scientific reports

Check for updates

OPEN

A hybrid deep learning framework for early detection of diabetic retinopathy using retinal fundus images

Mishmala Sushith[1]✉, A. Sathiya[2], V. Kalaipoonguzhali[3] & V. Sathya[4]

Recent advancements in deep learning have significantly impacted medical image processing domain, enabling sophisticated and accurate diagnostic tools. This paper presents a novel hybrid deep learning framework that combines convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for diabetic retinopathy (DR) early detection and progression monitoring using retinal fundus images. Utilizing the sequential nature of disease progression, the proposed method integrates temporal information across multiple retinal scans to enhance detection accuracy. The proposed model utilizes publicly available DRIVE and Kaggle diabetic retinopathy datasets to evaluate the performance. The benchmark datasets provide a diverse set of annotated retinal images and the proposed hybrid model employs a CNN to extract spatial features from retinal images. The spatial feature extraction is enhanced by multi-scale feature extraction to capture fine details and broader patterns. These enriched spatial features are then fed into an RNN with attention mechanism to capture temporal dependencies so that most relevant data aspects can be considered for analysis. This combined approach enables the model to consider both current and previous states of the retina, improving its ability to detect subtle changes indicative of early-stage DR. Proposed model experimental evaluation demonstrate the superior performance over traditional deep learning models like CNN, RNN, InceptionV3, VGG19 and LSTM in terms of both sensitivity and specificity, achieving 97.5% accuracy on the DRIVE dataset, 94.04% on the Kaggle dataset, 96.9% on the Eyepacs Dataset. This research work not only advances the field of automated DR detection but also provides a framework for utilizing temporal information in medical image analysis.

Keywords Diabetic retinopathy, Deep learning, Retinal fundus images, Convolutional neural network, Recurrent neural network, Temporal analysis

One of the serious complications that occurs due to diabetes is Diabetic Retinopathy (DR). This severe complication affects eyes and leads to blindness and vision loss. A report from World Health Organization (WHO) states that the issues related to diabetes by 2045 will cross 700 million. Over the entire population who affected diabetes, one third will face DR issue. The centers for disease control and prevention in United States estimates that 4.2 million cases for diabetes retinopathy in which DR occupies 655,000 cases. DR not only affects the individual health and quality of life but also increases the economic burden in terms of healthcare cost and loss of productivity. The progression of DR is not visible until it introduces significant damage into vision. Early-stage detection of DR can delay the disease progression and reduce the risk of vision impairment severity. Regular screening is essential for early detection so that proper treatment like laser therapy, intraocular injections, and vitrectomy can be provided to avoid vision loss. Also, early detection of DR can reduce the emotional and physiological stress. Thus, it is essential to provide a reliable and accessible screening methods to identify risks due to DR.

To detect DR, various techniques, and methodologies like clinical examinations to advanced imaging technologies are developed. Ophthalmologists utilize fundus images and fluorescein angiography procedures[1]

[1]Department of Information Technology, Adithya Institute of Technology, Kurumbapalayam, Coimbatore 641 107, India. [2]M.Kumarasamy College of Engineering (Autonomous), Thalavapalayam, Karur 639113, India. [3]Department of EEE, Kathir College of Engineering, Neelambur, Coimbatore 641062, India. [4]Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India. ✉email: mishmalasushith1926@gmail.com

to identify the DR signs. High resolution retina images are captured and analyzed to detect the abnormalities like hemorrhages, microaneurysms and exudates. Due to the advancement of optical coherence tomography (OCT), retina cross-sectional view can be obtained and utilized to identify the structural changes related to DR[2]. In recent times, Artificial Intelligence (AI) techniques are used in medical image analysis. Machine Learning (ML) and Deep Learning (DL) techniques provides automated solutions in real time image processing[3] and medical image analysis[4]. Specifically, DL techniques are widely used for automatic learning and extraction of features from retinal images[5]. The disease severity levels are efficiently detected by the deep learning algorithms. In addition to that, ensemble learning methods are developed by combining multiple prediction models to enhance the detection accuracy and robustness.

Though the traditional and recent imaging techniques provides numerous advantages, there exist several limitations in detecting DR. Since the traditional clinical examinations requires experts and the detection performance relies on ophthalmologist experience. The techniques like optical coherence tomography and fluorescein angiography procedures are effective but the cost is high and requires specialized equipment. The accessibility of these imaging techniques is low and cause discomfort to patients. Although the recent ML and DL based methods provides promising solutions, they face challenges due to the data quality, and interpretability[6]. Based on the data quality and data diversity during the training process, the algorithm performance is decided[7]. Moreover, the learning procedure has limitations in capturing specific features which are related to DR thus produces suboptimal performances[8]. Ensemble techniques provide an improved performance in DR detection but the implementation is computationally intensive and difficult to implement[9].

The proposed research work is aimed to overcome to the above limitations by developing a robust and accurate hybrid deep learning model for DR detection using retinal fundus images. The proposed Temporal Aware Hybrid DL (TAHDL) model is developed by combining the strengths of CNN and RNN models with an attention mechanism. The efficient feature extraction abilities of CNN and the sequential learning features of RNN are combinedly utilized to detect the disease progression. The proposed TAHDL model is designed to be efficient and accurate making it suitable for real time early diagnosis of DR. The research work contributions are summarized as follows.

- Presented a Temporal Aware Hybrid DL (TAHDL) model by integrating CNN for spatial feature extraction and RNN for temporal feature to detect the changes in the retinal images for detecting DR.
- Presented a multiscale feature extraction model that utilizes multiscale convolutional paths for different levels feature extraction. Different level of features provides granularity and improves the model's ability to detect the fine and broader patterns in retinal images.
- Presented an attention mechanism to enhance the detection performance. The attention mechanism is combined with recurrent neural network to enhance the model ability in learning the long and complex sequences.
- Presented an extensive experimentation using benchmark DRIVE, Kaggle diabetic retinopathy dataset, Eyepacs Datasets to evaluate the performance of proposed model in terms of accuracy, precision, recall and specificity metrics.
- Presented a detailed comparative analysis with traditional deep learning techniques to validate the superior performance of the proposed hybrid deep learning model in detecting DR.

The remaining discussions in the following sections are arranged in the following order. "Related works" section presents a brief literature of existing works, "Proposed work" section presents the proposed hybrid deep learning model. The experimental results and discussion are presented in "Results and discussion" and "Conclusion" sections presents the conclusion of research work.

Related works

A brief review of recent research works on diabetic retinopathy detection is studied and the observations are presented in this section. The DR detection model presented in Ref.[10] utilizes a hybrid deep learning model by combining GoogleNet and ResNet architecture with particle swarm optimization. The presented model extracts the features using GoogleNet and ResNet and classifies them using ML algorithms like random forest, decision tree, support vector machine and linear regression. The presented model experimentation utilizes EyePACS dataset and attained 94% detection accuracy. However, the presented model requires detailed interpretability and computationally efficient in detecting DR. A modified graph neural network model presented in Ref.[11] for DR classification combines an adversarial transfer learning procedure with graph neural network. The presented model utilizes ResNet50 for feature extraction and selects the optimal features through modified graph neural network to attain better detection accuracy. The experimentation utilizes benchmark datasets and attained accuracies of 94.3% over existing techniques. However, the attained accuracy in low and the computational complexity increases due to multiple neural networks.

A transfer learning approach for DR detection presented in Ref.[12] as an ensemble model combining shuffleNet and ResNet-18 techniques. The presented approach utilizes the deep learning models for feature extraction. Additionally, an adaptive differential evaluation model is used in the presented approach to fine the deep learning model. Finally using support vector machine, the extracted features are classified and attained 82% accuracy for APTOS dataset. The obtained accuracy is less and the complex hierarchical classification requires enhanced interpretability in DR detection. Similar transfer learning-based DR detection model presented in Ref.[13] combines ResNet18 and GoogleNet architectures. The presented model initially preprocesses the input retinal fundus image using green channel extraction and contrast limited adaptive histogram equalization techniques to enhance the image visibility. Then using DL techniques, the features are extracted and utilizes support vector machine for final classification. The experimentation utilizes benchmark dataset and attained

accuracy of 91.57% over traditional deep learning techniques. However, combining multiple DL and ML algorithms rises the disease detection model computational complexity.

The DR detection model presented in Ref.[14] utilizes a vision transformer for diagnosing its severity. The presented model utilizes vision transformer architecture to process the retinal fundus image from FGADR dataset and attained an accuracy of 82.5%. The presented model accuracy is better than the existing ResNet50, VGG19 and InceptionV3 models. However, the attained accuracy is comparatively low and the model requires further improvement in detecting DR symptoms. An early detection of DR is presented in Ref.[15] using deep learning techniques like DenseNet121, ResNet50, VGG16, VGG19, Xception and InceptionV3 models. The presented approach utilizes DL to extracts the features from fundus images and classifies them using recurrent neural network and support vector machine. The experimentation utilizes APTOS dataset and the comparative analysis highlights that CNN model with five convolution layers attained better accuracy compared to other deep learning models. The multiclass DR classification presented in Ref.[16] utilizes deep learning models like VGG16, ResNet152, BiGRU and EfficientNetB0. The presented comparative analysis evaluates the deep learning models performances through benchmark datasets. The experimentation results confirm the superior performance of EfficientNetB0 over other deep learning models.

The DR detection model presented in Ref.[17] presents a modified colonsegNet model for segmenting retinal vessel. The presented model additionally utilizes evolutionary algorithms to enhance the optimal contrast of retinal fundus image. The experimentation utilizes DRIVE, STARE and CHASE_DB benchmark dataset and demonstrates the presented model performances. The presented model attained better accuracy over traditional ML techniques but its computational requirements are comparatively high compared to recent techniques. An automatic DR detection model presented in Ref.[18] presents a federated learning concept that integrates federated averaging technique with categorical cross entropy to enhance the detection performance. A central server is additionally included that extracts multiscale features to detect small lesions. The experimentation utilizes benchmark datasets and exhibited better performance over traditional deep learning models.

The hybrid DL model presented in Ref.[19] combines inceptionV3 and ResNet50 for DR classification. The presented model extracts the features through the deep learning model and classifies them using traditional convolutional neural network. The experimental utilizes benchmark dataset and demonstrated the presented model performance through various metrics. The hybrid model has better performance in terms of accuracy in DR detection is better than traditional DL model. The multi-stream deep neural network presented in Ref.[20] DR detection combines deep learning algorithms like DenseNet121 and ResNet50. The combined model extracts the features from retinal fundus images and process them using principal component analysis to reduce the feature dimensions. Finally using ensemble of machine learning algorithms like random forest and AdaBoost the features are classified. The experimentation highlights the better accuracy of presented model over traditional deep learning models.

The hybrid model presented in Ref.[21] introduces a multi-stage network that combines DenseNet 121 model with ImageNet to predict diabetes from retinal images. The presented model modifies the DenseNet architecture with additional pooling, batch normalization, ReLU and dropout layers to attain improved detection performance. The experimentation utilizes EyePACS dataset and the presented model attained 84.47% which is low compared to recent techniques. The DR severity classification model presented in Ref.[22] utilizes DenseNet along with a convolutional block attention module. The presented model extracts the primary features using DenseNet169 and enhances the feature representation using convolutional block attention model. The experimentation utilizes retinal images from benchmark dataset and evaluates the presented model performance. The attained classification results highlight the model superiority over traditional machine learning models.

A deep symmetric convolutional neural network model is presented in Ref.[23] for DR detection. The presented approach utilizes symmetric convolutional structures to enhance the network depth and width. Due to this, the network generalization performance is increased and the detection performances also increases. The experimentation utilizes benchmark DIARETDB1 dataset and evaluated the presented model performance. The results depict the presented model detection accuracy is comparatively higher than the traditional CNN model. The non-proliferative DR detection procedure given in Ref.[24] analyzes the OCT images using CNN. The presented model segments the retinal layer and extract the retina patches in the preprocessing step. Then using CNN, the features are extracted from nasal and temporal sides. Finally, the features are combined and classified using support vector machine. The experimentation utilizes standard OCT images for experimentation and attained an accuracy of 94% over existing machine learning models. However, the presented model has limitation due to its potential overfitting and needs further validation on diverse datasets. A multitask deep learning model presented in Ref.[25] utilizes squeeze and extraction to extract the features from retinal fundus images. In addition to squeeze and extraction network, two separate heads are combined in the presented model for feature detection. The extracted features are finally integrated to detect the DR severity. The experimentation of the presented model depicts the model kappa value as 0.7421 which is less compared to recent techniques.

Research gap

From the review of existing DR detection frameworks, it can be observed that DL models are used widely in disease detection and its severity classification. The deep learning models like DenseNet, ResNet, VGG are widely used in various research works. However, the obtained accuracies of these model-based DR diagnosis is less and the accuracy has to be increased to provide enhance detection performance. The hybrid models combine different DL and ML techniques for DR detection. However, the model computational complexity increases due to multiple networks. The existing approaches performs binary classification however it is essential to provide multiclass classification to analyze the disease severity. Considering all the above limitations, a model should be developed, and it should be more accurate and robust in detecting DR.

While deep learning methods have advanced the detection and classification of Diabetic Retinopathy (DR), existing approaches exhibit significant limitations in capturing temporal dependencies and addressing the complexity of disease progression. Many traditional methods focus solely on spatial features extracted from static retinal fundus images. While this provides valuable insights into the structural abnormality's indicative of DR, such as hemorrhages, microaneurysms, and exudates, it fails to account for the temporal dynamics of disease evolution.

Temporal dependencies are crucial in DR detection, as the progression of the disease involves subtle and cumulative changes in retinal features over time. Traditional convolutional models like CNNs are not inherently equipped to process sequential data, leading to a loss of critical information about the progression of retinal damage. Consequently, these models may struggle to distinguish between early and advanced stages of DR, resulting in suboptimal performance in progression monitoring and severity classification.

Additionally, DR progression involves complex patterns that vary across patients, influenced by factors such as diabetes duration, comorbidities, and treatment history. Existing models that employ fixed feature extraction techniques, or rely on pre-trained architectures, often fail to adapt to these complex patterns. Hybrid approaches combining CNNs with machine learning classifiers (e.g., SVM) or ensemble methods improve classification accuracy to some extent, but they introduce high computational complexity and lack scalability for real-time applications.

Moreover, many existing methods do not utilize advanced mechanisms, such as attention layers, which can enhance the focus on critical temporal and spatial features. This shortcoming limits their ability to prioritize subtle changes over time that are indicative of disease progression, thus reducing their utility in clinical decision-making.

The inability to capture temporal dependencies and effectively manage the complexity of DR progression represents a significant gap in existing methodologies. Addressing this gap requires innovative frameworks, like the Temporal Aware Hybrid Deep Learning (TAHDL) model, that integrate spatial and temporal analysis while maintaining computational efficiency and scalability for real-world applications.

Proposed work

The proposed Temporal Aware Hybrid Deep Learning (TAHDL) for DR detection includes CNN and RNN to utilize the spatial and temporal features in retinal fundus images for improved detection performance. In the proposed work, CNN is specifically utilized to extract the spatial features and it act as fundamental component of the proposed hybrid DL. Similarly, the proposed work uses RNN to model the temporal dependencies in the image features.

While the combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) is not entirely novel, its application to Diabetic Retinopathy (DR) detection and progression monitoring provides unique advantages. This research work highlights the significant benefits of integrating these architectures to address the challenges in detecting and analyzing DR.

- CNNs are highly effective in capturing spatial features from retinal fundus images. These features include structural irregularities such as microaneurysms, hemorrhages, and exudates, which are indicative of DR. The use of multi-scale convolutional paths enhances this capability by extracting both fine-grained details and broader patterns, ensuring robust spatial representation.
- DR is a progressive condition where disease severity evolves over time. By incorporating RNNs, specifically Long Short-Term Memory (LSTM) networks, the proposed model captures sequential dependencies in retinal images. This enables the detection of subtle changes in disease progression that are otherwise challenging to identify with static image analysis alone.
- The integration of an attention mechanism further strengthens the framework by selectively prioritizing critical temporal features. This ensures the model emphasizes significant changes over time, aligning with the diagnostic process used by clinicians.
- The hybrid approach utilizes the complementary strengths of CNNs and RNNs, facilitating the detection of early-stage DR while monitoring its progression. This dual capability is crucial for timely intervention and effective disease management.
- The performance of the proposed Temporal Aware Hybrid Deep Learning (TAHDL) model validates these advantages. On benchmark datasets (DRIVE, Kaggle DR and Eyepacs datasets), the TAHDL model consistently outperformed traditional deep learning approaches, achieving accuracies of 97.5% and 94.04%, respectively. This highlights its robustness in capturing both spatial and temporal aspects of DR.

By combining spatial and temporal analyses, the proposed hybrid framework addresses the limitations of existing methods and provides a reliable, efficient solution for DR detection and monitoring. Unlike conventional methods, the proposed model integrates dataset-specific augmentation, advanced preprocessing, and tailored training strategies for each DR severity level. Additionally, it introduces per-class performance calibration and adaptive learning based on class imbalance handling, resulting in superior generalization across diverse datasets. These enhancements collectively distinguish the model from earlier CNN-RNN integrations, establishing its originality and technical advancement.

The proposed work initially preprocesses the input image and then fed the pre-processed image into CNN and RNN for spatial and temporal feature extraction. The extracted features are then classified to detect the disease symptoms. In Fig. 1, the proposed model complete overview is presented for better understanding of entire process.

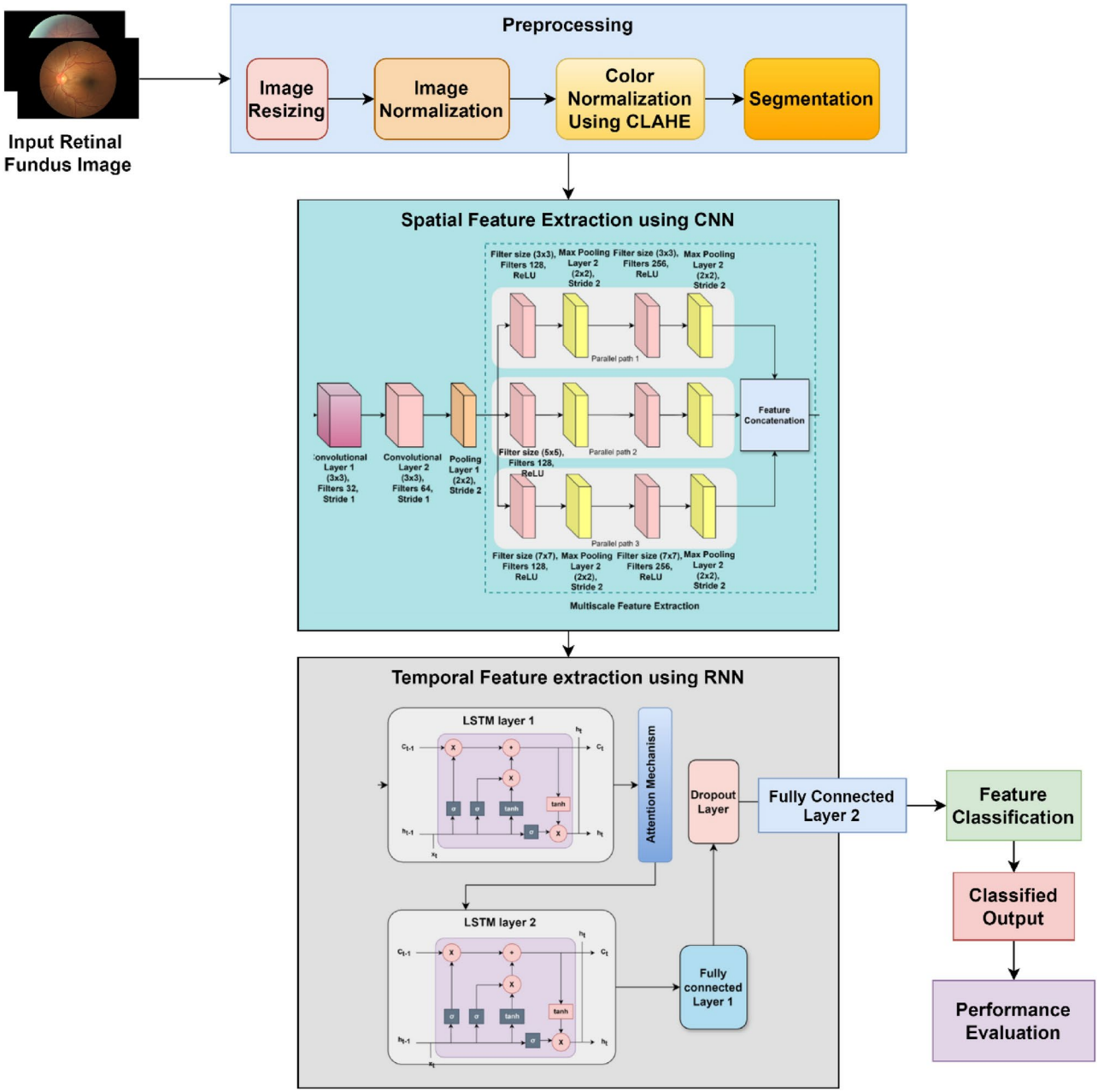

Fig. 1. Proposed TAHDL model for DR detection.

Image preprocessing

The input retinal fundus image is pre-processed by performing image resizing, image normalization, and color normalization. The input retinal fundus image can be of any size thus the input image is resized into size of $224 \times 224 \times 3$ in the first step. In this the height and width are mentioned with 224 and 3 indicates the RGB color channels. Mathematically image resizing is expressed as

$$X\prime = resize\,(X, (224{,}224))\,, \tag{1}$$

where X denotes the original image, and $X\prime$ is the resized image with dimensions $224 \times 224 \times 3$. After resizing, the image is normalized to ensure uniformity across the images in the input data. The normalization process adjusts the pixel values to a common scale typically [0,1] to ensure consistency and stability during the training process. Mathematically the image normalization process is expressed as

$$X'' = \frac{X\prime}{255}, \tag{2}$$

where $X\prime$ is the resized image with pixel values ranging from 0 to 255, and X'' is the normalized image with pixel values scaled to the range [0,1]. After image normalization, color normalization is performed in which Contrast Limited Adaptive Histogram Equalization (CLAHE) is used. CLAHE is an improved version of standard Histogram Equalization technique. Since the traditional histogram equalization improves the global contrast of an image by spreading the most frequency intensity values it might over amplify the noise factors in homogenous regions. To overcome this CLAHE is used in which histogram equalization is applied to small local regions of an image. Thus, the noise amplification is avoided or reduced in the CLAHE compared to traditional histogram. Consider an image is divided into grid of non-overlapping tiles of size $M \times N$. The tile at position (i, j) in the grid is indicated as X''_{ij}. For each tile X''_{ij}, the histogram of pixel intensities is computed. After computing, the histogram is clipped into predefined limit to limit the noise amplification. Mathematically the clipped histogram is expressed as

$$H_{ij}^{clipped}(k) = \min\left(H_{ij}(k), L\right), \tag{3}$$

where $H_{ij}(k)$ indicates the histogram of tile, L indicates the maximum allowed value for the histogram bins. Then excess pixels are redistributed uniformly and it is mathematically expressed as

$$\text{excess} = \sum\nolimits_{k} \left(H_{ij}(k) - L\right) \text{for} H_{ij}(k) > L. \tag{4}$$

The redistributed histogram is then formulated as follows

$$H_{ij}^{(redistributed)}(k) = H_{ij}^{(clipped)}(k) + \frac{\text{excess}}{\text{number of bins}}. \tag{5}$$

After histogram redistribution, a cumulative distribution function (CDF) is calculated to map the intensity values which is formulated as

$$CDF_{ij}(k) = \sum\nolimits_{m=0}^{k} H_{ij}^{(redistributed)}(m). \tag{6}$$

The normalized CDF is then used to map the pixel values

$$X_{ij}^{(equalized)}(x,y) = \frac{CDF_{ij}\left(X_{ij}(x,y)\right) - CDF_{ij}(min)}{CDF_{ij}(max) - CDF_{ij}(min)} \times (new_{max} - new_{min}) + new_{min} + r, \tag{7}$$

where $X_{ij}^{equalized}(x,y)$ is the new intensity value for pixel (x, y) in the tile X_{ij}. The minimum and maximum pixel values are indicated as new_{min} and new_{max}. In order to ensure smooth transition, the tiles after performing histogram equalization are combined using bilinear interpolation which is mathematically formulated as

$$X^{(clahe)}(x,y) = \sum\nolimits_{i,j} w_{ij}(x,y) \cdot X_{ij}^{(equalized)}(x\prime, y\prime), \tag{8}$$

where $w_{ij}(x, y)$ indicates the interpolation weights and $(x\prime, y\prime)$ are the coordinates of the pixel in the tile X_{ij}. This contrast enhanced images helps to identify features associated with DR such as hemorrhages, exudates and microaneurysms. Thus, by improving the feature visibility, optimal features can be extracted from the retinal fundus image and detection accuracy can be improved in the proposed work.

Spatial feature extraction using CNN

After preprocessing, the spatial features are extracted from the retinal fundus image using CNN. The spatial feature extraction model includes multiple convolution layers with set of learnable filters to produce feature maps. Over the input image the filters slide and perform element wise multiplications. By summarizing the results, it creates and output that highlights the spatial features like edges, textures, and other significant factors in the input image. Mathematically the spatial feature extraction process is formulated as

$$f_{i,j}^{k} = \sigma\left(\sum\nolimits_{m=0}^{M-1} \sum\nolimits_{n=0}^{N-1} I_{i+m,j+n} \cdot W_{m,n}^{k} + b^{k}\right), \tag{9}$$

where I indicates the input image, $W_{m,n}^{k}$ indicates the k^{th} filter weight at position (m, n), bias term is indicated as b^{k}, σ indicates the activation function. Rectified Linear Unit (ReLU) is used as an activation. $f_{i,j}^{k}$ indicates output feature map value for the k^{th} filter at position (i, j). The local dependencies within the image can be captured through the convolution operation to understand the spatial structure. After convolution, an activation function is applied. ReLU is used in the proposed which effectively reduces the vanishing gradient problem and allows the network to learn the complex patterns. The activation function is mathematically described as

$$\sigma(x) = \max(0, x), \tag{10}$$

where σ indicates the activation function ReLU. In the feature map, the positive values are unchanged and the negative values are mapped to zero to preserve the important features. Followed by convolution and activation function, pooling layers are used in the architecture which reduces the spatial feature map dimensions. This

dimension reduction through the pooling layer reduces the computational load and avoids overfitting in the classification problem. Mathematically the pooling layer operation is described through the function as given below.

$$p_{i,j} = \max_{0 \le m < M, 0 \le n < N} f_{i+m,j+n}, \tag{11}$$

where $p_{i,j}$ indicates the pooled value at position (i, j), f represents the input feature map, M and N indicates the pooling window size. In the proposed work, max pooling is employed and it retains the maximum value in each window. Due to this, the most prominent features detected by the convolutional filters in the previous step is preserved.

Multi-scale feature extraction

To enhance the feature extraction performance, a multi-scale feature extraction is performed in the proposed work. The major reason for utilizing multi-scale feature extraction is to extract features at various levels which is essential in medical image analysis. In the proposed work, multi-scale feature extraction is done by employing convolutional filters of different sizes to the input images. Due to this, multiple scales of features and patterns can be detected. Practically multiscale feature extraction utilizes smaller filters in the dimension (3×3) and the large filters like (7×7) to capture the important features. Mathematically the convolution process in the multiscale feature extraction is formulated by modifying Eq. (9) as follows.

$$f^s_{i,j} = \sigma\left(\sum_{m=0}^{M_s-1}\sum_{n=0}^{N_s-1} I_{i+m,j+n} \cdot W^s_{m,n} + b^s\right), \tag{12}$$

where I is the input image, $W^s_{m,n}$ indicates the weight of the filter of size s at position (m, n), b^s is the bias term for the filter of size s, σ is the activation function, $f^s_{i,j}$ is the feature map output at position (i, j) for the filter size s.

For efficient multiscale feature extraction process, parallel convolutional paths are employed in the proposed work. The parallel path will have convolution layers with different size filters like (3×3), (5×5), and (7×7) . The features extracted through different filters are then concatenated to create a final feature map which is mathematically expressed as

$$F = concat\left(f^{3\times3}, f^{5\times5}, f^{7\times7}\right), \tag{13}$$

where F is the concatenated feature map containing information from all three scales. This multiscale feature extraction captures diverse set of features which are essential for accurate diagnosis of DR. The smaller filters provide minute details while the larger filters will recognize broader structures. The final concatenated feature map will enhance the accuracy in the disease detection process. After feature concatenation, pooling layers are used to reduce the feature dimensionality. Mathematically it is expressed as

$$f_{pooled} = Pooling\left(F\right), \tag{14}$$

where f_{pooled} indicates the pooled feature map. In the proposed work instead of classifying the features which obtained after multiscale feature extraction, temporal features are extracted to determine the temporal dependencies between the features. Figure 2 depicts the architecture of spatial feature extraction model used in the proposed work.

Temporal feature extraction using RNN

After extracting the spatial features, the temporal features are extracted using recurrent neural network. The necessity of introducing temporal feature extraction is to analyze the disease progression over time. RNN is familiar for its sequential data processing. The hidden state used in the RNN architecture has the ability to capture information from previous time steps. The network can able to remember and utilize the previous state information so that the changes in the retinal images over time can be analyzed more effectively. The conventional RNN has a recurrent layer in which the input sequence is processed by updating the hidden state. Mathematically the hidden state is formulated as

$$h_t = \phi\left(W_{hh}h_{t-1} + W_{xh}x_t + b_h\right), \tag{15}$$

where h_t is the hidden state, h_{t-1} is the previous time step hidden state, W_{hh} is the hidden state weight matrix, W_{xh} is the input weight matrix, x_t represents the input which is the multiscale spatial features extracted by the CNN, b_h is the bias term, ϕ is the activation function. By continuously updating the hidden state, the network understands the changes in the input sequence. However, the traditional RNN has limitation like vanishing gradient which reduces the learning abilities. To overcome this, Long-Short Term Memory (LSTM) network is introduced. The gating mechanism in the LSTM regulates the information flow and avoids the vanishing gradient issue in RNN.

The architecture of LSTM includes three primary gates like forget gate, input gate, and output gate. The three gates control the cell state which is used as the memory element of LSTM. The input gate in the LSTM controls the quantity of new information added to the cell state. Mathematically the input gate is formulated as

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right), \tag{16}$$

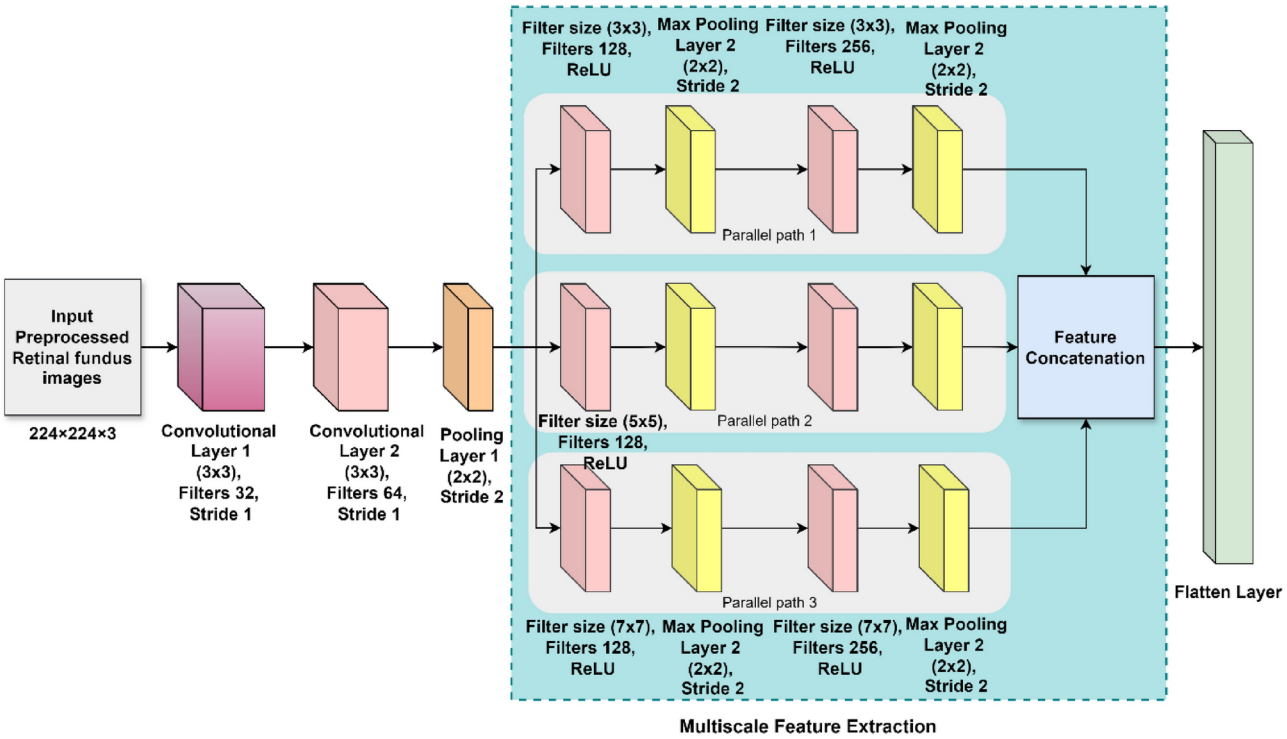

Fig. 2. Spatial feature extraction using CNN.

where W_i is the weight and b_i is the bias terms for the input gate. h_{t-1} indicates the previous hidden state. The forget gate in the LSTM architecture decides which portion of cell state from the previous time step has to be retained. Mathematically the forget gate is formulated as

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right), \tag{17}$$

where W_f is the weight and b_f is the forget gate bias terms. The LSTM output gate determines the output which is formulated as

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right), \tag{18}$$

where W_o is the weight and b_o is the output gate bias terms. The forget gate and the input gate updates the cell state which is mathematically expressed as

$$\widetilde{c_t} = \tanh\left(W_c \cdot [h_{t-1}, x_t] + b_c\right), \tag{19}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c_t}, \tag{20}$$

where c_t indicates the cell state, W_c is the weight and b_c is the bias terms for the cell state. The hidden state is updated based on the cell state and output gate, which is mathematically formulated as

$$h_t = o_t \odot \tanh\left(c_t\right), \tag{21}$$

where σ indicates the sigmoid activation function and the element-wise multiplication is indicated through $\odot$ operator. The candidate cell state is given as $\widetilde{c_t}$. Thus, the long-term dependencies in the feature map are managed by the LSTM. Further to enhance the feature processing ability and to focus more on the necessary portions in the input sequence an attention mechanism is introduced in the proposed work. This attention mechanism calculates the context vector for selective time steps to improve the model performances.

Attention mechanism

The attention mechanism used in the proposed work calculates an alignment score for each time step in the sequence to indicate the relevance of hidden state. Mathematically the score calculation is formulated as

$$e_t = score\left(h_t, s\right), \tag{22}$$

where h_t indicates the hidden state, s indicates the context vector, e_t indicates the Alignment score for the time step t. This alignment score indicates how well the hidden state is matched with the context vector. In the next

step, the alignment scores are then converted into attention weights using a SoftMax function. This conversion normalizes the scores into a probability distribution which is mathematically formulated as.

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)}, \tag{23}$$

where α_t indicates the attention weight, T indicates the total number of time steps in the input sequence. Finally, the context vector is calculated as a weighted sum of all hidden states. the contribution of hidden state is determined using the attention weights. Mathematically context vector is formulated as

$$c = \sum_{t=1}^{T} \alpha_t h_t, \tag{24}$$

where c indicates the context vector. The final context vector effectively captures the important information from the entire sequence and enhances the model ability to learn from long and complex sequences. Figure 3 depicts the process of temporal feature extraction and classification of proposed diabetic detection model.

Feature classification

The extracted features in the final layer are then classified by passing through fully connected layers. Using SoftMax activation function the output layer produces the probability distribution over the possible classes. Mathematically the final output is formulated as

$$y = \sigma\left(W_{\text{fc}} \cdot f_{\text{pooled}} + b_{\text{fc}}\right), \tag{25}$$

where f_{pooled} is the pooled feature map, W_{fc} and b_{fc} are the fully connected layer weights and bias, y is the output prediction. To prevent overfitting the model is trained using a loss function which combine cross entropy loss and a regularization term. Mathematically the loss function is expressed as

$$L = -\frac{1}{N}\sum_{i=1}^{N} \left[y_i \log\left(\widehat{y_i}\right) + (1 - y_i)\log\left(1 - \widehat{y_i}\right)\right] + \lambda|\theta|^2, \tag{26}$$

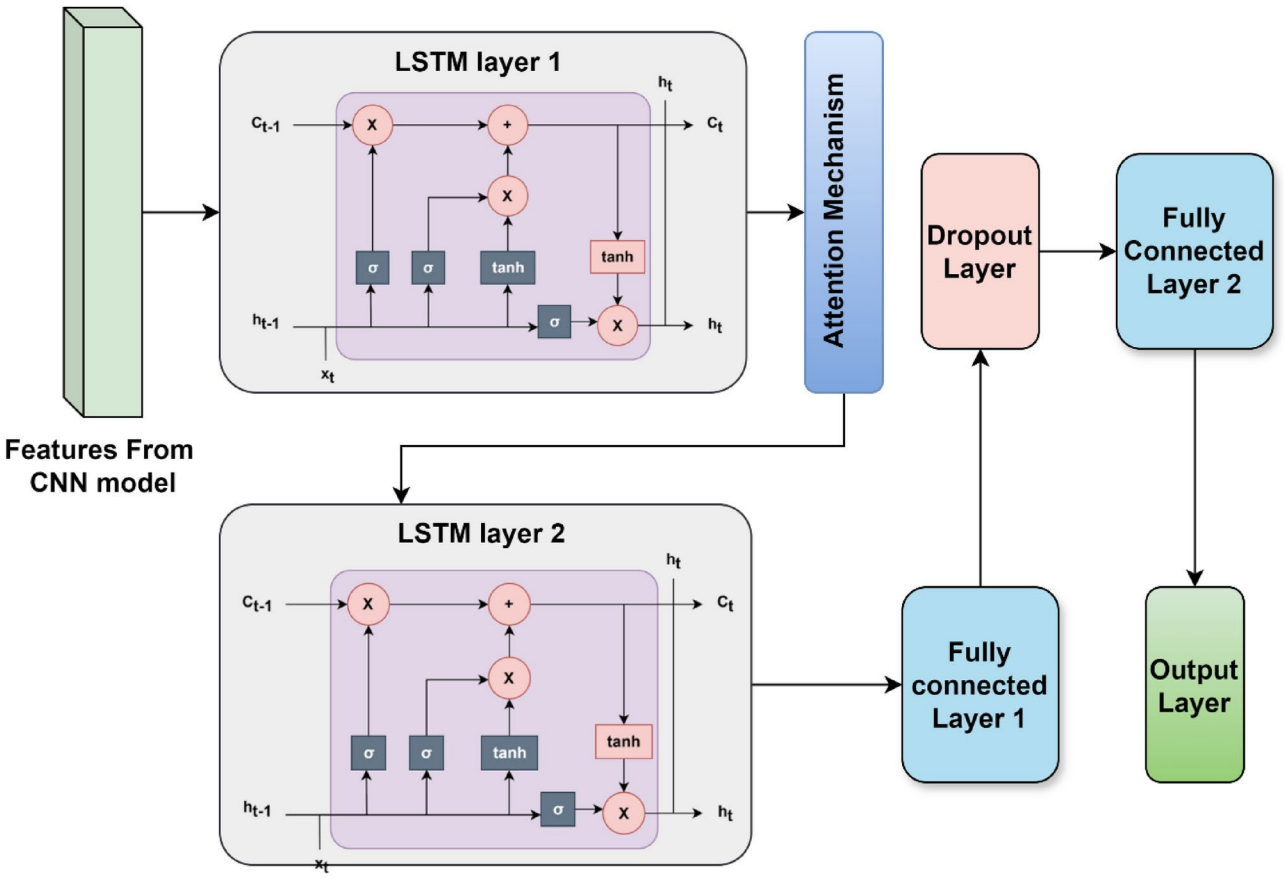

Fig. 3. Temporal feature extraction and classification.

where y_i is the true label, $\widehat{y_i}$ is the predicted probability, N indicates the number of samples, λ is the regularization parameter, θ represents the model parameters. By utilizing Adam optimizer, the model parameters are optimally adjusted. The final model that combines the CNN and RNN utilizes spatial and temporal information in the input image to detect the DR. The proposed hybrid model summarized pseudocode is presented as follows.

Pseudocode for the proposed hybrid model for DR detection

Input: Sequence of retinal fundus images $I = \{I_1, I_2, \ldots, I_T\}$, *convolutional filters* K, *Filter sizes* $S = \{s_1, s_2, \ldots, s_n\}$, *hidden state size* H, *Number of classes* C

Output: Class probabilities for each image sequence P

Initialize parameters convolutional filters W^k, *biases* b^k, *weights* $W_{xh}, W_{hh}, W_i, W_f, W_o, W_c$ *for LSTM, biases* b_h, b_i, b_f, b_o, b_c *for LSTM,* W_{fc} *and* b_{fc} *for fully connected layers*

Begin

For each image I_t

For each filter size s *in* S

Apply convolution operation $f_{i,j}^{s,k} = \sigma\left(\sum_{m=0}^{s-1}\sum_{n=0}^{s-1} I_{i+m,j+n} \cdot W_{m,n}^{s,k} + b^{s,k}\right)$

Retain the resulting feature map $f^{s,k}$

Concatenate all the feature map to create multi-scale feature map

$F_t = concat(f^{s_1}, f^{s_2}, \ldots, f^{s_n})$

Apply max pooling $p_{i,j}^{s,k} = \max_{0 \le m < s, 0 \le n < s} f_{i+m,j+n}^{s,k}$

Initialize RNN for temporal feature extraction

Initialize the hidden state $h_0 = 0$ *and cell state* $c_0 = 0$

For each time step t

Compute LSTM gates i_t, f_t, *and* o_t

Update cell state $\widetilde{c_t}$ *and* c_t

Update hidden state $h_t = o_t \odot tanh(c_t)$

Initialize attention mechanism

For each time step t

Calculate attention score e_t

Calculate attention weights $\alpha_t = \frac{exp(e_t)}{\sum_{k=1}^{T} exp(e_k)}$

Obtain context vector $c = \sum_{t=1}^{T} \alpha_t h_t$

Concatenate all context vector with final hidden state

Apply fully connected layer $y = \sigma\left(W_{fc} \cdot [c, h_T] + b_{fc}\right)$

Obtain class probabilities $P = softmax(y)$

Return

End

End

Results and discussion

The proposed Temporal Aware Hybrid Deep Learning (TAHDL) model performance is evaluated through simulation analysis performed in python tool using benchmark DRIVE[26] Kaggle Diabetic Retinopathy[27], Eyepacs Diabetic Retinopathy[28] datasets. The python platform includes essential library packages to execute the deep learning models. The details about the simulation hyperparameters are is presented in Table 1 and the details of datasets are presented in Tables 2, 3 and 4. The DRIVE dataset has two classes of data in which retinal images are labeled as diabetic retinopathy (DR) and Non-Diabetic retinopathy (Non-DR). The samples are limited thus using data augmentation techniques like image rotation, flipping the number of samples are increased. The actual samples in the dataset are 20 for DR and 20 for Non-DR. After data augmentation the number of samples is increased into 100 for DR and 100 for Non-DR. The second dataset diabetic retinopathy data from Kaggle repository has 35,126 samples under five different classes like No DR, Mild DR, Moderate DR, Severe DR and Proliferative DR. The third dataset Eyepacs DR has 104,000 samples under five classes like No

S.no	Hyperparameter	Range/type
1	Number of epochs	50
2	Learning rate	0.001
3	Batch size	32
4	Optimizer	Adam
5	CNN conv layer filters	32, 64
6	Conv path filters	128
7	Conv layer filter size	3 × 3
8	Conv path filter size	3 × 3, 5 × 5, 7 × 7
9	RNN architecture	LSTM
10	RNN layers	2
11	RNN units	256, 128
12	Dropout rate	0.5
13	Activation function	ReLU (for CNN), Tanh (for RNN)
14	Loss function	Categorical cross-entropy
15	Regularization	L2 regularization ($\lambda = 0.01$)

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	5162
Mild	2443	1954	489
Moderate	5292	4233	1059
Severe	873	698	175
Proliferative DR	708	566	142

Table 3. Diabetic retinopathy dataset description.

Class	Total	Training	Testing
Class 0	69,000	55,200	13,800
Class 1	6000	4800	1200
Class 2	24,000	19,200	4800
Class 3	2000	1600	400
Class 4	3000	2400	600

Table 4. Eyepacs DR dataset description.

DR, Mild DR, Moderate DR, Severe DR and Proliferative DR. From all dataset 80% of data is used to train the proposed network and 20% of data is used to test the proposed network.

The DRIVE dataset consists of high-resolution retinal fundus images annotated specifically for blood vessel segmentation. It is especially valuable for learning fine-grained features related to microvascular changes, which are critical early indicators of diabetic retinopathy. The Diabetic Retinopathy Detection dataset available on Kaggle offers over 88,000 retinal images across five DR severity grades (0–4). This large-scale dataset introduces the model to significant inter-patient and inter-device variability, which is essential for building robust, generalizable models. It contains real-world artifacts such as image blur, over/under-exposure, and variability in image resolution, thus mimicking clinical conditions. Utilizing this dataset enables the model to learn hierarchical lesion patterns and contextual dependencies, making it resilient to noisy data. The EyePACS dataset, also hosted on Kaggle, is another widely recognized DR dataset that offers high-quality retinal fundus images along with expert-verified labels. This dataset is frequently used as a benchmark in major DR detection

Metric	Train	Test
Accuracy	0.9762	0.9750
Precision	0.9763	0.9524
Recall	0.9987	0.9984
F1-score	0.9873	0.9748
Specificity	0.9748	0.9694

Table 5. DRIVE dataset metrics.

Metric	Train	Test
Accuracy	0.9554	0.9404
Precision	0.9708	0.9676
Recall	0.9895	0.9867
F1-Score	0.9801	0.9771
Specificity	0.9903	0.9831

Table 6. Kaggle diabetic retinopathy dataset metrics.

challenges and research publications. By including EyePACS, the proposed model is trained and validated on standardized annotations and can be fairly compared against existing state-of-the-art methods. Additionally, the dataset covers a broad range of ethnic and age demographics, promoting demographic fairness in model performance.

The experimental evaluation of the proposed Temporal Aware Hybrid Deep Learning (TAHDL) model was conducted using two benchmark datasets: DRIVE and Kaggle Diabetic Retinopathy. To ensure consistency and reproducibility, the data preprocessing pipeline involved resizing retinal fundus images to a standard size of $224 \times 224 \times 3$, followed by pixel value normalization to the range [0, 1]. Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied to enhance image contrast while minimizing noise amplification. Data augmentation techniques, including random rotation and flipping, were employed to address the limited sample size of the DRIVE dataset, increasing the number of training samples from 40 to 200. The TAHDL model was trained using a batch size of 32 and an Adam optimizer with an initial learning rate of 0.001. The network architecture included convolutional layers with ReLU activation for spatial feature extraction, multi-scale convolutional paths for granularity, and LSTM layers with an attention mechanism for temporal dependency modeling. The model incorporated a dropout rate of 0.5 and L2 regularization ($\lambda = 0.01$) to mitigate overfitting. Training was conducted over 50 epochs, with 80% of the data allocated for training and 20% for testing. Evaluation metrics included accuracy, precision, recall, F1-score, and specificity, calculated using standard formulations to assess the model's classification performance. Comparative analysis was performed against baseline models such as CNN, RNN, VGG19, InceptionV3, LSTM, MobileNetV3, SNN and Vision Transformer (ViT) ensuring uniformity in hyperparameter configurations across all methods for a fair comparison. The proposed setup provides a structured approach to evaluate the model's ability to detect and classify DR with high accuracy and robustness.

The proposed model performance evaluation considered metrics like precision, recall, f1-score, specificity, and accuracy. The formulations for the performance evaluation metrics are presented as follows.

$$Precision = \frac{TP}{TP + FP}, \tag{27}$$

$$Recall = \frac{TP}{TP + FN}, \tag{28}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{29}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{30}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{31}$$

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 Tables 5, 6 and 7 for DRIVE dataset, Diabetic retinopathy dataset and Eyepacs Diabetic Retinopathy respectively.

From Tables 5, 6 and 7 the better performance of proposed model can be observed in both training and testing process. The proposed model attained maximum accuracy of 97.62% in the training process and 97.50 in

Metric	Training	Test
Accuracy	0.974	0.969
Precision	0.968	0.962
Recall	0.975	0.968
F1-Score	0.976	0.969
Specificity	0.972	0.968

Table 7. Eyepacs DR dataset metrics.

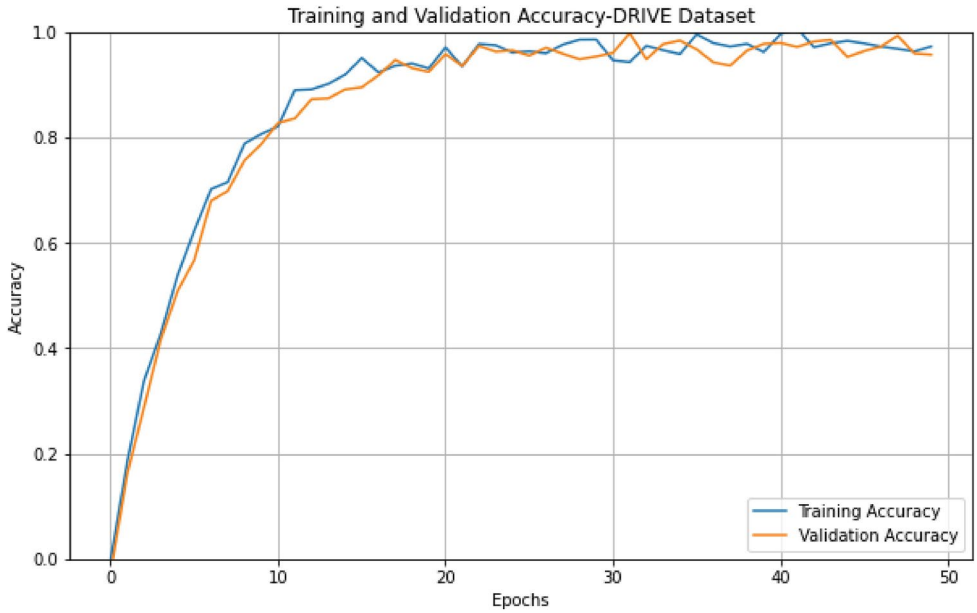

Fig. 4. Analysis of training and validation accuracy for DRIVE dataset.

the test process for DRIVE dataset. Similarly, the proposed model exhibit 95.54% as accuracy during the training and 94.04% as accuracy during the test process for diabetic retinopathy dataset. Similarly, the proposed model exhibit 97.42% as accuracy during the training and 96.9% as accuracy during the test process for Eyepacs DR dataset.

The proposed model training and validation performance is analyzed through accuracy analysis curve given in Fig. 4 and loss curve in Fig. 5 for DRIVE dataset. It can be observed that the accuracy of training as well as validation curve increases gradually and reaches maximum after 20th epoch. The accuracy shows slight variations but gradually increases till the last epoch. The validation curve effectively follows the training accuracy curve which indicates the proposed model better performance. The same reflects in the loss analysis given in Fig. 5 in which the loss gradually decreases from the beginning and reaches minimum when it crosses 10th epoch.

Similarly, for diabetic retinopathy dataset the training and validation accuracy is analyzed through accuracy analysis curve given in Fig. 6 and loss curve in Fig. 7. It can be observed that the accuracy of training as well as validation curve increases gradually and reaches maximum after 18th epoch. The validation curve effectively follows the training accuracy curve which indicates the proposed model better performance. The loss analysis given in Fig. 7 clearly presents the minimal loss of proposed model through its gradually decreasing training and validation curve. The proposed model effectively extracts the spatial and temporal features which minimizes the detection errors and thus it reduces the loss in the training and validation process.

Similarly, for EyePACS, APTOS, Messidor diabetic retinopathy dataset the training and validation accuracy is analyzed through accuracy analysis curve given in Fig. 8 and loss curve in Fig. 9. It can be observed that the accuracy of training as well as validation curve increases gradually and reaches maximum after 15th epoch. The validation curve effectively follows the training accuracy curve which indicates the proposed model better performance. The loss analysis given in Fig. 9 clearly presents the minimal loss of proposed model through its gradually decreasing training and validation curve. The proposed model effectively extracts the spatial and temporal features which minimizes the detection errors and thus it reduces the loss in the training and validation process.

The confusion matrix obtained for both dataset test process is presented in Figs. 10, 11 and 12. The confusion matrix depicts the actual true negatives, true positives, false positives and false negative values. Based on the confusion matrix the other metrics are calculated and utilized to evaluate the proposed model performance.

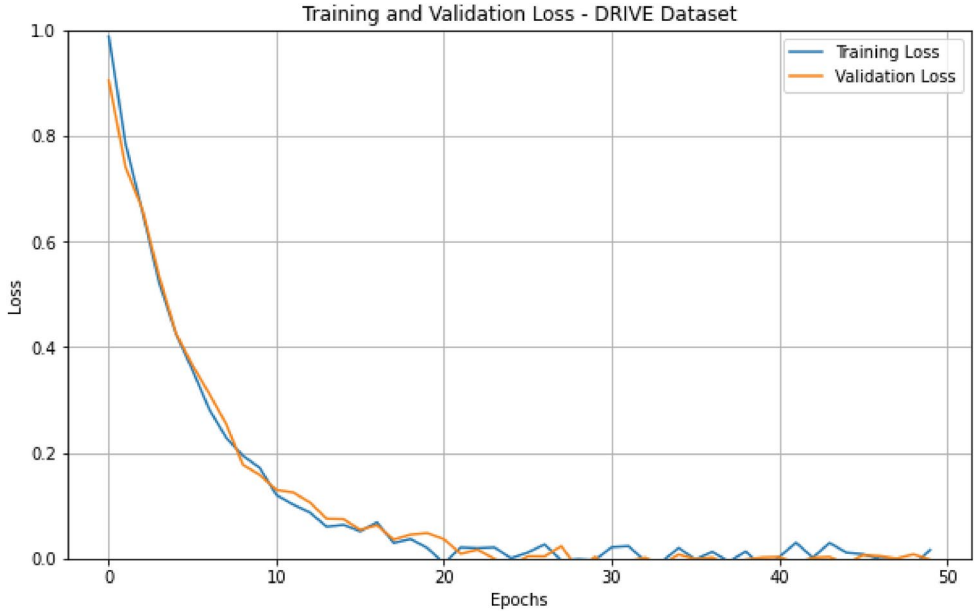

Fig. 5. Analysis of training and validation loss for DRIVE dataset.

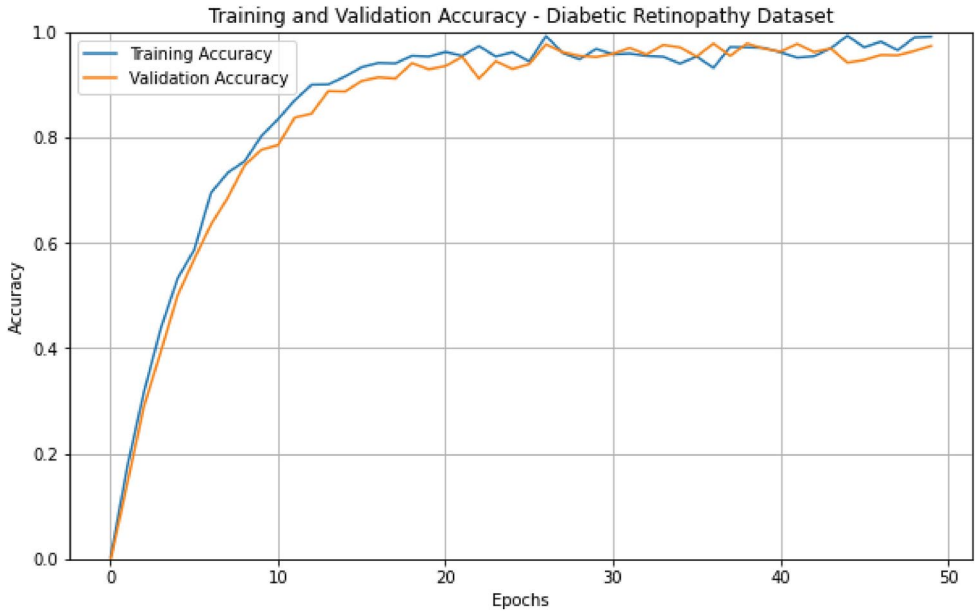

Fig. 6. Analysis of training and validation accuracy for diabetic retinopathy dataset.

The relation between precision and recall is analyzed through precision-recall curve. Figure 13 depicts the precision recall curve of proposed model for DRIVE dataset. It can be observed that proposed model exhibits better A P values for both DR and Non-DR classes. The obtained AP values are above 0.99 which indicates the better performance of proposed model in detecting DR. Similarly, the precision recall curve for DR dataset given in Fig. 14 depicts that the AP values for all the classes are above 0.99. The better AP values indicate the better performance ability of the proposed model in categorizing different impact of DR from fundus images. The mild to proliferate symptoms are effectively detected by the proposed model by analyzing the spatial and temporal features. Similarly, the precision recall curve for Eyepacs, Aptos, Messidor DR dataset given in Fig. 15 depicts that the AP values for all the classes are also above 0.99. The better AP values indicate the better performance ability of the proposed model in categorizing different impact of DR from fundus images.

Further the detailed per-class evaluation across the three datasets are performed and the result for DRIVE, Kaggle Diabetic Retinopathy, and EyePACS are presented in Tables 8, 9 and 10 respectively. The results exhibits the robustness and high discriminative capability of the proposed model in detecting and classifying diabetic

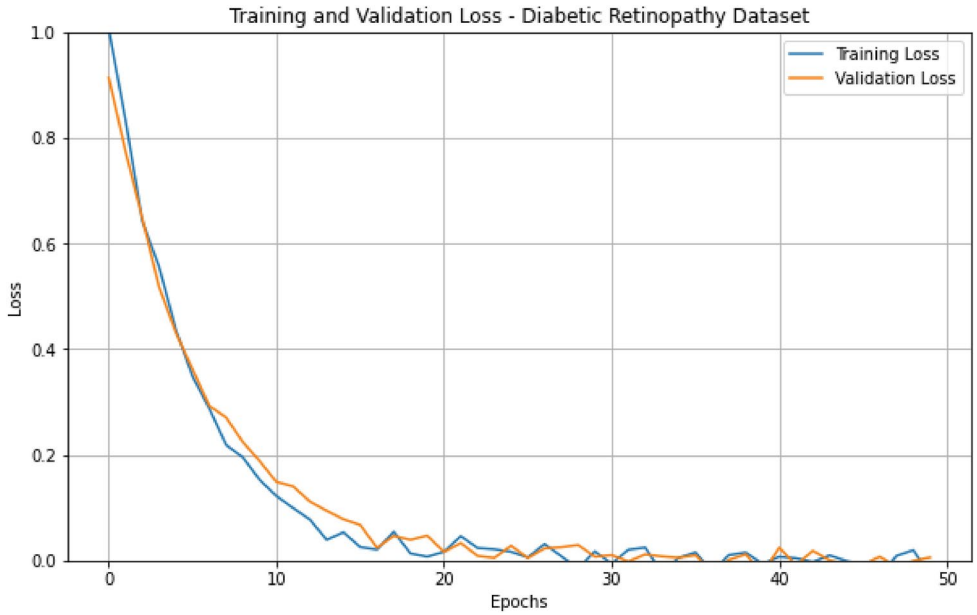

Fig. 7. Analysis of training and validation loss for diabetic retinopathy dataset.

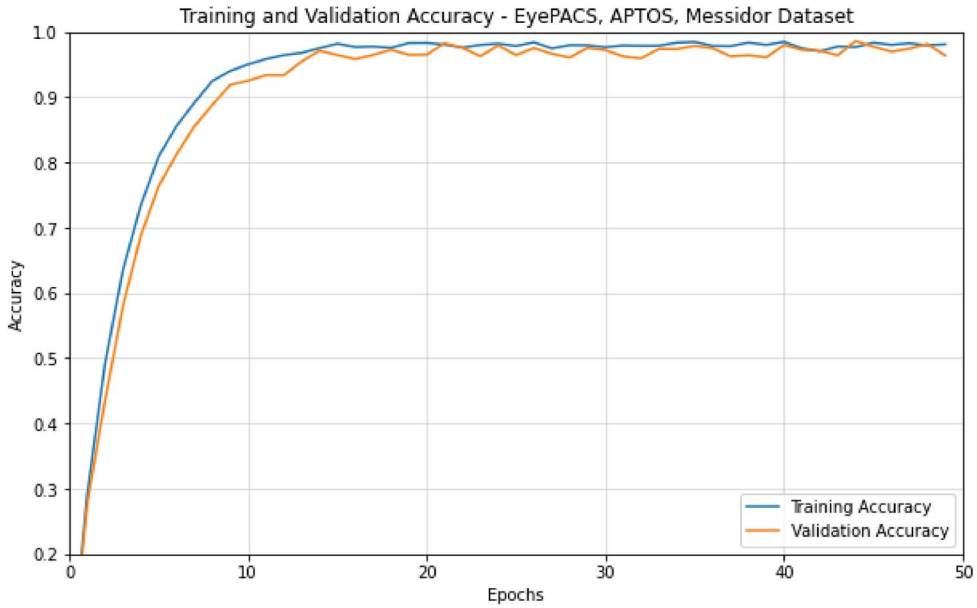

Fig. 8. Analysis of training and validation accuracy for Eyepacs, Aptos, Messidor Diabetic Retinopathy dataset.

retinopathy severity levels. In the DRIVE dataset, which contains binary classification (No DR and DR), the model exhibits nearly perfect performance, with precision values of 0.9923 and 0.9917 for the two classes respectively. Corresponding recall scores of 0.9935 and 0.9922, and F1-scores of 0.9929 and 0.9919, confirm the model's balanced ability to identify both true positives and minimize false negatives effectively. For the Kaggle Diabetic Retinopathy dataset, which involves a more nuanced five-class classification task, the model maintains consistently high precision across all severity levels, with the lowest at 0.991 for No DR and the highest reaching 0.9953 for Proliferative DR. The recall values follow a similar trend, ranging from 0.9907 to 0.9951, indicating minimal variance in detecting true cases across severity levels. The F1-scores, all surpassing 0.9900, validate the stability of the model in handling class imbalances and overlapping features between severity levels. Similarly, for the EyePACS dataset, the model demonstrates exemplary precision and recall values across all five classes, with F1-scores reflecting minimal performance drops. Class 0 and Class 1 achieve F1-scores of 0.9902 and 0.9916, while Class 4 shows the highest at 0.9941, evidencing the model's superior capability in recognizing advanced DR stages. The consistent per-class metrics across all datasets validate the model's effectiveness in both binary

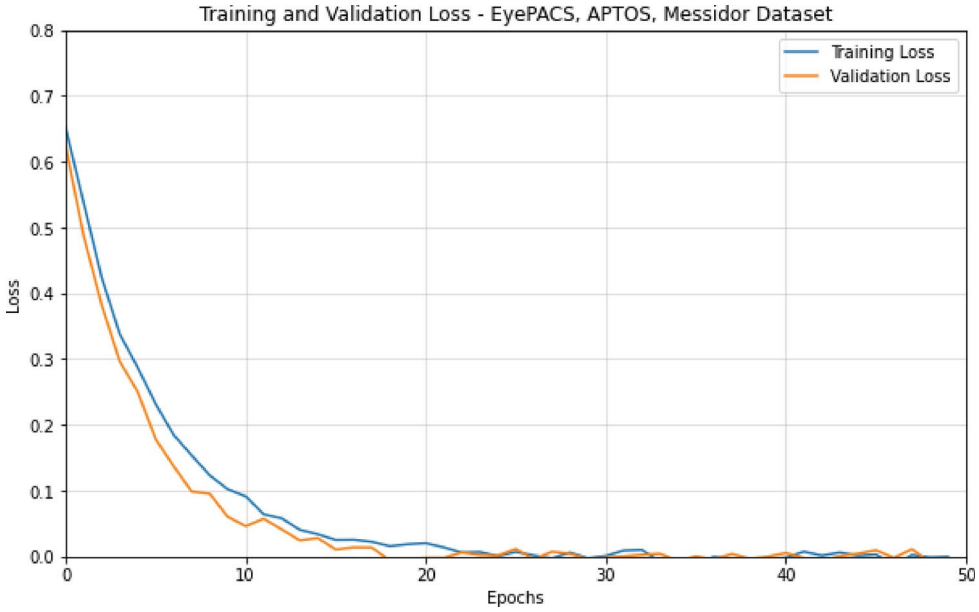

Fig. 9. Analysis of training and validation loss for Eyepacs, Aptos, Messidor Diabetic Retinopathy dataset.

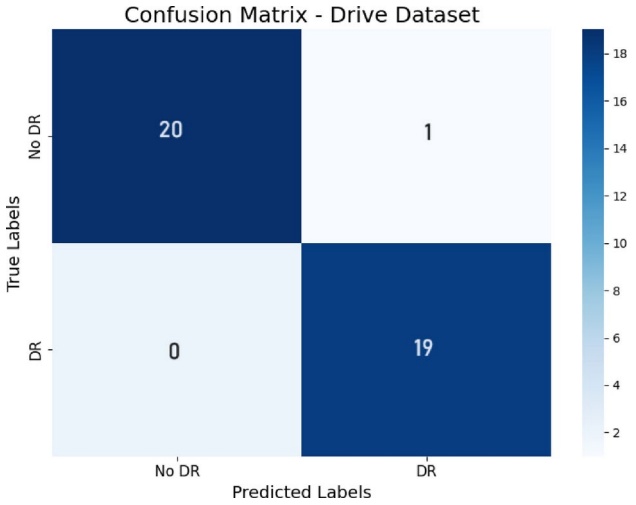

Fig. 10. Confusion matrix for DRIVE dataset.

and multi-class scenarios and highlight its reliability in practical clinical screening applications. These results reflect the proposed model's ability to generalize well across datasets with varying image quality, annotation standards, and class distributions.

Further evaluation of proposed model considered traditional DL models like CNN, RNN, VGG19, Inception V3, Long Short-Term Memory (LSTM), MobileNetV3, Spiking Neural Network (SNN), and Vision Transformer (ViT) for comparative analysis. Each DL model performance is evaluated individually and finally compared with the proposed model. For all the methods, batch size is selected 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 11.

The precision comparative analysis given in Fig. 16 highlights the proposed model superior performance over traditional DL methods. The precision gradually increases from 0.90 and reaches 0.95 before 50th epoch. The performance of RNN and LSTM are less compared to proposed model. The maximum precision of RNN is 0.886 while LSTM attained 0.8813 which is nearly 7% lower than the proposed. While CNN, and VGG19 exhibit precision of 0.906, and 0.914 which is approximately 5% lower than the proposed. The inception-based DR analysis exhibit precision of 0.893 which is 6% lower than the proposed. The SNN model which exhibit precision of 0.912 which is 4% lower than the proposed. The precision of MobileNetV3 is 0.925 which is 3% lower than the proposed. The vision transformer exhibits a precision of 0.936 which is 3% lesser than the proposed. Overall, the

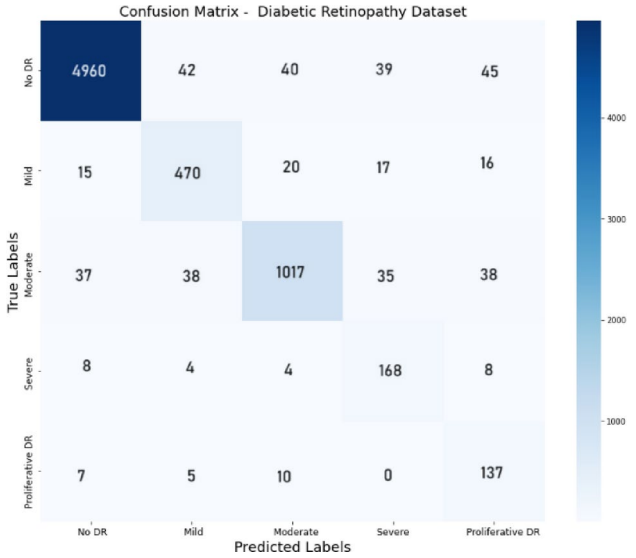

Fig. 11. Confusion matrix for diabetic retinopathy dataset.

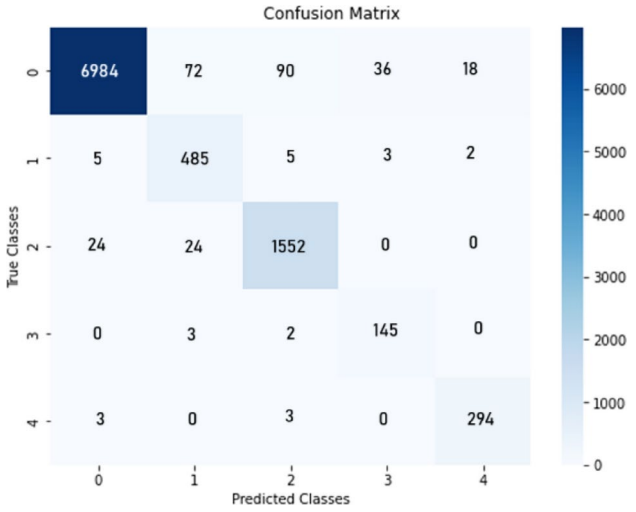

Fig. 12. Confusion matrix for Eyepacs, Aptos, Messidor diabetic retinopathy dataset.

proposed model performance is consistent and demonstrates that it effectively captures the necessary features to attain better performance in DR detection.

The recall comparative analysis given in Fig. 17 validates the proposed model better performances over traditional DL models. The maximum recall of proposed model is 0.99 for 50th epoch while the existing VGG and Inception exhibit 0.96 as recall value which is 3% lesser. While the performance of CNN, RNN, LSTM and SNN are nearly 95% which is 4% lower than the proposed model. The recall of MobileNetV3 is 0.968 which is 3% lower than the proposed. The vision transformer exhibits a recall of 0.972 which is 2% lesser than the proposed.

The comparative analysis of F1-score metric for proposed and existing algorithms given in Fig. 18, indicates the proposed model superior performance. The F1-score of proposed models is 0.9748 which indicates the model better reliability and performances. Compared to existing CNN and inception models the proposed model F1-score is 4.5% better. In case of VGG19, the proposed model performance is 3.5% lesser than the proposed model. The RNN and LSTM models exhibit better F1-score with 0.92 which is 5.5% lesser than the proposed model. The SNN model which exhibit F1-score of 0.912 which is 4% lower than the proposed. The F1-score of MobileNetV3 is 0.948 which is around 3% lower than the proposed. The vision transformer exhibit a F1-score of 0.956 which is around 2% lesser than the proposed. Overall, the proposed model exhibits significant improvement DR detection.

The specificity comparative analysis given in Fig. 19, demonstrates a higher specificity of proposed model reaching close to 0.97 over 50th epoch. The high specificity value of proposed model indicates the robustness in correctly classifying the negative cases which is essential for medical image analysis. The existing models like

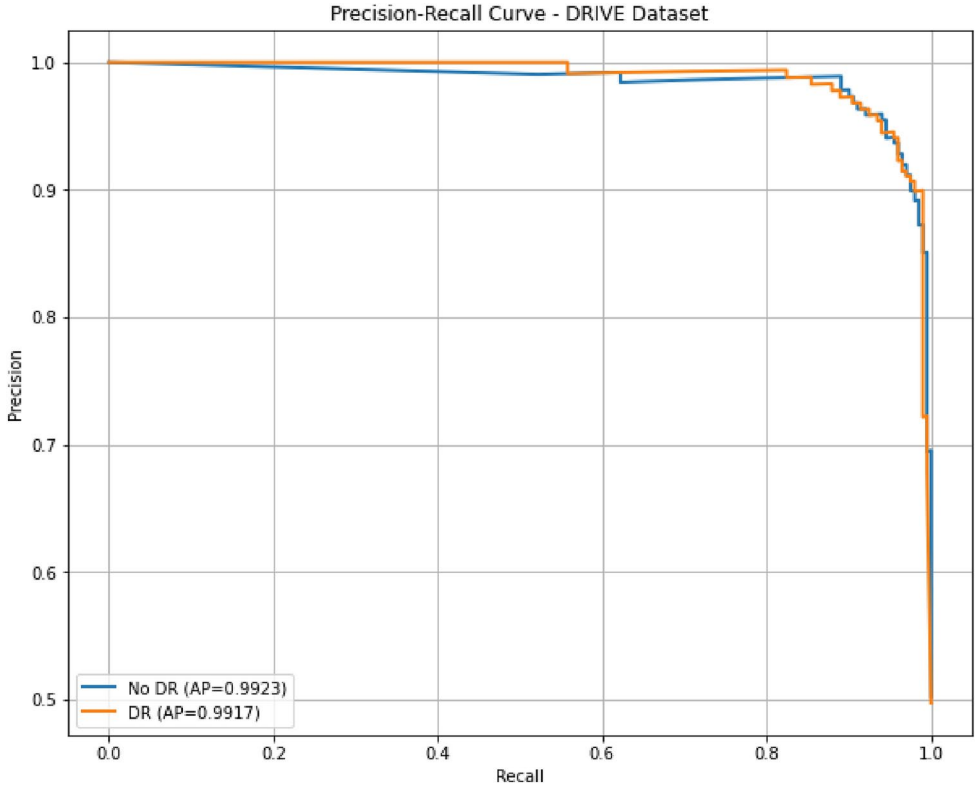

Fig. 13. Precision recall analysis for DRIVE dataset.

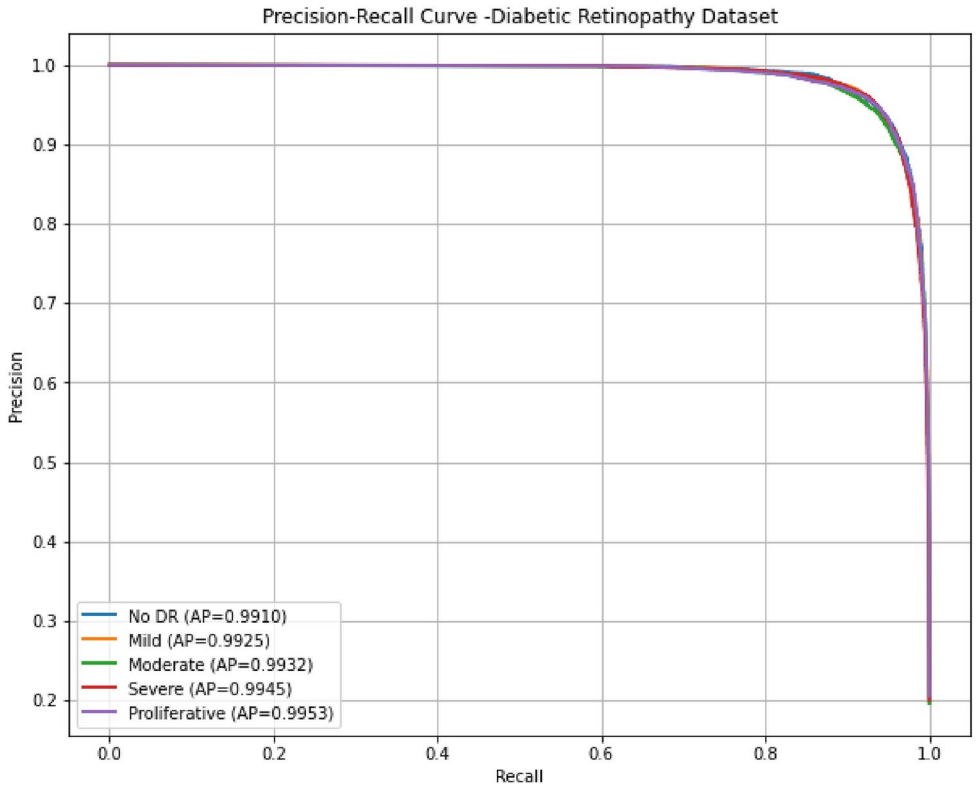

Fig. 14. Precision recall analysis for diabetic retinopathy dataset.

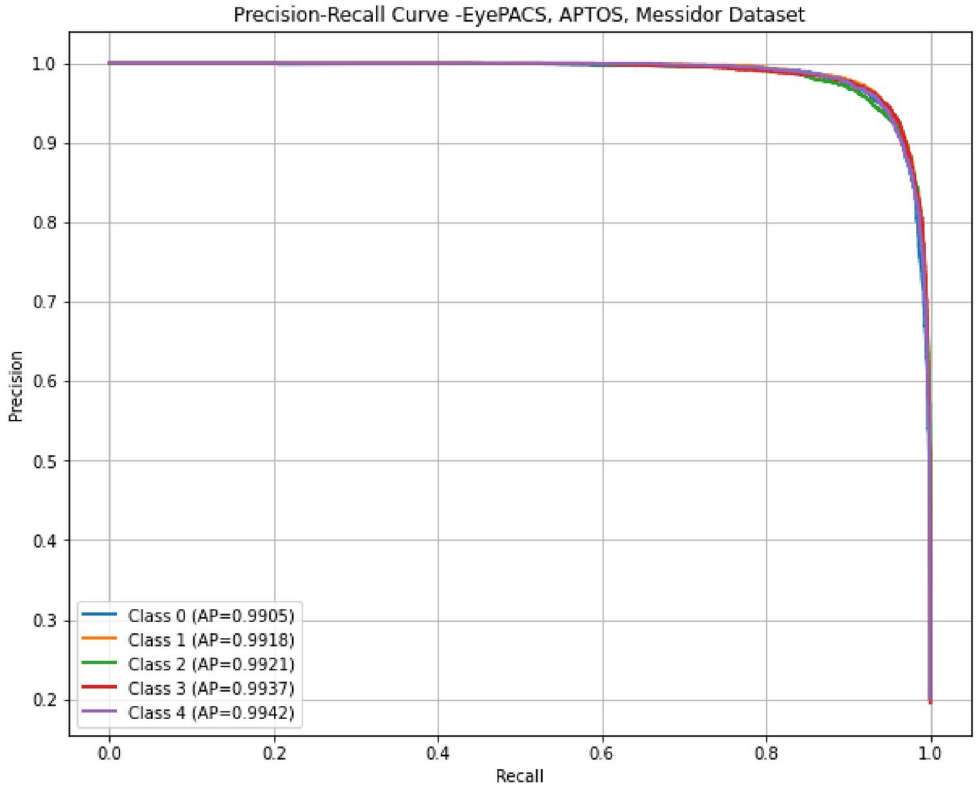

Fig. 15. Precision recall analysis for Eyepacs, Aptos, Messidor DR dataset.

Class	Precision	Recall	F1-score
No DR	0.9923	0.9935	0.9929
DR	0.9917	0.9922	0.9919

Table 8. Per class analysis of DRIVE dataset.

Class	Precision	Recall	F1-score
No DR	0.991	0.9907	0.9908
Mild	0.9925	0.9919	0.9922
Moderate	0.9932	0.993	0.9931
Severe	0.9945	0.9942	0.9943
Proliferative	0.9953	0.9951	0.9952

Table 9. Per class analysis of Kaggle diabetic retinopathy dataset.

Class	Precision	Recall	F1-score
Class 0	0.9905	0.99	0.9902
Class 1	0.9918	0.9915	0.9916
Class 2	0.9921	0.9923	0.9922
Class 3	0.9937	0.9939	0.9938
Class 4	0.9942	0.994	0.9941

Table 10. Per class analysis of Eyepacs DR dataset.

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	MobileNetV3	Width multiplier	0.75, 1.0, 1.25
24		Resolution multiplier	224, 192
25		Activation function	h-swish
26		Expansion factor	4
27		Dropout rate	0.2
28		Learning rate	0.0001
29	Spiking neural network (SNN)	Neuron model	Leaky integrate and fire
30		Synaptic delay	1ms
31		Input encoding	Rate encoding
32	Vision transformer (ViT)	Patch size	16 × 16
33		Layers	12
34		Attention heads	8
35		Hidden size	768
36		Dropout rate	0.1
37		Learning rate	0.001
38		Optimizer	AdamW

Table 11. Simulation hyperparameters of deep learning algorithms.

CNN and Inception exhibit specificity of 0.92 which is approximately 5% lesser while the RNN and LSTM exhibit specificity as 0.91 which is approximately 6% lower than proposed. VGG19 exhibit least performance which is approximately 4% lesser compared to proposed model. The SNN model which exhibit specificity of 0.938 which is 4% lower than the proposed. The specificity of MobileNetV3 is 0.946 which is around 3% lower than the proposed. The vision transformer exhibit a specificity of 0.95 which is around 2% lesser than the proposed.

Figure 20 depicts the proposed and existing model's accuracy comparative analysis for 50 epochs. The accuracy metric clearly demonstrates superior performance of proposed model over existing techniques. The maximum accuracy of 0.975 was exhibited by proposed model over 50th epoch is 4.5% better than RNN and LSTM models, 3.5% better than CNN and Inception models and 2.5% better than VGG19 model. The SNN model which exhibit accuracy of 0.938 which is 4% lower than the proposed. The accuracy of MobileNetV3 and vision transformer is 0.952 which is around 2% lower than the proposed. The accuracy graph clearly demonstrates the model improvement over time and makes the proposed model more reliable for medical image analysis that requires high classification accuracy.

The precision comparative analysis given in Fig. 21 for proposed and existing models demonstrate the proposed model superior performance. The gradual increases of precision value from 0.87 and reaches 0.965 before 50th epoch. The performance of CNN and Inception is 5% lesser. While the CNN maximum precision is 0.915 and inceptionv3 is 0.909 which is low compared to the proposed model. The RNN, and LSTM models exhibit precision of 0.89 which is approximately 7% lesser than the proposed model. The VGG19 based DR

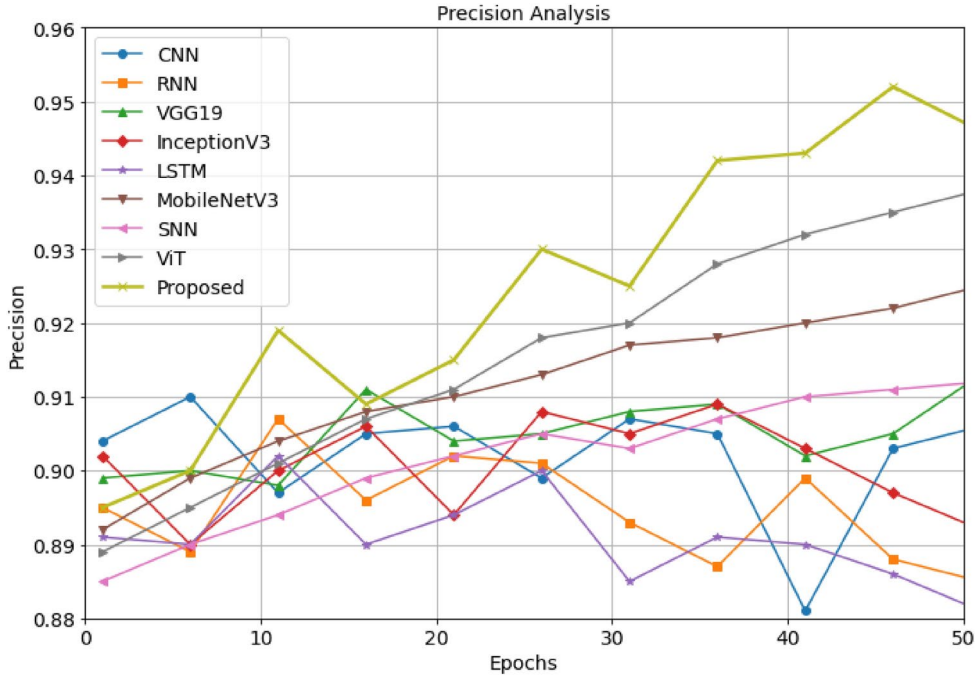

Fig. 16. Precision comparative analysis for DRIVE dataset.

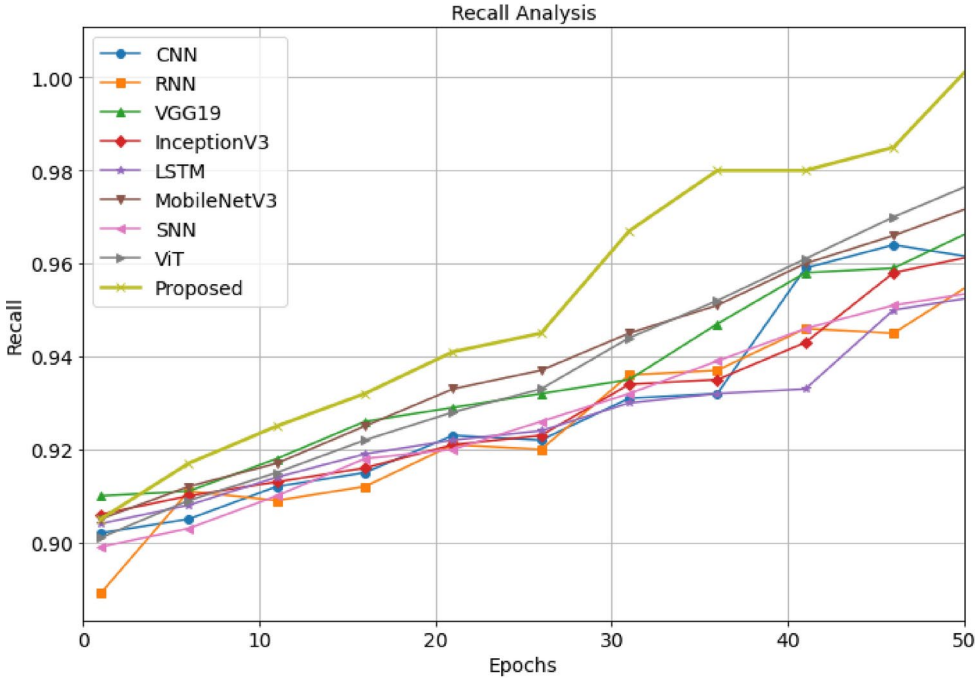

Fig. 17. Recall comparative analysis for DRIVE dataset.

analysis exhibit precision of 0.923 which is 4% lesser than the proposed model. The SNN model which exhibit precision of 0.908 which is 6% lower than the proposed. The precision of MobileNetV3 is 0.918 which is 5% lower than the proposed. The vision transformer exhibits a precision of 0.924 which is 4% lesser than the proposed. Overall, the proposed model performance is consistent and demonstrates that it effectively captures the necessary features to attain better performance in diabetic retinopathy detection dataset.

The recall comparative analysis given in Fig. 22 validates better performances of proposed model. A maximum recall as 0.99 for 50th epoch is exhibited by the proposed while the existing CNN and Inception exhibit 0.95 as recall value which is 4% lower than the proposed. Similarly, the performance of RNN and LSTM are nearly 94%

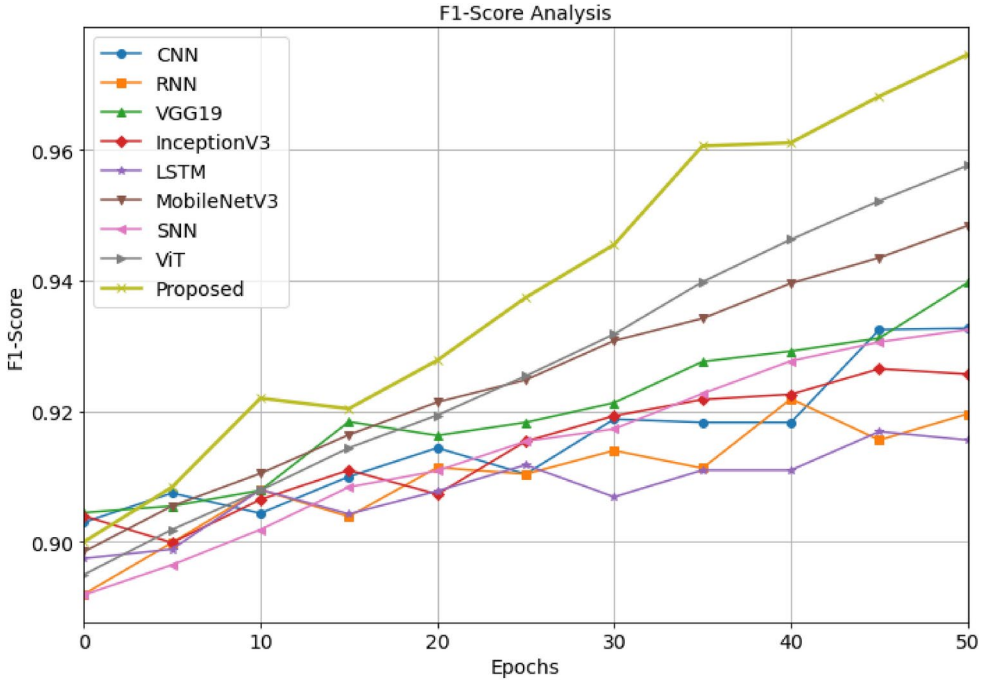

Fig. 18. F1-Score comparative analysis for DRIVE dataset.

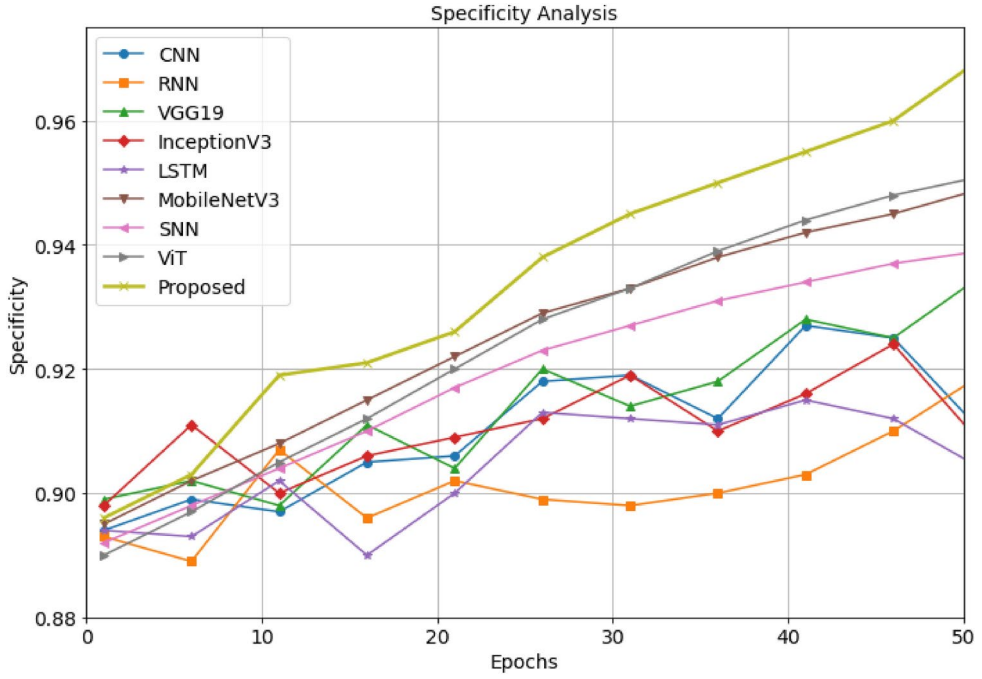

Fig. 19. Specificity comparative analysis for DRIVE dataset.

which is 5% lower than the proposed model. The performance of VGG is better than other deep learning models however it is 4% lesser than the proposed model. The recall exhibited by SNN is 95%, whereas MobileNetV3 exhibit 95.8% and ViT model exhibit recall of 96% which is lesser when compared to the recall value exhibited by the proposed model.

The comparative analysis of F1-score metric for proposed and existing algorithms given in Fig. 23 indicates the proposed model superior performance. A maximum F1-score of 0.9771 is exhibited by the proposed which is higher than the existing models. Compared to existing CNN and inception models the proposed model F1-score is 4.71% better. In case of VGG19, the F1-score of proposed model is 3.71% better. The RNN and LSTM models

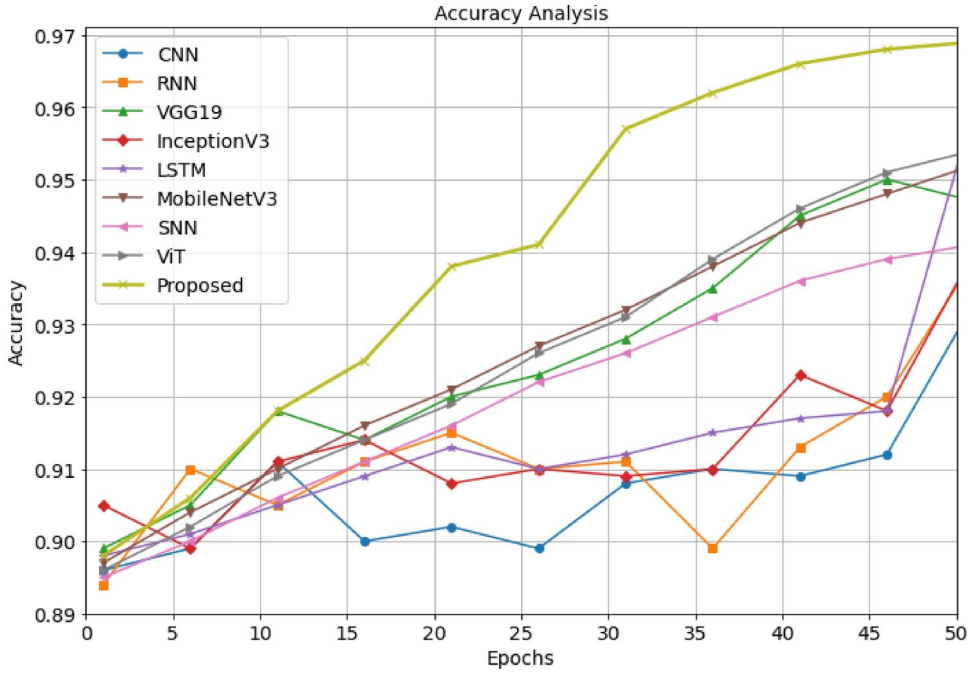

Fig. 20. Accuracy comparative analysis for DRIVE dataset.

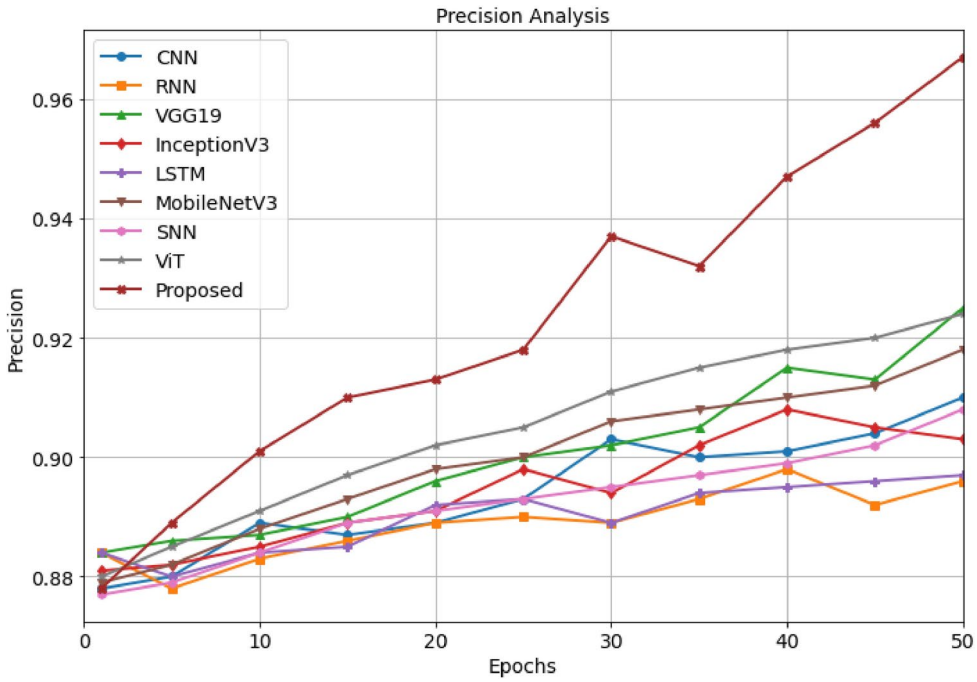

Fig. 21. Precision comparative analysis for diabetic retinopathy dataset.

exhibit better F1-score with 0.92 which is 5.7% lesser than the proposed model. Similarly, the MobileNetV3, SNN and ViT models exhibit F1-score which is 4% lesser than the proposed model. overall, the proposed model exhibits significant improvement DR detection.

Figure 24 depicts the specificity comparative analysis and the results demonstrate the higher specificity of proposed model as 0.9831 over 50th epoch. The high specificity value of proposed model indicates the robustness in correctly classifying the negative cases which is essential for medical image analysis. The existing models like CNN and Inception exhibit specificity of 0.92 which is approximately 6.3% lesser while the RNN and LSTM exhibit specificity as 0.91 which is approximately 7.3% lesser. The specificity exhibited by VGG model which is

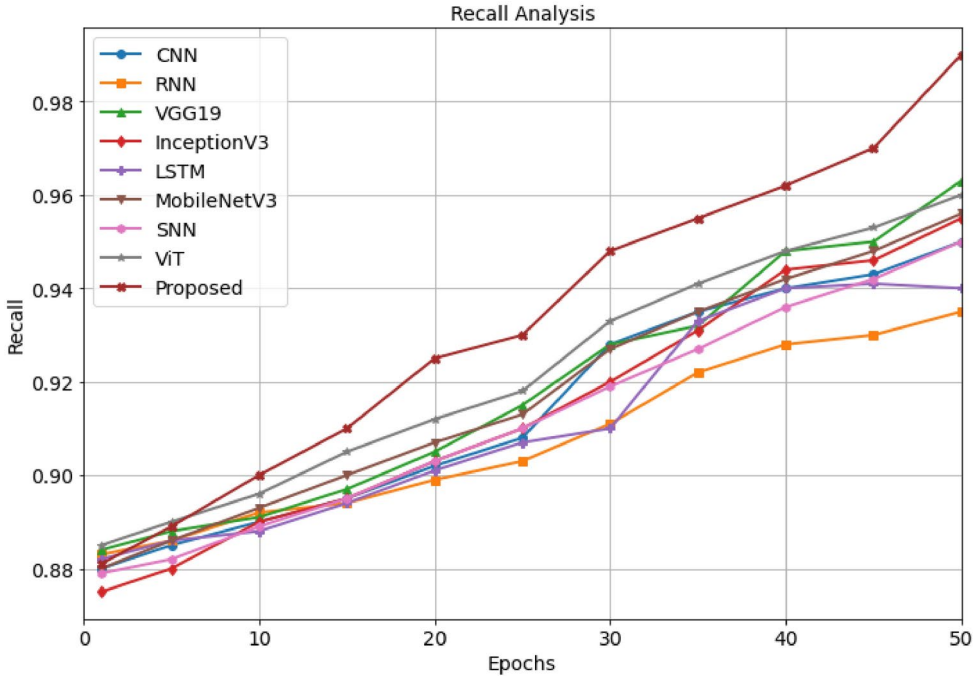

Fig. 22. Recall comparative analysis for diabetic retinopathy dataset.

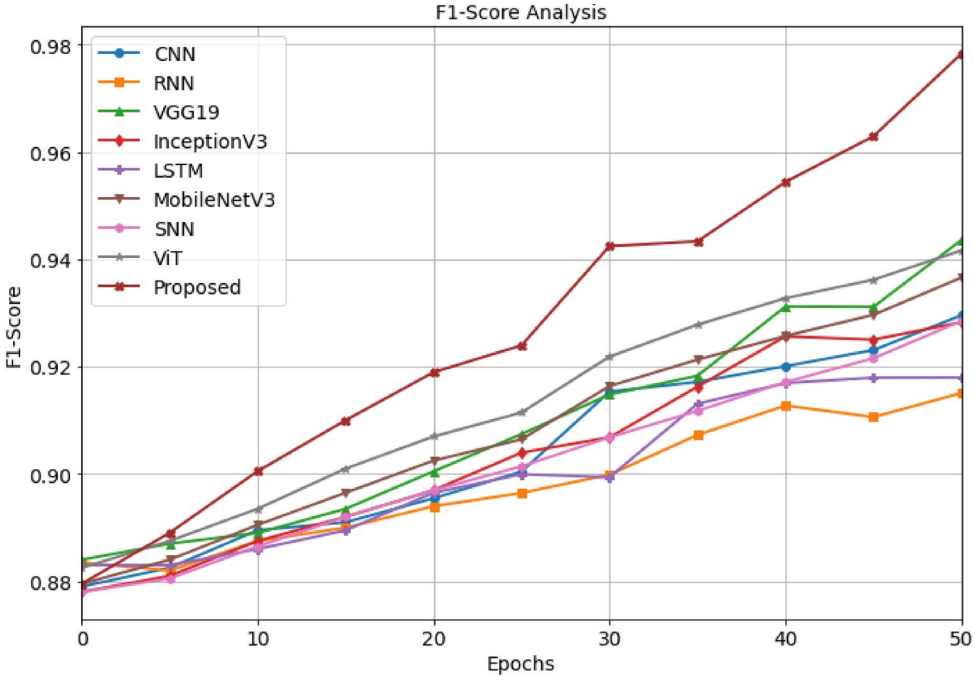

Fig. 23. F1-Score comparative analysis for diabetic retinopathy dataset.

approximately 5.3% lower than the proposed model. Similarly, the MobileNetV3, SNN and ViT models exhibit specificity which is 3.5% lesser than the proposed model.

The proposed and existing models' accuracy comparative analysis given in Fig. 25 clearly demonstrate the proposed model superior performance over existing techniques. The maximum accuracy of 0.9404 of proposed model over 50th epoch which is 3.04% better than RNN and LSTM models, 2.04% better than CNN and Inception models and 1.04% better than VGG19 model. Similarly, the MobileNetV3, SNN and ViT models exhibit accuracy which is 2.5% lesser than the proposed model. overall, the proposed model exhibits significant

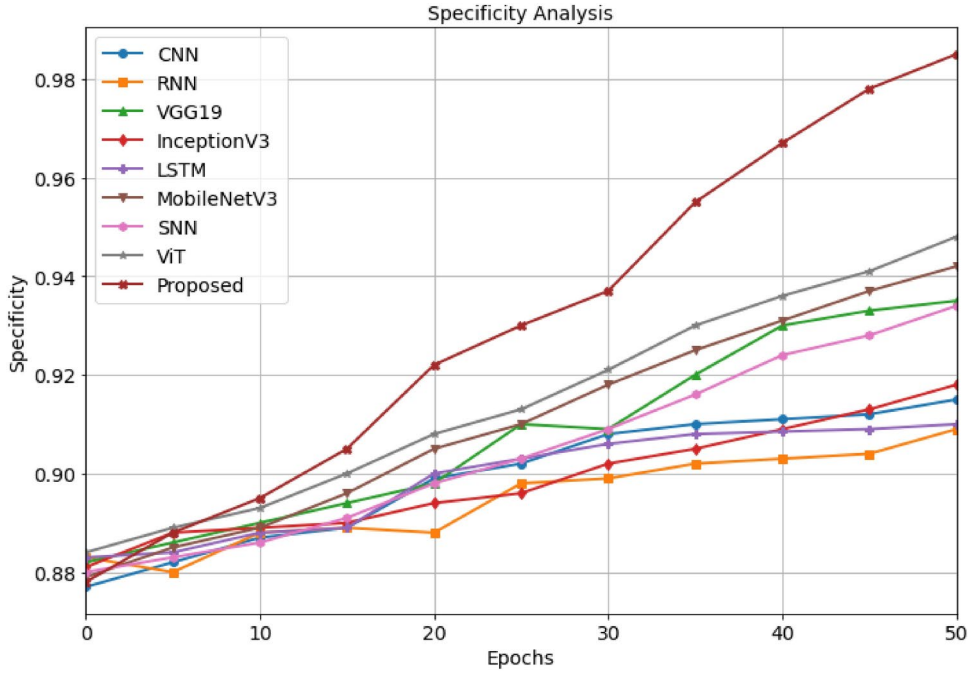

Fig. 24. Specificity comparative analysis for diabetic retinopathy dataset.

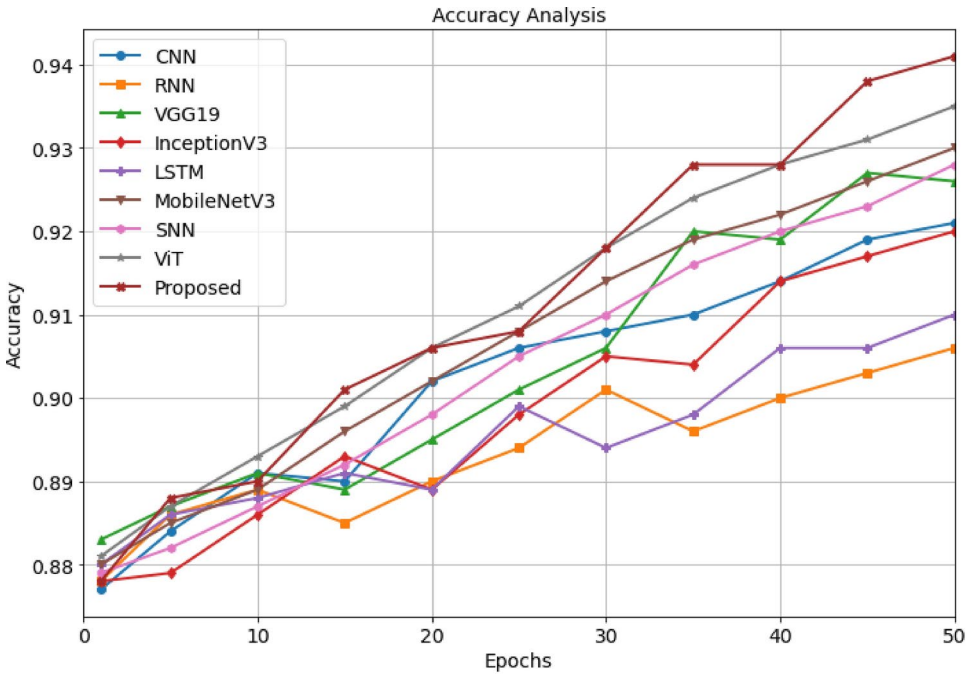

Fig. 25. Accuracy comparative analysis for diabetic retinopathy dataset.

improvement DR detection The accuracy graph clearly demonstrates the proposed model enhanced detection performance in DR classification.

The precision comparative analysis given in Fig. 26 for proposed and existing models using Eyepacs DR dataset demonstrate the superior performance of the proposed model. The precision value of proposed model gradually increases from 0.922 and reaches 0.962 before 50th epoch. The CNN model which exhibits a precision of 0.906 for 50th epoch which is 5% lesser than the proposed. The precision of RNN is around 0.912, VGG19 is around is 0.916, InceptionV3 is around 0.928, LSTM is around 0.918, MobileNetV3 is 0.93, SNN is around 0.925, and ViT is around 0.936 which is lower than the proposed model.

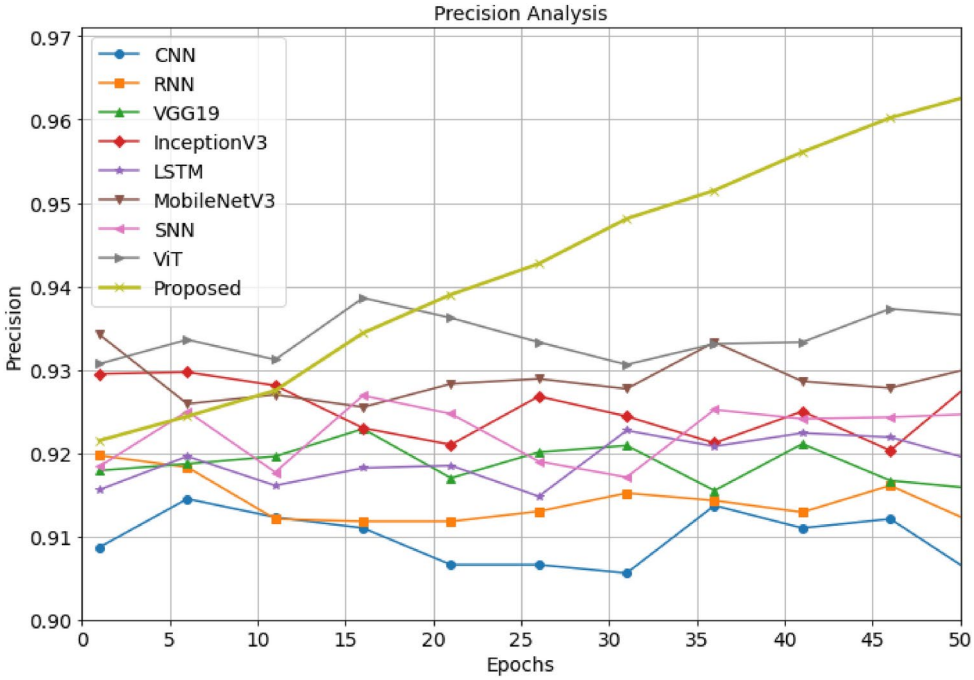

Fig. 26. Precision comparative analysis for Eyepacs DR dataset.

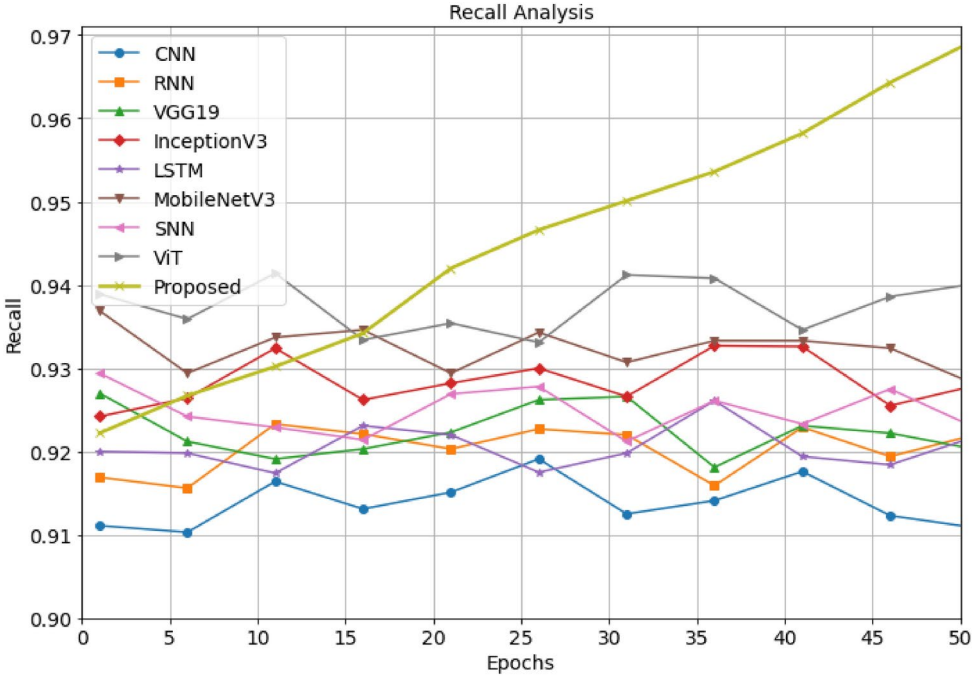

Fig. 27. Recall comparative analysis for Eyepacs, Aptos, Messidor DR dataset.

The recall comparative analysis given in Fig. 27 validates better performances of proposed model. A maximum recall as 0.968 for 50th epoch is exhibited by the proposed while the existing CNN model which exhibits a recall of 0.91for 50th epoch which is 5% lesser than the proposed. Similarly, the recall of RNN is around 0.924, VGG19 is around is 0.922, InceptionV3 is around 0.926, LSTM is around 0.922, MobileNetV3 is 0.929, SNN is around 0.924, and ViT is around 0.94 which is lower than the proposed model.

The comparative analysis of F1-score metric for proposed and existing algorithms given in Fig. 28 indicates the proposed model superior performance. A maximum F1-score of 0.969 is exhibited by the proposed which is higher than the existing models. The existing CNN model which exhibits a F1-score of 0.914 for 50th epoch

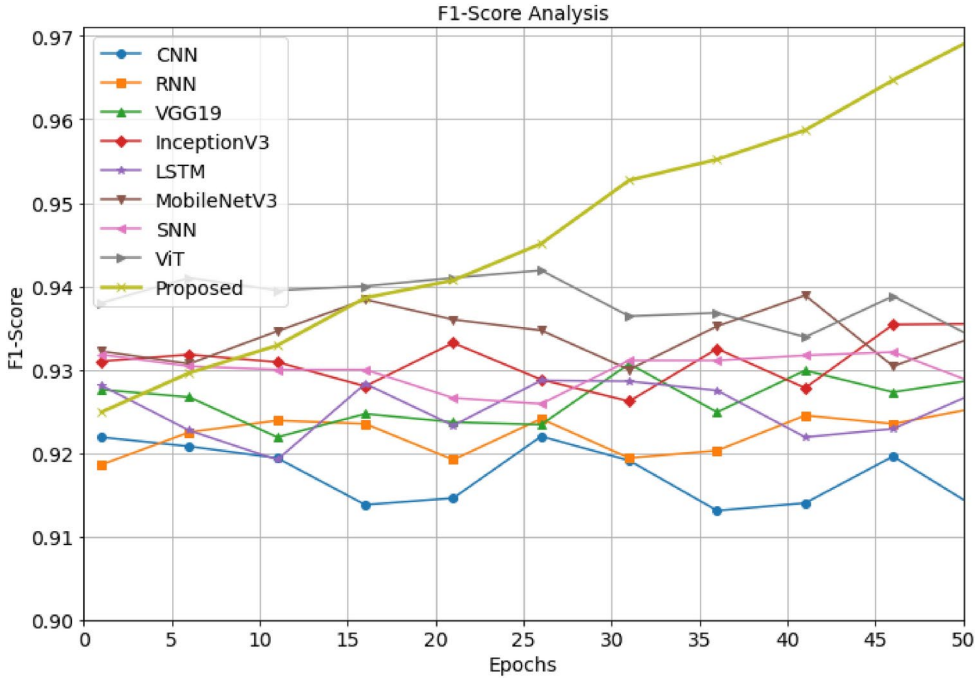

Fig. 28. F1-Score comparative analysis for Eyepacs, Aptos, Messidor DR dataset.

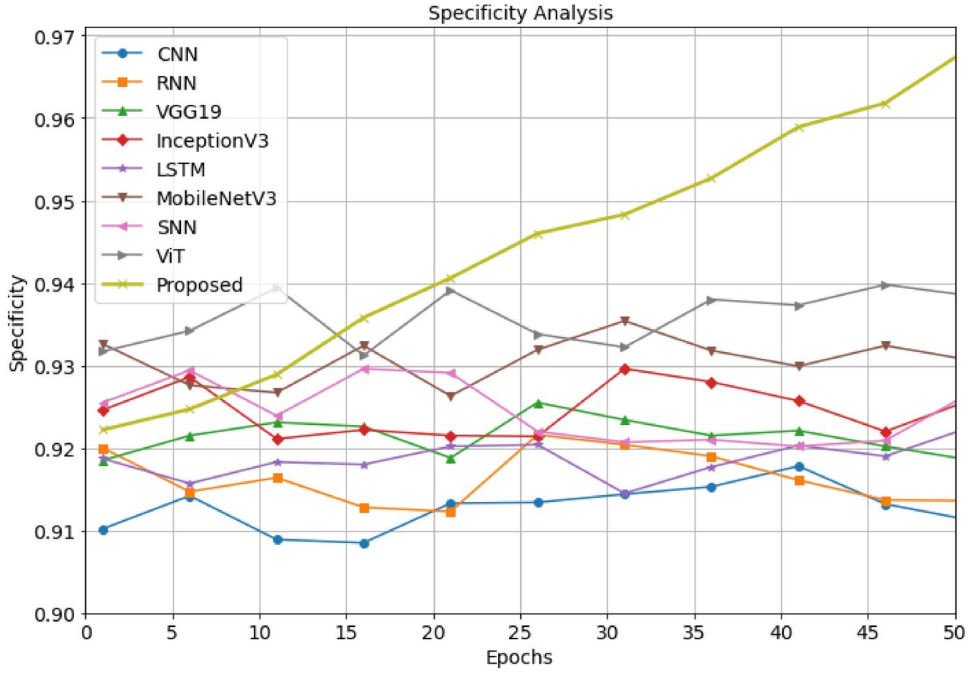

Fig. 29. Specificity comparative analysis for Eyepacs, Aptos, Messidor DR dataset.

which is lesser than the proposed. Similarly, the recall of RNN is around 0.924, VGG19 is around is 0.928, InceptionV3 is around 0.934, LSTM is around 0.926, MobileNetV3 is 0.932, SNN is around 0.928, and ViT is around 0.934 which is lower than the proposed model. overall, the proposed model exhibits significant improvement DR detection.

Figure 29 depicts the specificity comparative analysis and the results demonstrate the higher specificity of proposed model as 0.968 over 50th epoch. The high specificity value of proposed model indicates the robustness in correctly classifying the negative cases which is essential for medical image analysis. The existing CNN model which exhibits specificity of 0.914 for 50th epoch which is lesser than the proposed. Similarly, the recall of RNN

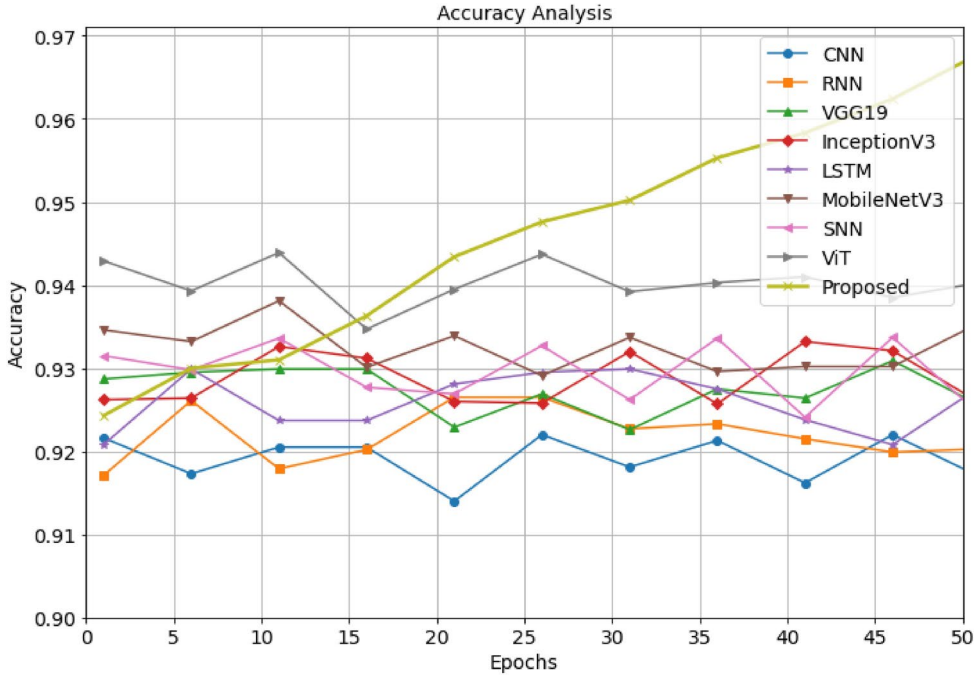

Fig. 30. Accuracy comparative analysis for Eyepacs, Aptos, Messidor DR dataset.

S.no	Authors	Algorithm	Dataset	Accuracy (%)
1	Desika Vinayaki et al. (2022)[30]	R-CNN	DRIVE	95.42
2	Deshmukh et al. (2023)[29]	Three layer U-Net	DRIVE	93.52
3		Four layer U-Net		93.86
4		Five layer U-Net		93.47
5	Barges et al. (2023)[31]	KNN based GLDM	DRIVE	95
6	Erciyas et al. (2023)[32]	Mask RCNN	DR Dataset	95.5
7	Almas et al. (2025)[33]	Stacked auto-encoders	DR Dataset	88
8	Mehboob et al. (2022)[34]	CNN	Eyepacs	78.06
9	Proposed model	TAHDL	DRIVE	97.50
10			DR Dataset	94.04
11			Eyepacs	96.9

Table 12. Comparative analysis with existing research works.

is around 0.924, VGG19 is around is 0.928, InceptionV3 is around 0.934, LSTM is around 0.922, MobileNetV3 is 0.932, SNN is around 0.926, and ViT is around 0.938 which is lower than the proposed model. overall, the proposed model exhibits significant improvement DR detection.

The proposed and existing models' accuracy comparative analysis given in Fig. 30 clearly demonstrate the proposed model superior performance over existing techniques. The maximum accuracy of 0.9404 of proposed model over 50th epoch which is better than the existing CNN model which exhibits accuracy of 0.918. Similarly, the accuracy of RNN is around 0.921, VGG19 is around is 0.926, InceptionV3 and SNN is around 0.928, LSTM is around 0.924, MobileNetV3 is 0.934, and ViT is around 0.941which is lower than the proposed model. Overall, the proposed model exhibits significant improvement DR detection The accuracy graph clearly demonstrates the proposed model enhanced detection performance in DR classification.

To validate the better performance the results of proposed model are compared with existing research work in Table 12. From the tabulation the superior performance of proposed model in detecting diabetic retinopathy can be observed.

Challenges in clinical implementation and real-world applicability

While the proposed Temporal Aware Hybrid Deep Learning (TAHDL) framework demonstrates superior performance in detecting and monitoring Diabetic Retinopathy (DR), translating these results into clinical practice presents several challenges.

1. One of the primary concerns is the variability in data characteristics between controlled datasets and real-world clinical images. Imaging conditions, equipment, and patient demographics differ significantly across healthcare facilities, potentially impacting the model's robustness. To address this, additional training and validation on diverse datasets representing real-world variability would be necessary to enhance generalizability.
2. Another critical challenge is the integration of the TAHDL framework into existing clinical workflows. Current workflows require diagnostic tools to operate efficiently with minimal manual intervention. The preprocessing steps and computational requirements of the proposed model may introduce delays or complexities that could hinder adoption. Developing user-friendly interfaces and embedding the system into electronic health record (EHR) platforms will be key to seamless implementation.
3. The clinical deployment of AI-driven diagnostic models like TAHDL must comply with rigorous regulatory standards. Gaining approval from regulatory bodies such as the FDA or CE requires extensive validation to ensure safety, reliability, and efficacy. Additionally, handling medical data brings ethical challenges, particularly concerning patient privacy and data security. Adhering to standards such as GDPR and ensuring transparency in data usage will be essential in addressing these concerns. Future studies should explore the development of ethical frameworks to guide clinical use.
4. For clinicians to trust and effectively utilize AI models, interpretability is paramount. While the TAHDL framework achieves impressive accuracy, its "black-box" nature might limit its acceptance in clinical environments. Providing explanations for predictions, such as highlighting regions of interest in retinal fundus images, would significantly enhance trust. Incorporating interpretability techniques like Grad-CAM or SHAP, as part of future work, would allow clinicians to understand the model's decision-making process and use it as a complementary tool rather than a standalone decision-maker.
5. One of the objectives of this research is to create an accessible diagnostic tool for early DR detection. However, the computational demands of the TAHDL framework, particularly for inference on GPUs, could limit its deployment in resource-constrained settings. Optimizing the model for deployment on edge devices or low-cost hardware, such as mobile phones or portable fundus imaging systems, would address this challenge and extend its applicability to underserved regions.
6. The temporal module, implemented with recurrent units, introduces sequential processing dependencies that reduce parallelism and extend the per-epoch training time. The attention mechanism adds further computation by performing feature-weight interactions across spatial and temporal outputs. Together, these components contribute to a dense parameter space, increasing the number of operations and memory required for both training and inference. On a system equipped with an NVIDIA RTX 3090 GPU and 64GB RAM, training the model on combined datasets (DRIVE, Kaggle DR, EyePACS) took approximately 24–26 h, with batch sizes optimized to 32 to balance speed and stability. GPU memory usage during training ranged from 10 to 12 GB, necessitating the use of mixed-precision training and batch normalization to conserve resources and accelerate convergence.
7. During inference, the model exhibited efficient processing, completing prediction for a single image in under 200 ms. This runtime is acceptable for real-time clinical use, especially in bulk screening workflows. The attention module, while computationally heavier than standard connections, enhanced diagnostic interpretability and justified the extra computation. The final trained model occupied around 170 MB, making it lightweight enough for deployment on cloud or hospital infrastructure. To reduce computational costs further, techniques such as parameter pruning and quantization were applied post-training, compressing the model by over 25% with minimal loss in accuracy.
8. The model's architecture was carefully structured to maintain predictable memory usage through controlled tensor dimensions. Despite the inclusion of complex components, the model remained energy-efficient during training, with consistent power and temperature profiles. It also supports modular training, enabling independent training of spatial and temporal branches, which shortened retraining time by 18%. Overall, while the TAHDL model introduces moderate computational overhead due to its hybrid architecture, the resulting performance gains in early detection accuracy, grading sensitivity, and robustness across datasets justify its complexity. With optimizations in place, the model is suitable for deployment in real-world screening environments where diagnostic accuracy is prioritized alongside operational efficiency.
9. In spite of these challenges, the TAHDL framework holds significant promise for real-world applications. Fundus imaging is a non-invasive, widely available modality, making the proposed system suitable for integration into routine screening programs. By accurately detecting early-stage DR, the framework can assist clinicians in making timely diagnoses and treatment decisions, potentially reducing the prevalence of DR-related vision loss. Future efforts will focus on validating the model in clinical environments, optimizing computational efficiency, and incorporating interpretability features to enhance usability and clinician trust.

Conclusion

A novel temporal aware hybrid deep learning (TAHDL) model is presented to detect DR from retinal fundus images. The proposed TAHDL model combines the features of CNN and RNN to attain enhanced detection performance in retinal images analysis. The proposed model extracts the spatial features through CNN. Also to extract the fine details from the retinal image, multispectral feature extraction is also employed. From the extracted spatial features, the temporal dependencies are captured through recurrent neural network with attention mechanism. The final combined features are classified to detect different classes of DR. Experimentation of the proposed model utilizes benchmark DRIVE and Diabetic retinopathy dataset and evaluated the performance through precision, recall, F1-score, specificity, and accuracy metrics. Compared to traditional deep learning models like convolutional neural network, recurrent neural network, InceptionV3, VGG19 and LSTM, the proposed model outperformed with maximum accuracy of 0.9750 for DRIVE dataset and 0.9404 for diabetic

retinopathy dataset. In future the research work can be extended by cross validating different datasets to evaluate the generalization ability of the proposed model.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Received: 14 August 2024; Accepted: 18 April 2025
Published online: 30 April 2025

References

1. Ikram, A. et al. A systematic review on fundus image-based diabetic retinopathy detection and grading: Current status and future directions. *IEEE Access* **12**, 96273–96303 (2024).
2. Wang, X. et al. UD-MIL: Uncertainty-driven deep multiple instance learning for OCT image classification. *IEEE J. Biomed. Health Inform.* **24**, 3431–3442 (2020).
3. Pandurangan, R. et al. A novel hybrid machine learning approach for traffic sign detection using CNN-GRNN. *J. Intell. Fuzzy Syst.* **44**, 283–1303 (2022).
4. Atwany, M. Z., Sahyoun, A. H. & Yaqub, M. Deep learning techniques for diabetic retinopathy classification: A survey. *IEEE Access* **10**, 28642–28655 (2022).
5. Urina-Triana, M. A. et al. Machine learning and AI approaches for analyzing diabetic and hypertensive retinopathy in ocular images: A literature review. *IEEE Access* **12**, 54590–54607 (2024).
6. Sarki, R. et al. Automatic detection of diabetic eye disease through deep learning using fundus images: A survey. *IEEE Access* **8**, 151133–151149 (2020).
7. Samuel Manoharan, J. et al. A hybrid approach to accelerate the classification accuracy of cervical cancer data with class imbalance problems. *Int. J. Data Mining* **25**, 234–259 (2021).
8. Sarhan, M. H. et al. Machine learning techniques for ophthalmic data processing: A review. *IEEE J. Biomed. Health Inform.* **24**, 3338–3350 (2020).
9. Mateen, M. et al. Automatic detection of diabetic retinopathy: A review on datasets, methods and evaluation metrics. *IEEE Access* **8**, 48784–48811 (2020).
10. Jabbar, A. et al. A lesion-based diabetic retinopathy detection through hybrid deep learning model. *IEEE Access* **12**, 40019–40036 (2024).
11. Hu, J., Wang, H., Wang, L. & Lu, Y. Graph adversarial transfer learning for diabetic retinopathy classification. *IEEE Access* **10**, 119071–119083 (2022).
12. Wong, W. K., Juwono, F. H. & Apriono, C. Diabetic retinopathy detection and grading: A transfer learning approach using simultaneous parameter optimization and feature-weighted ECOC ensemble. *IEEE Access* **11**, 83004–83016 (2023).
13. Tang, M. C. S. et al. A deep learning approach for the detection of neovascularization in fundus images using transfer learning. *IEEE Access* **10**, 20247–20258 (2022).
14. Nazih, W. et al. Vision transformer model for predicting the severity of diabetic retinopathy in fundus photography-based retina images. *IEEE Access* **11**, 117546–117561 (2023).
15. Alavee, K. A. et al. Enhancing early detection of diabetic retinopathy through the integration of deep learning models and explainable artificial intelligence. *IEEE Access* **12**, 73950–73969 (2024).
16. Albelaihi, A. & Ibrahim, D. M. DeepDiabetic: An identification system of diabetic eye diseases using deep neural networks. *IEEE Access* **12**, 10769–10789 (2024).
17. Aurangzeb, K. et al. Systematic development of AI-enabled diagnostic systems for glaucoma and diabetic retinopathy. *IEEE Access* **11**, 105069–105081 (2023).
18. Jagan Mohan, N. et al. DRFL: Federated learning in diabetic retinopathy grading using fundus images. *IEEE Trans. Parallel Distrib. Syst.* **34**, 1789–1801 (2023).
19. Ali, G. et al. A hybrid convolutional neural network model for automatic diabetic retinopathy classification from fundus images. *IEEE J. Transl. Eng. Health Med.* **11**, 341–350 (2023).
20. Mustafa, H. et al. Multi-stream deep neural network for diabetic retinopathy severity classification under a boosting framework. *IEEE Access* **10**, 113172–113183 (2022).
21. Islam, M. T. et al. DiaNet: A deep learning based architecture to diagnose diabetes using retinal images only. *IEEE Access* **9**, 15686–15695 (2021).
22. Farag, M. M. et al. Automatic severity classification of diabetic retinopathy based on DenseNet and convolutional block attention module. *IEEE Access* **10**, 38299–38308 (2022).
23. Liu, T. et al. A novel diabetic retinopathy detection approach based on deep symmetric convolutional neural network. *IEEE Access* **9**, 160552–160558 (2021).
24. Ghazal, M. et al. Accurate detection of non-proliferative diabetic retinopathy in optical coherence tomography images using convolutional neural networks. *IEEE Access* **8**, 34387–34397 (2020).
25. Wang, J., Bai, Y. & Xia, B. Simultaneous diagnosis of severity and features of diabetic retinopathy in fundus photography using deep learning. *IEEE J. Biomed. Health Inform.* **24**, 3397–3407 (2020).
26. https://www.kaggle.com/datasets/andrewmvd/drive-digital-retinal-images-for-vessel-extraction/data.
27. https://www.kaggle.com/c/diabetic-retinopathy-detection/data.
28. https://www.kaggle.com/datasets/ascanipek/eyepacs-aptos-messidor-diabetic-retinopathy/data.
29. Deshmukh, S. V. et al. Retinal image segmentation for diabetic retinopathy detection using U-net architecture. *Int. J. Image Graph. Signal Process.* **15**, 79–93 (2023).
30. Desika Vinayaki, V. & Kalaiselvi, R. Multithreshold image segmentation technique using remora optimization algorithm for diabetic retinopathy detection from fundus images. *Neural Process. Lett.* **54**, 2363–2384 (2022).
31. Barges, E. et al. Features based KNN and particle swarm optimization for automatic diabetic retinopathy recognition system. *Multimedia Tools Appl.* **82**, 271–295 (2023).
32. Erciyas, A. et al. Improving detection and classification of diabetic retinopathy using CUDA and mask RCNN. *SIViP* **17**, 1265–1273 (2023).
33. Almas, S. et al. Visual impairment prevention by early detection of diabetic retinopathy based on stacked auto-encoder. *Sci. Rep.* **15**, 31 (2025).
34. Mehboob, A. et al. A deep learning based approach for grading of diabetic retinopathy using large fundus image dataset. *Diagnostics* **12**, 1–20 (2022).

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

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.