achieved high accuracy and AUC compared to classifiers such as logistic regression (LR) and neural networks (NN). However, the presented model has limitations such as dependency on high-quality, labeled datasets and the need for significant computational power.

The deep neural network-based DR detection model presented in[17] initially preprocess the fundus images and extracts the features using Grey-Level Co-occurrence Matrix (GLCM). Further the extracted features are classified using DNN and compared with SVM through intense experimental analysis. The findings highlight the better performance of DNN over SVM in DR detection with better accuracy and AUC. An optimized DNN based DR detection is presented in[18]. The presented model initially utilizes performance dimensionality reduction and is performed using PCA. Then Firefly algorithm is used for feature extraction which identifies the most relevant features. Followed by DNN is used as classifier to identify DR stages. Experimental evaluations on benchmark dataset highlight that the hybrid model has significant performance over traditional machine learning models. The hybrid DL model presented in[19] for DR detection incorporates CNN and SVM for feature extraction and classification. The presented model initially preprocesses the fundus image and performs data augmentation to increase the number of samples. Then the increased samples are used to train and test the deep learning model, and the extracted features are finally classified using SVM model. The experimental results highlight the better accuracy and precision of presented hybrid model over traditional ML classifiers.

The DR detection model presented in[20] includes multi-Inception-v4 as an ensemble approach to attain improved performance. The Inception-v4 architecture with enhanced pooling and convolution layers used in the presented model provides efficient computation and better feature representation. The performance of ensemble model is evaluated through benchmark datasets and highlights its better sensitivity and specificity. However, the presented model has limitations such as high computational demands, and its performance depends on the high-resolution images. The dual channel fundus image analysis procedure reported in[21] for DR detection incorporates contrast-limited adaptive histogram equalization (CLAHE) and contrast-enhanced canny edge detection (CECED) fundus images. The preprocessed images are then processed through fine-tuned Inception V3 model and VGG-16 model. The outputs of both models are fused in a weighted approach to enhance the detection of critical DR features. Experimental results using benchmark data sets highlights the model better accuracy. However, the approach is limited by its computational requirements due to dual-channel processing.

The DR detection model presented in[22] utilizes hybrid deep learning models for automatic feature extraction and processing. The presented model incorporated U-Net for optic disc and blood vessel segmentation and CNN architectures such as Inception-V3 and ResNet for DR classification. The hybrid model experimental results have demonstrated improved accuracy, sensitivity, and specificity in DR detection. However, the presented model limitation is its requirement of extensive computational resources. A similar U-Net based DR detection model presented in[23] incorporates two U-Net models to segment the optic disc (OD) and blood vessels. Followed by segmentation, the presented model performs feature extraction using a hybrid model that combines CNN and SVD algorithms. Subsequently, an enhanced Inception-V3 model is incorporated through transfer learning to classify the severity of DR. Experimental evaluations highlight the model's high sensitivity and specificity. However, the presented face limitations handling complex feature extraction due to imbalanced class distributions within datasets.

The hybrid model presented in[24] includes deep learning with Harris Hawks Optimization (HHO) to attain improved detection performance in DR analysis. The presented model utilizes Principal Component Analysis (PCA) for dimensionality reduction in the first stage. In the second stage, the essential features are identified while discarding redundant information. In the third stage of the model, HHO is incorporated to optimize these selected features. The experimental analysis highlights the model's improved performance however the high computational cost of the presented model limits its applicability in real time applications. The hybrid deep learning model presented in[25] includes a simple CNN model with ResNet for DR detection. The combined model utilizes ResNet101 for feature extraction and classify the features using CNN model. The experimental results of the presented model highlight its better accuracy and specificity compared to conventional machine learning classifiers. The hybrid model presented in[26] for DR detection incorporates Inception and DenseNet models. The presented model extracts different level features through InceptionV3 and DenseNet121 models and fuse them to attain improved detection performance in DR detection. The experimental results of the presented model highlight the superior detection accuracy of presented model over traditional classifiers. However, the presented limitations include its dependency on high-quality image datasets and potential challenges in processing images with low contrast or occlusions.

Research gap

From the analysis of existing research works given above, several research gaps in diabetic retinopathy (DR) detection approaches are identified, and it is summarized to highlight the need for adaptive methodologies. Existing methodologies largely focus on CNN, SVM, and ensemble methods which are effective but struggle with feature extraction and classification. The consistency across diverse retinal images and varying DR stages limits model performances. Many algorithms depend heavily on high-quality datasets and involve extensive preprocessing steps which lead to feature loss or dependency on resource-intensive image enhancement techniques. Additionally, few methodologies utilize transfer learning but the adaptation of pre-trained models to DR-specific features remains limited which leads to issues in providing better sensitivity and specificity. Existing models also face challenges with class imbalance and generalizability in DR analysis. The proposed approach

addresses these gaps by integrating advanced feature extraction with adaptive attention mechanisms to enhance spatial–temporal detail and by minimizing computational demands. Thereby offering a more robust solution for real-time DR detection, the proposed solution aims to enhance accuracy, reduce false positives, and improve diagnostic performance across diverse datasets and DR stages.

Proposed work

The proposed ADTATC model is designed to enhance the detection performance in diabetic retinopathy (DR) analysis by integrating advanced spatial and temporal feature extraction techniques. The proposed model is developed to address the limitations such as insufficient attention to spatial details and challenges with long-term temporal dependencies which are present in traditional DL architectures. By combining dual transformers for multi-scale spatial attention with adaptive temporal convolutional memory units, ADTATC effectively captures both the complex spatial features of retinal abnormalities. The Dual Spatial Transformer Block (DSTB) in the proposed model is used for spatial feature extraction. The DSTB includes two parallel transformer networks in which the first transformer focusses on fine-grained features and the other focus on larger relevant patterns.

The multi-headed self-attention mechanism within each transformer enables a more comprehensive representation by examining the image from different perspectives. The Adaptive Temporal Convolutional (ATC) Memory Units in the proposed model is used for temporal feature extraction. Unlike recurrent networks, which require sequential processing and face vanishing gradient issues over long-time intervals, ATC provides an efficient path to capture dependencies over short and extended time spans. Dilated temporal convolutions used in the ATC cover a wide range of temporal dependencies with enhanced computational efficiency. Additionally, the adaptive gating mechanism in ATC includes a memory component to retain temporal information based on feature importance. This ensures that complex temporal patterns are preserved across time even though the DR progression varies over time. Incorporating temporal convolutions with adaptive gating the model provides memory-enhanced temporal representation which are essential in the dynamic progression of DR.

The Hierarchical Cross-Attention Module (HCAM) in the proposed work integrates the extracted spatial and temporal features. By using a cross-attention mechanism, HCAM refines the fusion of spatial and temporal data which allows the model to focus on regions that strongly indicate the temporal progression patterns of DR. The hierarchical structure involves multiple cross-attention stages which enables progressive refinement of the fused representation and ensures the minimization of irrelevant information and highlighting of critical DR indicators. Thus, the multi-stage fusion process strengthens the proposed model ability in capturing complex spatial–temporal relationships and also improves interpretability as the model selectively highlights key regions and time points relevant to the diagnosis. The final step in the proposed model involves a fully connected classification layer that processes the fused spatial–temporal features and outputs class probabilities corresponding to various DR stages. Using a sequence of fully connected layers followed by a SoftMax function, the model converts the spatial–temporal features into actionable DR stage detection. The complete overview of the proposed model is presented in Fig. 1. The novelty of proposed ADTATC is present in its hybrid use of attention transformers and adaptive temporal memory as an integrated spatial–temporal diagnostic tool. The process flow from DSTB to ATC and HCAM, followed by classification and is designed to retain both spatial and temporal intricacies which effectively address the needs of a progressive disease like DR. Each module contributes to capturing different aspects specifically DSTB for spatial feature diversity, Adaptive Temporal Convolutional Memory Unit (ATCMU) for temporal continuity, and HCAM for selective fusion. Together, these combinations create a robust model that outperforms traditional methods and offers a better level of precision and reliability in automated DR analysis.

Dual spatial transformer block (DSTB) for spatial feature processing

The DSTB in the proposed ADTATC model is used to capture the small-scale and large-scale spatial features. Multi-scale spatial feature extraction is performed through two parallel transformer networks. Each transformer unit is incorporated specifically to recognize patterns at different scales within the retinal fundus images. This novel combination helps to detect various DR-related abnormalities which appear at different scales. The mathematical model considers the retinal image $I \in R^{(H \times W \times C))}$ as input in which H, W indicates the image height, width and C represents the color channels. The input image is first passed through a learnable embedding function $f(\cdot)$ before processing the image through transformer layers to transform the image into a lower-dimensional representation. The embed function is used as a series of convolutional layers which reduces the spatial dimensions and increases the channel depth. The embedded function is mathematically represented as $E = f\,(I) \in R^{(N \times D))}$. Here N indicates the number of patches, and D indicates the embedding dimension. The image is then divided into a series of patches and each patch is encoded into a vector of dimension D.

In the proposed ADTATC model, the DSTB is designed to extract complementary spatial features by employing two parallel transformer networks that operate on different scales. The small-scale transformer processes high-resolution patches (e.g., 16×16) to capture fine-grained details such as microaneurysms and small hemorrhages. In contrast, the large-scale transformer processes coarser patches (e.g., 64×64) to capture broader anatomical context such as vascular structures.

To avoid confusion regarding parameter sharing, we explicitly differentiate the weight parameters used by each transformer. The small-scale transformer utilizes weight matrices denoted as $W_{Q_{small}}, W_{K\,small}, W_{v\,small}$, while the large-scale transformer employs its own distinct parameters: $W_{Q_{large}}, W_{Klarge}, W_{vlarge}$. This separation ensures that each transformer path is optimally tuned to its respective patch scale, thereby enhancing the overall feature extraction process.

Once the embedded representation E is generated it is linearly projected into three separate spaces like query Q, key K, and value V. These projections are fundamental for the transformer attention mechanism and allow

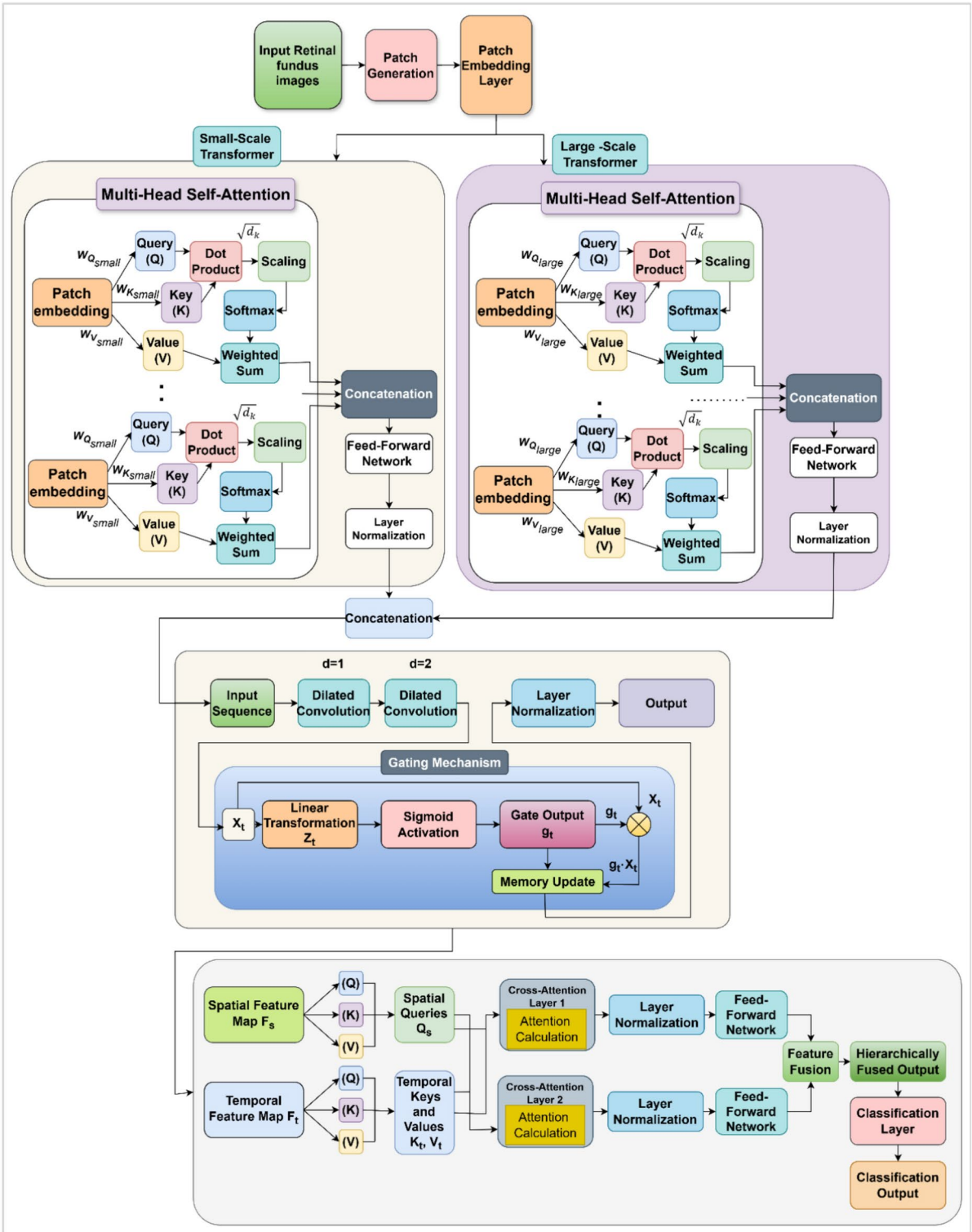

Fig. 1. Proposed model overview.

it to selectively focus on specific regions within the image. The query, key, and value matrices are mathematically formulated as

$$Q = EW_Q, \quad K = EW_K, \quad V = EW_V \tag{1}$$

where $W_Q = R^{D \times d_k}$, $W_K = R^{D \times d_k}$, and $W_V = R^{D \times d_V}$ are the learnable projection matrices. d_k and d_v indicates the dimensions of the query, key, and value vectors, respectively. These dimensions are selected

to balance computational efficiency and feature representations. The query Q, key K, and value V matrices represent transformations of the input data which captures the unique aspects present in retinal image. The DSTB includes a multi-head self-attention mechanism to simultaneously focus on various regions in the image. Figure 2 depicts an illustration of multi-head self-attention mechanism.

The self-attention scores which represent the relationship between each pair of patches are computed by performing dot product of Q and K and scaling by $\sqrt{d_k}$ to maintain stability in the gradients. Mathematically it is represented as follows.

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

where $softmax\,(\cdot)$ normalizes the attention scores across the patches and ensures that the attention weights sum to one for each query. The scaling factor $\sqrt{d_k}$ is used to stabilize the dot product values and prevents excessively large values that could reduce the model learning process. This attention mechanism allows the DSTB to assign higher weights to patches that contain important DR indicators such as lesions or blood vessel abnormalities.

DSTB incorporates two separate transformer networks optimized for process different spatial scales. The first transformer path is configured to capture small-scale features that are essential for detecting fine-grained structures which appear as small, high-detail patterns. The second transformer path is used to detect larger-scale features which require broader contextual understanding. Each transformer path individually applies multi-head self-attention and follows up with a position-wise feed-forward network. This dual-path structure ensures that the model recognizes patterns at both granular and extensive levels. For each transformer path the output from the multi-head attention layer is processed through a feed-forward network (FFN) which is mathematically formulated as

$$\mathrm{FFN}\,(x) = \mathrm{ReLU}\,(xW_1 + b_1)\,W_2 + b_2 \tag{3}$$

where $W_1 \in R^{d_v \times d_f}$ and $W_2 \in R^{d_f \times d_v}$ indicates the learnable weight matrices, b_1 and b_2 indicates the bias terms, d_f indicates the hidden layer dimension, and $\mathrm{ReLU}\,(\cdot)$ is the activation function. The feed-forward network applies non-linearity and enables each path to refine its focus on specific scales. After processing

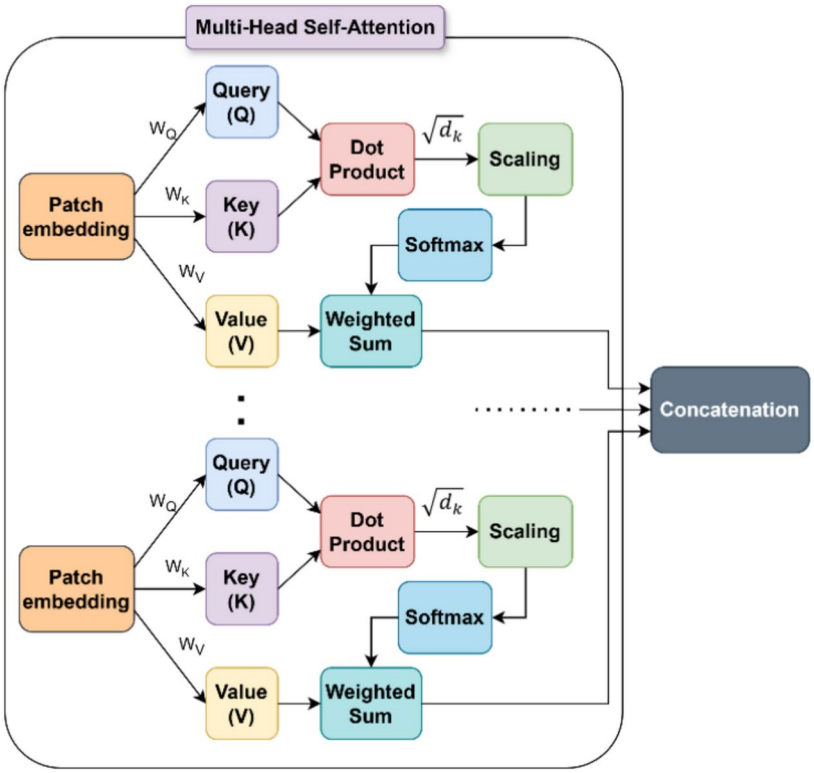

Fig. 2. Multi-Head Self attention module.

the image through both transformer paths the outputs are concatenated to form a combined feature map F_s . Mathematically it is formulated as

$$F_s = concat\left(O_{small}, O_{large}\right) \tag{4}$$

where O_{small} and O_{large} indicates the outputs of small-scale and large-scale transformer paths, respectively. By combining these outputs, the DSTB captures diverse spatial details which are essential for detecting DR. The complete overview of DSTB is presented in Fig. 3.

Adaptive temporal convolutional (ATC) memory unit for temporal feature processing

The Adaptive Temporal Convolutional (ATC) Memory Unit in the proposed model captures temporal dependencies. This helps to retain relevant information over varying time intervals in retinal image sequences. Unlike traditional recurrent networks, ATC utilizes dilated convolutions and adaptive gating to retain complex temporal features. ATC effectively handles short, long-term temporal patterns and processes the sequence of spatial feature maps $\{\mathrm{F}_{(\mathrm{s,t})}\}_{(\mathrm{t}=1)}^{\mathrm{T}}$. Here $F_{s,t} \in R^{N\times D}$ represents the spatial features extracted from the t^{th} retinal image. N indicates the number of spatial patches or regions, and D indicates the feature dimension for each patch. This sequence of spatial features allows the model to process changes over time. To capture temporal dependencies, ATC utilizes dilated convolutions along the time axis. Dilated convolution operation increases the receptive field without increasing the number of parameters. This enables the model to capture temporal dependencies in the feature sequence. The output of the dilated convolution at time t is formulated as

$$y_t = \sum_{i=0}^{k-1} w_i \cdot F_{s,t-d\cdot i} \tag{5}$$

where y_t indicates the output, k indicates the convolutional kernel size, w_i indicates the convolutional filter learnable weights, and d indicates the dilation rate. The dilation rate d controls the spacing between the temporal positions considered by the convolution. This allows the model to cover a larger range of time steps without increasing the computational complexity. By adjusting d, ATC adapts to various time scales and captures the rapid changes and long-term variations in DR progression. To dynamically control the retention of relevant temporal information a gating mechanism is used. The gating mechanism assigns a temporal importance score

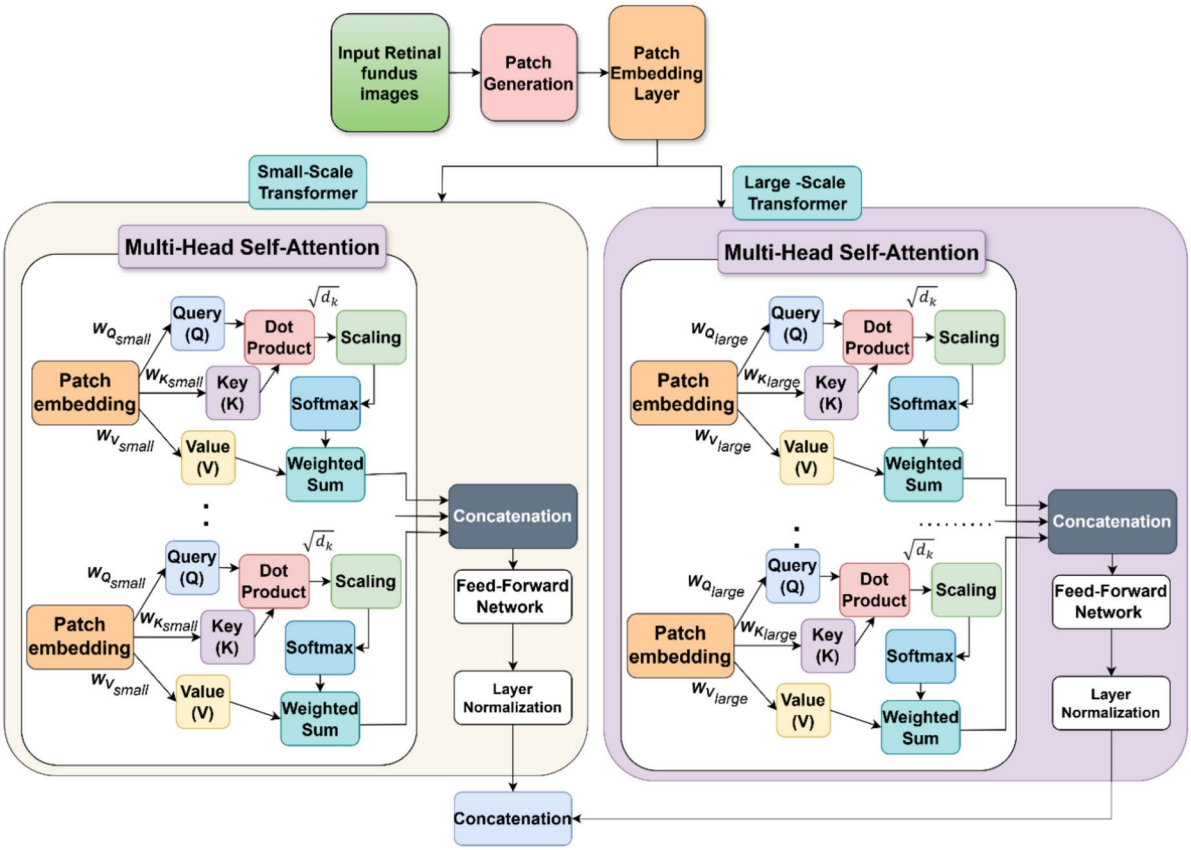

Fig. 3. Dual Spatial Transformer.

to each time step and allows the model to determine which information from that step should be retained. The gate at each time step t is mathematically formulated as

$$g_t = \sigma\left(W_g F_{s,t} + b_g\right) \tag{6}$$

where $g_t \in [0,1]$ indicates the gate value, $W_g \in R^{D\times 1}$ indicates the learnable weight matrix, $b_g \in R$ indicates the bias term, and $\sigma(\cdot)$ indicates the sigmoid activation function. The sigmoid function ensures that the gate values remain between 0 and 1. The value close to 1 indicates the strong retention of temporal information at that step, and the value close to 0 suggests discarding the information. The gating mechanism enables ATC to selectively retain features that are important for DR progression and allows to focus on critical time intervals and minimize noise from irrelevant frames.

Figure 4 depicts the elements of the ATC module. Once the gated output is computed, ATC updates the temporal memory representation by combining the dilated convolution output y_t and the gated input $g_t \cdot F_{s,t}$. This update is formulated as:

$$M_t = g_t \cdot F_{s,t} + (1 - g_t) \cdot y_t \tag{7}$$

where M_t indicates the updated memory at time t. $g_t \cdot F_{s,t}$ highlights the newly processed information based on the gate importance score, while $(1 - g_t) \cdot y_t$ retains past temporal information from the dilated convolution output.

This combination allows ATC to adaptively incorporate both present, past data, and provide a robust temporal representation that aligns with the nature of DR which progresses in severity over time. The memory sequence $\{M_t\}_{t=1}^{T}$ summarizes both recent and long-term temporal dependencies. This temporal representation integrates the dynamic evolution of DR features across sequential images and allows the model to analyze progression and make accurate classifications.

Hierarchical cross-attention model (HCAM) for feature fusion

The Hierarchical Cross-Attention model (HCAM) in the proposed ADTATC fuses the spatial and temporal features to highlight the relevant diabetic retinopathy (DR). HCAM achieves this through a hierarchical cross-attention mechanism that selectively integrates spatial features from retinal images with their corresponding temporal features across multiple stages. This creates a fused feature map that includes spatial and temporal dynamics which are essential for DR detection and progression analysis. The HCAM process two primary inputs, one is from the DSTM model which provides the spatial features $F_s \in R^{N\times D_s}$. The second input to HCAM is from the ATC model which provides the temporal features $F_t \in R^{T\times D_t}$. Where F_s indicates the spatial feature map extracted from a retinal image, where N indicates the number of spatial regions or patches in the image, and D_s indicates the dimensionality of the spatial feature vector for each patch. The temporal feature sequence indicated as F_t with temporal feature dimensionality of D_t. For cross-attention to function both spatial and temporal features are projected into three components as query Q, key K, and value V matrices. These projections enable the cross-attention mechanism to identify relevant spatial regions in relation to temporal changes. The query, key, and value matrices for spatial and temporal features are mathematically formulated as

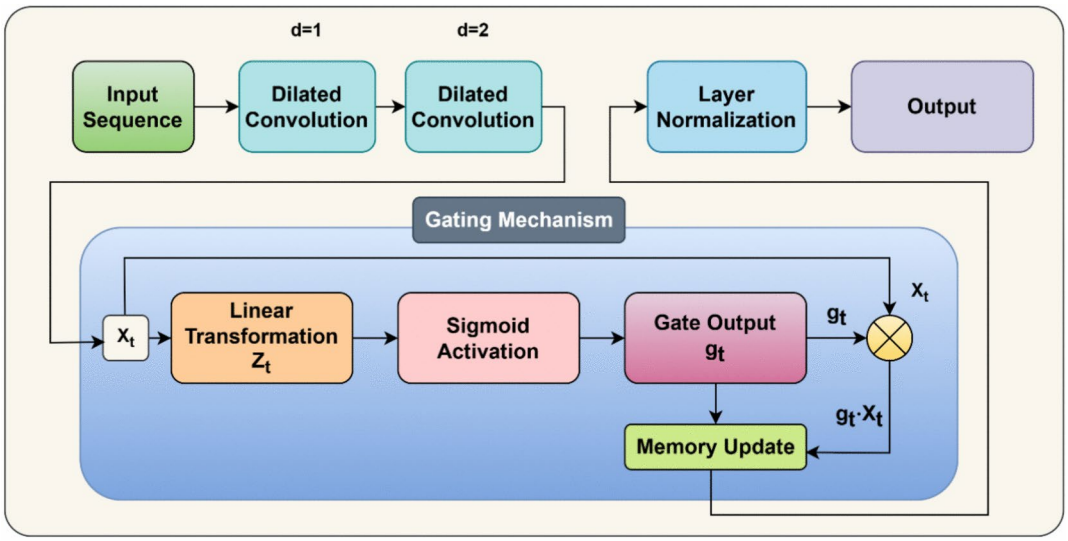

Fig. 4. Adaptive Temporal Convolutional (ATC) Memory Unit.

$$Q_s = F_s W_Q^s \tag{8}$$

$$K_t = F_t W_K^t \tag{9}$$

$$V_t = F_t W_V^t \tag{10}$$

where $Q_s \in R^{N \times d}$ indicates the query matrix for spatial features, $K_t \in R^{T \times d}$ and $V_t \in R^{T \times d}$ indicates the key and value matrices for temporal features. $W_Q^s \in R^{D_s \times d}$, $W_K^t \in R^{D_t \times d}$, and $W_V^t \in R^{D_t \times d}$ indicates the learnable weight matrices that transform spatial and temporal features into a common dimension d which enables attention alignment. Using the query, key, and value matrices, HCAM calculates the cross-attention scores between spatial and temporal features. These scores measure the relevance of each spatial region in relation to each temporal feature and allow the model to focus on spatial regions that are most aligned with DR progression. The mathematical formulation to compute the cross-attention scores are given as follows.

$$CrossAttention\left(Q_s, K_t, V_t\right) = softmax\left(\frac{Q_s K_t^T}{\sqrt{d}}\right) V_t \tag{11}$$

In the above, the $softmax\left(\frac{Q_s K_t^T}{\sqrt{d}}\right)$ produces a normalized attention score matrix and ensures that the attention weights for each query sum to one. The scaling factor $\sqrt{d}$ stabilizes gradients and prevents large values that interrupt the learning process.

This cross-attention operation produces an output matrix where each spatial region's representation is highlighted with relevant temporal context and combinedly creating a temporally aware spatial feature map. Figure 5 depicts the complete process of hierarchical cross attention model used in the proposed work.

The hierarchical aspect of HCAM is attained by applying multiple layers of cross-attention in which each focuses on progressively refined levels of spatial–temporal interaction. Each stage s in HCAM refines the fused spatial–temporal features by reapplying the cross-attention operation with updated spatial and temporal projections. For stage s, the spatial feature $F_s^{(s)}$ is updated which is mathematically formulated as follows

$$F_s^{(s+1)} = CrossAttention\left(Q_s^{(s)}, K_t^{(s)}, V_t^{(s)}\right) \tag{12}$$

where $Q_s^{(s)}$, $K_t^{(s)}$, and $V_t^{(s)}$ indicates the query, key, and value matrices for stage s obtained from the previously updated features. Each stage further refines the spatial–temporal alignment by recalculating attention scores and gradually enhances the focus on critical DR patterns and discards the irrelevant details. This hierarchical attention approach allows HCAM to refine the spatial and temporal integration progressively and leads to a fused representation that captures complex multi-level relationships in the data. After the final stage of cross-attention, HCAM produces a fused feature map F_{st}, which integrates both the spatial layout and temporal progression of DR-related features. The aggregated features at each stage in HCAM provides a highly informative representation that captures the disease spatial characteristics in addition to temporal evolution.

Classification layer

The Final Classification Layer in the ADTATC model is responsible for classifying the fused spatial–temporal features generated by the HCAM to make accurate results on the stage of diabetic retinopathy (DR). Using a fully connected network the fused feature maps are processed and applied non-linear transformations to generate class probabilities corresponding to DR stages. The fused spatial–temporal feature map from HCAM is $F_{st} \in R^{N \times d}$ in which N indicates the number of spatial–temporal patches and d indicates the feature dimensionality. The fused features need to be transformed into a one-dimensional vector before passing it through the fully connected layers. Flattening the feature map into a single vector enables the fully connected layer to handle all the features as a single input, integrating information across all patches. Mathematically it is formulated as

$$x = \text{flatten}\left(F_{st}\right) \in R^{N \times d} \tag{13}$$

The flatten operation reshapes F_{st} into a vector $x \in R^{N \times d}$ and ensures that each spatial–temporal feature is consolidated for the classification process. The flattened feature vector x is then passed through fully connected layers. Each layer applies a linear transformation followed by a non-linear activation function to capture complex patterns in the fused spatial–temporal data that correspond to different DR stages. The fully connected layer output h_i is formulated as

$$h_i = \text{ReLU}\left(h_{i-1} W_i + b_i\right) \tag{14}$$

where $W_i \in R^{h_{i-1} \times h_i}$ and $b_i \in R^{h_i}$ are the weights and biases for layer i. $\text{ReLU}\left(\cdot\right)$ indicates the activation function Rectified Linear Unit which is defined as $\text{ReLU}\left(z\right) = \max\left(0, z\right)$. This activation function introduces non-linearity and helps the model learn complex mappings. These layers enhance the model ability in recognizing higher-order patterns and refine the feature representations. This makes it easier to classify DR stages accurately. The last fully connected layer reduces the output dimensionality to match the number of DR classes C. This

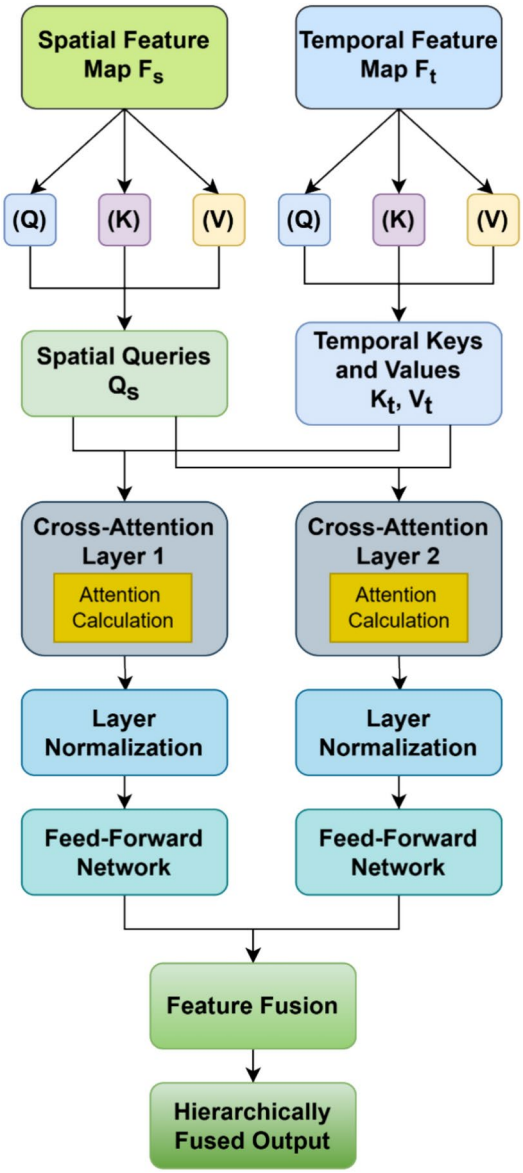

Fig. 5. Hierarchical Cross-Attention Model.

transformation is required to project the output into a space where each dimension represents the likelihood of a particular DR stage. The output z is mathematically formulated as

$$z = h_L W_{fc} + b_{fc} \tag{15}$$

where h_L indicates the output of the last hidden layer, $W_{fc} \in R^{h_L \times C}$ indicates the weight matrix mapping the final hidden layer to the class space. $b_{fc} \in R^C$ indicates the bias vector for the final layer. This linear transformation z produces a score for each DR class which represents the unnormalized likelihood of each class. The output z is then passed through the softmax activation function to convert the scores into probabilities and ensures that each score is non-negative and the sum of probabilities across all classes equals one. The SoftMax function for class j is mathematically formulated as

$$P(j) = \frac{e^{z_j}}{\sum_{c=1}^{C} e^{z_c}} \tag{16}$$

where z_j indicates the score for class j, and the denominator sums the exponentiated scores of all classes C. This transformation provides a probability distribution $P(j)$ across all DR stages in which each probability reflects the model's ability in classifying the input sequence into that specific DR stage. The model's final prediction is obtained by selecting the class with the highest probability

$$Class = \underset{j}{argmax} P(j) \tag{17}$$

where $Class$ indicates the DR stage with the highest probability score, indicating the model's decision on the stage of DR for the input sequence. This classification output provides an interpretable and clinically relevant decision helping to determine the progression level of DR based on the fused spatial–temporal features.

Algorithm: ADTATC model for diabetic retinopathy detection

Input: Retinal image sequence $I = \{I_1, I_2, \ldots, I_T\}$.
Output: Predicted DR stage $(Class)$ *representing the most probable stage of diabetic retinopathy.*

Initialize Parameters: projection matrices W_Q, W_K, W_V *for queries, keys, and values in the transformer layers, weights* w_i *with kernel size* k *and dilation rate* d *for dilated convolutions in ATC, and parameters for the fully connected layers.*
For each image I_t *in* I
Divide I_t *into patches and obtain spatial representation* E_t.
Compute $Q_s = E_t W_Q^s$, $K_s = E_t W_K^s$, *and* $V_s = E_t W_V^s$ *for the spatial transformers*
Apply multi-head self-attention for each transformer

$$\text{Attention}(Q_s, K_s, V_s) = \text{softmax}\left(\frac{Q_s K_s^T}{\sqrt{d_k}}\right) V_s$$

Combine outputs from the small-scale and large-scale transformer paths to form $F_{s,t}$, *the spatial feature map for* I_t
End
For each time step t *from 1 to* T
Apply dilated convolution across spatial features $\{F_{s,t-d\cdot i}\}$ *using* w_i *with* k *and* d

$$y_t = \sum_{i=0}^{k-1} w_i \cdot F_{s,t-d\cdot i}$$

Compute gate $g_t = \sigma(W_g F_{s,t} + b_g)$ *to selectively retain relevant information.*
Update the temporal memory using

$$M_t = g_t \cdot F_{s,t} + (1 - g_t) \cdot y_t$$

Store M_t *as the temporal feature for time step* t
End
Initialize HCAM
Set initial spatial and temporal features for cross-attention $F_s^{(1)} = F_s$ *and* $F_t^{(1)} = F_t$.
For each hierarchical stage s
Project spatial and temporal features to query, key, and value spaces: $Q_s^{(s)}, K_t^{(s)}, V_t^{(s)}$.
Compute cross-attention to integrate spatial and temporal features

$$F_s^{(s+1)} = \text{CrossAttention}(Q_s^{(s)}, K_t^{(s)}, V_t^{(s)}) = \text{softmax}\left(\frac{Q_s^{(s)} (K_t^{(s)})^T}{\sqrt{d}}\right) V_t^{(s)}$$

Set $F_{st} = F_s^{(S)}$, *the final fused spatial-temporal feature map after* S *stages.*
End
Flatten F_{st} *to create vector* x
For each fully connected layer i
$h_i = \text{ReLU}(h_{i-1} W_i + b_i)$
Set $(h_0 = x)$, *and compute each hidden layer output* h_i *sequentially*
Compute final scores z *using* $z = h_L W_{fc} + b_{fc}$
Convert (z) *to class probabilities* $P(j) = \frac{e^{z_j}}{\sum_{c=1}^{C} e^{z_c}}$
Obtain the DR stage as the class with the highest probability $\text{Class} = \arg\max_j P(j)$
If termination condition is reached
End
Else
Repeat
End all
End

S.No	Parameter	Value
1	Image Resolution	512 × 512 pixels
2	Batch Size	32
3	Learning Rate	0.001 (with decay)
4	Optimizer	Adam with Lookahead
5	Loss Function	Cross-entropy with Focal Loss
6	Training—Testing Split	80–20%
7	Number of Epochs	100
8	Activation Functions	ReLU, SoftMax

Table 1. Simulation Hyperparameters.

Class	Total	Training	Testing
DR	100	80	20
Non-DR	100	80	20

Table 2. DRIVE Dataset description.

Class	Total	Training	Testing
No DR	25,810	20,648	5,162
Mild	2,443	1,954	489
Moderate	5,292	4,233	1,059
Severe	873	698	175
Proliferative DR	708	566	142

Table 3. Diabetic Retinopathy Dataset description.

Results and discussion

The experimentation for the proposed ADTATC model evaluates its effectiveness in detecting diabetic retinopathy (DR) through simulation conducted using the Python programming environment with TensorFlow and Keras libraries for deep learning model implementation. This setup allows comprehensive integration of multi-module architecture, enabling efficient execution of dual transformers, temporal convolutions, and cross-attention modules as a unified framework. The experimentations utilize benchmark datasets such as DRIVE[27] and Diabetic Retinopathy datasets[28] as the primary sources of retinal fundus images which includes a wide range of DR stages from healthy to advanced. These datasets provide high-quality images and reliable ground truth labels which are essential for evaluating classification accuracy across varying stages of DR progression. The simulation hyperparameters of the proposed ADTATC model is presented in the Table. 1. The key parameters in the experimentation included the batch size which is set to 32 to balance memory usage and computational efficiency. The learning rate is initialized at 0.001 with a decay strategy to improve convergence over epochs. The Adam optimizer with Lookahead was selected for its stability and adaptive learning capabilities so that smooth gradient updates during training can be obtained. For model training, 80% of the data was allocated as the training set, 20% was used for testing to ensure that the model had sufficient data to learn from while also being evaluated for generalization on unseen images. Table 1 summarizes the simulation hyperparameters. The complete details about the dataset are presented in Table 2 and Table 3 for DRIVE and Diabetic Retinopathy Dataset, respectively.

The performance metrics utilized in the proposed model evaluation are accuracy, precision, recall, F1-score, and specificity to verify the model's capability across all severity levels. Cross-entropy loss with focal loss modification was employed to mitigate the class imbalance often observed in medical datasets, ensuring that the model did not overlook underrepresented DR stages. Training was conducted for 50 epochs, with early stopping to prevent overfitting. Throughout the experimentation, GPU acceleration was utilized to handle the computational demands of the model's transformer-based architecture and large dataset sizes. The formulations for the performance evaluation metrics are presented as follows.

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

Metric	Train	Test
Accuracy	0.9864	0.9826
Precision	0.9906	0.9822
Recall	0.9989	0.9986
F1-Score	0.9947	0.9903
Specificity	0.9846	0.9821

Table 4. DRIVE dataset metrics.

Metric	Train	Test
Accuracy	0.9815	0.9744
Precision	0.9912	0.9869
Recall	0.9918	0.9884
F1-Score	0.9915	0.9876
Specificity	0.9924	0.9868

Table 5. Kaggle diabetic retinopathy dataset metrics.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{20}$$

$$Specificity = \frac{TN}{TN + FP} \tag{21}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{22}$$

where the true positive is indicated as TP, the true negatives are indicated as TN, the false positives are indicated as FP and false negatives are indicated as FN. The proposed model performance in training and testing is depicted in Table 4 and 5 for DRIVE dataset and Diabetic retinopathy dataset, respectively.

From Tables 4 and 5 the better performance of proposed model can be observed in both training and testing process. The proposed model attained maximum accuracy of 98.64% in the training process and 98.26 in the test process for DRIVE dataset. Similarly, the proposed model exhibits 98.15% as accuracy during the training and 97.44% as accuracy during the test process for diabetic retinopathy dataset.

The training and validation accuracy in Fig. 6 and the corresponding loss is depicted in Fig. 7 for the DRIVE dataset exhibit the proposed model learning progression in diabetic retinopathy (DR) detection. The accuracy graph exhibits a rapid increase within the initial epochs with training accuracy reaching over 96% around the 20th epoch. The validation accuracy closely follows the training which indicates the model generalization ability. The model performance stabilizes beyond the 30th epoch, achieving near-optimal accuracy around 98% demonstrating the proposed model effective learning of DR features.

The loss graph given in Fig. 7 shows a decrease in both training and validation loss, dropping below 0.1 by the 20th epoch. This indicates the model's efficiency in minimizing errors. The proposed ADTATC model utilizes dual transformers and adaptive temporal convolutional memory which allows to capture spatial patterns and progressive DR stages more effectively. The minimal gap between training and validation curves across both graphs highlights the model generalization ability and it further enhanced by adaptive gating mechanisms, cross-attention modules that selectively focus on DR-relevant features, reduces the irrelevant noise and preventing overfitting.

The training and validation accuracy in Fig. 8 and the corresponding loss are depicted in Fig. 9 for the diabetic retinopathy dataset show the effectiveness of the proposed model in the DR detection process. The accuracy graph exhibits a rapid rise in both training and validation accuracy with the initial epochs and reaching over 95% by the 20th epoch. This high accuracy indicates that the model ability in learning complex DR features. Training accuracy stabilized close to 98% by the end and the validation accuracy closely follows at approximately 97%. The loss graph shows in Fig. 9 exhibit rapid decline with both training and validation loss dropping below 0.1 by the 15th epoch. This decrease in loss reflects the model's efficiency in minimizing prediction errors. The low training and validation loss signify effective learning without significant overfitting. The proposed ADTATC model's superior performance can be attributed by this minimal error and demonstrates the model robustness and reliability in DR detection.

The Precision-Recall analysis for the DRIVE dataset given in Fig. 10 and the diabetic retinopathy dataset given in Fig. 11 highlights the proposed ADTATC model superior performance in detecting diabetic retinopathy (DR). In Fig. 8, the PR curves for No DR and DR categories exhibit better AP values of 0.9948 and 0.9932 respectively indicate their high precision and recall across both classes. The high AP values demonstrate the

Fig. 6. Analysis of Training and validation accuracy for DRIVE dataset.

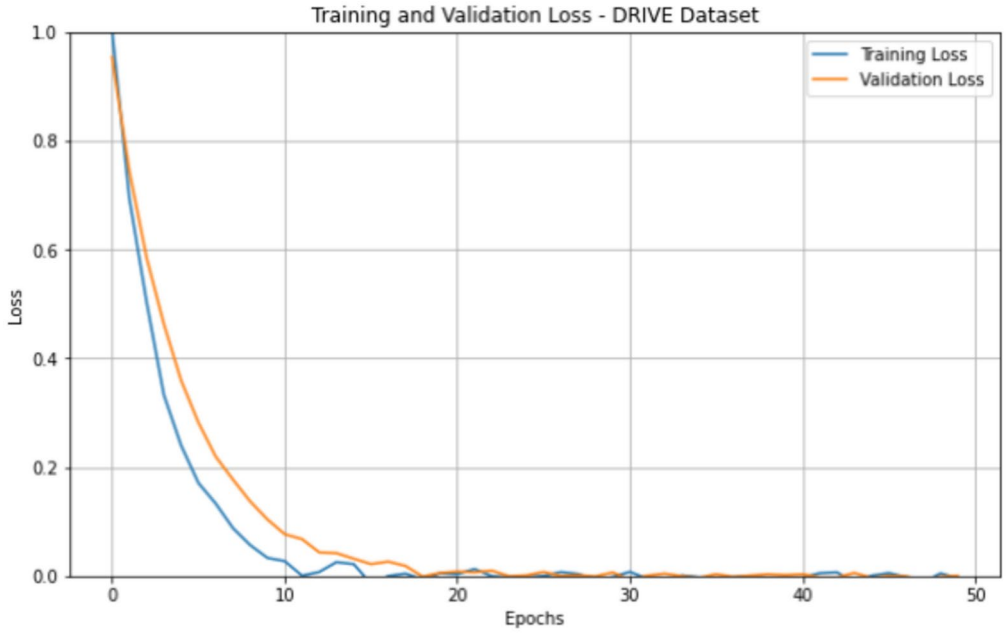

Fig. 7. Analysis of Training and validation Loss for DRIVE dataset.

proposed model's effectiveness in accurately identifying DR cases while minimizing false positives. For the diabetic retinopathy dataset PR curve exhibited in Fig. 9, the DR stages including No DR, Mild, Moderate, Severe, and Proliferative are tightly aligned with AP values close to 0.99 for each category. For Proliferative DR achieving an AP of 0.9958. The model ability to maintain high precision across stages reflects its robustness in handling varying severity levels which is essential for accurate DR staging. The slightly lower AP in earlier DR stages such as Mild (0.9941) is due to overlapping features with normal cases which is a challenge in DR detection.

Further evaluation of proposed model considered traditional DL models like CNN, RNN, VGG19, Inception V3, Long Short-Term Memory (LSTM) networks and temporal aware hybrid deep learning model (TAHDL) for comparative analysis. The TAHDL model is the early-stage experimentation of our research in diabetic retinopathy which combines CNN and RNN for temporal feature processing. Each DL model's performance is evaluated individually and finally compared with the proposed model. For all the methods, batch size is selected

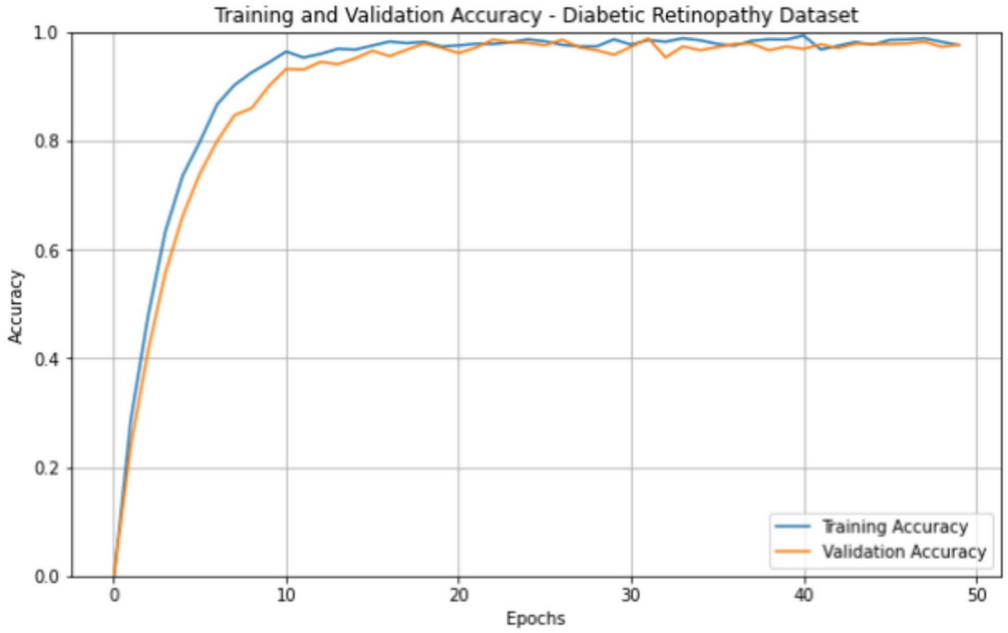

Fig. 8. Analysis of Training and validation accuracy for diabetic retinopathy dataset.

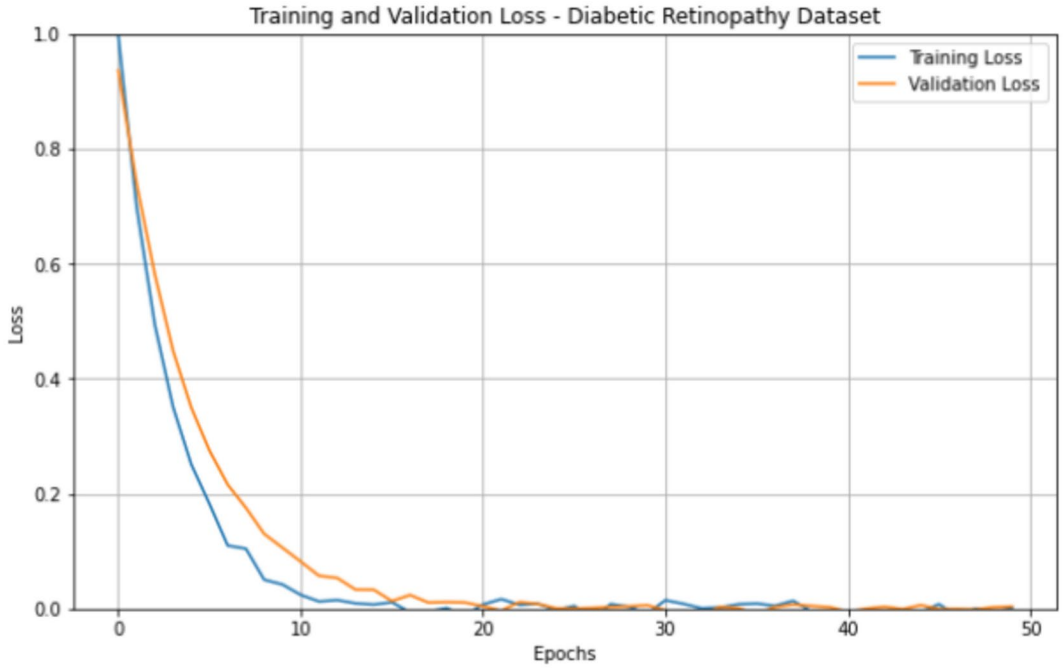

Fig. 9. Analysis of Training and validation loss for diabetic retinopathy dataset.

as 32, epoch as 50 and loss function is categorical entropy loss. The dropout rate is selected as 0.5 for all models. The simulation hyperparameters for the existing DL models are presented in Table 6.

The precision comparative analysis graphs given in Fig. 12 for the DRIVE dataset and Fig. 13 for the diabetic retinopathy dataset highlight the better performance of the proposed ADTATC model over existing models across 50 epochs. In Fig. 12, the precision of the proposed ADTATC model exhibits its significant improvement starting from 0.90 and reaching over 0.98 by the final epoch. This demonstrates the superior performance compared to traditional models such as CNN, RNN, VGG19, and LSTM, which obtained precision in the range of 0.88 to 0.92 over time. The existing TAHDL model exhibits some better improvement, compared to others with a maximum of 0.94 by the 50th epoch but it is still lesser than the proposed ADTATC. Similarly, in Fig. 13 for the diabetic retinopathy dataset, the precision of the proposed ADTATC model overcomes the other models

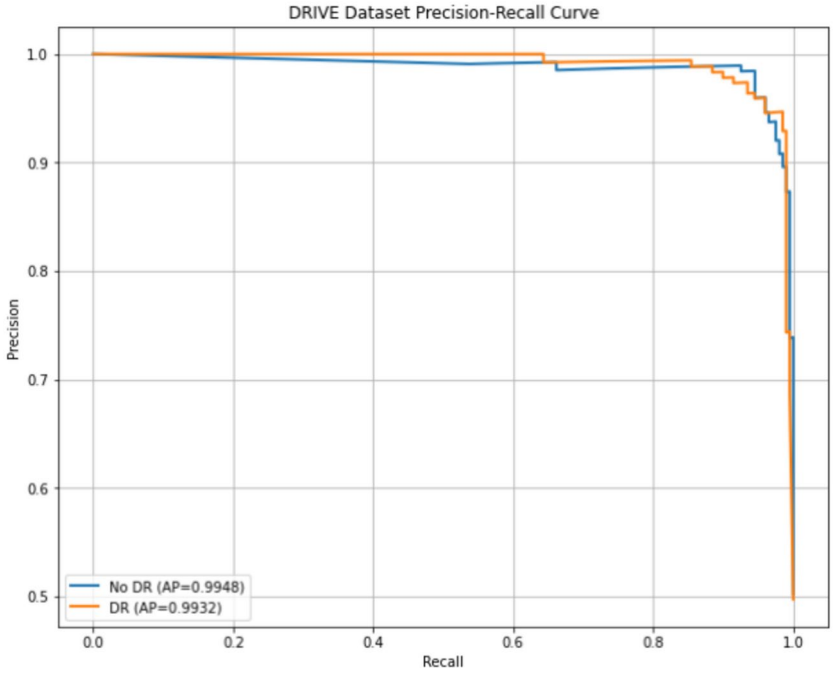

Fig. 10. Precision Recall analysis for DRIVE dataset.

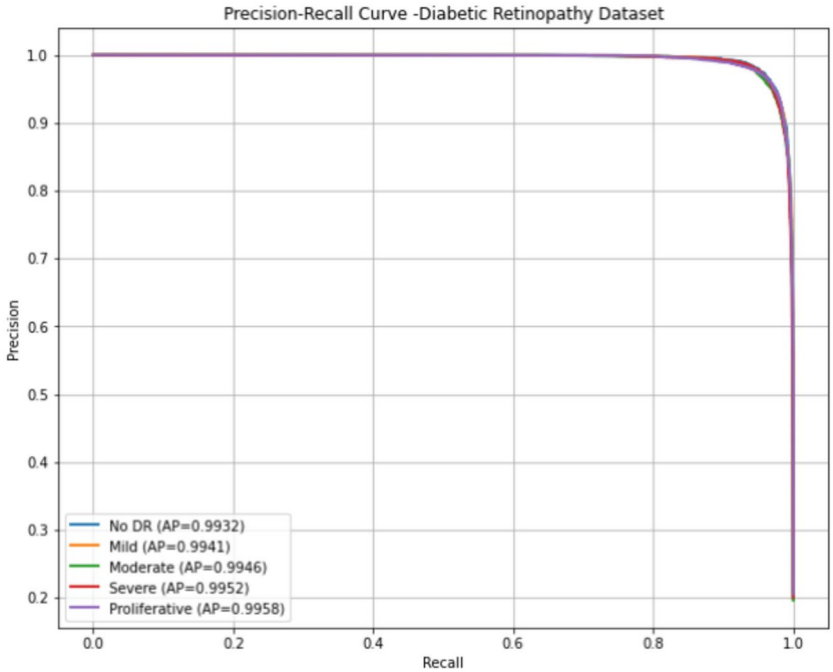

Fig. 11. Precision Recall analysis for diabetic retinopathy dataset.

with clear increment in precision from 0.90 initially and reaches maximum 0.986 by 50th epoch. Conventional models like Inception and LSTM exhibit less precision and reach around 0.91. The proposed model's superior performance can be attributed to its dual transformer architecture that captures complex spatial features and the adaptive temporal memory that enhances temporal feature retention.

The recall comparative analysis graphs given in Fig. 14 for the DRIVE dataset and Fig. 15 for the diabetic retinopathy dataset highlight the better performance of the proposed ADTATC model over existing models across 50 epochs. In Fig. 14, the recall of the proposed ADTATC model exhibits its significant improvement starting from 0.92 and reaching almost 1.0 by the 50th epoch. This increased performance of the proposed

S.No	Model	Hyperparameters	Range/Type
1	CNN	Conv Layer Filter	64
2		Conv Layer Filter Size	3 × 3
3		Conv Layer Activation	ReLU
4		Pooling Layer Type	Max Pooling
5		Pooling Layer Size	2 × 2
6		Pooling Layer Stride	2
7		Output Layer Activation	SoftMax
8		Optimizer	Adam
9		Learning Rate	0.001
10		RNN Layer Units	128
11	RNN	RNN Layer Activation	tanh
12		Optimizer	Adam
13		Learning Rate	0.001
14	VGG19	Optimizer	SGD
15		Learning Rate	0.0001
16	InceptionV3	Optimizer	RMSprop
17		Learning Rate	0.0001
18	LSTM	LSTM Layer Units	256
19		LSTM Layer Activation	tanh
20		LSTM Layer Rec. Activation	sigmoid
21		Optimizer	Adam
22		Learning Rate	0.001
23	TAHDL	Number of Epochs	50
24		Learning Rate	0.001
25		Batch Size	32
26		Optimizer	Adam
27		Dropout Rate	0.5
28		Activation Function	ReLU (for CNN), Tanh (for RNN)
29		Loss Function	Categorical Cross-Entropy
30		Regularization	L2 Regularization (λ = 0.01)

Table 6. Simulation hyperparameters of deep learning algorithms.

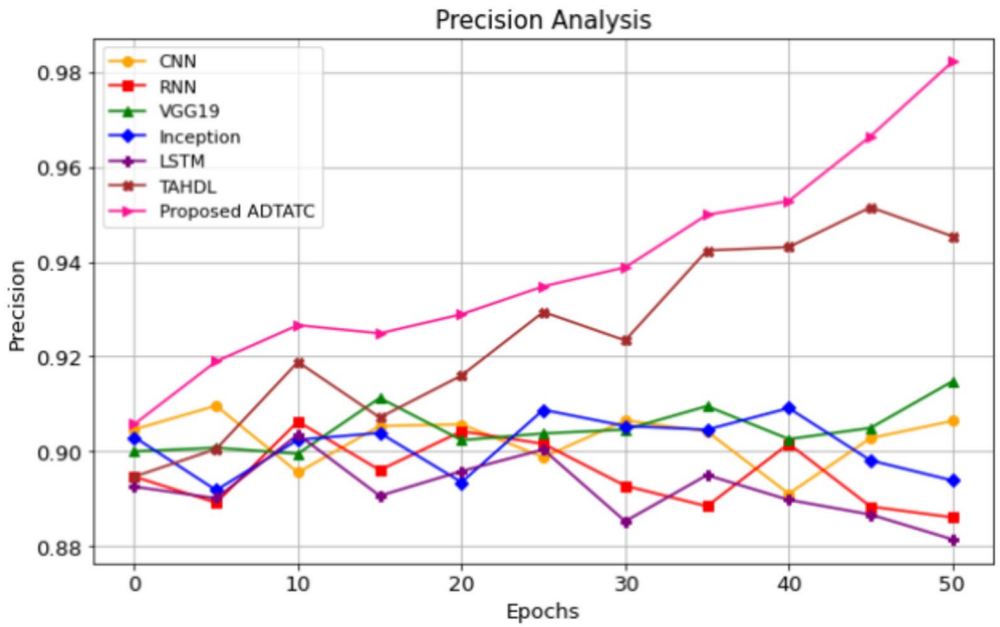

Fig. 12. Precision comparative analysis for DRIVE dataset.

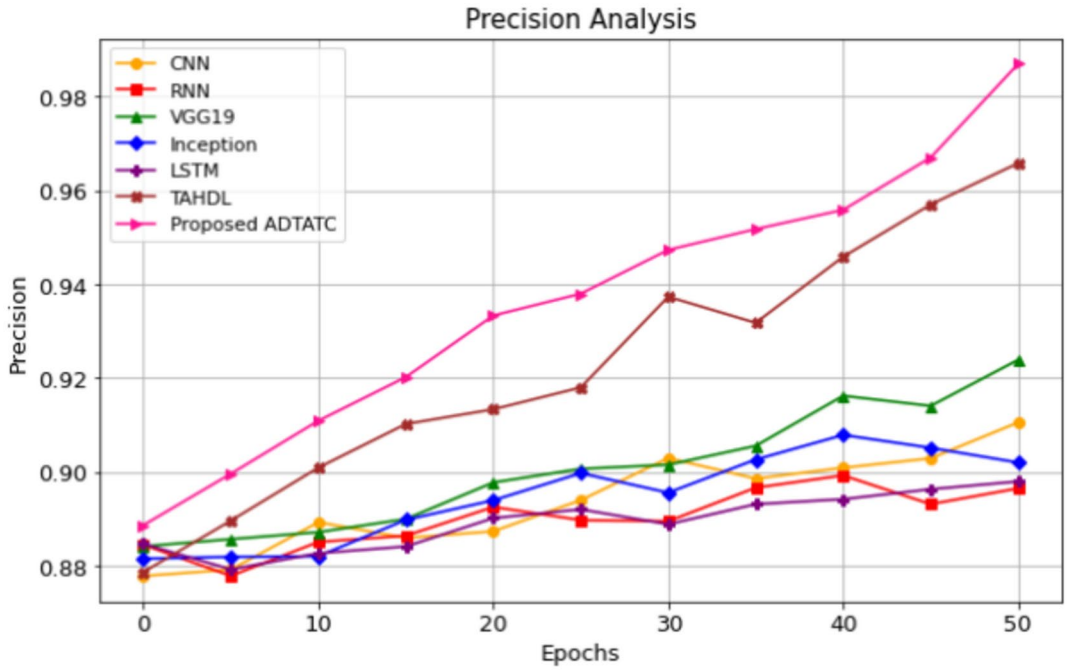

Fig. 13. Precision comparative analysis for Diabetic Retinopathy dataset.

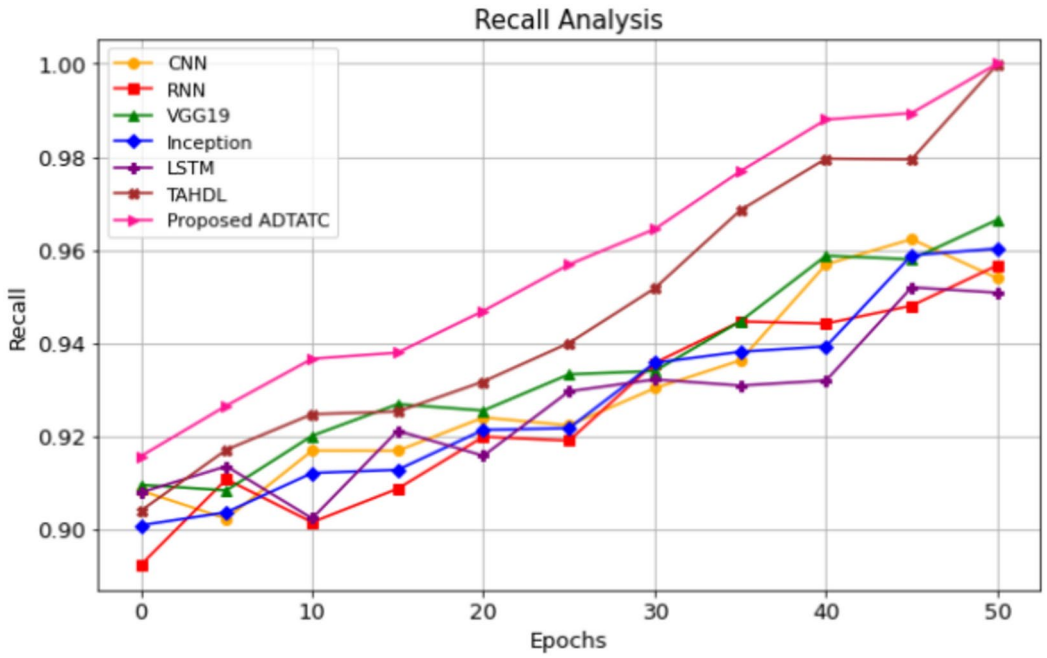

Fig. 14. Recall comparative analysis for DRIVE dataset.

model highlights its ability to capture true positives in DR detection. The proposed model outperformed other models like CNN, RNN, and Inception, which fluctuated around the 0.90 to 0.94 range. The existing TAHDL model exhibits some better improvement, compared to others with a maximum of 0.97 by the 50th epoch but it is still lesser than the proposed ADTATC. Similarly, in Fig. 15 for the diabetic retinopathy dataset, the recall of the proposed ADTATC model overcomes the other models with clear increment in precision from 0.90 initially and reaches maximum 0.98 by 50th epoch. This superior recall performance of the ADTATC model is attributed due to its advanced architecture and its ability to identify subtle DR features across various stages ensures fewer false negatives yields better recall compared to existing methods.

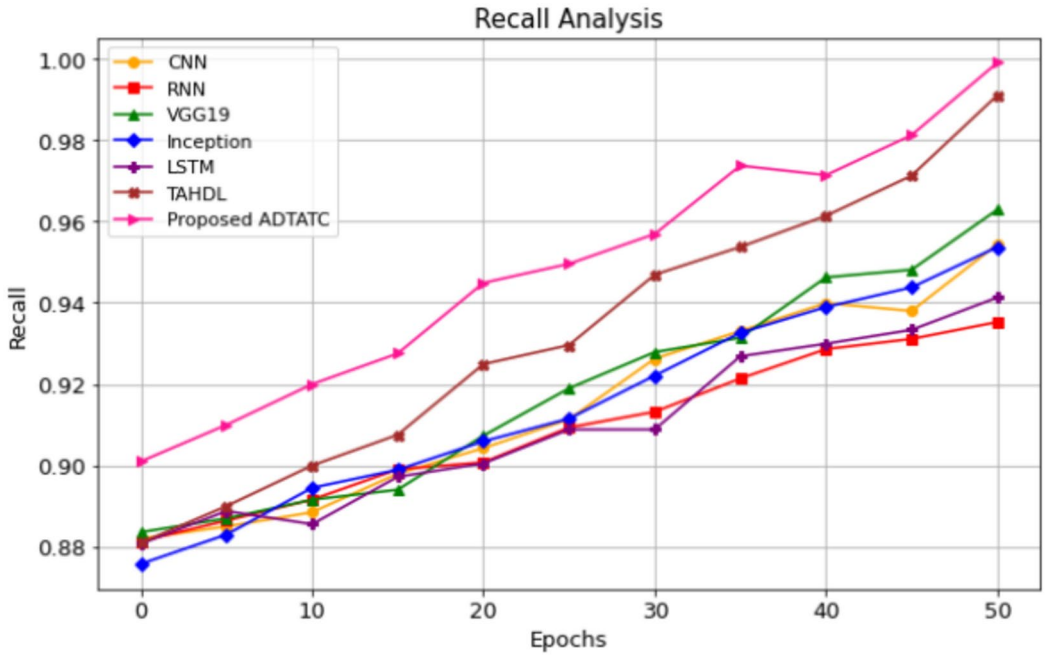

Fig. 15. Recall comparative analysis for Diabetic Retinopathy dataset.

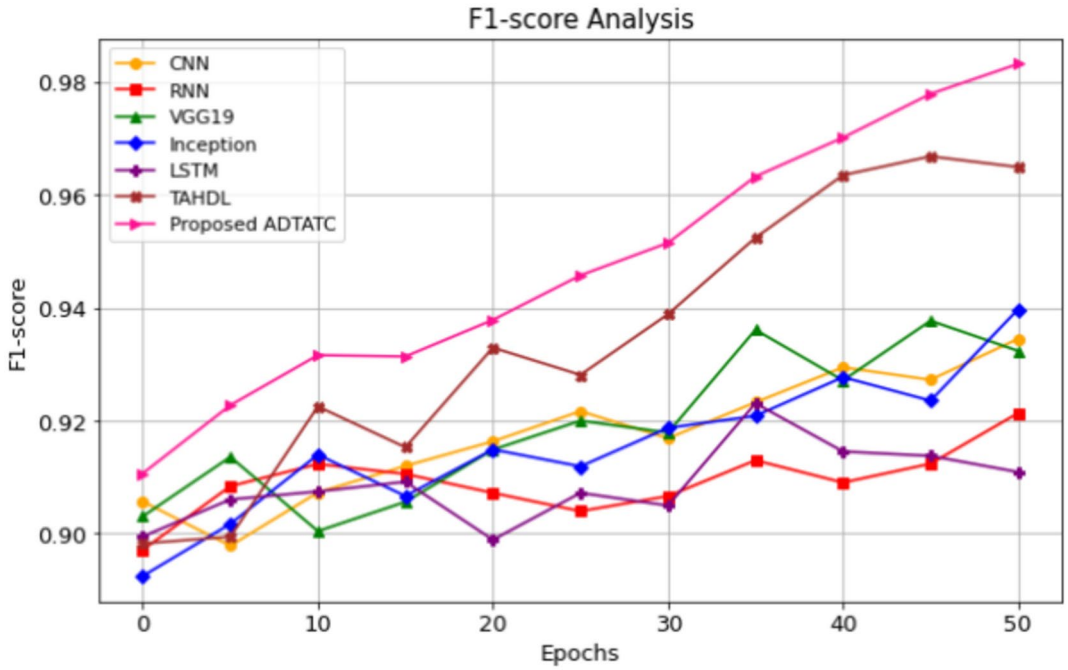

Fig. 16. F1-Score comparative analysis for DRIVE dataset.

The F1-score comparative analysis given in Fig. 16 for the DRIVE dataset and Fig. 17 for the diabetic retinopathy dataset highlights the superior performance of the proposed ADTATC model compared to existing models across multiple epochs. In Fig. 16, the F1-score of the proposed ADTATC model is exhibited and it starts around 0.92 and reached maximum of 0.98 by the 50th epoch. This increased f1-score highlights the model's ability to maintain a balanced precision and recall. Also, the increased performance ensures the model high reliability in DR detection. Traditional models like CNN, RNN, and LSTM show minimal improvement in F1-scores in the range of 0.90 to 0.94, indicating their limitations in accurately handling complex DR patterns. The TAHDL model reaches up to 0.96 but it is lesser than proposed ADTATC model performance. In Fig. 17, The proposed ADTATC F1-score on the diabetic retinopathy dataset exhibits a similar increase which starts nearly

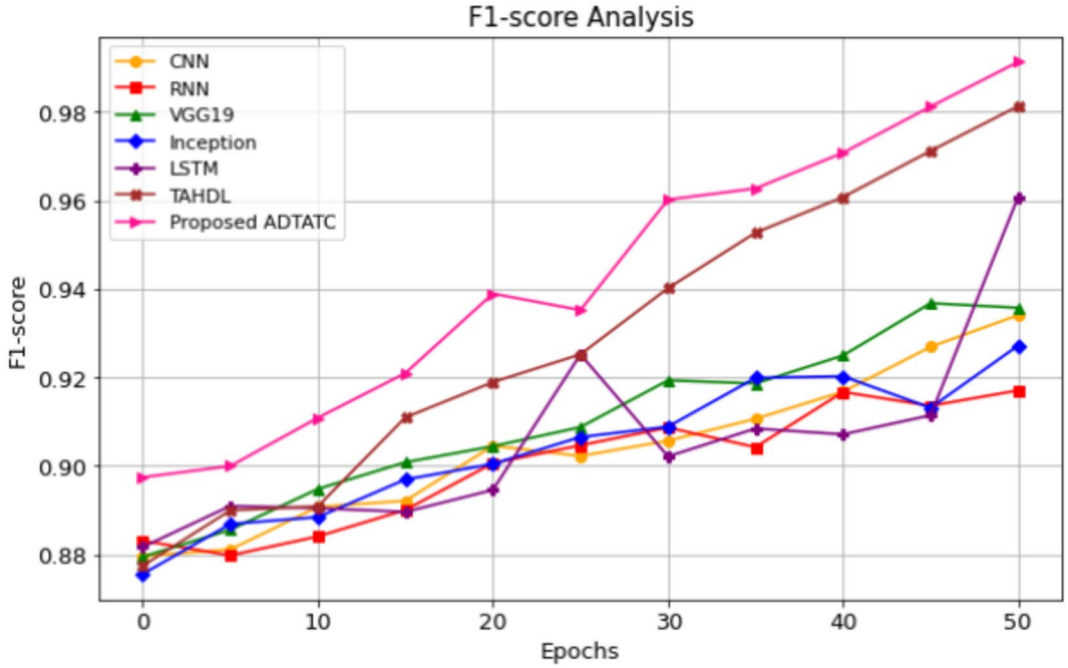

Fig. 17. F1-Score comparative analysis for Diabetic Retinopathy dataset.

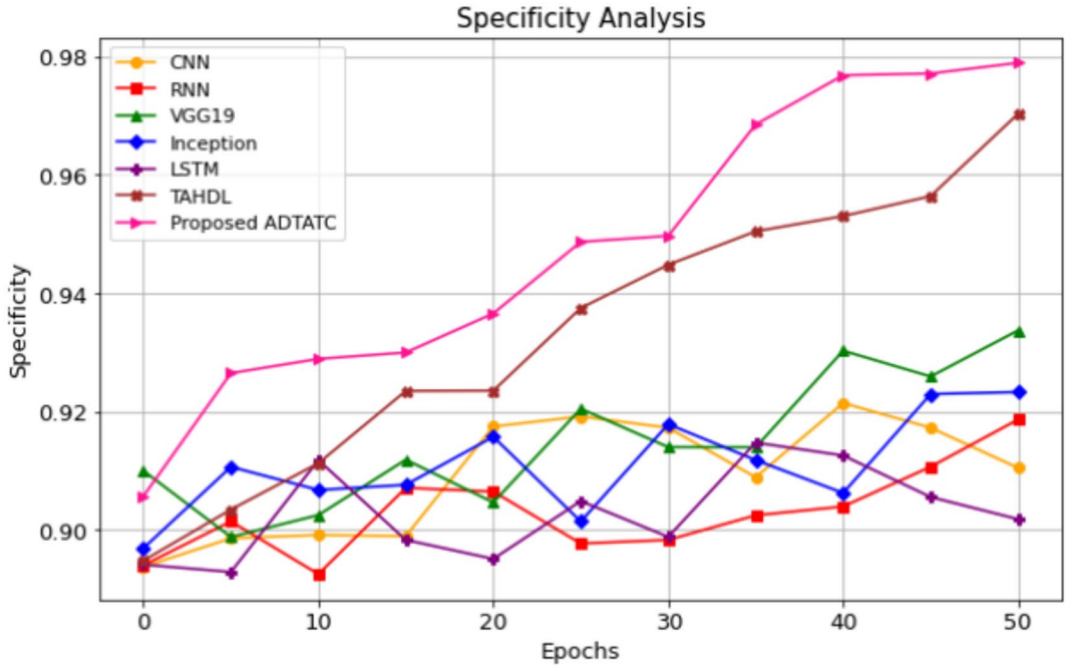

Fig. 18. Specificity comparative analysis for DRIVE dataset.

0.90 and reaches maximum of 0.987 by the end. The improvement in F1-score reflects the proposed ADTATC model ability in minimizing the false positives and false negatives across various DR stages. In contrast, other models like Inception and VGG19 exhibit lesser performance which reaches maximum in the range of 0.94 exhibit their limited feature extraction capabilities in a complex medical imaging context. The higher F1-score of ADTATC compared to other existing methods demonstrates its robustness in detecting and classifying DR with a balanced precision-recall, but still, it is lesser than the proposed ADTATC model.

The specificity comparative analysis given in Fig. 18 for the DRIVE dataset and Fig. 19 for the diabetic retinopathy dataset presents the specificity attained by the proposed ADTATC model compared to other models across 50 epochs. In Fig. 18, the specificity of proposed ADTATC exhibits 0.91 in the beginning and reaches

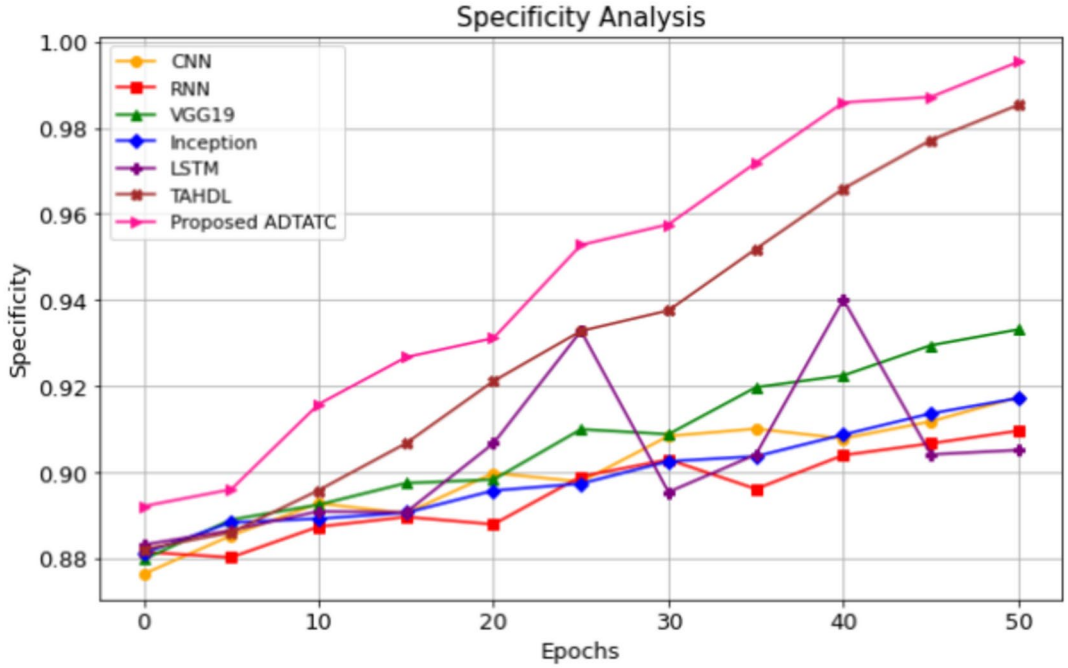

Fig. 19. Specificity comparative analysis for Diabetic Retinopathy dataset.

maximum of close to 0.98 by the final epoch. This performance is higher than that of traditional models like CNN, RNN, and LSTM, which exhibit specificity in the range of 0.90 to 0.93. The lesser performance of existing methods demonstrates their limitations in accurately differentiating true negatives. The existing TAHDL model performs better than these traditional models with 0.96 by epoch 50, but still, it is lesser than the proposed ADTATC specificity. In Fig. 19, the proposed ADTATC model specificity on the diabetic retinopathy dataset shows similar increased performance from beginning with 0.90 as specificity and reaches maximum of 0.986 by the last epoch. This improvement in specificity highlights the model ability in minimizing false positives and ensuring that non-DR cases are correctly identified as non-DR. This is essential for clinical applications where an accurate exclusion of healthy cases is essential to reduce unnecessary follow-ups. The proposed ADTATC model enhanced specificity validates its precise classification whereas the traditional models lack this capability which results in less stable and lower specificity scores in DR detection.

The accuracy comparative analysis given in Fig. 20 for the DRIVE dataset and Fig. 21 for the diabetic retinopathy dataset highlight the superior performance of the proposed ADTATC model compared to existing models for 50 epochs. In Fig. 20, the accuracy of ADTATC starts around 0.91 and increases steadily till 50th epoch and reaches maximum of 0.982 by the 50th epoch. This consistent increase in accuracy reflects the proposed model effective learning and ability to generalize over traditional models like CNN, RNN, and Inception which exhibits more fluctuating accuracy scores in the range of 0.92 to 0.94. The existing TAHDL model performs better than the traditional models with an accuracy close to 0.96 by the last epoch but it is lesser than the proposed ADTATC maximum accuracy. Similarly, in Fig. 21, the ADTATC model exhibits an increased accuracy on the diabetic retinopathy dataset which starts from 0.89 and reaches approximately 0.978 by 50th epoch. This improvement over time highlights the proposed ADTATC performance which was attained due to the combination of dual transformers for enhanced spatial representation and adaptive temporal memory units for effective tracking of temporal DR features. The advanced attention mechanisms enable ADTATC to focus on relevant DR patterns which improve accuracy significantly over other existing models.

The experimental results clearly demonstrate the superiority of the proposed ADTATC model over existing methods across performance metrics such as accuracy, precision, recall, F1-score, and specificity. The proposed ADTATC consistently achieved higher values reaching nearly 0.9826 as accuracy for DRIVE dataset and 0.9744 as accuracy for diabetic retinopathy dataset. The proposed model performance is better and outperforms traditional models like CNN, RNN, and TAHDL for all the metrics. The proposed model architecture ensures better generalization, higher detection accuracy, and fewer false positives making it highly suitable for clinical applications. The proposed model has been compared with recent state-of-art hybrid models and the outcomes are projected in Table 7.

Table 8 presents the additional validation of proposed model done on DDR dataset. The DDR dataset[27] is a large and diverse dataset specifically designed for diabetic retinopathy (DR) analysis, comprising a total of 13,673 fundus images collected from 147 hospitals across 23 provinces in China. This diversity in geographical and clinical sources ensures a wide variety of image characteristics, including differences in quality, resolution, and disease representation, making the dataset particularly challenging and comprehensive. The dataset classifies images into five distinct categories based on the severity of DR: none, mild, moderate, severe, and proliferative DR. Additionally, a sixth category is included to label images of poor quality, which adds another layer of

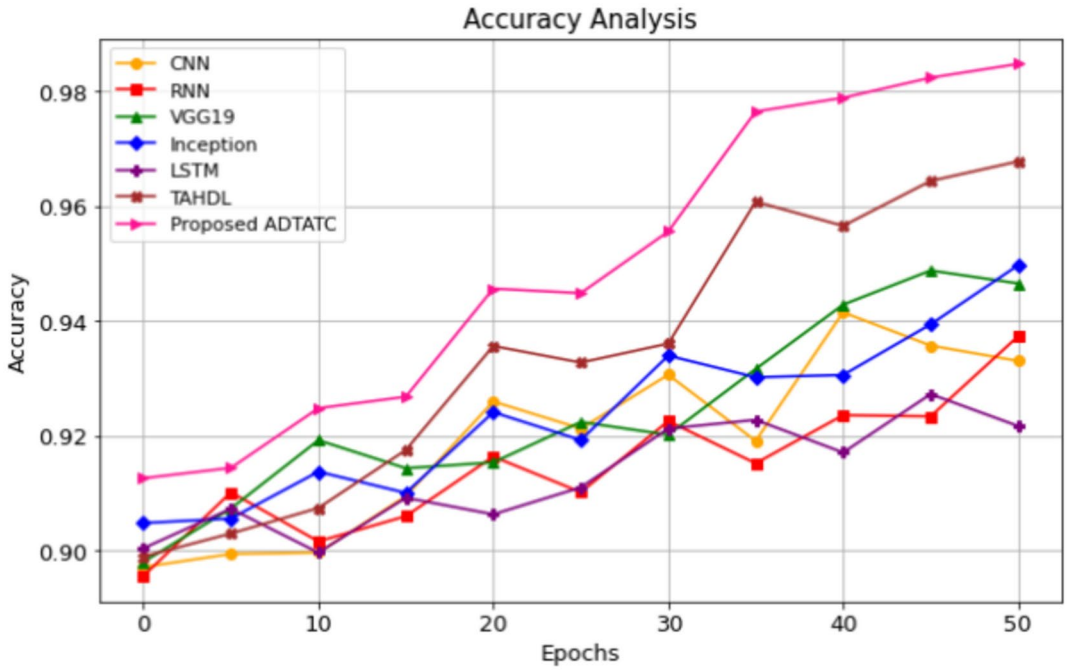

Fig. 20. Accuracy comparative analysis for DRIVE dataset.

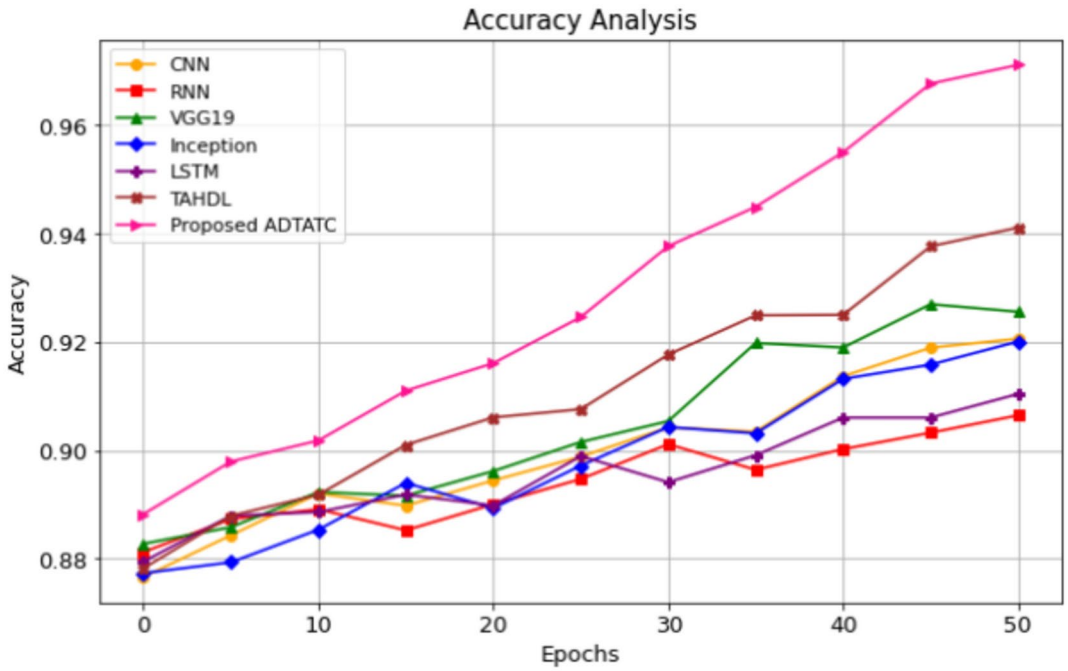

Fig. 21. Accuracy comparative analysis for Diabetic Retinopathy dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)
Proposed ADTATC	98.2	98.1	99.8
DNN-PCA-Firefly[28] (2020)	97	96	96
WFDLN[29] (2022)	96.5	97.8	98.7
DNN-PCA-Harris Hawks[30] (2023)	96.9	96.8	97.1

Table 7. Outcomes with existing Hybrid Models.

Model	Accuracy (%)	Precision (%)	Recall (%)
Proposed ADTATC	91.5	92.1	91.8
TAHDL	87.2	86.4	87.1
LSTM	85.4	84.8	85

Table 8. Validation on DDR Dataset.

Configuration	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)
Single Small-Scale Transformer	88.2	88.6	88.4	88.1
Single Large-Scale Transformer	89.3	89.7	89.5	89.2
Medium Scale	89	89.4	89.2	89
Small Scale + Medium Scale	91	91.5	91.2	90.2
Medium Scale + Large Scale	91.2	91.7	91.4	91
Dual-Scale Transformer (Proposed)	91.5	92.1	91.8	91.2
Dual + Medium-Scale Transformer	92.0	92.4	92.2	91.8

Table 9. Outcome on Ablation Studies (DDR Dataset).

complexity as the model must not only classify DR severity but also identify and disregard suboptimal images. These features make the DDR dataset a robust benchmark for evaluating the adaptability and effectiveness of DR detection models.

The observed decrease in accuracy from 98% on the DRIVE and Diabetic Retinopathy datasets to 91.5% on the DDR dataset can be attributed to several factors. First, the large volume and diversity of the DDR dataset introduce significant variability in image quality, illumination conditions, and patient demographics, challenging the model's ability to generalize. Unlike smaller datasets with controlled imaging environments, DDR contains images of varying resolutions and noise levels, including the category of poor-quality images, which makes classification more difficult. Second, the inclusion of five severity classes and an additional poor-quality category increases the complexity of the classification task. In particular, the overlapping features among early DR stages such as mild and moderate can lead to higher false positives or false negatives, affecting overall performance metrics.

Moreover, the DDR dataset's class imbalance, where certain DR severity levels may be underrepresented compared to others, poses additional challenges. Despite the use of focal loss and adaptive temporal mechanisms in the proposed model to mitigate this issue, some degradation in accuracy is expected when dealing with real-world, imbalanced datasets. Finally, the higher variability and complexity inherent in the DDR dataset align closely with the challenges encountered in clinical settings, making it a more realistic but demanding benchmark. The proposed ADTATC model's accuracy of 91.5% still represents a significant improvement over prior approaches and demonstrates its ability to handle challenging datasets effectively.

For the proposed ADTATC model, the kappa coefficient was calculated based on the classification results across the five DR severity levels (none, mild, moderate, severe, proliferative) and the poor-quality category. On the DDR dataset, the kappa value achieved by the ADTATC model is approximately 0.89, which reflects a strong level of agreement between the predicted and actual classifications. This result further substantiates the model's robustness and reliability in handling diverse and complex retinal fundus images.

In comparison, the TAHDL and LSTM models achieved kappa values of 0.82 and 0.78, respectively. These lower values indicate that the proposed ADTATC model offers a significant improvement in multi-class classification consistency and agreement, validating its advanced spatial and temporal feature extraction capabilities.

To validate the effectiveness and applicability of the Dual Spatial Transformer Block (DSTB) in the proposed ADTATC model, an ablation study was conducted to analyze the impact of its components on experimental accuracy and performance. The study investigated three configurations: a single small-scale transformer, a single large-scale transformer, and the combination of both as proposed in the DSTB. Additionally, an extended configuration that incorporated a medium-scale transformer was evaluated to examine potential improvements.

The results as shown in Table 9 revealed that the small-scale transformer, which processes high-resolution patches (e.g., 16×16), performed well at detecting localized features such as microaneurysms and small hemorrhages, achieving an accuracy of 88.2%. However, it lacked the contextual understanding required for capturing broader patterns, limiting its effectiveness in detecting larger anatomical changes. Conversely, the large-scale transformer, which processes coarser patches (e.g., 64×64), demonstrated an improved accuracy of 89.3%, as it captured broader contextual patterns such as vascular structures. However, it struggled with fine-grained feature extraction, which is critical for detecting early-stage DR.

When both transformers were combined in the DSTB, the performance improved significantly, achieving an accuracy of 91.5%. This improvement highlights the complementary nature of the two transformers, with the small-scale transformer excelling at extracting fine-grained details and the large-scale transformer capturing global contextual features. The hierarchical integration of these outputs ensured a comprehensive spatial analysis, enhancing the model's ability to detect and classify DR severity levels effectively.

Furthermore, the addition of a medium-scale transformer, processing intermediate patches (e.g., 32×32), resulted in a marginal accuracy improvement to 92.0%. While this configuration demonstrated the potential benefits of capturing intermediate-scale features, the computational overhead introduced by a third transformer needs careful consideration. The incremental performance gain suggests that the proposed dual-scale configuration already achieves an optimal balance between feature extraction and computational efficiency.

In summary, the ablation study underscores the critical role of the dual transformer architecture in the proposed DSTB. The combination of small-scale and large-scale transformers provides a synergistic advantage in capturing diverse spatial features, making it highly effective for DR detection. The findings also indicate that while extending the architecture to include a medium-scale transformer offers some benefits, the trade-off between accuracy and computational cost must be evaluated for practical applications.

Conclusion

A novel Attention Dual Transformer with Adaptive Temporal Convolutional (ADTATC) model is proposed in this research work as an innovative approach for detection of diabetic retinopathy (DR) . The proposed ADTATC incorporates dual transformers for enhanced spatial attention with adaptive temporal convolutional memory units to capture disease progression. The proposed model effectively addresses the limitations of traditional CNN, RNN, and even advanced TAHDL models by providing a detailed spatial–temporal analysis in complex medical imaging process. Experimental analysis confirms the proposed model superiority with an accuracy of 98.2% and 97.4% for DRIVE and diabetic retinopathy datasets. The proposed model precision of 96.6%, recall up to 99.8%, and specificity nearing 98.2% for the DRIVE datasets outperforms existing learning algorithms. Similarly, the proposed model precision of 98.6%, recall up to 98.8%, and specifically nearing 98.6% for the diabetic retinopathy dataset exhibit its superior performance over existing learning algorithms. Though the proposed model is highly efficient however it has certain limitations such as increased computational complexity due to the dual transformer and adaptive memory architecture. Additionally, the model performs exceptionally well on labeled datasets and further validation on real time clinical data would strengthen its robustness in DR detection. Future work could explore optimizing the model architecture for faster inference and possibilities in incorporating unsupervised learning algorithms to adapt the model to broader clinical settings.

Data availability

The datasets analyzed during the current study are available in the Kaggle repository, [https://www.kaggle.com/datasets/andrewmvd/drive-digital-retinal-images-for-vessel-extraction/data]. [https://www.kaggle.com/c/diabetic-retinopathy-detection/data]

Received: 10 December 2024; Accepted: 27 February 2025
Published online: 05 March 2025

References

1. Nawaz, F. et al. Early detection of diabetic retinopathy using machine intelligence through deep transfer and representational learning. *Comput. Mater. Contin.* https://doi.org/10.32604/cmc.2020.012887 (2021).
2. Islam, M., Yang, H.-C., Poly, T. N., Jian, W.-S. & Yu-Chuan,. "Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis. *Comput. Methods Progr. Biomed.* **191**, 1–16. https://doi.org/10.1016/j.cmpb.2020.105320 (2020).
3. Samuel Manoharan, J. & Jayaseelan, G. Single Image Dehazing Using Deep Belief Neural Networks to Reduce Computational Complexity. In *New Trends in Computational Vision and Bio-inspired Computing* (eds Smys, S. et al.) (Springer, Cham, 2020).
4. Samuel Manoharan, J., Braveen, M. & Ganesan Subramanian, G. A hybrid approach to accelerate the classification accuracy of cervical cancer data with class imbalance problems. *Int. J. Data Min.* **25**(3/4), 234–259. https://doi.org/10.1504/IJDMB.2021.122865 (2021).
5. Li, W. et al. Predictive model and risk analysis for diabetic retinopathy using machine learning: a retrospective cohort study in China. *BMJ open* **11**, 1–11. https://doi.org/10.1136/bmjopen-2021-050989 (2021).
6. Zhang, X. et al. Automated detection of severe diabetic retinopathy using deep learning method. *Graefe's Arch. Clin. Exp. Ophthalmol.* **260**, 849–856. https://doi.org/10.1007/s00417-021-05402-x (2022).
7. Mushtaq, G. & Siddiqui, F. Detection of diabetic retinopathy using deep learning methodology. *Mater. Sci. Eng.* **1070**, 1–14. https://doi.org/10.1088/1757-899X/1070/1/012049 (2021).
8. Pinedo-Diaz, G. et al. Suitability classification of retinal fundus images for diabetic retinopathy using deep learning. *Electronics* **11**(16), 1–21. https://doi.org/10.3390/electronics11162564 (2022).
9. Shekar, S., Satpute, N. & Guptaa, A. Review on diabetic retinopathy with deep learning methods. *J. Med. Imag.* **8**(6), 1–32. https://doi.org/10.1117/1.JMI.8.6.060901 (2021).
10. Das, D., Biswas, S. K. & Bandyopadhyay, S. A critical review on diagnosis of diabetic retinopathy using machine learning and deep learning. *Multimed. Tools Appl.* **81**, 25613–25655. https://doi.org/10.1007/s11042-022-12642-4 (2022).
11. Bora, A. et al. Predicting the risk of developing diabetic retinopathy using deep learning. *The Lancet Digital Health* **3**(1), 10–19. https://doi.org/10.1016/S2589-7500(20)30250-8 (2021).
12. Alyoubi, W. L., Shalash, W. M. & Abulkhair, M. F. Diabetic retinopathy detection through deep learning techniques: A review. *Inf. Med. Unlocked* **20**, 1–11. https://doi.org/10.1016/j.imu.2020.100377 (2020).
13. Sumathy, B. et al. Prediction of diabetic retinopathy using health records with machine learning classifiers and data science. *Int. J. Reliab. Qual. E-Healthcare* **11**(2), 1–8. https://doi.org/10.4018/IJRQEH.299959 (2022).
14. Mahmoud, M. H., Alamery, S., Fouad, H., Altinawi, A. & Youssef, A. E. An automatic detection system of diabetic retinopathy using a hybrid inductive machine learning algorithm. *Personal Ubiquit. Comput.* **27**, 751–765. https://doi.org/10.1007/s00779-020-01519-8 (2023).
15. ArunSampaul Thomas, G. et al. Intelligent prediction approach for diabetic retinopathy using deep learning based convolutional neural networks algorithm by means of retina photographs. *Comput. Mater. Contin.* **66**(2), 1613–1629. https://doi.org/10.32604/cmc.2020.013443 (2021).
16. Goel, S. et al. Deep learning approach for stages of severity classification in diabetic retinopathy using color fundus retinal images. *Math. Probl. Eng.* **2021**, 1–8. https://doi.org/10.1155/2021/7627566 (2021).

17. Mujeeb Rahman, K. K., Nasor, M. & Imran, A. Automatic screening of diabetic retinopathy using fundus images and machine learning algorithms. *Diagnostics* **12**(9), 1–18. https://doi.org/10.3390/diagnostics12092262 (2022).
18. Gadekallu, T. R. et al. Early detection of diabetic retinopathy using PCA-firefly based deep learning model. *Electronics.* **9**(2), 1–16. https://doi.org/10.3390/electronics9020274 (2020).
19. Bhimavarapu, U. & Battinen, G. Deep learning for the detection and classification of diabetic retinopathy with an improved activation function. *Healthcare* **11**(1), 1–13. https://doi.org/10.3390/healthcare11010097 (2023).
20. Li, F. et al. Deep learning-based automated detection for diabetic retinopathy and diabetic macular oedema in retinal fundus photographs. *Eye* **36**, 1433–1441. https://doi.org/10.1038/s41433-021-01552-8 (2022).
21. Nneji, G. U. et al. Identification of diabetic retinopathy using weighted fusion deep learning based on dual-channel fundus scans. *Diagnostics* **12**(2), 1–19. https://doi.org/10.3390/diagnostics12020540 (2022).
22. Bilal, A., Zhu, L., Deng, A., Huihui, Lu. & Ning, Wu. AI-based automatic detection and classification of diabetic retinopathy using U-Net and deep learning. *Symmetry* **14**(7), 1–19. https://doi.org/10.3390/sym14071427 (2022).
23. Alwakid, G., Gouda, W. & Humayun, M. Deep learning-based prediction of diabetic retinopathy using CLAHE and ESRGAN for enhancement. *Healthcare* **11**(6), 1–17. https://doi.org/10.3390/healthcare11060863 (2023).
24. Gundluru, N. et al. Enhancement of detection of diabetic retinopathy using harris hawks optimization with deep learning model. *Comput. Intell. Neurosci.* **2022**, 1–13. https://doi.org/10.1155/2022/8512469 (2022).
25. Ryu, G., Lee, K., Park, D., Park, S. H. & Sagong, M. A deep learning model for identifying diabetic retinopathy using optical coherence tomography angiography. *Sci. Rep.* **11**, 1–9. https://doi.org/10.1038/s41598-021-02479-6 (2021).
26. Khan, M. B. et al. Automated diagnosis of diabetic retinopathy using deep learning: On the search of segmented retinal blood vessel images for better performance. *Bioengineering* **10**(4), 1–17. https://doi.org/10.3390/bioengineering10040413 (2023).
27. Tao, Li. et al. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Inf. Sci.* **501**, 511–522 (2019).
28. Gadekallu, T. R. et al. Early detection of diabetic retinopathy using PCA-firefly based deep learning model. *Electronics* https://doi.org/10.3390/electronics9020274 (2020).
29. Nneji, G. U. et al. Identification of diabetic retinopathy using weighted fusion deep learning based on dual-channel fundus scans. *Diagnostics* https://doi.org/10.3390/diagnostics12020540 (2022).
30. Gundluru, N. et al. Enhancement of detection of diabetic retinopathy using Harris Hawks Optimization with Deep Learning Model. *Computational Intelligence and Neuroscience* https://doi.org/10.1155/2022/8512469 (2022).

Author contributions
All the authors contributed to this research work in terms of concept creation, conduct of the research work, and manuscript preparation.

Funding
None declared.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to M.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.