scientific reports

Check for updates

OPEN

A multi model deep net with an explainable AI based framework for diabetic retinopathy segmentation and classification

Neeraj Sharma✉ & Praveen Lalwani

Diabetic Retinopathy (DR) is a serious condition affecting diabetes people caused by hemorrhage in the light-sensitive retinal area. DR sufferers should receive urgent therapy to avoid vision loss. The intelligent medical diagnosis system for DR is evolving as a result of the inclusion of Artificial Intelligence (AI) technology. The current AI-driven DR diagnosis techniques are hampered by numerous serious challenges, which undermine their performance. Model performance is generally degraded by low contrast, inhomogeneous lighting, and noise, rendering accurate results hard to achieve. As a result, the proposed study has created an Adaptive Gabor Filter (AGF) based on the Chaotic Map to improve filtering performance. The multi-folded features like Local Binary Patterns (LBP), Speeded-Up Robust Features (SURF), and Texture Energy Measurement (TEM) are extracted and fed into classification phase. The classification phase has the Attention layer, the dense block of DenseNet, and Optimized Gated Recurrent Unit (OGRU) based on a Self-Adaptive Northern Goshawk Optimization (SANGO) algorithm for enhancing classification performance. The system was evaluated using three datasets: DiaRetDB1, APTOS 2019, and EyePacs, which demonstrated its robustness and reliability. Furthermore, the Grad Cam in the suggested technique assures the effective implementation of segmentation and classification performance. Performance is demonstrated by the use of Intersection over Union (IoU), Accuracy, Precision, Recall, F1-Measure, Dice Similarity Coefficient (DSC), and other metrics. In addition, the five-fold categorization is used to analyse the outcome performance. The suggested model achieved an accuracy of 99.01% on the DiaRetDB1 dataset, 98.98% on the APTOS 2019 dataset, and 99.12% on the EyePacs dataset.

Keywords Diabetic Retinopathy, Adaptive Gabor filter, Modified U-Net, Multi-folded features, Self-Adaptive Northern Goshawk Optimization, Explainable AI and Grad Cam

Chronic hyperglycemia is the primary cause of diabetes mellitus, a serious public health concern that has increased in both industrialized and developing countries[1]. Diabetes is now very common worldwide, which suggests that older people account for a larger share of all deaths. As per the 2021 report by the International Diabetic Federation, 537 million people worldwide suffered with diabetes. By 2030, it is expected to reach 643 million, and by 2045, it will reach 783 million[2]. Patients with diabetes may experience DR, an eye condition that can cause vision loss and blindness. The destruction of the retina's tiny blood vessels by high blood sugar levels brought on by DR alters the retina's blood flow[3,4]. Vision loss may result from internal bleeding brought on by damage to these blood arteries. Initial symptoms such as blurred or fluctuating vision, dark or empty spots in vision, hemorrhages into the eye, or strain in the eyes are indicative of the source of vision loss. If these early signs are disregarded, a diabetic's diabetes may progress to diabetic retinopathy[5]. Non-Proliferative DR (NPDR) and Proliferative DR (PDR) are the two main categories. PDR is the term used to define abnormal blood vessel growth, also known as neovascularization, when it is seen in the retina. NPDR is also known as retina without neovascularization[6,7].

While the signs that characterize PDR include neovascularization and vitreous HEMorrhage (HEM), hard exudates, microaneurysms, hemorrhages, and soft exudates are some of the symptoms that characterize NPDR. However, to prevent loss of vision, early detection of DR is important[8]. The first signs of DR are small, red swellings of the capillary walls, called Microaneurysms (MAs), due to the depleting capillaries in the retina.

School of Computing Sciences and Engineering, VIT Bhopal University, Kothrikalan, Sehore 466114, Madhya Pradesh, India. ✉email: neeraj.sharma2019@vitbhopal.ac.in

When less robust blood vessels break open, blood and fluid swim into the retina and cause HEMs[9]. Exudates (EX) are amorphous, glittering, bright yellowish-white patches resulting from lipid proteins leaking out and fats accumulating in the abnormally friable blood vessels. Depending on their colour and look, they can be either soft or firm. Hard exudates are bright, yellowish circles on the outer layer of the retina with sharp edges[10].

The DR severity stage is determined by how these lesions are distributed throughout the retina. In addition, risk factors for the development of diabetic retinopathy include genetic predisposition, hypertension, hyperlipidemia, long-term diabetes, and insufficient glycemic management. Diabetes sufferers are increasingly concerned about DR because it can have a serious impact on their vision and quality of life[11,12]. To enable the most suitable clinical intervention, it is necessary to identify and categorize the severity of DR. This is due to the fact that prompt analysis and care can greatly drop the chance of blindness and visual impairment[13,14]. Early intervention can lessen the disease's effects on vision and aid to limit its progression. Traditional approaches used by ophthalmologists to diagnose diabetic retinopathy involve manual operations. These include ocular coherence tomography, which involves the use of light waves to make an image of the retina; fundus photography; dilation of the pupil; visual acuity testing; tonometry to read the pressure on the eyes; and many others[15]. Although these tests are cumbersome, require knowledgeable interpretation of results, and most likely will not find the early stages of DR, they are very much useful. The manual detection of DR may turn out cumbersome and very subjective, which may end up delaying therapy and resulting in suboptimal outcomes for the patient[16].

Interest in using AI approaches to automated diagnosis of DR is growing quickly, especially when it comes to leveraging Machine Learning (ML) and Deep Learning (DL) technologies[17,18]. The most widely used DL models for ocular image examination use Convolutional Neural Networks (CNN). These models are capable of learning intricate patterns and problems within retinal pictures; hence helping in early detection of DR. Deep Neural Networks (DNN) have a variety of hyperparameters. The hyperparameters have to be tuned for DNN performance, making it a painful process in network training[19,20]. These methods are based on metaheuristic optimization algorithms for the selection of the best deep learning methods parameters in which the model's performance would be optimized. Meta-heuristics repeatedly explore and use the whole search space to come up with an optimal solution[21].

The following are the work's major contributions:

- The Chebyshev chaotic map has been incorporated within an Adaptive Gabor Filter to upgrade the quality of images by handling noise, low contrast, and illumination problems of the retina in an image.
- For efficient segmentation of retinal structures, the architecture of the U-Net has been modified with adaptive batch normalization and efficient net layers.
- The combination of TEM, SURF, and LBP in the feature extraction method, with multiple folders, gives a very comprehensive representation to the retinal pictures for classification.
- This involves the design of an OGRU with a SANGO algorithm. This optimization strategy, through its dynamic component, improves the learning by the model of intrinsic patterns, hence improving classification performance.
- The findings of segmentation and classification can be expressed visually with the help of Grad Cam. With the inclusion of this feature of grad cam the model becomes more transparent, reliable and easily explainable.

The paper is organized as follows: An overview of pertinent literature is provided in Section "Literature review". Section "Proposed methodology" delves into the proposed method, while Section "Result and discussion" reviews the outcomes. Section "Conclusion" contains the paper's conclusion.

Literature review

This section reviews some of the most recent research work regarding DR using deep learning.

Authors in[22] have established an automated ensemble DL model for DR recognition and categorization. Two DL models are combined to identify diabetic retinopathy: ResNeXt and modified DenseNet101. Compared to the previous ResNet models, the ResNeXt model is superior. The concept incorporates stacking layers, modifying the split-transform-merge technique, and providing a shortcut from one block to the next. The dense blocks in the DenseNet model undergo concatenation, which improves feature utilization efficiency. The final class label is computed by ensemble of these two models using maximum a posteriori over the outputs from the classes and normalization across them. The two datasets employed in the experiments are DIARETDB1 and APTOS19. In[23] an effective, well-optimized deep neural network is suggested that uses the Chronological Tunicate Swarm Algorithm (CTSA) to categorize DR severity. The segmentation procedure is applied after preprocessing the retinal pictures obtained from the low-quality fundus imaging. Initially, a sparse fuzzy C-means and U-Net based hybrid entropy approach is used to segment the blood vessels and the optic disc. The Gabor filter banks are then used to identify the lesion region, after which the characteristics are retrieved. Applying a bio-inspired TSA predicated on the chronological notion along with a deep Stacked Autoencoder (SAE), the final classification procedure is carried out. DIARETDB0 and DIARETDB1, two benchmark datasets, were used to assess the model.

Authors in[24] employed DL techniques for the classification of the fundus photos into five DR classes with maximum accuracy and minimum execution time. 5,819 raw photos are obtained from the merging of three different DR datasets: These include Messidor2, APTOS and IDRiD. To enhance the quality of the image and to eliminate all the unnecessary noise and artifact from the images, some preprocessing techniques are applied before training the model. This shallow CNN is built from three convolutional layer blocks, and maxpool layers and has a categorical cross entropy loss function. In[25] HADL-DR is suggested as a Hybrid Adaptive DL classifier for the early analysis of DR. As for the problem of blood vessel segmentation, introduced an improved Multichannel-based Generative Adversarial Network (MGAN) with semi-maintenance. To reduce reliance on

the encoded information, the subsequent resolution of images can be used to identify the individual components of a specific partially hidden MGAN reference that cannot be divided. The Scale Invariant Feature Transform (SIFT) function is then downloaded and the optimal function is determined using the enhanced Sequential Approximation Optimization SAO. Following that, DR classification is performed in a manner through a hybrid recurrent neural network with Long Short-Term Memory layer (LSTM). This LSTM classifier was tested using the Messidor and Kaggle benchmark datasets.

Authors in[26] have come up with accurate identification of diabetic retinopathy through a hybrid system that increments classification by a CNN that has been optimized in ensemble. When it comes to fine-tune a pretrained ResNet50 on DR images, it is proposed to use a new GraphNet124 to extract features. The suggested method of feature fusion and selection comprises of the following steps. It is important to apply Shannon Entropy to choose and integrate GraphNet124's as well as ResNet50's features. The features vector for ensemble was created using LBP and DL features used in this paper. Subsequently, the optimisation was carried out using the Sine Cosine Algorithm (SCA) and the Binary Dragonfly Algorithm (BDA). It has to be pointed out that this refined feature vector had been provided to the machine learning classifiers. The performance evaluation of the recommended hybrid architecture is based on a public and unified data set, Kaggle EyePACS. In[27] authors have presented a hybrid method for detecting DR early using the GoogleNet model based on Adaptive Particle Swarm Optimization (APSO) and transferring learning from ResNet-16 to extract features from images. The hybrid model's collected characteristics are then applied to several ML models, including random forests, linear regression, support vector machines and decision trees. These attributes are then fed into a number of classifiers for multiclass DR classification using the EyePACs dataset.

An ensemble of deep CNN for DR grading and detection is developed in[28] and trained using fundus images. InceptionV3, Xception and other pre-trained CNN were applied on each of the four patches that make up each input in the first stage. There is obviously a better understanding by the model of important information from DR pictures when narrow and profuse layer features are incorporated. The second phase deals with the training of an artificial neural network-based classifier, the input of which is the fused probability vectors of the four patches. At the last stage, the outcomes of each CNN model are accumulated to tender the best option. Three major categorisation methods are proposed in the research work, and the most efficient of them is termed as Multistage Patch based Deep CNN (MPDCNN) which integrates both local patch based as well as the generalised information of the fundus image. Authors in[29] have put forward an idea that consists of using the CNN model for DR diagnosis. The proposed process is an end-to-end one that extracts features from the publicly available images of the diabetic fundus and it utilizes Resnet50 and Inceptionv3. The suggested model IR-CNN accepts the join of feature maps which is produced by both the models to classify DR. For the improvement of the suggested model, several experiments are as follows: For the addition of a substantial quantity of data, the current approach for data augmentation is used and along with a number of images enhancing techniques. The suggested approach is then evaluated via a public database of patients' fundus images.

In[30] the features are extracted with the help of many DL models such as Resnet50, InceptionV3 and VGG19 on publicly accessible fundus images dataset from Kaggle. All these features are combined and passed through the CNN algorithm as a way of categorizing them. DL and AI are used on DR to show how CNN models and Ensemble models such as, VGG19 + Inception V3, and VGG19 + Inception V3 + Resnet 50 could improve medical image interpretation. The Ensemble Model (VGG19 + Inception V3 + Resnet 50) is somewhat beneficial for the diagnosis of DR because merits of the several models to improve robustness and performance. In[31] a modified Capsule Network (CapsNet) is developed for the identification and categorization of DR. Their capsule network comprises a convolution layer, a primary capsule layer, and a class capsule layer for the diagnosis of DR. Here, the class capsule layer measures the probability in regards to a certain class, and the former two layers were used for feature extraction. The resulting CapsNet correctly detects the issue at each of the four phases. The suggested reformed network's effectiveness is confirmed in terms of four performance metrics by taking the Messidor dataset into account. Table 1 provide the summary of the literature with dataset used, technique and its limitation.

The development of AI and DL methods presents encouraging opportunities for automating DR diagnosis. However, low contrast, uneven lighting, and noise are problems that these AI models must deal with because they impair automated diagnosis systems' performance. To consistently detect and categorize DR in retinal images, the research aims to develop an understandable and dependable AI-based framework that can get beyond the challenges brought on by low image quality. To improve the diagnosis of DR, the proposed multi-model deep Net incorporates sophisticated preprocessing approaches, segmentation strategies, and classification algorithms enhanced by metaheuristic algorithms.

Proposed methodology

The proposed multi-model deep net greatly improves the detection of diabetic retinopathy by fusing a reliable classification model with sophisticated image processing methods. The proposed model focuses on enhancing the quality of fundus images by removing noise, correcting uneven illumination, and improving low contrast. It uses techniques like AGF and Contrast-Limited Adaptive Histogram Equalization (CLAHE) to achieve this. The modified U-Net architecture is used for segmentation, incorporating adaptive batch normalization and efficient net layers. Various features are extracted from the segmented images using a combination of TEM, SURF, and LBP, providing a comprehensive representation of retinal images. The classification phase uses an attention layer, a dense block of DenseNet, and an OGRU enhanced by the SANGO algorithm to accurately classify diabetic retinopathy stages. The model's performance is evaluated using metrics such as IoU, accuracy, precision, recall, F1-measure, and DSC. Gradient-based Class Activation Map (Grad-CAM) is used for visual explanations of the model's predictions.

Ref., Year	Dataset	Technique	Limitation
[27], 2024	Kaggle EyePACS	Hybrid model with GoogleNet, APSO, ResNet-16.	High complexity with multiple model combinations; potential overfitting due to the extensive use of classifiers.
[30], 2024	Publicly accessible fundus images dataset from Kaggle	Ensemble Model (VGG19 + InceptionV3 + ResNet50) for feature extraction and classification.	The ensemble approach may increase model complexity and computational overhead.
[25], 2023	Messidor, Kaggle benchmark datasets.	Hybrid adaptive deep learning classifier (HADL-DR), RNN-LSTM, MGAN with semi-maintenance.	High complexity due to multiple stages and semi-maintenance GAN.
[29], 2023	Publicly available fundus images.	End-to-end CNN (IR-CNN) using InceptionV3, ResNet50, data augmentation, image enhancing techniques.	Limited generalizability due to the use of a single public dataset.
[24], 2023	Merged dataset of APTOS, Messidor2, IDRiD	Shallow CNN (RetNet-10) with three convolutional layer blocks, maxpool layers.	Potentially less robust due to shallow network; may lack depth for complex feature extraction.
[26], 2023	Kaggle EyePACS	Hybrid method with GraphNet124, ResNet50.	High complexity due to multiple algorithms and optimizations; may be difficult to replicate.
[31], 2023	Messidor	Modified capsule network (CapsNet) with convolution, primary capsule, and class capsule layers.	Capsule networks may require more data for effective training; potential computational challenges.
[22], 2022	DIARETDB1, APTOS19	Ensemble of ResNeXt and modified DenseNet101.	Computationally intensive; may require high processing power and resources.
[28], 2022	Dataset consists of 2290 fundus images.	MPDCNN using InceptionV3, Xception, ANN-based classifier.	Dataset not specified; potential overfitting due to complex ensemble model.
[23], 2022	DIARETDB0, DIARETDB1	Deep neural network with CTSA, U-Net, hybrid entropy model.	Limited dataset diversity; potential overfitting due to optimization-based approach.

Table 1. An overview of the literature review.

The Chebyshev chaotic map is embedded into the AGF to improve the latter's performance so that it can better cope with noise and contrast problems of retinal images. The U-Net structure is altered by adding adaptive batch normalization and Efficient-Net layers to enhance the segmentation of retinal structures to ensure a more accurate delineation of key features. Multi-fold feature extraction with LBP, SURF, and TEM methods is used in combination to offer a complete representation of retinal images for classification, representing varied textural and structural features. OGRU is implemented using the SANGO algorithm, which improves the model's learning of inherent patterns and enhances classification performance. With this combined method, a potent tool for precise and early DR detection is produced. The block diagram of the proposed multimodal Deep Net DR model is elucidated in the Fig. 1.

Preprocessing

Preprocessing involves removing noise, uneven illumination, and low contrast from pictures of the fundus to improve the quality of the images. The process starts with data augmentation, a method of enhancing the diversity of the training dataset. The step is essential in making the model more robust and capable of generalizing over a broader variety of variations in retinal images. The datasets are first augmented and then the fundus images are now transformed to a grayscale image. RGB data, which has 24 bits, must be converted into an 8-bit grayscale value in order to convert a digital image to a grayscale image. However, the original chrominance qualities, structure, and brightness are all well preserved in the grayscale image. To eliminate noise and uphold crisp edges in the fundus image, picture de-noising is executed via the Gabor filter. CLAHE (Contrast-Limited Adaptive Histogram Equalization), which is a histogram based computer image processing technique for improving the contrast of an image while reducing noise, is then used for contrast enhancement.

Pre-processing is applied to each image before any further steps. Pre-processing includes converting the image to a grayscale format and scaling it to a standard size. Typically, the connection is used to estimate the greyscale image from the R, G, and B variables.

$$G = \lfloor 0.299R + 0.587G + 0.114B \rfloor \tag{1}$$

Adaptive Chaotic Gabor filter

The Gabor filter is a sort of linear filter utilized in image processing applications for texture analysis. This filter basically determines whether a particular frequency is present in an image at a local zone surrounding a place of interest or an analysis region, along a specific orientation. Chaotic map-enhanced AGF is a new method applied in the pre-processing phase to solve noise and contrast problems of retinal images. The chaotic map-enhanced AGF is an integration of the texture analysis potential of the Gabor filter with the non-linear dynamics of the Chebyshev chaotic map. The Gabor filter is highly efficient in picking up particular frequencies in an image at a local scale and, therefore, proves to be effective in the identification of various textures in retinal images that denote various structures or pathologies. The Gabor transformation function is a particularly helpful tool for extracting and evaluating texture information as well as for detecting image edges because of its versatile distinguishing and resolving capabilities in the position and frequency domains. In the spatial domain, the complex form of a two-dimensional Gabor filter is adapted with the chaotic Chebyshev map (d_{t+1}) in the range (-1,1) and can be written as in Eqs. (2),

$$g\left(x, y; \lambda, \theta, \psi, \sigma, \gamma\right) = \exp\left(-\frac{x^{2'} + \gamma^2 y^{2'}}{2\sigma^2}\right) \exp\left(i\left(2\pi \frac{x'}{\lambda} + \psi\right)\right) + d_{t+1} \tag{2}$$

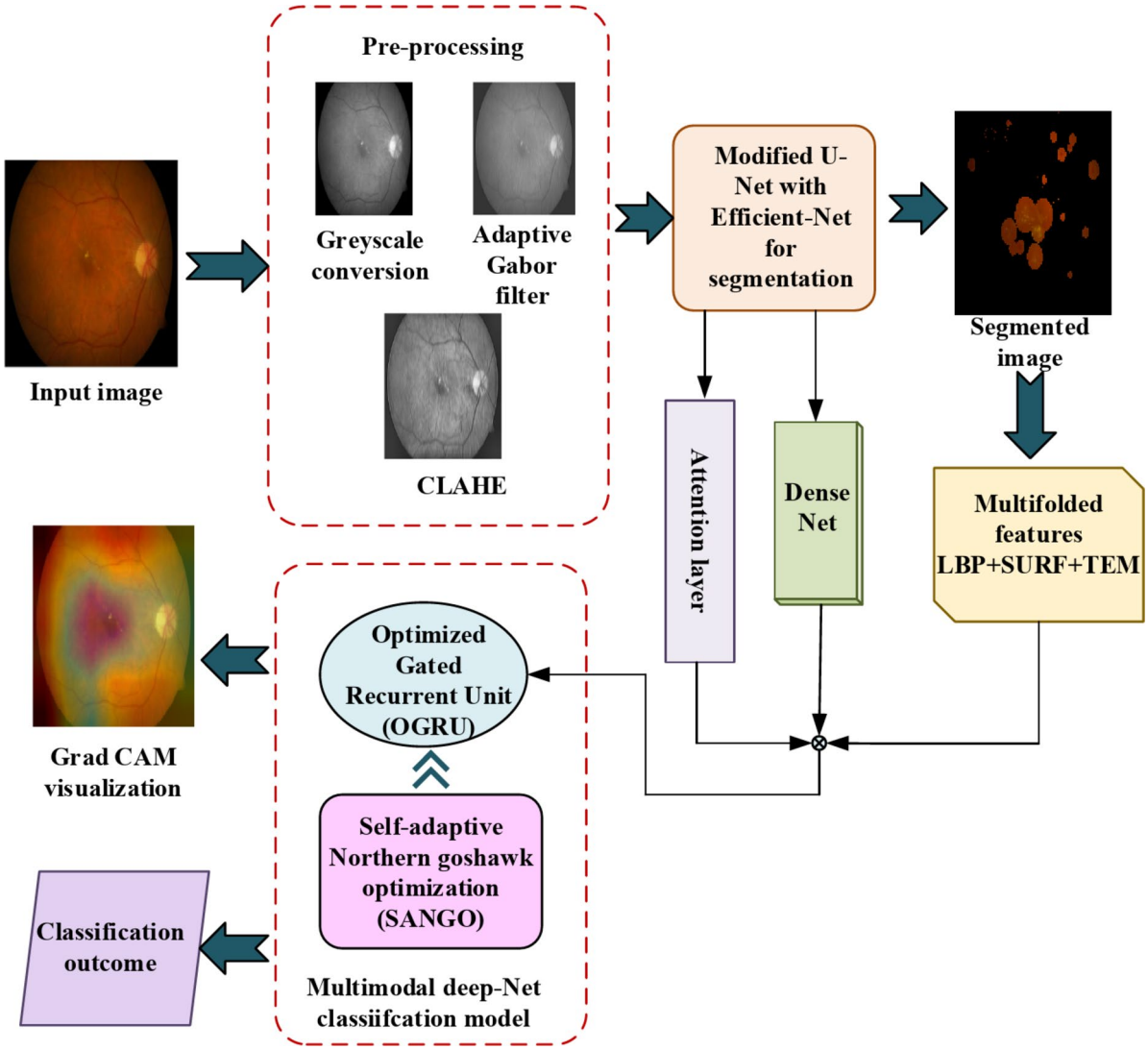

Fig. 1. Block diagram of the proposed DR classification model.

Where

$$x' = xcos\theta + ysin\theta$$
$$y' = -xcos\theta + ysin\theta$$
$$d_{t+1} = \cos\left(0.5cos^{-1}d_t\right) \quad (3)$$

where λ is the sinusoidal component's wavelength, γ is the spatial aspect ratio that represents the elliptical support of the Gabor function, σ is the standard deviation of the Gaussian envelope, ψ is the phase offset, and θ is the orientation of the normal to the parallel stripes of the Gabor function. This technique is unique because it addresses the issues of poor contrast, erratic illumination, and noise in retinal images by fusing the non-linear dynamics of the chaotic map with the texture analysis powers of the Gabor filter. The sample images of conventional and Adaptive Gabor Filter are displayed in the Fig. 2.

Through this integration of these two methods, the filter boosts its performance to maintain key structural information and repress noise, normalizing the contrast to result in higher overall quality for the retinal images. This enables subsequent diagnostic procedures to be more accurate and less time-consuming. Overall, this filter is remarkable for its unique incorporation of chaotic systems with standard filtering methods, producing an efficacious tool for overcoming the complex problems of retinal image preprocessing.

CLAHE contrast enhancement

The CLAHE pipeline consists of five primary steps. The image is first separated into rectangular blocks of identical size, and each block's histogram is adjusted. Creating, cutting, and redistributing histograms are all included in histogram correction. Next, the clipped histogram's cumulative distribution function (or CDF)

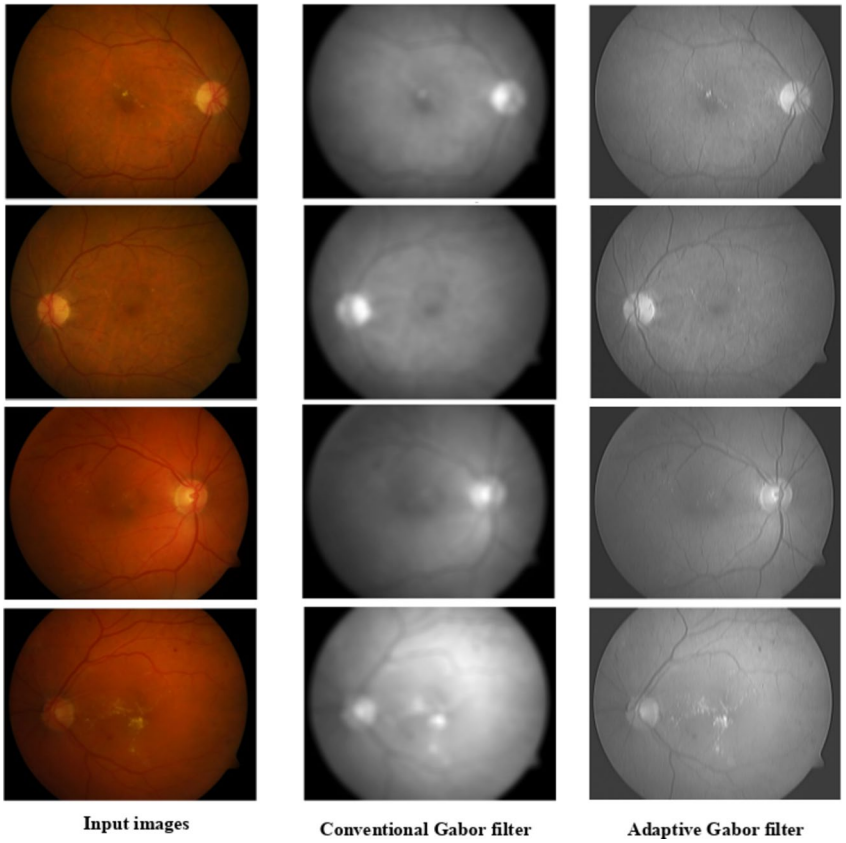

Fig. 2. Sample images of conventional and Gabor filter.

yields the mapping function. Lastly, to eliminate any potential block artifacts, bilinear interpolation is carried out in between the blocks. Unlike the conventional Histogram Equalization (HE), CLAHE limits the contrast by cutting off the peak value in each block's histogram at a clip point. Every gray level receives a new distribution of the clipped pixels. The contrast is amplified to a greater extent with increasing clip point. The clip point is found using formula (4) as follows:

$$\beta = \frac{M}{N}\left(1 + \frac{\alpha}{100} S_{max}\right) \tag{4}$$

where N is the block's dynamic range, M is the number of pixels in each block, α is the clip factor, and S_{max} is the maximum slope. The pixel in this block would be a constant when α is closed to 0. This would result in M/N being the clip point. When α gets closer to 100, there is a noticeable increase in contrast. To modify the contrast enhancement, the clip point is therefore crucial. To effectively adjust the brightness of the image, specify the clip limits threshold parameter in order to clip the histograms. Higher clip limits increase the brightness of the local image; therefore, it is best to set it to the lowest possible value. The transformation functions that are selected alter each histogram. Every histogram is adjusted so as to stay within the selected clip limit. Equation (5) provides a numerical representation of the changed gray levels for the baseline CLAHE technique with uniform distribution. It states that gr_{max} represents the maximum pixel value, gr_{min} the minimum pixel value, $PD\,(f)$ is the CPD and gr is the measured pixel value.

$$gr = [gr_{max} - gr_{min}] * PD\,(f) + gr_{min} \tag{5}$$

The gray level is established by Eq. (6), in order to enhance the exponential distribution.

$$gr = gr_{min} - \left(\frac{1}{\alpha}\right) * \ln\left[1 - PD\,(f)\right] \tag{6}$$

Bilinear interpolation was used to combine adjacent tiles and alter the image's grayscale values in relation to adjusted histograms. This method improves the preprocessing stage by providing better edge detection and texture analysis, which is crucial for subsequent segmentation and classification tasks. The pre-processed sample images of greyscale conversion, AGF and CLAHE contrast enhancement is shown in the Fig. 3.

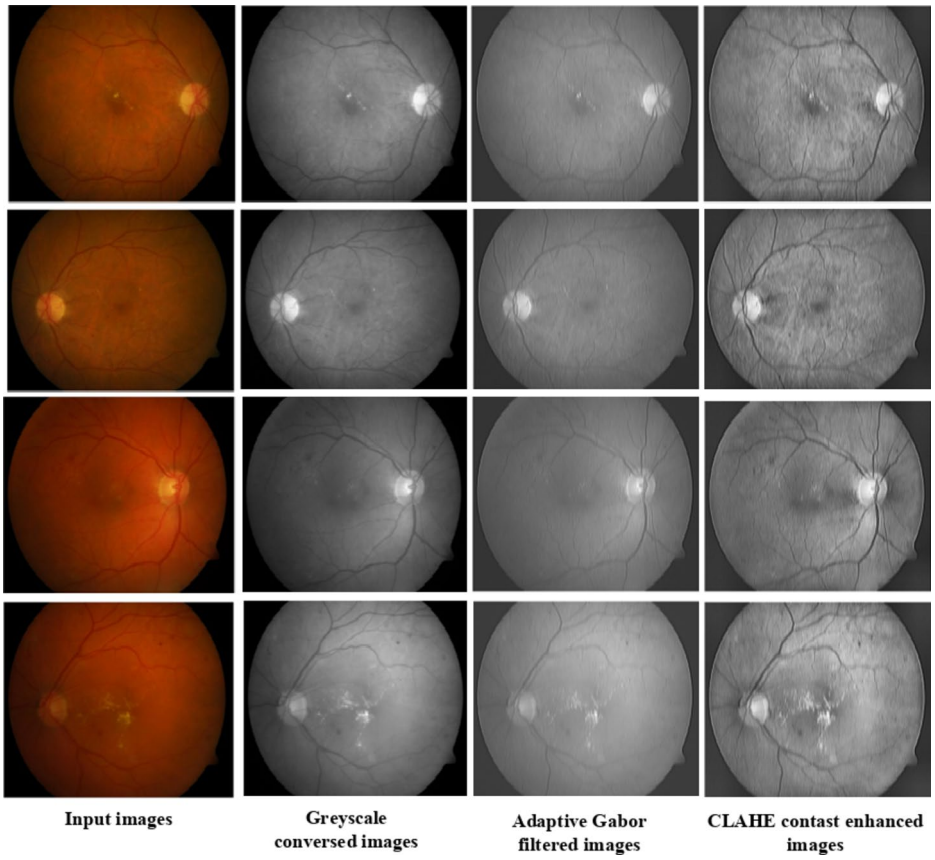

Fig. 3. Sample pre-processed images for DR.

To maintain consistency in the model's input dimensions, all the images from the datasets were resized to a common resolution of 224 × 224 pixels at the preprocessing level. This standardization process helps maintain compatibility and allows for effective processing across the various datasets used in the present study.

The robustness of the model to low-quality input images or variations in preprocessing is mostly due to its adaptive preprocessing pipeline, which incorporates a number of important techniques. AGF improves image quality by filtering noise and enhancing contrast, such that changes in image quality do not have a profound effect on performance. CLAHE further enhances image contrast without compromising vital details, allowing the model to better recognize major retinal features. Data augmentation is also used to synthetically increase the dataset by performing different transformations, enhancing the model's generalization across images of varying qualities. The use of these preprocessing techniques ensures that the model performs well regardless of variations in input images.

Modified U-Net with Efficient-Net for segmentation

The characteristic structure of the U-Net comprises of an expanding path and a contracting path that, when seen, resemble a U-shape. Through a series of pooling and convolutional layers, the contracting route extracts high-level information while gradually down sampling the input image. The method captures more abstract representations of the image content while decreasing the spatial dimensionality. The image is gradually rebuilt to its original spatial resolution while the expanded path up samples the feature maps concurrently. Figure 4 illustrates the U-Net which is modified with the Efficient-Net layers.

The encoder is composed of nine stages: a 3 × 3 mobile reversible bottleneck convolutional (MBConv) structure, max pooling of 2 × 2 and a stride of 2. Four upsampling and a number of convolution processes make up the decoder. The contracting method steadily downsamples the input image while extracting high-level information. These convolutional layers, referred to as MBConv 3 × 3, use 3 × 3 kernels to record local patterns, and then non-linearity is introduced with Leaky ReLU activation functions. The network can capture increasingly abstract representations of image information due to max pooling layers, which are symbolized by downward arrows and minimize the spatial dimensionality of the feature maps.

The elements that make up the MBConv structure are a depthwise convolution, a 1 × 1 convolution for dimension reduction, a dropout layer, and a Squeeze-and-Excitation (SE) module. Adaptive Batch Normalization (BN) and Swish activation operations are performed after the first 1 × 1 convolution and Depthwise Convolution, while adaptive BN processes are only performed after the second 1 × 1 convolution. Because every layer of the network's normalized data must line up with the covariance matrix. When there is a negative standardization

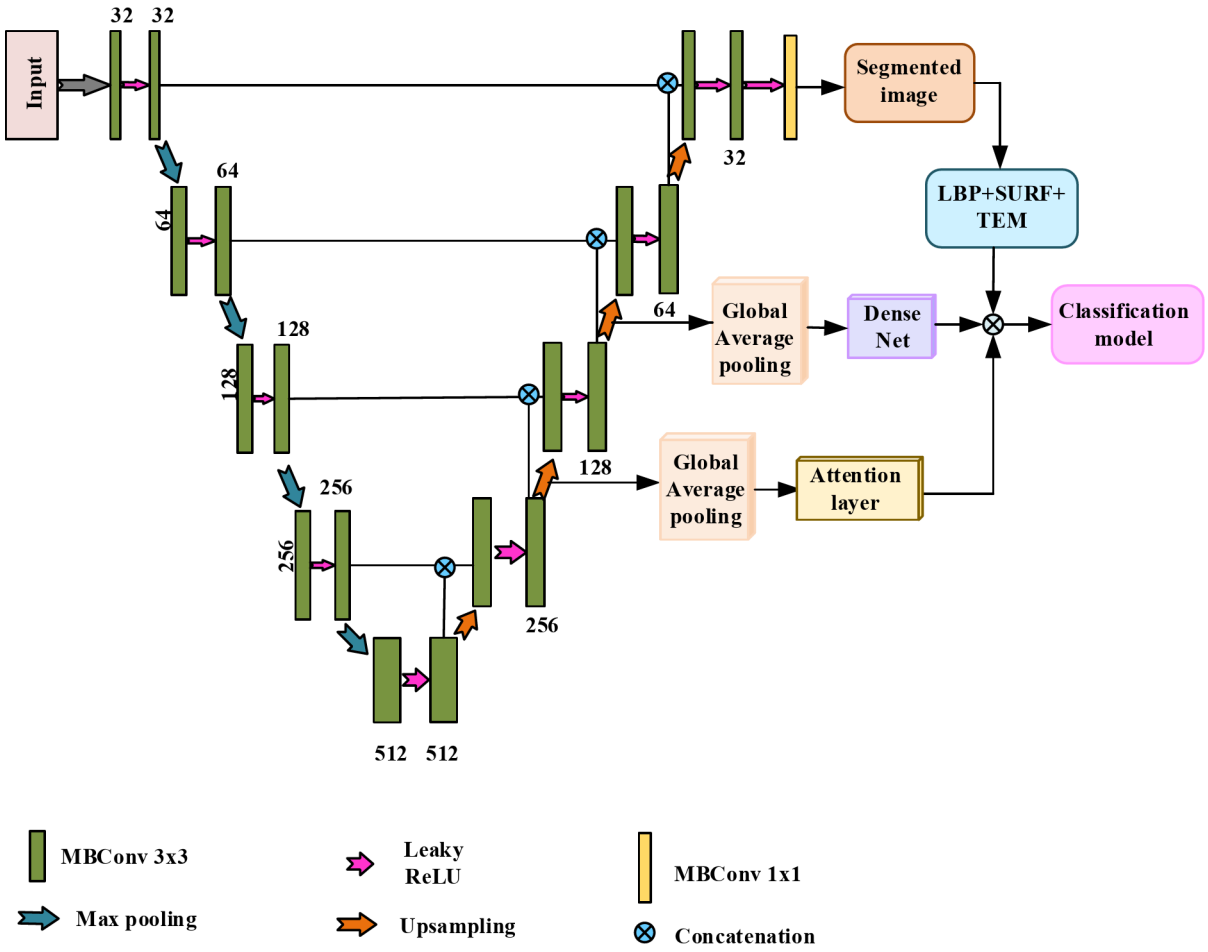

Fig. 4. Modified U-Net with Efficient –Net.

value in this matrix, normalizing is necessary. It identifies similarities and correlation values between the values of the training constraints. The structure of MBconv and SE block is displayed in the Fig. 5.

Creating a singular covariance matrix is necessary since there's a probability the mini-batch size will be less than the number of components in the layer whose activation needs normalizing. Therefore, it is suggested that the activation vectors parameter for normalization be assigned using the variance σ^2. The normalized input data $\widehat{x}_i$

$$\widehat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{7}$$

The σ^2 and μ are calculated for the micro batch during training in order to identify the optimal values for testing. The input variables may only be in the linear part of the function because the normalization process is completed before the activation functions' linearity or nonlinearity. Two more learnable parameters, γ and β , are incorporated to maintain numerical stability and prevent this behaviour. Finally, the calculated value of $\widehat{x}_i$ is scaled and shifted using the γ in the batch normalization process.

$$BN_{\gamma,\beta} = \gamma . \widehat{x}_i + \beta + E^b \tag{8}$$

Where E^b is the embedding features. To produce features with desired geometric qualities, the network is trained, adding another layer of regularization. The network is motivated to increase its generalization abilities, which enhances its performance in segmentation tasks, by applying a contrastive constraint to these embedding features. To regularize the network, embedding features that are subject to the BN which is obtained from taking the Pearson correlative coefficient of the batch. The shared network is trained to produce features with acceptable geometric qualities for optimal generalization behaviour by imposing a contrastive constraint on E^b.A shortcut connection is included in order to fuse more feature data.

The shortcut link is only present when the input MBConv structure's feature matrix and the output feature matrix have the same shape. The accuracy of target identification, picture segmentation, and image classification has significantly increased due to the SE module. The employed SE module consists of two fully connected

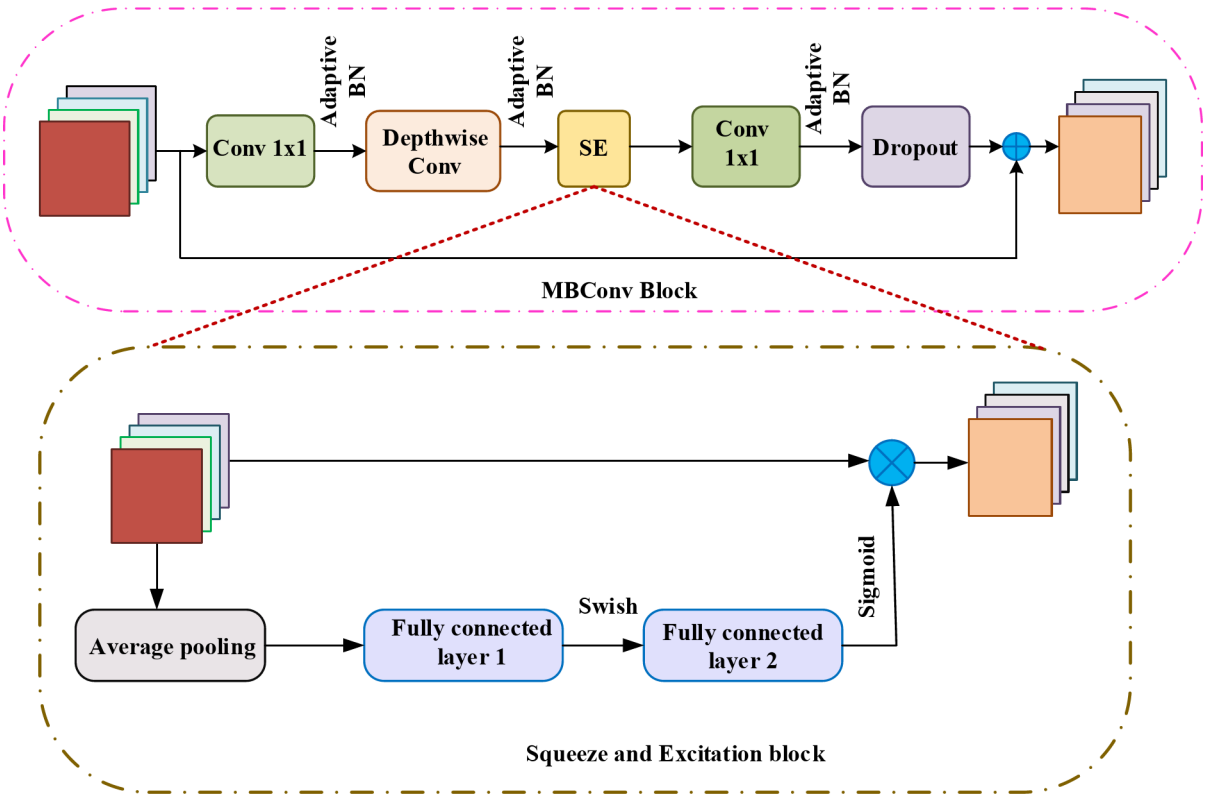

Fig. 5. Structure of MBConv and SE block.

layers, global average pooling, and Sigmoid activation function. Between the two full connection levels, there is also the Swish activation function.

The segmentation map is rebuilt from the encoded characteristics via the expanding path. First, feature maps are upsampled via methods such as bilinear interpolation or transposed convolutions. In the process of every upsampling in the expansion path, feature maps from the contraction path are concatenated by "concatenation", relevant here for endowing the upsampled features with fine detailed information. This makes it possible for the model to generate accurate segmentation results through both global context and fine-grained data.

This integrated output allows the decoder to have access to all levels of features, hence improving the model's precision in locating and segmenting objects. Additional convolutional layers called MBConv 1×1 are applied on the concatenated feature maps to enhance the segmentation map. Finally, this modified U-Net generates the final segmentation map through its output layer. The result of the segmentation module is a segmented image where different parts highlight different retinal structures. To learn complicated patterns for additional categorization, feature maps from the U-Net are also fed into DenseNet and simultaneously to the fully connected layers with Swish activation functions via global average pooling. On intermediate feature maps, global average pooling is employed to reduce each channel to a single number that summarizes the feature's presence throughout the whole image.

The sample segmented images with and without adaptive BN and masked image (ground truth) is displayed in the Fig. 6. The achieved results evidence that BN helps to increase the quality of obtained segmentations, to decrease the effect of overfitting and speed up training. From the Fig. 6, it is observed that when adaptive BN is used the model results in sharper boundaries, less noisy, and leads to higher accurate segmentations as compared to when BN is not used. Adaptive BN offers these advantages via minimizing covariate shift as well as working as a better regularizer hence improving the efficiency of the model for the tasks of image segmentation.

To improve segmentation performance with decreased computational complexity in the enhanced U-Net architecture, a number of enhancements can be made. To begin with, light-weight convolutional modules can be added by substituting normal convolutions with depth wise separable convolutions or group convolutions. This decreases the parameters and computational cost while maintaining performance. Moreover, incorporating Ghost-Net modules can effectively produce redundant feature maps, enhancing feature representation with fewer computations. Adaptive BN improvements can also enhance performance, with choices such as Instance Normalization or Group Normalization to minimize batch size dependency, or Batch Renormalization, which makes training more stable and minimizes mini-batch dependency. For feature fusion, Efficient Net-Lite layers can be incorporated into the encoder to preserve high accuracy while minimizing FLOPs, and employing MobileNetV3-based encoders as feature extractors can further improve efficiency.

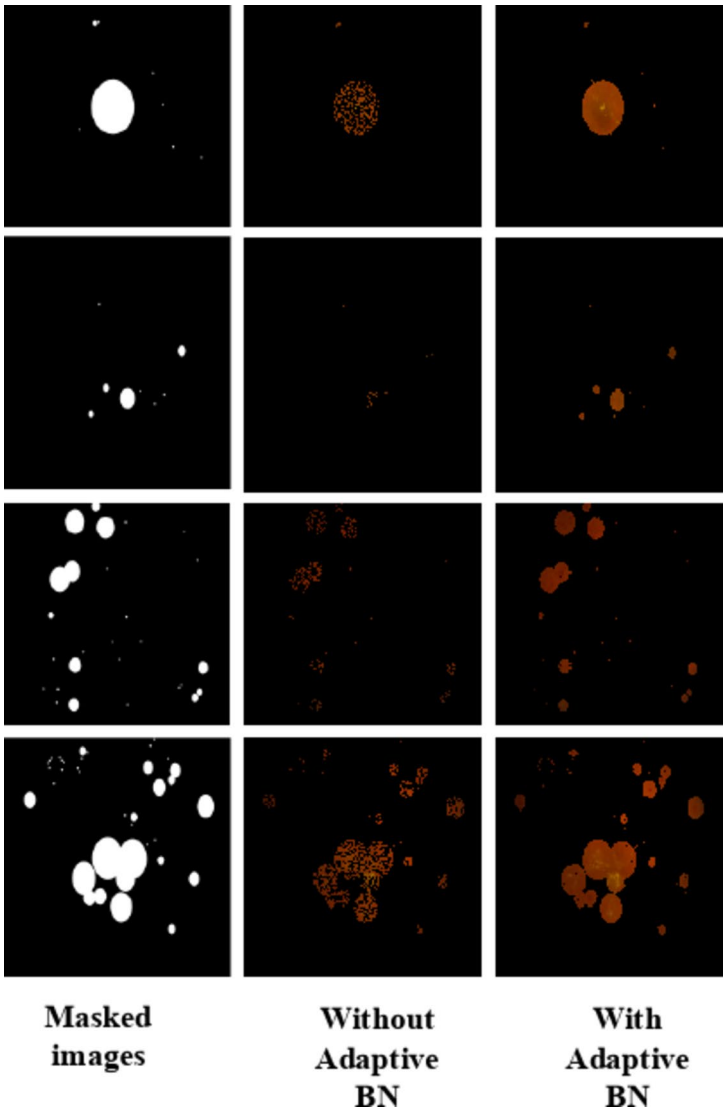

Fig. 6. Sample segmented images.

Multi-folded feature extraction

The segmented image is then processed using Texture Energy Measurement (TEM), Speeded-Up Robust Features (SURF), and Local Binary Patterns (LBP) to improve feature representation. These methods supplement the information extracted in previous phases by capturing a variety of textural and shape qualities.

LBP, SURF, and TEM are very strong feature extraction methods that have been extensively used in image classification processes, such as DR classification. LBP extracts the local texture patterns of an image by comparing the intensity at each pixel with the intensities of its neighbouring pixels. It is computationally efficient and insensitive to changes in illumination levels and hence is especially suitable for DR images, which tend to have different illumination levels. SURF, however, identifies interest points by the determinant of the Hessian matrix, and is robust and efficient to image rotation and scaling. This renders it especially well-suited to identifying DR lesions that can have varying sizes and orientations. TEM estimates the energy of an image across scales and orientations, recording the textural properties of DR lesions, including shape, size, and contrast. Combining LBP, SURF, and TEM features offers a new and synergistic solution for feature extraction, since each method measures different aspects of the image. Fusing these methods results in a stronger and more accurate model for classification, enhancing the overall performance of DR detection and diagnosis.

LBP feature extraction

To extract texture-oriented information, Local Binary Patterns (LBP) is employed. LBP is a crucial method for finding and classifying objects. LBP properties are two bitwise variations from 0 to 1 and 1 to 0, correspondingly. Taking a greyscale picture as input, LBP determines the mean and variance of the intensity of each pixel. The mathematical representation of LBP is as follows in Eq. (9):

$$LBP_{features}(\phi, \mathcal{R}) = \sum_{\phi=0}^{\phi-1} \mathcal{S}(\mathfrak{U}_{\phi} - \mathfrak{U}_{c}) 2^{\phi} \tag{9}$$

$\mathfrak{U}_{\phi}$ represents the variance of the nearby pixel intensity, $\mathcal{R}$ stands for the radius, ϕ represents the number of neighborhood intensities, and $\mathfrak{U}_{c}$ represents the intensity contrast calculated from $(\phi, \mathcal{R})$.

$$\mathcal{S}_{n}(\phi) = \begin{cases} 1 & if\ \phi \geq t \\ 0 & otherwise \end{cases} \tag{10}$$

where the surrounding pixels $\mathcal{S}_{n}(\phi)$ and the central pixel t are contrasted.

SURF

The primary concept of the SURF detection algorithm is to identify keypoints, or points of interest, in a picture at the locations where the Hessian matrix's determinant has a maximum value, such as corners or blob-like shapes[26]. The descriptor creates the feature vectors of the discovered keypoints and defines their features while the detector finds the keypoints. Keypoint matching, keypoint describing, and keypoint detecting are the three steps that make up the SURF algorithm steps. The determinant of the Hessian matrix can be used to identify the keypoints in the first step. Formally, at a point and scale μ, the Hessian matrix $H(a, \mu)$ in a particular image I can be defined as follows in Eq. (11):

$$H(a, \mu) = \begin{bmatrix} L_{aa}(a, \mu) & L_{ab}(a, \mu) \\ L_{ab}(a, \mu) & L_{bb}(a, \mu) \end{bmatrix} \tag{11}$$

Where

$$L_{aa}(a, \mu) = I(a) * \frac{\partial^{2}}{\partial a^{2}} g(\mu);\ L_{ab}(a, \mu) = I(a) * \frac{\partial^{2}}{\partial ab} g(\mu)$$
$$L_{bb}(a, \mu) = I(a) * \frac{\partial^{2}}{\partial b^{2}} g(\mu) \tag{12}$$

where $L_{aa}(a, \mu)$is the image's convolution at scale s with the Gaussian's second derivative, $g(\mu)$. To make Hessian matrix estimations simpler, Gaussian second order derivatives are converted into box filters, and a Gaussian second order derivative approximation is produced. To obtain a good approximation, low contrast key points and points on or near edges are excluded from the determinant by weighting it:

$$\det(H_{approx} = D_{aa} D_{bb} - (\omega\ D_{ab})^{2} \tag{13}$$

where the discrete and approximation kernels for L_{bb} and L_{ab}, respectively, are denoted by D_{bb} and D_{ab}. Although the w term is potentially sensitive to scale, it is nearly certainly constant at 0.9. Scale-varied box filters were used in the development of the image pyramid.

The first filter for SURF was a 9*9 size filter, whose dimensions were determined by,

$$L = 3 * (2^{octave} * interval + 1) \tag{14}$$

wherein Octave's value is set to 1. Gaussian kernel characteristics are attained using Eq. (15) when using integral pictures of a given size to approximate Gaussian kernel filtering.

$$\mu = \mu_{0} * \frac{L}{9} \tag{15}$$

An image pyramid was created using non-maximal suppression in a 3*3*3 neighbourhood. A pixel needs to be compared against additional pixels from scale layers and residual pixels, which are eight pixels from the scale layer to which it belongs. A pixel value is considered to be an interest point when it is superior than the values of the surrounding pixels. Finally, a SURF descriptor ($SURF_{features}$) has been created. It is necessary to determine the dominant orientation of the interest points and use the Haar wavelet transform to estimate the SURF descriptor.

TEM

TEM is a widely used texture descriptor that finds use in various fields, including medical picture analysis. The complete masks are obtained from a five-pixel-long, 1D vector. Level detection is denoted by L5, spot detection by S5, edge detection by E5, ripple detection by R5. Then, for feature extraction, one-dimensional filters with vector length l = 5 are used: L5 (Level) = [1 4 6 4 1], E5 (Edge) = [-1 -2 0 2 1], S5 (Spot) = [-1 0 2 0 1], R5 (Ripple) = [1–4 6 – 4 1]. A column vector multiplied by a row vector of identical length yields a l × l filter. As a result, several rows and columns are used to achieve varied sized filters. The new image that is obtained is called a "energy image" because textural information is extracted from it by convolving these filters with the image. The statistical values (such as entropy, mean, and standard deviation) of the acquired energy images are ultimately used to create feature vectors. Equation 16 gives the TEM for filter in mathematical Equation.

$$TEM_{features}(a,b) = \sum_{j=b-7}^{b+7} \sum_{i=a-7}^{a+7} F_m(i,j) \tag{16}$$

In this case, the m^{th}-filtered picture at pixel $(i,\ j)$, the filtered image sizes, and the energy map sizes are represented, respectively, by $F_m(i,j)$, $(i,\ j)$, and $(a,\ b)$. Ultimately, a feature vector is formed by applying a first-order statistic, or mean, to the energy map, and parameters are produced from each image to feed the classification stage.

The final term for the retrieved overall characteristics is $Ext_{features}$, and it is represented in Eq. (17)

$$Ext_{features} = LBP_{features}\left(\phi\ ,\mathcal{R}\right) + SURF_{features} + TEM_{features}(a,b) \tag{17}$$

These methods yield a comprehensive representation of the retinal pictures by capturing a broad variety of texture and form properties. The model is guaranteed to capture the comprehensive data required for precise classification due to the multi-folded feature extraction approach.

Multimodal deep-Net DR classification model

The collected features from the U-Net architecture's upsampled feature maps and segmented image are processed in parallel. First, global average pooling is applied to these features, which turns them into feature vectors while retaining important spatial information. By channeling one of these feature vectors into a DenseNet block and then parallelly pass-through attention module that comprises two fully connected layers (FC1 and FC2) that have sigmoid and swish activation functions, respectively, in order to capture feature correlations and further reduce dimensionality.

The pooled features are fed into a DenseNet block in one of the parallel routes. The highly connected layers of DenseNet make it possible to reuse data efficiently and improve gradient flow both of which are essential for learning intricate patterns within the features. The network's high level of connectivity makes it possible to record minute details that are essential for precisely categorizing the several stages of disaster recovery. Therefore, by utilizing the rich feature representations in an efficient manner, the DenseNet block is essential in increasing the accuracy of classification. The other path uses a series of two fully linked layers, designated FC1 and FC2, to process the pooled features. These layers help to capture complex correlations between the feature vectors and further reduce their dimensionality. FC1 makes use of the non-linear Swish activation function, which has been demonstrated to enhance model performance by preserving smooth gradients throughout training. After that, FC2 uses a Sigmoid activation function to change the feature representations by producing probabilities between 0 and 1. The architecture of multimodal deep DR classification model is illustrated in the Fig. 7.

DenseNet output and FC2 output, the processed features from both parallel routes, are combined with the features extracted from the segmented images. This integration combines the strengths of both processing pathways, potentially leading to a more comprehensive feature representation for classification. Following feature concatenation, the optimized Gated Recurrent Unit (GRU), a recurrent neural network type that performs exceptionally well with sequential input, receives the features. A variation of RNN is called GRU. The issue that RNN struggles with long-distance information acquisition is resolved by the introduction of gating structures. GRU is more straightforward than LSTM, requiring only the introduction of the update gate (up_t) and reset gate ($reset_t$). In GRU, the reset gate controls how much of the historical data to forget, while the update (or input) gate determines how much input (in_t) and previous output (hid_{t-1}) to transfer to the next cell. The content of the present memory makes sure that, depending on the weight W, only pertinent data needs to be passed on to the following iteration. The following formulas control the primary functions of GRU.

Update gate

$$up_t = \sigma\ \left(W_{up} * [hid_{t-1}, in_t]\right) \tag{18}$$

Reset gate

$$reset_t = \sigma\ \left(W_{reset} * [hid_{t-1}, in_t]\right) \tag{19}$$

The input at time t is represented by in_t, the hidden state at the previous time by hid_{t-1}, the weight matrix by W_{up} and W_{reset}, the sigmoid function is denoted by σ .

$$\tilde{hid}_t = tanh\ \left(W_{\underset{hid}{\sim}} * [reset_t * hid_{t-1}, in_t]\right) \tag{20}$$

where the weight matrix is represented by $W_{\underset{hid}{\sim}}$ and the tanh activation function is denoted by $tanh(\cdot)$.The tanh activation function generates vectors with all possible values in accordance with the new input after it receives the updated state information through the update gate. Following gate resetting and updating, the GRU unit's candidate status value is $\tilde{hid}_t$ and the final output status value is hid_t:

$$hid_t = (1 - up_t) * hid_{t-1} + up_t * \tilde{hid}_t \tag{21}$$

$$o_t = \sigma\ \left(W_o * hid_t\right) \tag{22}$$

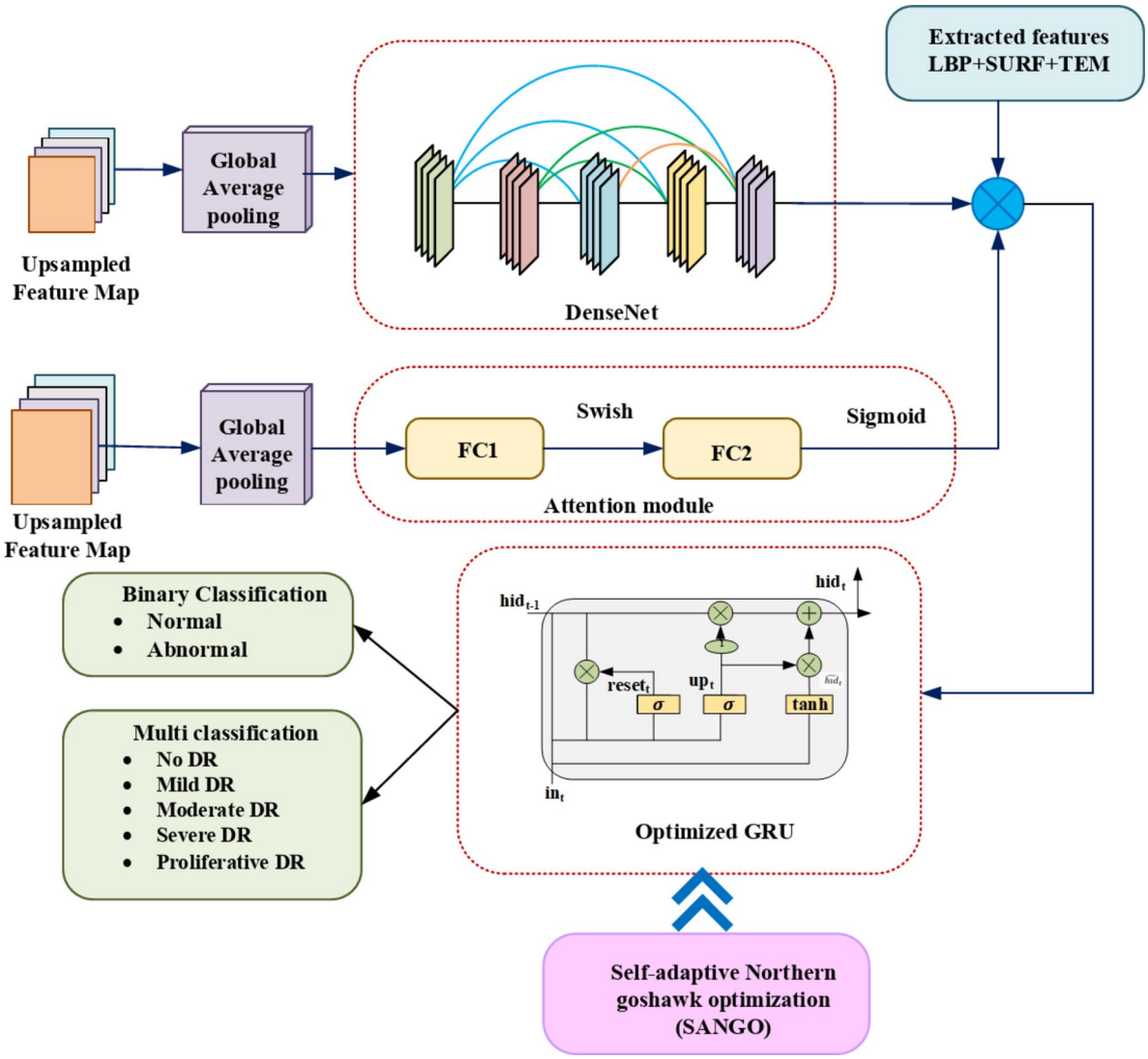

Fig. 7. Proposed Multimodal deep DR classification model.

The calculation formula above states that GRU uses two gates to store and filter data, gate functions to preserve key features, and learning to identify dependencies in order to provide the optimal output value. To optimize the structure of the neuron, the update gate is modified using the reset gate's output. To optimize the weight of the GRU, the proposed SANGO optimization is employed. In the last classification layer, a SoftMax activation function is utilized to the output of the optimized GRU. The model can now accurately and consistently classify the diabetic retinopathy stage through this step, which transforms the GRU's output into the probability distribution among the several categories. The cross entropy (CE) loss function is reshaped to introduce the focal loss, which downplays simple examples and concentrates training on hard negatives. Focal loss reduces the weightage of easy-to-classify examples and gives higher weightage to hard, misclassified examples, an approach tailored specifically to handle the problem of class imbalance. This modification guarantees that the model puts more concentration on learning from minority class instances, thereby enhancing its capability to differentiate between varying phases of diabetic retinopathy, especially in underrepresented instances. The following Eq. (23) is the focal loss formula:

$$Loss_{fl} = \begin{cases} -\omega\,(1-\widehat{y})^{\gamma}\, log\widehat{y}, & y=1 \\ -(1-\omega)\;\widehat{y}^{\gamma}\, log\,(1-\widehat{y}) & y=0 \end{cases} \tag{23}$$

The CE loss function was modified to include a tuneable focusing parameter, $\gamma \geq 0$, which can reduce the loss of samples that are easily classified and concentrate more of the classification process on difficult and incorrectly classified samples. There is also an addition of a weighting factor $\omega \in [0,1]$ to correct for the imbalanced proportion of both negative and positive samples. The proposed model combines feature integration, parallel processing, and efficient sequential analysis in a complex way to obtain improved performance in the difficult task of classifying diabetic retinopathy.

Self-adaptive Northern goshawk optimization

The northern goshawk and other members of the Accipiter genus pursue a variety of prey, such as birds, mammals, and larger animals like foxes and raccoons[32]. The northern goshawk hunts in two stages, which are as follows: In the first, it swiftly approaches its victim after detecting it, and in the second, it goes on a fast tail-chase hunt. The Northern GOshawk (NGO) algorithm's efficient prey hunting and catching procedure serves as the basis for its search mechanism. Prey identification, prey capture, and population initialization make up the three stages of the method.

(i) **Initialization**.

First, matrix NG presents the initialization population of the NGO as in Eq. (24).

$$NG = \begin{bmatrix} NG_1 \\ NG_2 \\ \vdots \\ NG_N \end{bmatrix} = \begin{bmatrix} NG_{1,1} & NG_{1,2} & \cdots & NG_{1,M} \\ NG_{2,1} & NG_{2,2} & \cdots & NG_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ NG_{N,1} & NG_{N,2} & \cdots & NG_{N,M} \end{bmatrix} \tag{24}$$

where the NG_i, $1 \leq i \leq N$, denotes the i^{th} individual in the whole population. N and M stand for the size of the population and the dimension of the objective function, correspondingly. With upper bound U_{bound} and lower bound L_{bound}, the components of NG_i, can be calculated for a single objective optimization problem by,

$$NG_{i,j} = L_{bound} + rand\,(U_{bound} - L_{bound}),\ 1 \leq i \leq N; 1 \leq j \leq M \tag{25}$$

(ii) **Prey Identification**.

Initially, the northern goshawk would select its target and try to attack it. The random selection of the prey may suggest that this behaviour represents the algorithm's ability to explore the whole feasible space globally. In the event that the $prey_i$ is the target selected by the individual NG_i, as implied by Eq. (26), Eq. (27) represents the northern goshawk striking its victim.

$$prey_i = NG_p,\ i = 1,2,\ldots,N; p = 1,2,\ldots.,i-1,i+1,\ldots.,N \tag{26}$$

$$NG_{i,j}^{new1} = \begin{cases} NG_i + r\,(prey_i - ING_i), & F\,(prey_i) < F\,(NG_i) \\ NG_i + r\,(NG_i - prey_i), & F\,(prey_i) \geq F\,(NG_i) \end{cases} \tag{27}$$

where I is a vector made up of 1 or 2 and r is a random vector with numbers in the range of [0, 1]. The algorithm's unpredictability is increased with the help of r and I in order to thoroughly search the space. After that, Eq. (28) will update each individual NG_i.

$$NG_i = \begin{cases} NG_i^{new1} & F\left(NG_i^{new1}\right) < F\,(NG_i) \\ NG_i & F\left(NG_i^{new1}\right) \geq F\,(NG_i) \end{cases} \tag{28}$$

(iii) **Prey Capture based on dynamic factor**.

The northern goshawk will latch onto its prey and launch an attack that will startle it and cause it to start to run. This is the time for the northern goshawk to keep chasing its prey. The northern goshawk's quick pursuit speed allows it to hunt and ultimately capture prey in almost any situation. A new Dynamic Factor (DF) on the original basis to include disturbance factors and enhance the algorithm's random walkability during the exploration stage; provide the population with the ability to explore local regions gradually; lower the likelihood of individuals being affected by fluctuations and falling into the local extremum; and increase the algorithm's optimization accuracy. Equation (30) is used to determine the new DF. Equation (29) can be used to replicate this stage when the chasing behaviour is within a circle of radius R.

$$NG_i^{new2} = NG_i + R\,(2r-1)\,NG_i * DF \tag{29}$$

$$DF = 0.4 * (2 * rand - 1) * e^{\left(-\frac{t}{T}\right)^2} \tag{30}$$

where $R = 0.02(1 - \mathrm{t/T})$. T is the maximum number of iterations, and t is the current iteration. A random number between 0 and 1 is called a rand and t is the number of iterations that are currently being performed. Next, Eq. (31) updates each individual NG_i.

$$NG_i = \begin{cases} NG_i^{new2}, & F\left(NG_i^{new1}\right) < F\,(NG_i) \\ NG_i, & F\left(NG_i^{new1}\right) \geq F\,(NG_i) \end{cases} \tag{31}$$

The new status for the i^{th} solution is denoted by NG_i^{new2}, and the objective function value is $F\left(NG_i^{new1}\right)$ which is derived from the second phase of the NGO. Upon completion of an iteration of the NGO algorithm, which involves updating all population members based on the first and second stages, the optimal proposed solution, the goal function, and the new population member values are identified. After that, the algorithm moves on to the following iteration, updating the population's members based on Eq. (3) through (8) until the

program reaches its final iteration. The "prey capture" phase, especially through the addition of the dynamic factor (DF), enhances the algorithm's local search so that the chosen features are fine-tuned more accurately. The integration of global and local search abilities of the algorithm maintains a balance between exploration and exploitation, preventing the algorithm from converging to a local optimum solution and instead, converging towards the global optimum solution. Due to this, the SANGO algorithm converges faster compared to conventional optimization approaches, enhancing efficiency.

Grad-CAM visualization

The gradient-based class activation map (Grad-CAM), a class-discriminative localization map that highlights relevant areas of the image, determines the gradient of the class score for every category of the map of feature activations of a convolutional layer. DR and normal images are generated using the Grad-CAM visualization with different emphasizing schemes as shown in Fig. 8. While a normal eye's class activation map emphasizes the entire image with a focus on the middle region, DR images highlight the upper portion of the image more densely. The portion of the image that is highlighted in the class activation map is the significant area that the model uses to forecast the concept.

Result and Discussion

The proposed work is done using the PYTHON tool on a Windows 10 computer equipped with a 64-bit Intel(R) Core i5 CPU. The proposed model was implemented for three datasets DiaRetDB1, APTOS 2019 and Kaggle EyePacs dataset. The performance of the three datasets are evaluated and contrasted with the existing techniques like CNN, DNN, RNN, GRU and Proposed model and also for the five fold classification. The performance are evaluated for the metrics like accuracy, precision, sensitivity, specificity, False Positive Ratio (FPR), False Negative Ratio (FNR), Matthew Correlation Coefficient (MCC), F1-score, Negative Predictive Value (NPV), Intersection over Union (IoU) and Dice Similarity Coefficient (DSC).

Dataset description

Three distinct benchmark datasets are used to verify the proposed approach.

(i) **Dataset 1: DiaRetDB1 dataset**.

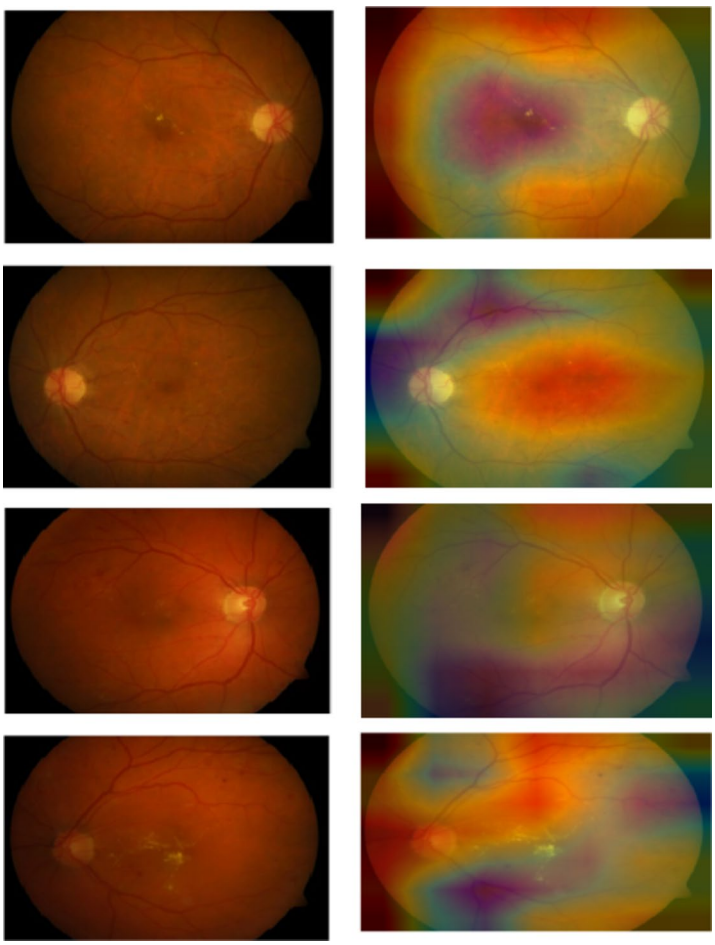

Fig. 8. Grad-CAM visualization.

Datasets	Training samples	Testing samples
Dataset 1	705	303
Dataset 2	5019	2151
Dataset 3	15,126	6484

Table 2. Training and testing sample of three datasets.

Hyperparameter	Optimal Value	Effect on Model	Sensitivity
Population Size	10	Balanced performance with reasonable computational cost.	Moderate
Dimensionality	2	Adequate for GRU model optimization (2 parameters for GRU layers).	Low
LB	(16, 16)	Restricting the lower bounds limited the model's flexibility in the search space.	High
UB	(128, 128)	Larger upper bound allowed the algorithm to explore deeper architectures.	High
Maximum Iterations	100	More iterations increased convergence towards the best solution.	Moderate
Learning Rate	0.02	Controlled convergence speed. A higher rate caused instability.	Moderate
Prey Capture Dynamic Factor (DF)	0.4	A higher value allowed for more dramatic updates to individual solutions.	Moderate
Prey Identification Factor (R)	0.02	Helped balance the exploration and exploitation phases.	Low

Table 3. Hyperparameters of SANGO optimization.

The retina fundus taken at the 50° field of view is included in the DIARETDB1 dataset[33]. The collection comprises five normal photos and eighty-four DR images, all with a resolution of 1500×1152 pixels. Four medical professionals annotate the dataset images for the presence of HEM, MAs, hard exudates, and soft exudates. The data includes twenty high-resolution fundus (HRF) images each of which is divided into two groups: healthy and DR. The database contains 89 colour fundus images, of which 84 show at least mild non-proliferative signs of DR. Five, on the other hand, are deemed normal by all experts involved in the evaluation and do not show any signs of the disease.

(ii) **Dataset 2: APTOS 2019 dataset**.

The APTOS 2019 diabetic retinopathy dataset[19] is where the data were gathered. The Kaggle dataset is accessible. The dataset consists of a sizable collection of retinal photos obtained in various imaging scenarios utilizing fundus imaging. Nearly 5584 high-resolution fundus photos were chosen from the Kaggle dataset of 5597 images captured by various models to include in the "APTOS2019" dataset, which was used for testing. Out of all the datasets, 640×480 is the smallest native size. Based on the clinician's assessment, a number is assigned to each image in this study corresponding to the severity of DR, ranging from 0 to 4. Professionals supply the labels, which indicate "No DR," "Mild DR," "Moderate DR," "Severe DR," and "PDR" on a scale of 0, 1, 2, 3, and 4, respectively.

(iii) **Dataset 3: Kaggle EyePACS**.

The study made use of the EyePACS dataset[34], which is freely accessible via Kaggle.com. Fundus pictures of the retina, taken and labeled by ophthalmologists, are included in this dataset. A total of 35,126 retinal pictures, classified into five severity levels of DR, are included in this dataset. The photographs in the dataset are from various camera models and types, which may have an impact on how left and right seem visually. Certain images are displayed in the anatomical position of the retina (macula on the left, optic nerve on the right for the right eye). Fundus images are named as follows: mild NPDR, moderate NPDR, severe NPDR, PDR, and healthy or normal image. These grades (0, 1, 2, 3, 4) are based on the severity of DR. The three datasets are initially augmented and the training and testing samples for three datasets are mentioned in the Tables 2 and 3 depicts the hyperparameters of the proposed SANGO algorithm.

The robustness of the model has been tested on three datasets: DiaRetDB1, APTOS 2019, and EyePacs. The datasets probably have differences in image quality, population, and disease severity, and thus provide some diversity for the purpose of testing robustness. Also, using 5-fold cross-validation is a sign of trying to check the performance of the model on different subsets of the data, which can be interpreted as simulating testing on unseen data. fundus images. The average time required for the model to process a single retinal image from preprocessing to classification is 0.8 s.

Performance metrics

The proposed approach is assessed using the following evaluation metrics:

(i) **Accuracy**: The ratio of samples that were successfully identified to all samples is used to determine accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{32}$$

(ii) **Precision**: The percentage of samples that were correctly identified as positive out of all samples that were predicted to be positive is known as precision.

$$Precision = \frac{TP}{TP + FP} \tag{33}$$

(iii) **Recall**: Recall is a classification problem evaluation metric that measures a model's accuracy in correctly identifying all relevant instances from the dataset.

$$Recall = \frac{TP}{TP + FN} \tag{34}$$

(iv) **F-Measure**: It is the harmonic mean of recall and precision.

$$F - Measure = \frac{2 * precision * recall}{Precision + recall} \tag{35}$$

(v) **Sensitivity**: The percentage of actual positive samples that were correctly classified as positive is measured by sensitivity.

$$Sensitivity = \frac{TP}{TP + FN} \tag{36}$$

(vi) **Specificity**: Specificity is determined by how many actual negative samples are correctly categorised as negative.

$$Specificity = \frac{TN}{TN + FP} \tag{37}$$

(vii) **NPV**: NPV calculates the percentage of real negative samples that were correctly identified as being negative out of all samples that were expected to be negative.

$$NPV = \frac{TN}{TN + FN} \tag{38}$$

(viii) **MCC**: MCC spans from − 1 to + 1 and integrates data regarding true and false positives and negatives into a single value, where + 1 denotes a perfect classification, 0 denotes random categorization, and − 1 denotes the full discrepancy between prediction and observation.

$$MCC = \frac{((TP * TN) - (FP * FN))}{\sqrt{((TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{39}$$

(ix) **IoU**: The IOU value indicates how comparable the ground scene region of the objects in the set of images is to the forecast region.

$$IoU = \frac{TP}{FP + TP + FN} \tag{40}$$

(x) **DSC**: A statistical measure for comparing the similarity of two samples is called the DSC. The DSC, which is defined as the measurement of the spatial overlap between two segmentations, A and B target regions,

$$DSC = \frac{2\left|A \bigcap B\right|}{A + B} \tag{41}$$

where the number of positively classified samples that were mistakenly categorized is called FN (False Negatives), the number of negatively classified samples that were wrongly classified is called FP (False Positives), the number of positively categorized samples that were correctly categorized is called TP (True Positives), and the number of negatively classified samples that were correctly categorized is called TN (True Negatives).

Performance analysis and comparison

The performance analysis for the proposed DR classification model (OGRU-SANGO) are analysed and compared with the existing techniques like CNN, DNN, RNN and GRU for the above-mentioned metrics for the three datasets. Figures 9, 10 and 11 illustrates the graphical performance analysis comparison for various metrics.

The comparison of sensitivity, specificity, accuracy, and precision is displayed in Fig. 9. All the results were obtained on test set. The proposed model obtains an accuracy of 99.01%, Specificity of 0.9946, Sensitivity of 0.9854, and precision of 99.12% for Dataset 1, while CNN and DNN come in second and third, with corresponding accuracies of 87.23% and 90.23%. With 98.99% accuracy, 0.9957 of specificity, 0.9846 of sensitivity, and 99.10% of precision in Dataset 2, the proposed model outperforms CNN with 92.65% accuracy and DNN with 93.46% accuracy. Similarly, with an accuracy of 99.12%, specificity of 0.9965, sensitivity of 0.9865, and precision of 99.23% in Dataset 3, the proposed model outperforms CNN and DNN, which have accuracy of 89.23% and

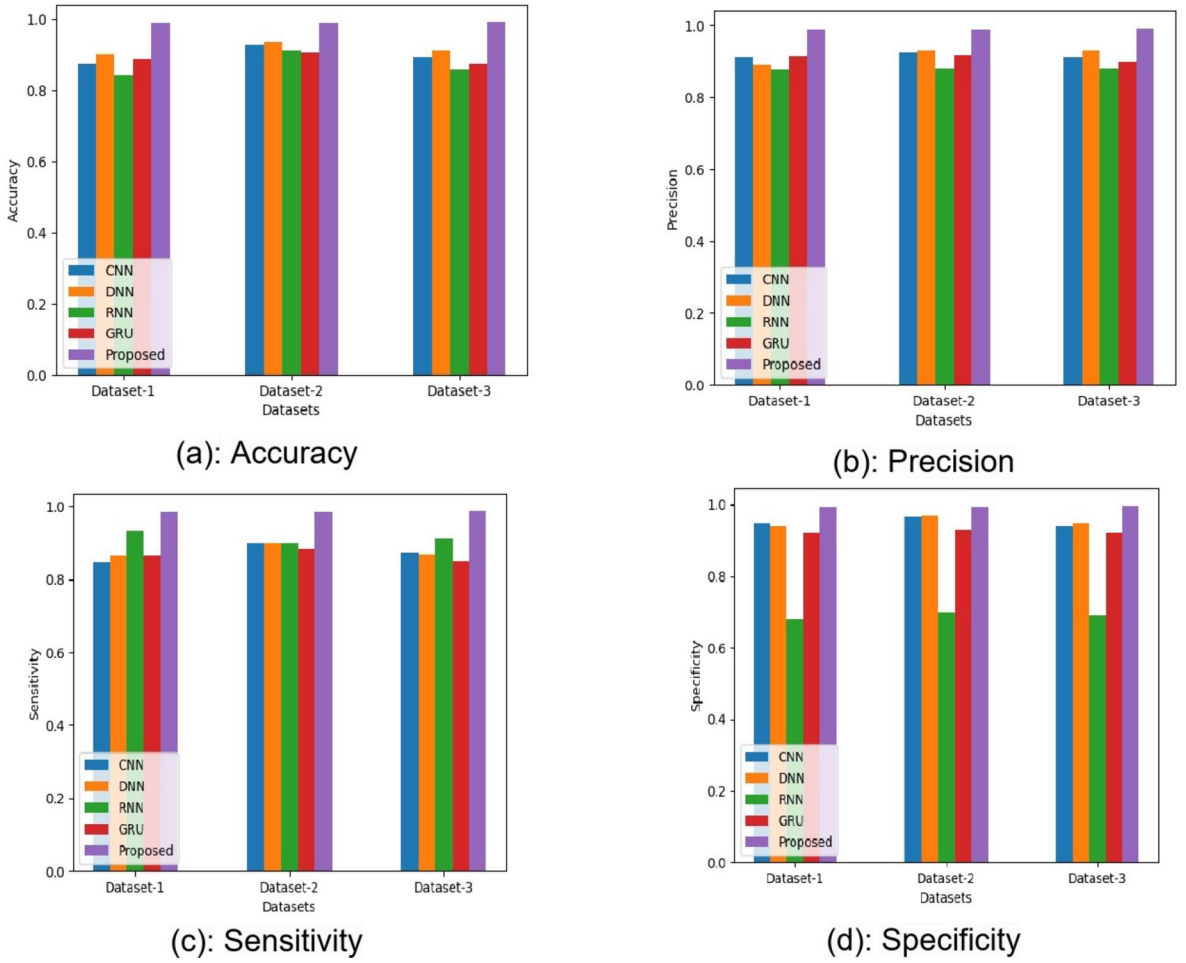

Fig. 9. Performance comparison for (**a**): Accuracy; (**b**): Precision; (**c**): Sensitivity and (**d**): Specificity.

91.10%, respectively. Although the RNN has decent sensitivity of 0.9332 in Dataset 1, it shows notably poorer specificity, especially in Dataset 1 of 0.6799 and in Dataset 2 of 0.6999, suggesting a larger rate of false positives. Although GRU performs better overall, it still lags behind the proposed OGRU-SANGO method, demonstrating the latter's higher efficacy in all datasets and metrics.

The comparison of the F1 score, MCC, and NPV is displayed in Fig. 10. Across all datasets, the proposed model consistently outperforms all other classifiers, obtaining the greatest F1 Score, NPV, and MCC. In Dataset 1, it achieves an F1 Score of 0.9873, an NPV of 0.9812, and an MCC of 0.9795. RNN, on the other hand, performs worse, especially in MCC, as evidenced by values like 0.4965 in Dataset 1 and 0.5257 in Dataset 2, which indicate a less successful prediction balance. Across all datasets, CNN and DNN perform similarly, with DNN typically outperforming CNN. This is shown in Dataset 2, where DNN has an F1 Score of 0.9139 and an MCC of 0.8010. The proposed model performs exceptionally well, outperforming the other models in NPV and MCC, highlighting its resilience and dependability in class prediction for both positive and negative outcomes.

Figure 11 shows the comparison of FPR and FNR. Across all datasets, the proposed model consistently obtains the lowest FNR and FPR, with Dataset 1 showing FNR of 0.0146 and FPR of 0.0054, Dataset 2 showing FNR of 0.0154, FPR of 0.0043, and Dataset 3 exhibiting FNR of 0.0135 and FPR of 0.0035. Though RNN's FNR of 0.0665 in Dataset 1 is lower, it shows high FPR of 0.1201 and 0.1110 in Datasets 1 and 3. While CNN and DNN both perform moderately well, DNN often outperforms CNN in FPR across all datasets, with FPR of 0.0299 in Dataset 2. Since it minimizes both FNR and FPR, the proposed model performs better, proving that it can reliably distinguish between positive and negative cases in all datasets.

The performance of the proposed model on the DiaRetDB1, APTOS 2019, and EyePACS datasets is measured in terms of accuracy, precision, sensitivity, specificity, F1-score, MCC, NPV, FPR, and FNR. On the DiaRetDB1 dataset, the model performs with an accuracy of 99.01%, sensitivity of 0.9854, specificity of 0.9946, and precision of 99.12%. On the APTOS 2019 dataset, the model provides an accuracy of 98.99%, sensitivity of 0.9846, specificity of 0.9957, and precision of 99.10%. For the EyePACS dataset, the model provides an accuracy of 99.12%, sensitivity of 0.9865, specificity of 0.9965, and precision of 99.23%. This consistency implies a high generalizability to new, unseen data in the same domain of diabetic retinopathy detection. Nevertheless, the generalizability of the model needs to be confirmed with additional diverse datasets and real-world clinical environments to determine its robustness and reliability for wider applications in the future.

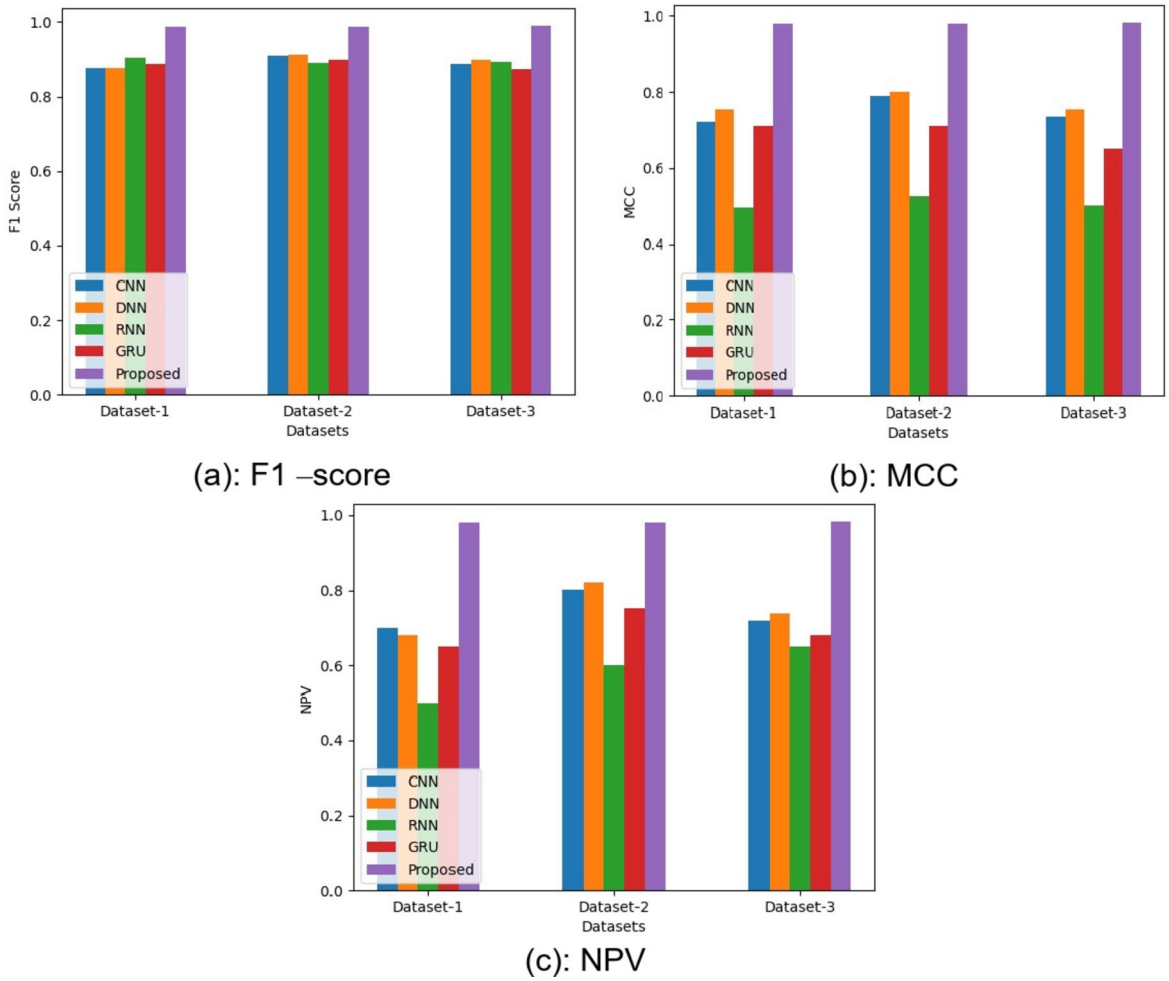

Fig. 10. Performance comparison for (**a**): F1 -score; (**b**): MCC and (**c**): NPV.

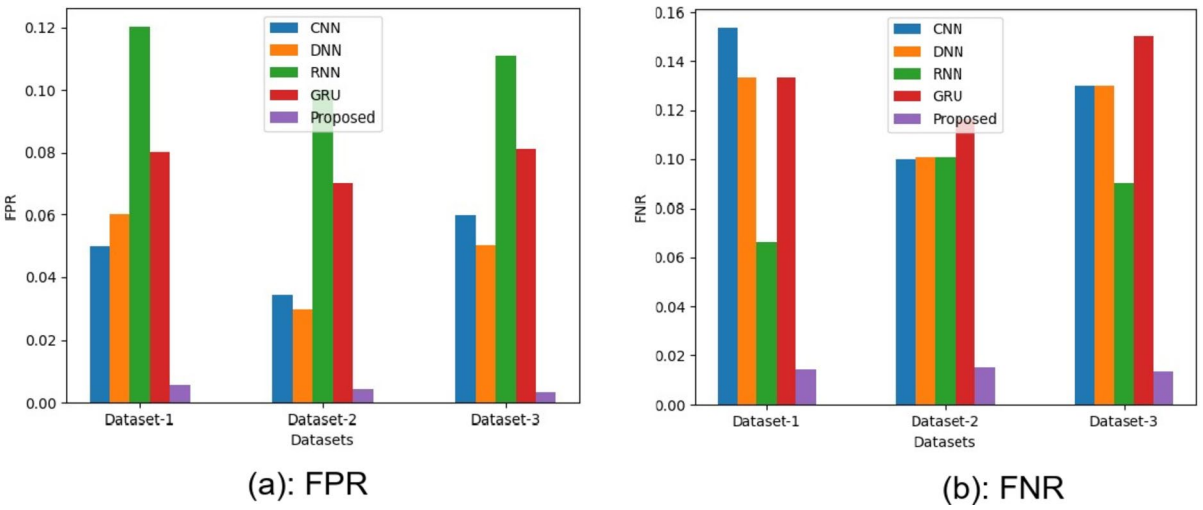

Fig. 11. Performance comparison for (**a**): FPR and (**b**): FNR.

Dataset	Method	Accuracy	Precision	FNR	FPR
Dataset 1	Without LBP, SURF, and TEM	0.8554	0.8801	0.1899	0.0699
	Proposed	0.9801	0.9823	0.0346	0.0112
Dataset 2	Without LBP, SURF, and TEM	0.9110	0.9110	0.1299	0.0446
	Proposed	0.9799	0.9812	0.0299	0.0099
Dataset 3	Without LBP, SURF, and TEM	0.8710	0.8810	0.1410	0.0543
	Proposed	0.9910	0.9923	0.0146	0.0043

Table 4. Comparison of datasets with and without features.

Technique Used	IoU	DSC
Segmentation without BN	0.2915	0.4359
Segmentation with BN	0.8193	0.9006
Segmentation with adaptive BN	0.8272	0.9054

Table 5. Comparison of IoU and DSC for segmentation.

Performance analysis on multi model features and segmentation

The proposed model and a baseline model without LBP, SURF, and TEM characteristics are compared for accuracy, precision, FNR, and FPR for the identification of DR across three datasets are analysed and contrasted in the Table 4.

The proposed method's accuracy for Dataset 1 is 98.01%, a noteworthy gain of 12.47% points above the baseline method's 85.54% accuracy. Additionally, precision increases to 98.23%, up 10.22% from 88.01% accuracy of baseline method. Furthermore, there is a decrease of 15.53% in the FNR and a decrease of 5.87% in the FPR from 6.99 to 1.12%. The accuracy obtained by the proposed strategy in Dataset 2 is 97.99%, which is greater by a margin of 6.89% compared to 91.10% attained by the method lacking LBP, SURF, and TEM. From 91.10 to 98.12%, there is a 7.02% improvement in precision. The FPR falls from 4.46 to 0.99%, a loss of 3.47%, and the FNR falls from 12.99 to 2.99%, representing a 10% decline. The accuracy of the proposed approach for Dataset 3 is 99.10%, a 12% improvement above the baseline method's accuracy of 87.10%. Also, precision increases dramatically, rising from 88.10 to 99.23%, an increase of 11.13%. The FPR falls from 5.43 to 0.43%, a drop of 5%, and the FNR falls from 14.10 to 1.46%, a drop of 12.64%.

The results show a significant improvement in performance with the proposed model, achieving an accuracy of 98.01% on Dataset 1, compared to 85.54% for the baseline model. Precision also improves from 88.01 to 98.23%. Additionally, the FNR and FPR are significantly reduced, indicating better classification performance. However, there are trade-offs associated with this improvement. While the inclusion of LBP, SURF, and TEM features enhances accuracy, it also increases the computational cost. In contrast, the baseline model, although less accurate, is computationally more efficient but fails to provide high accuracy which is the primary requirement of an automated medical diagnosis system.

The overlap between the expected segmentation and the actual segmentation is measured using the IoU method. A greater agreement between the two is indicated by a higher IoU. or DSC, likewise measures the overlap between the expected and ground truth segmentations; however, the overlap is weighted more heavily. IoU and DSC values for segmentation with and without BN are contrasted in the Table 5.

When BN is used, the IoU score rises significantly from 0.2915 without BN to 0.8193. Similarly, the DSC increases somewhat with BN, going from 0.4359 to 0.9006. With an IoU of 0.8272 and a DSC of 0.9054, the segmented image significantly outperforms the performance. These findings highlight how important batch normalization is to improving the model's IoU and DSC performance and producing more precise segmentation results. Both IoU and DSC measure segmentation performance, which is critical for diabetic retinopathy detection since it requires segmenting the diseased regions of the retina. Both directly quantify how well the model's segmentation matches the ground truth, and therefore, they are very important for accurate diagnosis and treatment planning. Accurate segmentation enables clinicians to determine the severity and extent of the disease, and consequently, better-informed decisions regarding treatment plans and possible interventions. Likewise, the MCC is a balanced performance measure but does not emphasize segmentation as much as IoU and DSC. In brief, IoU and DSC are most important to assess the model's capability to segment retinal structures related to DR correctly.

K-Fold cross validation

K-Fold cross-validation is a validation test that employs the training and test data sets. First, the data set is split up into many k-folds. Five folds are employed in this work as $K = 5$ produced the overall best accuracies in the experimental setup. This method enables the impact of bias, unpredictability, and variability to be observed. A discrepancy between the actual and expected accuracy serves as a sign of bias. It is used to assess the stability, robustness, and dependability of our models. Figure 12 (a), (b) and (c) depicts the K-fold ($K = 5$) cross validation accuracy for Dataset 1, 2 and 3 respectively.

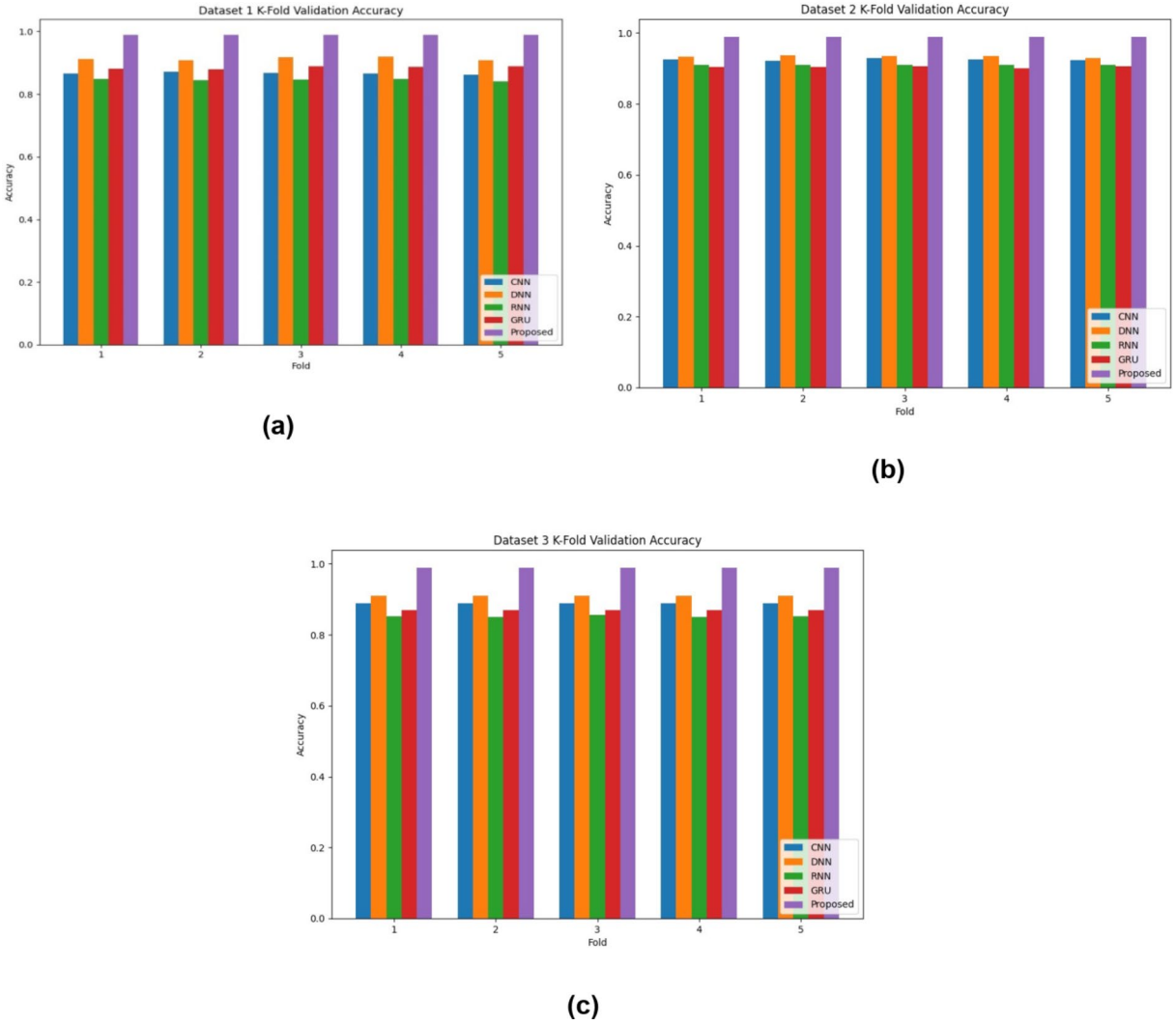

Fig. 12. (**a**) Comparison of K-Fold cross validation accuracy for Dataset (1) (**b**) Comparison of K-Fold cross validation accuracy for Dataset (2) (**c**) Comparison of K-Fold cross validation accuracy for Dataset 3.

The proposed approach demonstrates its higher accuracy and robustness by consistently outperforming the conventional classifiers (CNN, DNN, RNN, and GRU) across all three datasets. The performance of existing models varies, reflecting their susceptibility to variations in the data, however the proposed model maintains excellent accuracy levels across various data folds. Higher overall accuracy values indicate that Dataset 2 is presumably easier to classify than Dataset 1. The proposed model's supremacy, however, remains consistent across all datasets, demonstrating how well it can handle a variety of data distributions. The confusion matrices for the three datasets for the proposed model is shown in the Fig. 13 (a), (b) and (c) which depicts the Dataset 1,2 and 3 confusion matrices correspondingly.

Figure 14 illustrates the convergence rate of various optimization algorithms of the proposed SANGO, NGO, HHO (Harris Hawk Optimization), PSO (Particle Swarm Optimization) on three datasets. SANGO always reports the lowest computation cost on all datasets, i.e., 1.7s for Dataset 1, 4.8s for Dataset 2, and 11.9s for Dataset 3. NGO and HHO have larger computation times of 2.4s and 2.0s for Dataset 1, 5.5s and 5.3s for Dataset 2, and 14.2s and 14.0s for Dataset 3, respectively. PSO takes the most computation time when it comes to Dataset 1 (2.8s) and is consistently high in Dataset 3 (13.1s). While increasing dataset size, computation time increases drastically with all methods with SANGO emerging as the best overall.

Experimental evaluation measures the computational efficiency of the proposed method in terms of overall detection time on three datasets and compares it with conventional deep learning models CNN, DNN, RNN, and GRU is depicted in the Fig. 15. In Dataset 1, the lowest detection time of 2.9 s was accomplished by the proposed method compared with CNN (3.6s), DNN (4.2s), RNN (4.1s), and GRU (3.6s), where it performed at a better efficiency for small-sized datasets. In Dataset 2, the presented method retained its lead in detection time of 7.2 s, being quicker than CNN (8.8s), RNN (8.4s), and GRU (8s), yet slightly slower than DNN (7.5s), reflecting the optimal compromise between feature extraction and classification speed. For Dataset 3, the computationally most demanding, the proposed approach was still efficient with a total detection time of 16.1 s, as opposed

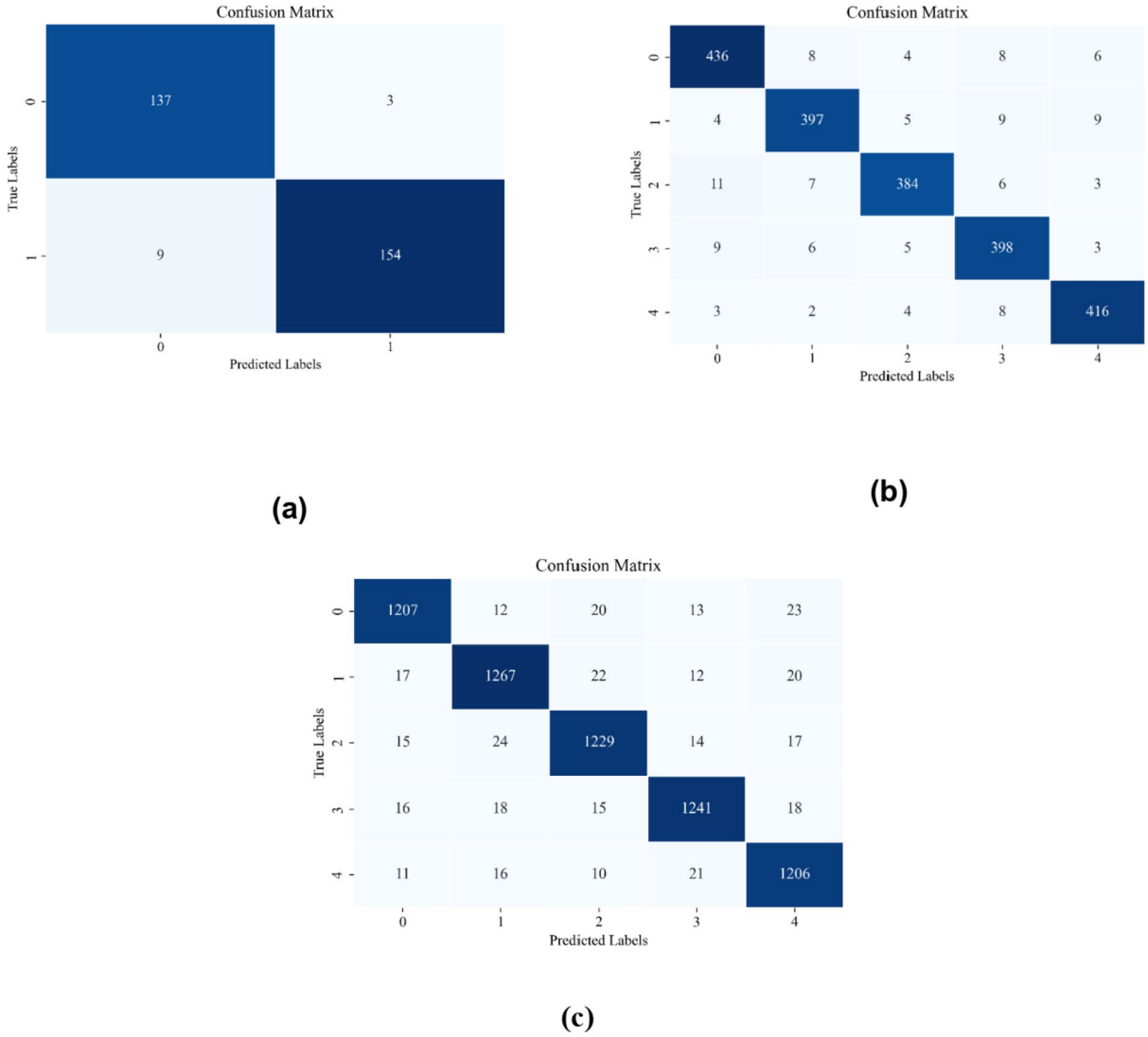

Fig. 13. (**a**): Confusion matrix for Dataset (1) (**b**): Confusion matrix for Dataset (2) (**c**): Confusion matrix for Dataset 3.

to CNN (16.7s), DNN (17.5s), RNN (17.6s), and GRU (17.4s). The decrease in detection time across datasets indicates the proposed model's streamlined feature extraction and classification pipeline.

Comparison analysis for the proposed and the existing approaches

Table 6 presents a tabular comparison between the proposed strategy and existing DR approaches. In[23] a deep neural network is used with CTSA, U-Net, and a hybrid entropy model for high accuracy on DIARETDB0 and DIARETDB1 datasets. In[24] a shallow CNN is used with 98.65% accuracy on a merged dataset. In[30] an ensemble model is created by combining VGG19, InceptionV3, and ResNet50 for feature extraction and classification, achieving 98.47% accuracy on the Kaggle DR dataset and in[31] 97.98% accuracy is achieved using CapsNet.

The proposed strategy performs exceptionally well in terms of accuracy compared to all other approaches on the Kaggle EyePacs dataset of 99.12%, DiaRetDB1 of 99.01%, and the APTOS 2019 dataset of 98.98%. The proposed model is more accurate because it uses higher-level techniques such as adaptive BN, attention, and SANGO optimization. It employs a modified U-Net architecture with an attention mechanism, DenseNet, OGRU, SoftMax Activation Function, and a Modified Cross-Entropy Loss Function, achieving exceptional accuracy on DiaRetDB1, APTOS 2019, and Kaggle EyePacs datasets. The proposed DR classifier offers new techniques, which include AGF, modified U-Net, multifaceted attention feature, OGRU, Grad-CAM, and modified loss function, and it enhances accuracy to a greater degree, along with robustness and explainability. It achieved 99.10% accuracy on Dataset 3, which is 12% better than the baseline of 87.10%, and precision increased to 98.23%, which enhanced by 10.22% over the earlier precision of 88.01%. This model improves diagnostic quality, enhances image quality, and incorporates Grad-CAM (AI), while it suffers from additional computational complexity, and the model becomes more resource-consuming compared to less complex models such as shallow CNNs or one-architecture deep learning models. The reduction in complexity, enhancement

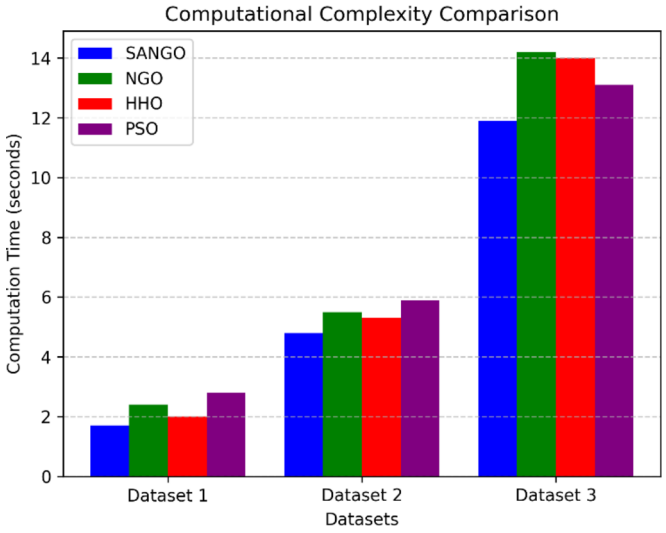

Fig. 14. Computational complexity of the proposed algorithm.

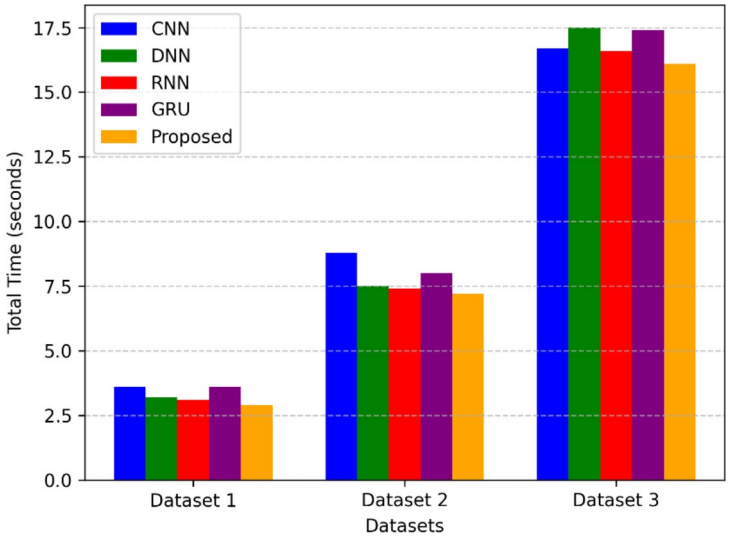

Fig. 15. Comparison of total detection time.

Ref.	Techniques	Accuracy
[23]	Deep neural network with CTSA, U-Net, hybrid entropy model	DIARETDB0 :95.9% DIARETDB1:95.48%
[24]	Shallow CNN (RetNet-10) with three convolutional layer blocks, maxpool layers.	Merged dataset of APTOS, Messidor2, IDRiD 98.65%
[30]	Ensemble Model (VGG19 + InceptionV3 + ResNet50) for feature extraction and classification.	Kaggle DR dataset :98.47%
[31]	Modified capsule network (CapsNet) with convolution, primary capsule, and class capsule layers	Messidor: 97.98%
Proposed work	Modified U-Net architecture with attention mechanism, DenseNet, OGRU, SoftMax Activation Function, Modified Cross-Entropy Loss Function, Grad-CAM, and Multi-Fold Cross-Validation.	DiaRetDB1 :99.01% APTOS 2019 : 98.98% Kaggle EyePacs: 99.12%

Table 6. Comparison analysis of the proposed and existing method.

of explainability, and integration with real-world AI should be directed for future work. A future clinical study may be performed where the model is utilized to process retinal images from patients in an actual clinical environment. This is something that can be addressed efficiently in future by using a number of validation strategies to validate the real-world utility of the model.

For severely blurred images, the preprocessing pipeline improves contrast and denoises to recover informative image details. But in the event of extremely poor image quality that is unsuitable for credible classification, Grad-CAM assists in identifying such instances by highlighting whether the model is concentrating on informative areas. The deep learning feature extraction technique also assists in identifying intricate patterns outside of the explicitly trained examples, enhancing generalization to unusual retinal anomalies. In addition, the integration of multi-folded feature extraction and OGRU with SANGO optimization improves the model's capacity to identify complex or rare retinal diseases, which makes it extremely versatile in real-world clinical applications. Light-weight modules are employed by substituting regular convolutions with depth-wise separable convolutions or group convolutions, which decrease the number of parameters and computational cost. Moreover, Ghost-Net modules are embedded to produce redundant feature maps with reduced computations, enhancing feature representation. The encoder further employs Efficient-Net Lite layers to fuse features, which reduces FLOPs while preserving high accuracy. In order to further increase efficiency, MobileNetV3-based encoders are employed as feature extractors, providing further computational performance improvements. These measures collaborate to decrease computational overhead without reducing the effectiveness of the model.

Conclusion

A new multi-model deep learning framework for DR identification has been developed. On several datasets, the multi-model deep Net demonstrated great accuracy, precision, sensitivity, and specificity values against other approaches for the segmentation and classification of DR. A modified U-Net is used for segmentation, an AGF with a Chaotic Map for preprocessing, a phase of classification, and feature folding, coupled with a SANGO-optimized OGRU. The developed model's validity and dependability were testified with strict evaluation by evaluation indices such as IoU and DSC and using Grad-CAM as an explainable AI. In our experiments, the proposed model gave accuracy of 99.01% on the DiaRetDB1 dataset, 98.99% on the APTOS 2019 dataset and 99.12% in Kaggle EyePacs dataset. For all data sets, the model has an impressive F1-score, specificity, accuracy, and sensitivity rate that unsubstantiated any doubt on the diagnostic capability of the proposed model. The method of segmentation was effective, and this is proved by the high IoU and DSC values across the three datasets. Further studies can be directed toward developing the usage of the proposed model for other retinal illnesses or integrating the described approach into a system that includes other sorts of medical images.

Data availability

The datasets generated and/or analysed during the current study are available in the Kaggle repository: https://www.kaggle.com/datasets/nguyenhung1903/diaretdb1-standard-diabetic-retinopathy-database/data, https://www.kaggle.com/c/aptos2019-blindness-detection, https://www.kaggle.com/competitions/diabetic-retinopathy-detection.

Received: 10 January 2025; Accepted: 6 March 2025
Published online: 13 March 2025

References

1. Talukder, M. S. H., Sarkar, A. K., Akter, S. & Nuhi-Alamin, M. An Improved Model for Diabetic Retinopathy Detection by using Transfer Learning and Ensemble Learning. *arXiv preprint arXiv:2308.05178*. (2023).
2. Hossain, M. J., Al-Mamun, M. & Islam, M. R. Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. *Health Science Reports*, *7*(3), e2004. (2024).
3. Albadr, M. A. A., Ayob, M., Tiun, S., Al-Dhief, F. T. & Hasan, M. K. Gray Wolf optimization-extreme learning machine approach for diabetic retinopathy detection. *Front. Public. Health*. **10**, 925901 (2022).
4. Sungheetha, A. & Sharma, R. Design an early detection and classification for diabetic retinopathy by deep feature extraction-based Convolution neural network. *J. Trends Comput. Sci. Smart Technol. (TCSST)*. **3** (02), 81–94 (2021).
5. Raja Sarobin, M. & Panjanathan, R. V., Diabetic retinopathy classification using CNN and hybrid deep convolutional neural networks. *Symmetry*, *14*(9), 1932. (2022).
6. Bilal, A., Sun, G., Li, Y., Mazhar, S. & Khan, A. Q. Diabetic retinopathy detection and classification using mixed models for a disease grading database. *IEEE Access*. **9**, 23544–23553 (2021).
7. Fayyaz, A. M., Sharif, M. I., Azam, S., Karim, A. & El-Den, J. Analysis of diabetic retinopathy (DR) based on the deep learning. *Information* **14** (1), 30 (2023).
8. Qureshi, I., Ma, J. & Abbas, Q. Diabetic retinopathy detection and stage classification in eye fundus images using active deep learning. *Multimedia Tools Appl.* **80** (8), 11691–11721 (2021).
9. Erciyas, A. & Barışçı, N. An effective method for detecting and classifying diabetic retinopathy lesions based on deep learning. *Comput. Math. Methods Med.* **2021** (1), 9928899 (2021).
10. Aatila, M., Lachgar, M., Hrimech, H. & Kartit, A. Diabetic retinopathy classification using ResNet50 and VGG-16 pretrained networks. *Int. J. Comput. Eng. Data Sci. (IJCEDS)*. **1** (1), 1–7 (2021).
11. Ullah, N. et al. Diabetic retinopathy detection using genetic Algorithm-Based CNN features and error correction output code SVM framework classification model. *Wirel. Commun. Mob. Comput.* **2022** (1), 7095528 (2022).
12. Mohanty, C. et al. Using deep learning architectures for detection and classification of diabetic retinopathy. *Sensors* **23** (12), 5726 (2023).
13. Goel, S. et al. Deep learning approach for stages of severity classification in diabetic retinopathy using color fundus retinal images. *Math. Probl. Eng.* **2021** (1), 7627566 (2021).
14. Katada, Y. et al. Automatic screening for diabetic retinopathy in interracial fundus images using artificial intelligence. *Intelligence-Based Med.* **3**, 100024 (2020).
15. Taifa, I. A., Setu, D. M., Islam, T., Dey, S. K. & Rahman, T. A hybrid approach with customized machine learning classifiers and multiple feature extractors for enhancing diabetic retinopathy detection. *Healthc. Analytics* **5**, 100346. (2024).
16. Nadda, R., Singh, J. & Shrivastava, U., Automatic diabetic retinopathy detection using an ensemble learning approach and classifiers with self-adjusting weights, PREPRINT (Version 1) available at Research Square. https://doi.org/10.21203/rs.3.rs-4376163/v1 (2024).
17. Sikder, N. et al. Severity classification of diabetic retinopathy using an ensemble learning algorithm through analyzing retinal images. *Symmetry* **13** (4), 670 (2021).

18. Meshram, A. & Dembla, D. MCBM: implementation of multiclass and transfer learning algorithm based on deep learning model for early detection of diabetic retinopathy. *ASEAN Eng. J.* **13** (3), 107–116 (2023).
19. Menaouer, B., Dermane, Z., Houda Kebir, E., Matta, N. & N., & Diabetic retinopathy classification using hybrid deep learning approach. *SN Comput. Sci.* **3** (5), 357 (2022).
20. Jayanthi, J. et al. An intelligent particle swarm optimization with convolutional neural network for diabetic retinopathy classification model. *J. Med. Imaging Health Inf.* **11** (3), 803–809 (2021).
21. Jadhav, A. S., Patil, P. B. & Biradar, S. Optimal feature selection-based diabetic retinopathy detection using improved rider optimization algorithm enabled with deep learning. *Evol. Intel.* **14**, 1431–1448 (2021).
22. Mondal, S. S., Mandal, N., Singh, K. K., Singh, A. & Izonin, I. Edldr: an ensemble deep learning technique for detection and classification of diabetic retinopathy. *Diagnostics* **13** (1), 124 (2022).
23. Dayana, A. M. & Emmanuel, W. S. An enhanced swarm optimization-based deep neural network for diabetic retinopathy classification in fundus images. *Multimedia Tools Appl.* **81** (15), 20611–20642 (2022).
24. Raiaan, M. A. K., Fatema, K., Khan, I. U., Azam, S., Rashid, M. R. U., Mukta, M. S.H., … De Boer, F. (2023). A lightweight robust deep learning model gained high accuracy in classifying a wide range of diabetic retinopathy images. IEEE Access, 11, 42361–42388.
25. Hemanth, S. V. & Alagarsamy, S. Hybrid adaptive deep learning classifier for early detection of diabetic retinopathy using optimal feature extraction and classification. *J. Diabetes Metabolic Disorders*. **22** (1), 881–895 (2023).
26. Ishtiaq, U., Abdullah, E. R. M. F. & Ishtiaque, Z. A hybrid technique for diabetic retinopathy detection based on ensemble-optimized CNN and texture features. *Diagnostics* **13** (10), 1816 (2023).
27. Jabbar, A. et al. A Lesion-Based Diabetic Retinopathy Detection Through Hybrid Deep Learning Model. *IEEE Access*. (2024).
28. Deepa, V., Kumar, C. S. & Cherian, T. Ensemble of multi-stage deep convolutional neural networks for automated grading of diabetic retinopathy using image patches. *J. King Saud University-Computer Inform. Sci.* **34** (8), 6255–6265 (2022).
29. Ali, G., Dastgir, A., Iqbal, M. W., Anwar, M. & Faheem, M. A hybrid convolutional neural network model for automatic diabetic retinopathy classification from fundus images. *IEEE J. Translational Eng. Health Med.* **11**, 341–350 (2023).
30. Pavithra, S., Jaladi, D. & Tamilarasi, K. Optical imaging for diabetic retinopathy diagnosis and detection using ensemble models. *Photodiagn. Photodyn. Ther.* **48**, 104259 (2024).
31. Kalyani, G., Janakiramaiah, B., Karuna, A. & Prasad, L. N. Diabetic retinopathy detection and classification using capsule networks. *Complex. Intell. Syst.* **9** (3), 2651–2664 (2023).
32. Dehghani, M., Hubálovský, Š. & Trojovský, P. Northern goshawk optimization: a new swarm-based algorithm for solving optimization problems. *Ieee Access.* **9**, 162059–162080 (2021).
33. Das, S., Kharbanda, K., Suchetha, M., Raman, R. & Dhas, E. Deep learning architecture based on segmented fundus image features for classification of diabetic retinopathy. *Biomed. Signal Process. Control.* **68**, 102600 (2021).
34. Jabbar, M. K., Yan, J., Xu, H., Ur Rehman, Z. & Jabbar, A. Transfer learning-based model for diabetic retinopathy diagnosis using retinal images. *Brain Sci.* **12** (5), 535 (2022).

Author contributions

Neeraj Sharma is primarily responsible for the drafting of the article. Praveen Lalwani is involved (alongwith Neeraj Sharma) in generation of the experimental results and the formulation of the framework described in the article.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to N.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.