Electronic Letters on Computer Vision and Image Analysis 24(2):158-182, 2025

Selective Feature Map Fusion Architecture for Enhanced Diabetic Retinopathy Detection and Grading using ResNet50 and EfficientNetV2L Models in Fundus Images

Vinya Vijayan* Salim A+

* *College of Engineering Trivandrum, Affiliated to APJ Abdul Kalam Technological University, Kerala, India*
+ *College of Engineering Trivandrum, Affiliated to APJ Abdul Kalam Technological University, Kerala, India*

Received 11th of june, 2024; accepted 2nd of december, 2025

Abstract

Diabetic Retinopathy caused by Diabetes Mellitus is a major vision threatening condition across the global population. Early detection and grading of diabetic retinopathy are pivotal to avoiding the associ ated vision impairment, and retinal image analysis serves as an effective method for the screening process. Analysing retinal images manually to detect diabetic retinopathy and grade them is a time-consuming pro cess that necessitates the involvement of experts to perform the classification. Computer-aided diagnosis of diabetic retinopathy through retinal image analysis is an effective tool to reduce the time involved in the screening process. This study proposed a convolutional neural network based Selective Feature Map Fusion architecture that leverages the fused feature maps of ResNet50 and EfficientNetV2L networks for the detection and grading of diabetic retinopathy. Feature set of ResNet50 and EfficientNetV2L captures different aspects of the underlying lesion distribution, resulting in a more comprehensive representation of the features. This approach helps the model generalize better to unseen data by incorporating diverse data aspects, thereby preventing overfitting to a specific feature set and developing a model that is more robust to noise. Selective feature fusion helps to reduce the computational overhead of the feature fusion computation. Comparative analysis against the ResNet50 and EfficientNetV2L models individually revealed the superior performance of the fused model in both detection and grading of diabetic retinopathy. The proposed model attained an excellent level of accuracy of 95.2% in diabetic retinopathy detection with a sensitivity of 95.2% and a specificity of 95.2%, when tested on the IDRiD dataset. The model achieved a classification accuracy of 92% for diabetic retinopathy grading with a sensitivity of 88%, a specificity of 98% and an F1-score of 0.89. Moreover, the fused framework attained superior performance when compared against existing methodologies on the IDRiD dataset. The developed model exhibited robust performance also in the DeepDRiD dataset, and it can be employed successfully in the screening processes for the diagnosis and grading of diabetic retinopathy.

Key Words: Computer aided Diagnosis, Convolutional Neural Network, Deep Learning, Diabetic Retinopathy Detection and Grading, EfficientNetV2L, Image Classification, Retinal Images, Resnet50, Transfer Learning.

Correspondence to: $vinya.vijayan@cet.ac.in$

Recommended for acceptance by $AngelD.Sappa$
https://doi.org/10.5565/rev/elcvia.1954
ELCVIA ISSN:15108-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

1 Introduction

Diabetic Retinopathy is one of the leading causes of vision impairment in people and early detection and treatment are advised to minimise the impact on affected individuals. This complication arises from Diabetes Mellitus, resulting in damage to retinal blood vessels often leading to leakage or rupture[**?**]. As of 2020, an estimated 3.2 million individuals worldwide were affected by diabetic retinopathy and associated vision impairment or blindness[**?**]. Effective screening methods are important, as early-stage vision impairment due to diabetic retinopathy can be managed with appropriate medication. Currently, laboratory screening involves the systematic evaluation of retinal images by ophthalmologists. Computer-aided diagnosis of diabetic retinopathy can assist in the early screening process by reducing manual analysis time and minimizing human intervention. Automated systems serve as efficient screening mechanisms, particularly beneficial for helping individuals in remote and rural areas where access to clinical facilities is limited.

Different imaging modalities are used to capture the retinal images and the extensively used and cost effective one is the fundus imaging technique. The automated systems for diabetic retinopathy detection and grading mainly depends on the analysis of retinal fundus images using various image processing and machine learning algorithms. There are different solutions available for detection of diabetic retinopathy by applying machine learning algorithms[**?**].

The introduction of deep learning methods significantly enhances classification accuracy in all fields includ ing medical image analysis[**?**]. Deep learning algorithms exhibited superior performance compared to conventional machine learning and image processing methods for detecting diabetic retinopathy. Convolutional Neural Network (CNN) is a type of neural network commonly used in deep learning for image classification and recognition tasks. CNN systems improve the accuracy of the detection and grading process further by the use of effective self-generated feature maps from the images. These algorithms avoid an explicit feature extraction process and thereby reducing the overhead in detection.

Different retinal fundus image databases are publicly available to evaluate the performance of the methods for detecting and grading diabetic retinopathy. The IDRiD dataset that contains verified fundus images are used for this study[**?**]. Detection involves a binary classification of images into normal or abnormal based on retinal image analysis with the help of various algorithms. Grading diabetic retinopathy is a more advanced classification that involves identifying the severity of the condition, which is associated with various types of lesions in the image. Datasets of fundus images for grading have been categorized into five groups based on the severity of diabetic retinopathy using the classification standard adopted by the Global Diabetic Retinopathy Project Group [**?**]. These categories include *No DR, Mild Non proliferative diabetic retinopathy (NPDR), Moderate NPDR, Severe NPDR*, and *Proliferative diabetic retinopathy*.

The fused model is constructed by integrating the feature maps of ResNet and EfficientNetV2L models. The integration of the feature maps from different models provide an enhanced feature representation and correspondingly better lesion identification. In addition to feature selection the model exploits the possibility of different preprocessing techniques including contrast-limited adaptive histogram equalization for obtaining better performance.

The remaining content of this article is organized into five sections. The current status of computer-aided diagnosis and grading of diabetic retinopathy is discussed in Section 2, along with the main drawbacks and potential solutions. The dataset used in this work is described in Section 3 and the architecture of the proposed CNN-based Feature Fusion model is described in Section 4. Section 5 provides the experiment results, and Section 6 uses several performance indicators to compare the performance of the proposed system with that of the existing approaches. The results of this investigation are summarised in Section 7, together with their limitations and some prospective directions.

2 Related Work

In recent decades, computer-aided diagnosis of diabetic retinopathy has been an explored area of study in medical image analysis, with a large number of methods suggested to develop an efficient automated system for this purpose.

Deep learning based computer-aided diagnosis serves as an effective screening method for the early detection of diabetic retinopathy. Numerous techniques have been proposed to distinguish the retinal images affected by diabetic retinopathy from the normal images, using various datasets designed for this task. The effectiveness and applicability of these detection methods vary depending on the specific technique employed and the dataset being used.

Before implementing deep learning algorithms, image pre-processing techniques are utilized to enhance the quality of the images. In most cases, retinal images are subjected to resizing, contrast enhancement, and noise reduction as part of the pre-processing. To address the issue of class imbalance during detection and grading, data augmentation is commonly employed as a pre-processing step in many deep learning algorithms[**?**]. To prevent the negative impact of class imbalance on classification performance, Anas et al. utilised data augmen tation techniques such as rotation, shearing, flipping, and cropping on images [**?**]. The study by Hameed et al. revealed that using an enhancement approach significantly improved the results of classification[**?**]. The algorithm aimed to enhance the quality of colour fundus images through noise reduction and contrast enhancement.

Asra et al. proved that the use of contrast-limited adaptive histogram equalisation as a contrast enhancement method for optical disc localization improved the performance of lesion segmentation-based diabetic retinopa thy detection [**?**]. In this article, both data augmentation and contrast enhancement are applied as image preprocessing steps before creating the feature maps. Li et al. developed a lesion-attention pyramid network for both grading and associated lesion segmentation[**?**]. The lesion attention module is used to develop a lesion activation map which is helpful in the segmentation of lesions. Evaluation on the Messidor and EyePACS datasets demonstrated that the suggested model consistently delivered reliable performance.

Several deep learning-based algorithms have been suggested for the assessment of diabetic retinopathy through the grading of retinal images. Eman et al. introduced the E-DenseNet model, which integrates EyeNet and DenseNet models for this purpose[**?**]. By adding dense blocks, they modified the conventional EyeNet and then adjusted the hyperparameters of the hybrid E-DensNet model to improve the performance. Elswah et al. [**?**] utilize a pre-existing Resnet50 CNN model for grading the severity of diabetic retinopathy and achieved an accuracy rate of 86.67% when applied to the IDRiD dataset. To ensure a balanced representation of each class during the training process, they applied various image preprocessing and augmentation techniques.

Ling et al. apply self-knowledge distillation idea in CNN models to conduct grading, with a particular emphasis on different lesions within the image data [**?**]. Another CNN based architecture InceptionResnet-V2 is suggested by Charu et al. and the method provides reliable classification accuracy in the IDRiD dataset for diabetic retinopathy detection [**?**]. The use of transfer learning is also beneficial in enhancing classification performance in this field. In their study on the IDRiD dataset, Das et al. [**?**] investigated over 15 pre-trained CNN models, such as VGG-19, DenseNet121, EfficientNetB0, and ResNet50, using transfer learning. They determined that the EfficientNetB4 model achieved the highest performance among the pre-trained models, attaining a testing accuracy of 79.11% on the IDRiD dataset. The study also pointed out that the efficientnet family of models consistently delivers reliable classification performance. Lin et al. proposed an improved ResNet-50 model for detecting diabetic retinopathy, which produced better results than the standard ResNet-50[**?**].

Optimization algorithms are also used to boost classification performance in detection and grading. Elwin et al. introduced a deep learning technique based on Autoregressive-Henry Gas Sailfish Optimization (Ar-HGSO) for the detection of diabetic retinopathy and the classification of severity levels using retinal fundus images [**?**]. Data augmentation techniques are used to increase the number of images representing each class and the lesion segmentation is applied to find the lesions associated with differnt stages of diabetic retinopathy.

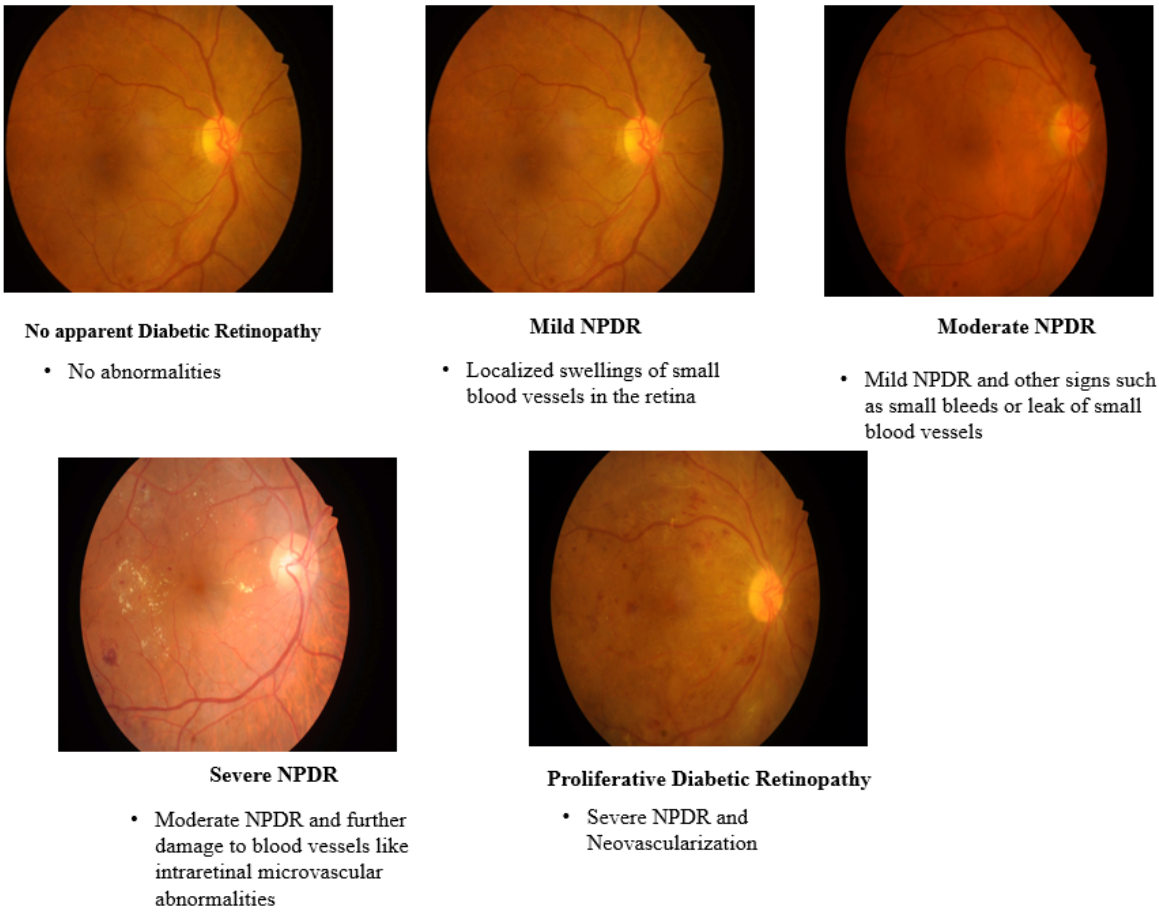


Figure 1: International Grading of Diabetic Retinopathy stages based on findings observable on dilated opthalmoscopy.[**?**]

The analysis of the existing methods clearly shows that both the efficientnet family of models and Resnet50 provide high performance in diabetic retinopathy detection and grading. The fusion of these models leverages the strengths of both architectures, which provides enhanced feature representation. ResNet is renowned for its depth and rich feature representation but can be computationally expensive. EfficientNet, on the other hand, is designed to achieve better performance with fewer parameters and FLOPs. Concatenating these models allows for a more efficient utilization of computational resources while improving performance.

Most of the studies referenced above employed lesion-based diabetic retinopathy detection and grading. One limitation of this lesion-based diabetic grading involves identifying related lesions associated with different grades, which necessitates the application of numerous image preprocessing techniques for the identification and segmentation of lesions. Additionally, lesion-based detection and grading also require image quality en hancement. Consequently, this leads to increased computational demands during both the development and real-time testing stages of computer-aided diagnosis systems. However, it is feasible to conduct grading without the need for complex lesion segmentation and quantification tasks, thereby reducing testing time in real-time scenarios.

The aforementioned studies have considerably enhanced the detection and grading of diabetic retinopathy. In order to further improve the classification performance and to perform grading without lesion segmentation, a CNN-based feature fusion model that is evaluated on the IDRiD dataset is suggested in this article. This suggested model combines the deep hierarchical features learned by ResNet50 and EfficientNetV2L models, potentially leading to improved accuracy and generalization. The feature fusion process provides accurate and

efficient feature map generation and robust detection and grading.

3 Dataset Used

The Indian Diabetic Retinopathy Image Dataset (IDRiD) is a publicly available dataset that aims to develop and evaluate algorithms for the detection and grading of diabetic retinopathy using retinal images. The dataset includes deatils of common lesions associated with diabetic retinopathy and information on the severity based on which each image is categorized into different grades.

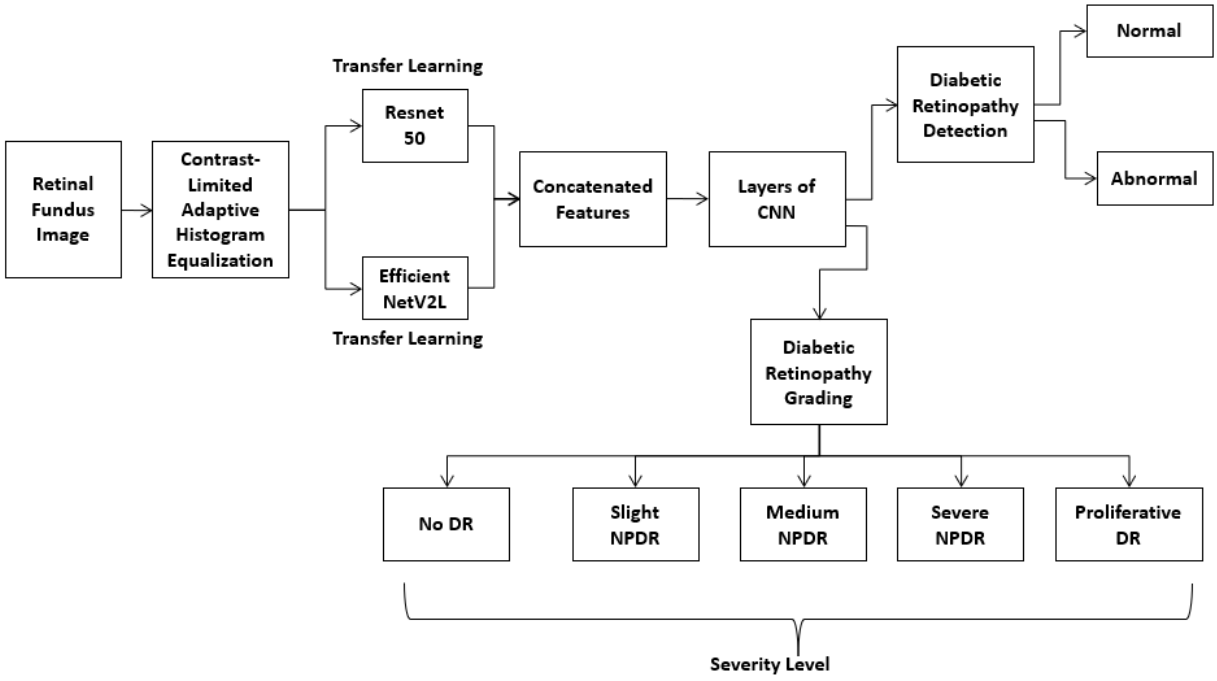


Figure 2: Schematic illustration of the suggested CNN based Feature Fusion Diabetic Retinopathy detection and Grading model.

The database contains 516 images, divided into retinal images with diabetic retinopathy and normal retinal images without diabetic retinopathy. Further, the medical experts graded the retinal images of diabetic retinopathy in the range of 0 to 4 based on their severity. Grade 0 represents normal retinal images, while grade 4 indicates images with severe diabetic retinopathy. Ophthalmologists partitioned these labelled images into training and testing sets, ensuring diversity within the dataset.

The training set consists of 413 images and the testing set consists of 103 images which are captured using the retinal fundus camera model kowa VX-10α This specialized high-resolution fundus camera is designed to capture the interior surface of the eye including the retina, optic disc, macula and posterior pole. The training set includes images representing a wide range of diabetic retinopathy stages, ensuring that the model can generalize well across different severities of the abnormality. The test set is a separate subset of retinal images not seen by the model during training, and it provides an unbiased evaluation of model performance and generalization capability. Fig.**??** shows the retinal images of different categories of diabetic retinopathy and their corresponding symptoms as identified by human experts.

The model is also evaluated on the challenging DeepDRiD dataset to establish the generalizability of the model[**?**]. This dataset includes 2,000 regular fundus images obtained from 500 patients, with detailed grading information. Each patient in the dataset has four fundus images, centered on the macula and optic disc. To ensure diagnostic accuracy, two certified ophthalmologists initially grade the images independently, and a senior

supervisor then reviews their assessments. The final grade is determined based on the agreement of three experts. Clinically, DR is classified into five stages according to the International Clinical DR (ICDR) scale as no retinopathy (grade 0), mild NPDR (grade 1), moderate NPDR (grade 2), severe NPDR (grade 3), and PDR (grade 4) based on the presence of lesions[**?**].

4 Methodology

CNN-based feature fusion model generated from Resnet50 and EfficientNetV2L models are suggested for dia betic retinopathy detection and grading. The retinal fundus images are given as input to the system and image pre-processing techniques are applied to increase the quality of the images. Feature combination is an efficient method to improve the classification performance of the system. The repeated application of convolutional and pooling operations in the network is equivalent to adding infintely strong prior to a density model which progressively refined and enhanced the robustness of the learned features. As the input passes through multiple layers, the features become more refined and discriminative. Combining features from multiple sources can result in more accurate models, as each feature set captures different aspects of the underlying data distribution, leading to a more comprehensive representation and improved classification. In the proposed model, ResNet50 and EfficientNetV2L networks are employed for feature extraction, and the resulting feature maps are concate nated and given to subsequent CNN layers. The architecture diagram of the diabetic retinopathy detection and grading module is depicted in Fig. **??**. A comprehensive explanation of each stage is provided in the subsequent sections.

4.1 Image Pre-processing

Image resizing, data augmentation and contrast enhancement techniques are employed to enhance the quality of the retinal images. The input images are resized to size of 224 x 224 in order to be compliant with the input size of the trained network models used for the feature extraction process. The datasets are highly imbalanced, necessitating strategies to handle the underrepresented classes. Techniques such as data augmentation, class weighting, and balanced batch creation are employed to mitigate this imbalance. Data augmentation techniques are used to ensure the proper stratification of each grade and to avoid the class imbalance issue in the training process. The IDRiD dataset used in this study contains a limited number of images, which is insufficient for training a deep learning model. This issue was solved by using data augmentation to increase the number of images. Horizontal flipping, random rotation and color contrast methods are used for augmentation and this approach helped mitigate the imbalance by generating additional samples for underrepresented classes. Class weights are also used to reduce the class imbalance and assigned higher weights to the minority classes to ensure that the model adequately learns to recognize them, which can significantly enhance performance across all classes. The balanced batch sampling is also utalized during training to ensure that each batch contains a representative number of samples from each class, thus promoting better learning outcomes.

The quality of the images is enhanced by using the contrast-limited adaptive histogram equalization. It enhances the contrast of the images by using a variant of adaptive histogram equalisation. The equalization operates on small regions in the image with grid size 8 x 8 rather than the entire image and the results are combined using bilinear interpolation. For contrast limiting, a clip limit value of 40 is set as the threshold. Adaptive histogram equalization can amplify noise excessively in areas of an image that are uniform, but contrast-limited adaptive histogram equalization addresses this problem by restricting the amount of amplification. Clip value for each region is used for this purpose. After clipping, the Cumulative Distribution Function (CDF) of the histogram are computed for each section. The CDF is used to map the old pixel values to new values for con trast enhancement. Images that undergo contrast enhancement contribute to a more efficient feature extraction process. Applying the contrast enhancement technique improves the RMS contrast of the images and enhanced their overall quality.[**?**]

4.2 ResNet50 and EfficientNetV2L models

Transfer learning is a machine learning technique that uses information learned from one task to improve performance on a similar task. This technique is widely used in image classification and object detection tasks to increase the efficiency of the model. CNN is a popular choice for image classification, using pretrained models trained on extensive datasets like ImageNet[**?**].

ResNet50 is a pre-trained deep CNN model developed for image classification tasks which uses the residual connections that helps to solve the vanishing gradient problem that may arise in deep networks. It has 50 layers and takes as input an RGB image with a fixed resolution of 224 x 224 pixels. The core part of Resnet50 architecture is the residual block which helps the network to learn a better approximation of the image by skipping of layers during the processing. ResNet50 consists of a set of convolutional layers, followed by max pooling and global average pooling layers, and then fully connected layers for classification. The residual block implements the identity skip connection which is a type of residual connection that involves connecting the input of one layer to the output of the same layer[**?**]. This enables the layer to learn a residual mapping that requires only minor changes to the input to produce the desired output. The structure of the residual block is represented in Fig.**??**. The first layer of the architecture is a convolutional layer that uses a set of filters to

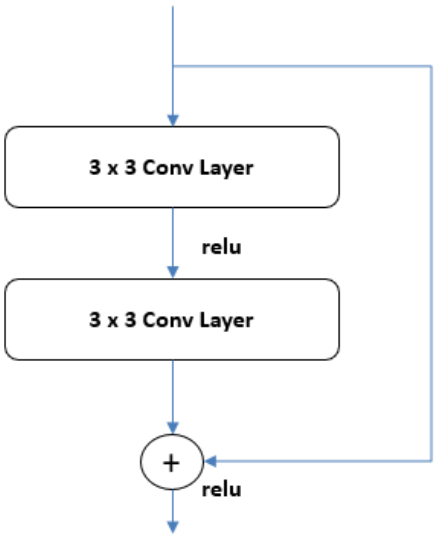


Figure 3: Schematic representation of residual block in Resnet50.

extract features from the image. This output of the layer is passed through a batch normalisation layer and followed by a ReLU activation function, which performs a nonlinear transformation on it. The architecture shows four repeating blocks, each with multiple convolutional layers. The stride of the first convolutional layer in each block is 2, which reduces the spatial dimensions of the input feature map. This stride 2 convolutional layer is followed by a series of convolutional layers, each of which is followed by batch normalisation and ReLU activation. Using the identity skip connection, the output of each set of convolutional layers is added to the input feature map and this helps to preserve important information and gradients during training.

The final layer of the network is a global average pooling layer that computes the average of each feature map across all spatial locations based on the results of the previous convoluted layer. The output generated is a feature vector that represents the entire image and it is passed through a fully connected layer. In the classification task the estimated probabilities of the input image for each class are output from a fully connected layer with a softmax activation function.

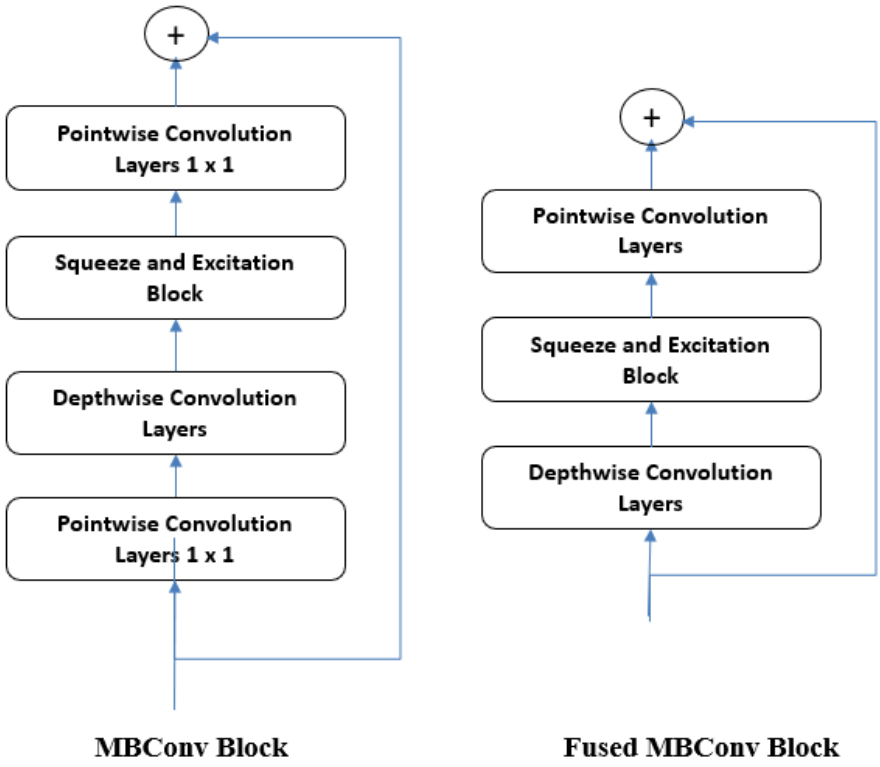


Figure 4: Schematic representation of MBConv block in EfficientNetV2L.

EfficientNetV2 is a type of CNN that outperforms earlier models in terms of parameter efficiency and training rate[**?**]. In the early layers, EfficientNetV2 prefers smaller 3 x 3 kernel sizes and makes considerable use of both Mobile Inverted Residual Bottleneck(MBConv) and the recently added fused-MBConv. The structure of MBConv block and the fused MBConv in the EfficientNetV2L is represented in the Fig. **??**. The MBConv block includes a 1x1 pointwise convolutional layer, a depthwise convolutional layer, a batch normalization layer, a non-linear activation function, and another 1x1 pointwise convolutional layer. The input data is subjected to a spatial filter by the depthwise convolutional layer while the channel dimension is altered by the pointwise convolutional layers. The batch normalization layer normalizes the output of the convolutional layers and the non-linear activation function introduces non-linearity into the output.

EfficientNetV2L design also uses a number of optimization strategies, such as Squeeze-and-Excitation (SE) blocks, Swish activation functions, and drop-connect regularisation. By recalibrating each block's output, the SE blocks assist the network in concentrating on its most crucial features, while the Swish activation function enhances the non-linearity of the network. Overfitting is prevented through drop-connect regularisation, which randomly removes connections between neurons during training.

Progressive learning with adaptive regularization, which modifies the regularization in accordance with im age size is the crtical part in the training of EfficientNetV2. EfficientNetV2-L is the largest model variant of EfficientNetV2 in terms of number of layers and it has a scaling factor of 4.3 with 27 layers. In classifi cation EfficientNetV2-L outperforms the ViT-L/16(21k), a much larger transformer model pretrained on the ImageNet21k dataset.

4.3 Selective Feature Fusion Model Architecture using ResNet50 and EfficientNetV2L

Feature fusion is a crucial technique for classification tasks and it involves combining multiple sets of features from different sources or modalities to improve the performance of a classifier. The architecture of the suggested

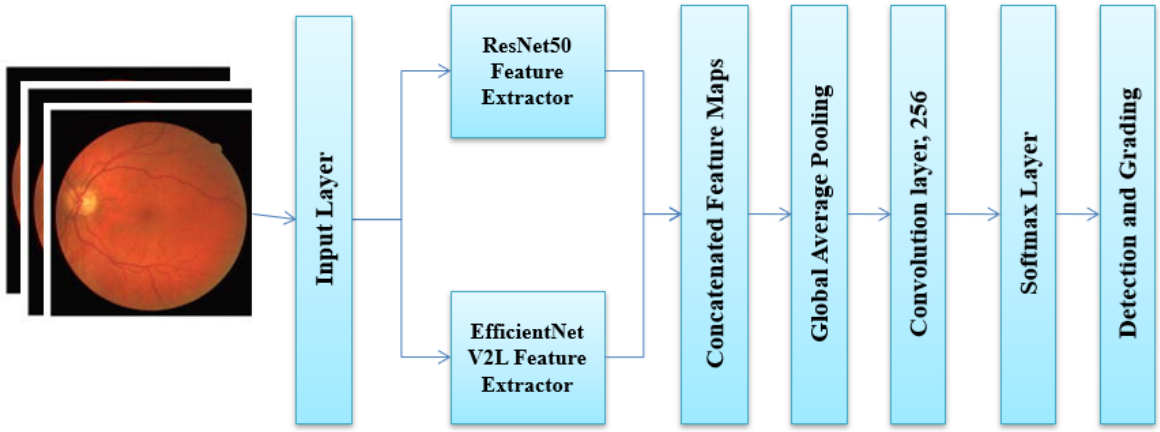


Figure 5: Architecture diagram of the suggested Feature Map Fusion model.

CNN-based feature fusion model is given in Fig. **??**. An instance of the EfficientNetV2L model is created, which uses the weights that were pre-trained in the ImageNet dataset. Similarly, another instance of ResNet50, pretrained on ImageNet, is instantiated. These networks are then fine-tuned to efficiently generate feature maps of retinal images. During training, unique feature maps are generated from each of these models by freezing the weights of the pretrained layers. ResNet50 and EfficientNetV2L generate different feature maps due to their architectural differences. ResNet50 maps are influenced by its depth and residual connections, while EfficientNetV2L maps benefit from compound scaling and efficient convolutions, leading to potentially richer and more detailed feature representations.

The feature maps capture the results of applying the filters to the image, and image-specific information is preserved depending on the filter used. Generally, the feature maps at the lower layers preserve the fine-grained features and the higher layers preserve more image-specific features. Through the visualization of the feature maps, an idea of what information is preserved in each of the layers of the CNN model is obtained.

4.3.1 Feature Maps of ResNet50

The feature maps in ResNet50 progressively reduce in spatial dimensions by the use of pooling layers and strides and play a key role in extracting discriminative features essential for distinguishing between various grades of diabetic retinopathy. As we progress to higher layers of the model, the feature maps encapsulate more semantic information and high-level patterns, which are crucial for making the ultimate classification decision. Additionally, some intermediate layers contain feature maps that highlight specific regions of interest in the image, such as lesions, and this helps the network focus on relevant areas and ignore irrelevant background information. This results in lower-resolution but more abstract feature representations in the deeper layers. Feature maps generated from some of the layers of ResNet50 are represented in Fig.**??**.

Consider input image is represented as a tensor (X with shape (B, C, H, W), in that B is the batch size, C is the number of channels, H is the height, and W is the width. The feature maps at each layer of the ResNet-50 model can be represented as follows:

Initial Convolutional Layer:

$$F_{conv1} = ReLU(BatchNorm(Conv2d(X, W_{conv1}, b_{conv1}))) \quad (1)$$

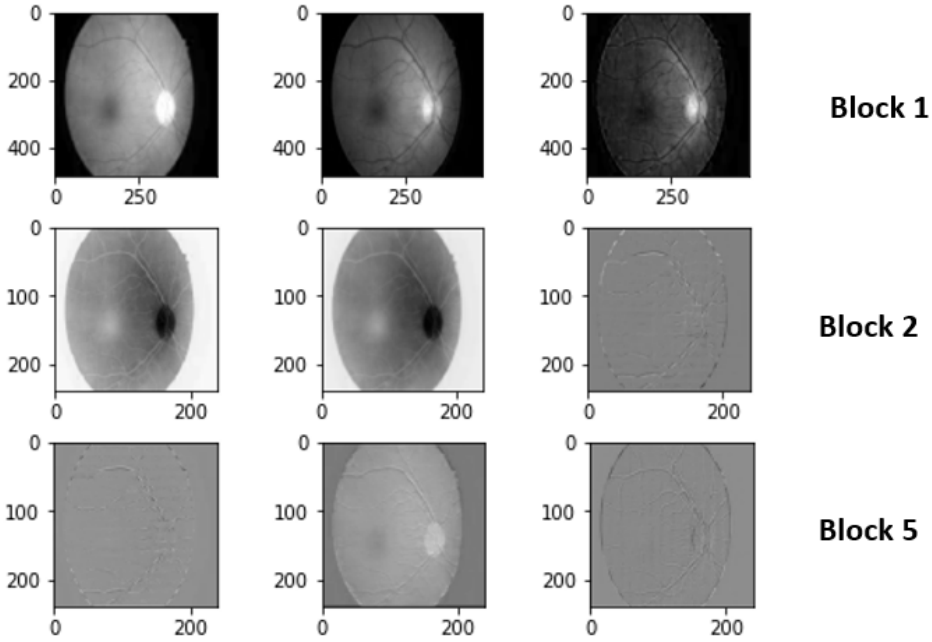


Figure 6: Feature maps created by the various blocks of EfficientNetV2L

In equation 1, W_{conv1} and b_{conv1} are the weights and biases of the initial convolutional layer, respectively. The feature map F_{conv1} has shape $(B, 64, H/2, W/2)$ after the max-pooling layer.

Residual Blocks: The ResNet-50 model consists of 4 stages, each containing multiple residual blocks.

For stage 1:

$$F_1 = ResidualBlock(F_{conv1}, W_{res1}, b_{res1}) \tag{2}$$

The feature map F_1 has shape $(B, 256, H/2, W/2)$. For stage 2:

$$F_2 = ResidualBlock(F_1, W_{res2}, b_{res2}) \tag{3}$$

The feature map F_2 has shape $(B, 512, H/4, W/4)$. For stage 3:

$$F_3 = ResidualBlock(F_2, W_{res3}, b_{res3}) \tag{4}$$

The feature map F_3 has shape $(B, 1024, H/8, W/8)$. For stage 4:

$$F_4 = ResidualBlock(F_3, W_{res4}, b_{res4}) \tag{5}$$

The feature map F_4 has shape $(B, 2048, H/16, W/16)$.

4.3.2 Feature Maps of EfficientNetV2L

EfficientNetV2L uses a combination of depthwise separable convolutions and inverted residuals, which help in maintaining a good balance between feature map size and computational efficiency. Compact representa tion of the feature maps is the main advantage of using EfficientNetV2L model. Despite the complexity of the input images, the feature maps provide a compact representation of the relevant information required for classification. This enables efficient processing and inference with the help of the discriminative features. EfficientNetV2L maintains higher resolution in its feature maps for a longer range within the network and retain more spatial information. The feature maps are generally more diverse and detailed due to the ability

of the model to learn complex patterns with its scaled architecture. The combination of different types of convolutions helps in capturing finer details and more abstract features simultaneously. Feature maps generated from some of the layers of EfficientNetV2L is represented in Fig.**??**.

Consider input image is represented as a tensor (X with shape (B, C, H, W),in that B is the batch size, C is the number of channels, H is the height, and W is the width. The feature maps at each stage of the EfficientNetV2-L model can be represented as follows Initial Convolutional Layer:

$$F_{conv1} = ReLU(BatchNorm(Conv2d(X, W_{conv1}, b_{conv1}))) \tag{6}$$

In equation W_{conv1} and b_{conv1} are the weights and biases of the initial convolutional layer, respectively. The feature map F_{conv1} has shape $(B, 32, H/2, W/2)$ after the max-pooling layer.

MBConv and Fused-MBConv Blocks: EfficientNetV2-L consists of multiple stages, each containing multiple MBConv and Fused-MBConv blocks. The feature maps at the end of each stage i is represented as F_i.

For stage 1:

$$F_1 = MBConvBlock(F_{conv1}, W_{mbconv1}, b_{mbconv1}) \tag{7}$$

The feature map F_1 has shape $(B, 16, H/2, W/2)$.

For stage 2:

$$F_2 = FusedMBConvBlk(F_1, W_{fmbconv2}, b_{fmbconv2}) \tag{8}$$

The feature map F_2 has shape $(B, 24, H/4, W/4)$.

For stage 3:

$$F_3 = MBConvBlk(F_2, W_{mbconv3}, b_{mbconv3}) \tag{9}$$

The feature map F_3 has shape $(B, 40, H/8, W/8)$. For stage 4:

$$F_4 = FusedMBConvBlk(F_3, W_{fmbconv4}, b_{fmbconv4}) \tag{10}$$

The feature map F_4 has shape $(B, 80, H/16, W/16)$.

For stage 5:

$$F_5 = MBConvBlock(F_4, W_{mbconv5}, b_{mbconv5}) \tag{11}$$

The feature map F_5 has shape (B, 112, H/16, W/16).

For stage 6:

$$F_6 = FusedMBConvBlock(F_5, W_{fmbconv6}, b_{fmbconv6}) \tag{12}$$

The feature map F_6 has shape (B, 192, H/32, W/32).

For stage 7:

$$F_7 = MBConvBlock(F_6, W_{mbconv7}, b_{mbconv7}) \tag{13}$$

The feature map F_7 has shape (B, 320, H/32, W/32).

4.3.3 Selective Feature Fusion of Feature Maps

The feature fusion process involves linearly combining the feature maps from ResNet and EfficientNet models using appropriate weights. This allows the fused feature map to capture complementary information from both models, enhancing the overall representation and performance of the model. The feature fusion process can be represented mathematically as follows

$$F_{fusion} = \alpha \cdot F_R + \beta \cdot F_E \tag{14}$$

In the above equation, F_R denotes the feature maps produced by the Resnet50 model, F_E represents the feature maps generated by the EfficientNetV2L model and F_{fusion} is the fused feature map. α and β are weights assigned to the feature maps of Resnet50 and EfficientNetV2L, respectively.

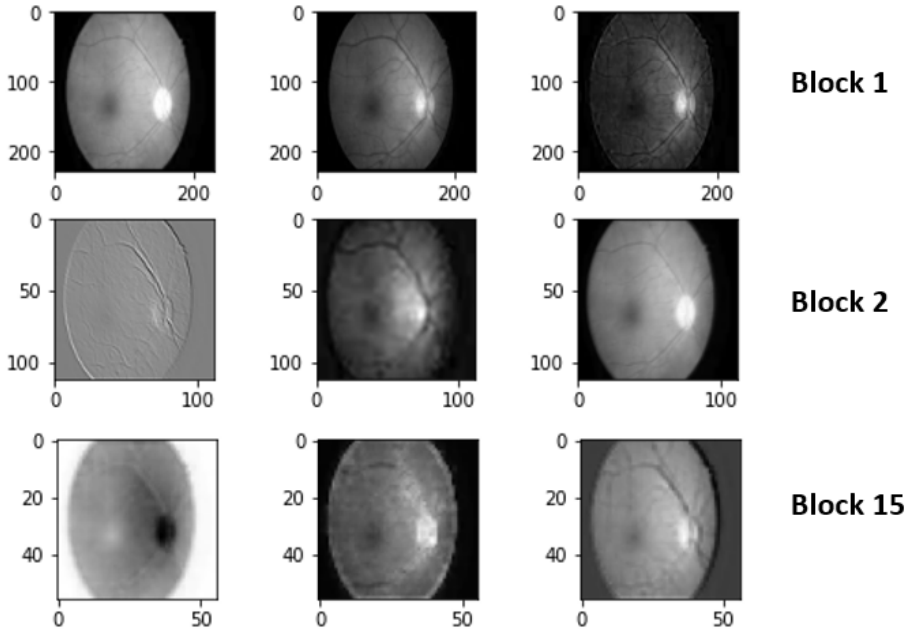


Figure 7: Feature maps created by the various blocks of Resnet50

The mathematical representation of concatenating the feature maps of ResNet50 and EfficientNetV2L mod els can be described as follows: Let $\mathbf{F}_{ResNet50} \in R^{H_1 \times W_1 \times D_1}$ represent the feature maps extracted from the ResNet50 model, where H_1 is the height, W_1 is the width, and D_1 is the depth (number of channels) of the feature maps. $\mathbf{F}_{EffNetV2L} \in R^{H_2 \times W_2 \times D_2}$ represent the feature maps extracted from the EfficientNetV2L model, where H_2 is the height, W_2 is the width, and D_2 is the depth (number of channels) of the feature maps. Before concatenation, the feature maps need to be aligned in terms of spatial dimensions. This can be achieved using resizing or pooling operations to ensure that $H_1 = H_2$ and $W_1 = W_2$.

Generally fusion models improve the performance but often comes at the cost of high memory usage and increased latency. A selective feature fusion method is proposed to addresses these challenges by reducing the computational overhead while preserving essential information. The output feature maps are projected to a smaller number of channels using 1×1 convolutions and which reduces the size of intermediate feature maps, lowering memory requirements and computational cost in subsequent fusion layers. Let $\mathbf{F}'_{ResNet50}$ and $\mathbf{F}'_{EffNetV2L}$ be the resized or pooled feature maps with dimensions $H \times W$. The hyperparameter $\mathbf{C}_{reduced}$ controls the compression ratio of the feature maps and used the value of 12 to align feature dimensions before fusion.

$$\mathbf{F}'_{ResNet50} = Conv2D(\mathbf{C}_{reduced}, 1 \times 1)(\mathbf{F}_{ResNet50}) \tag{15}$$

$$\mathbf{F}'_{EffNetV2L} = Conv2D(\mathbf{C}_{reduced}, 1 \times 1)(\mathbf{F}_{EffNetV2L}) \tag{16}$$

The projected feature maps are combined through element-wise addition, integrating information from both sources and maintains the number of channels constant. This reduces memory overhead and improves the realtime perofrmance.

The fused feature map $\mathbf{F}_{fused}$ can be represented as:

$$\mathbf{F}_{fused} = add(\mathbf{F}'_{ResNet50}, \mathbf{F}'_{EffNetV2L}) \tag{17}$$

Here, add denotes the element-wise addition operation and $\mathbf{F}_{fused} \in R^{H \times W \times (D_1)}$.

Global average pooling layer is applied to the fused convolutional feature map $\mathbf{F}_{fused}$ to reduce the spatial dimension to 1x1. This will help reduce the parameters and overfitting, which is a major challenge during the concatenation of feature maps generated by different CNN models.

$$\mathbf{F}_{gap} = GlobalAveragePooling2D()(\mathbf{F}_{fused}) \tag{18}$$

The output of the global average pooling layer is a 1D tensor with shape $(batchsize, 1280)$ and it is followed by a fully connected dense layer with 256 units and a ReLu activation function. The dense layer acts as the classification head and converts the feature maps to a 256-dimensional feature space. This transition of fused feature maps to a lower-dimensional feature space improves the performance of the classification. A dropout layer is also applied to avoid overfitting using regularization.

In the case of diabetic retinopathy detection, an additional dense layer with 2 units is placed on top of the previous dense layer. This layer maps the features to a 2-dimensional output space using the softmax activation function. The 2-dimensional output represents the normal and abnormal classes in the classification. Mathematically, the output of this final layer can be expressed as:

$$y = softmax(Wx + b) \tag{19}$$

In the above equation, y is the 2D output tensor of shape $(batchsize, 2)$, W is the weight matrix of shape $(2, 256)$ and b is the bias vector of shape $(2,)$. Subsequently, a new end-to-end model is created for training the fundus image dataset and performing the binary classification.

In the scenario of diabetic retinopathy grading, a dense layer containing 5 units is added on top of the preceding dense layer in the final stage. Utilizing softmax as the activation function, this layer maps the features into a 5-dimensional output space, representing the various grades of diabetic retinopathy. The output of this final layer can be expressed as:

$$y = softmax(Wx + b) \tag{20}$$

In the above equation, y is the 2D output tensor of shape$(batchsize, 5)$, W is the weight matrix of shape $(5, 256)$ and b is the bias vector of shape $(5,)$. After this, a new end-to-end model is created for training the fundus image dataset and performing the grading. The Adam optimizer is used for both binary and multiclass classification, with a learning rate of 0.001.

5 Results

The implementation and evaluation of the model are carried out by using the Python language environment on the challenging IDRiD dataset, which consists of 516 retinal images. Of these, 413 images are allocated for training the network and the remaining 103 images are used for validation purposes. Extensive experiments are carried out to evaluate the ability of the suggested model to detect diabetic retinopathy and determine its severity in fundus images. Additionally, the model is also evaluated on the DeepDRiD dataset to ensure the generalizability. High-end NVIDIA DGX A100 640 GB (8 x 80 GB GPUs), is used for training and evaluation. The experiments conducted in a system with sufficient memory capacity and GPUs ensured the proper training of the model. The trained model can be transfered to a system with less memory and this helps in the deployment of the model in clinical settings. The results are analysed using various key performance indicators to evaluate the effectiveness of the model in grading diabetic retinopathy.

5.1 Performance Evaluation Metrics

The efficiency of the model on the dataset is assessed by computing metrics such as accuracy, sensitivity, precision, and AUC. The number of correctly categorised normal and abnormal images by the model out of all

the images is called accuracy and it is typically employed as a performance measure. Categorical accuracy is used to assess the effectiveness of grading, whereas binary accuracy is used to assess the diagnosis of diabetic retinopathy. The percentage of anticipated values that coincide with actual values for binary labels is defined as binary accuracy. For one-hot labels, categorical accuracy determines the proportion of predicted values that coincide with real values.

The other often used measures are sensitivity and specificity which is helpful for comparison since many retinal image datasets have a problem with class imbalance. Sensitivity provides insight into correctly detected positive findings, and a high sensitivity rating reveals the reliability of the system. Another measure specificity identifies the accurately detected negative results. ROC is a curve that plots sensitivity against the false positive rate and the Area under the curve AUC is a measure of how well the model performs in classification. Precision and recall are important performance indicators in binary classification problems. Precision represents the proportion of correct positive findings among all positive predictions, whereas recall indicate the proportion of correct positive findings among all actual positives. These performance indicators are significant because they show how well the model is performing in terms of identifying true positives and minimising false positives or false negatives. The F1 score is a metric that combines precision and recall and it provides a balanced assessment of the performance of the model. The Cohen's Kappa (κ) and Geometric Mean are also computed to evaluate model performance in imbalanced datasets. The Kappa Score assesses the agreement between two classifiers while accounting for chance-based agreement. Meanwhile, the Geometric Mean ensures balanced performance across all classes, making it particularly useful when accuracy alone may be misleading. The formulas for computing the performance metrics mentioned above are as follows

$$Sensitivity = \frac{TP}{TP + FN} \tag{21}$$

$$Specificity = \frac{TN}{TN + FP} \tag{22}$$

$$Precision = \frac{TP}{TP + FP} \tag{23}$$

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{24}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{25}$$

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{26}$$

$$G - Mean = \sqrt{Sensitivity * Specificity} \tag{27}$$

True positive (TP) is the proportion of correctly classified abnormal images, and True negative (TN) is the proportion of correctly classfied normal images. False positive (FP) refers to the proportion of normal images that are incorrectly labelled as abnormal, and False negative (FN) refers to the proportion of pathological images that are incorrectly labelled as normal. In the kappa score, the observed agreement is denoted as p_o and p_e denotes the agreement by chance.

5.2 Diabetic Retinopathy Detection

The binary accuracy, sensitivity, Specificity, and AUC metrics are used to evaluate the classification performance for the binary classification problem. For training, the final network achieved an AUC value of 0.9705 and exhibited a binary accuracy of 95.24%, along with 95% precision and 95% recall. Subsequently, when validated against the 103 images in the test dataset, the system achieved an accuracy of 95.24%.

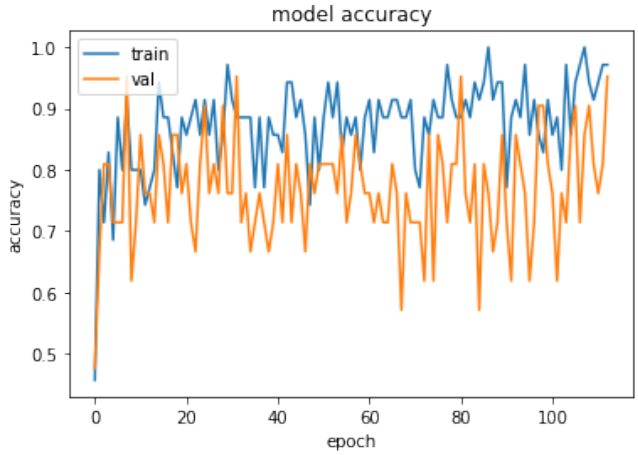


Figure 8: Line plots of training and validation accuracy values for the suggested CNN based Feature Fusion model in detection of diabetic retinopathy

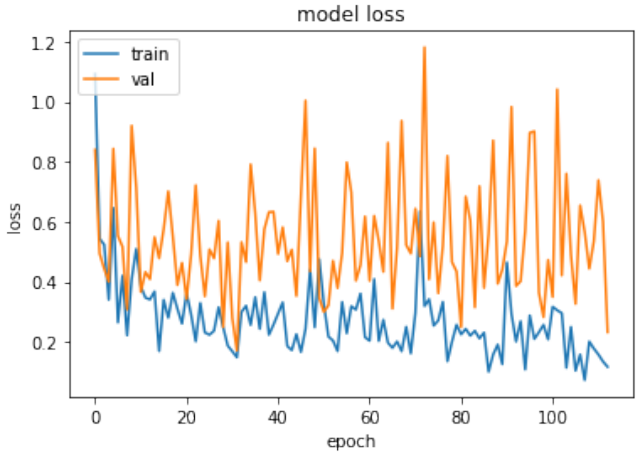


Figure 9: Line plots of the training and validation loss for the suggested CNN based Feature Fusion model in detection of diabetic rerinopathy

The AUC score of 0.9705, along with validation precision and sensitivity values of 95% each, indicates the robustness of the model. Furthermore, the model yields an F1-score of 0.95 upon evaluation on the IDRiD dataset.

The accuracy and loss data of the model for each epoch are recorded and plotted as a graph for analysis. The Fig.**??** illustrates the training and validation accuracy plotted against the number of epochs, while Fig.**??** displays the training and validation loss over the course of the epochs.

5.3 Diabetic Retinopathy Grading

For this multiclass classification task, we computed accuracy, sensitivity, and AUC values to assess the perfor mance of the system. For diabetic retinopathy grading, the model achieved a classification accuracy of 91.78% during training and achieved a validation accuracy of 92%. During validation on the IDRiD dataset, the model demonstrated a sensitivity of 88%, a specificity of 98%, and an F1-score of 0.89.

The Fig.**??** depicts the plot of training and validation accuracy against the number of epochs for the model. Meanwhile, Fig.**??** illustrates the plot of training and validation loss throughout the epochs for the model.

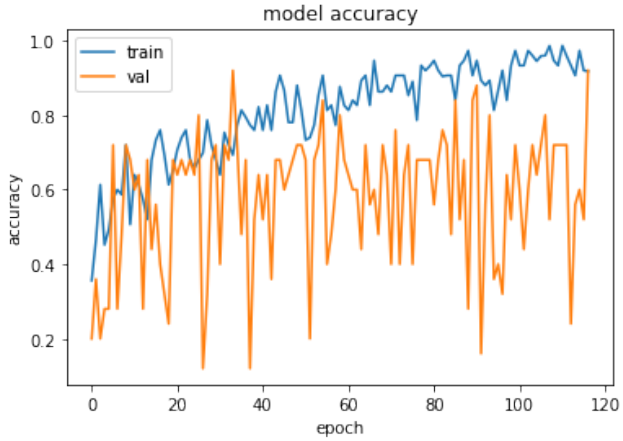


Figure 10: Line plots of the training and validation loss for the suggested CNN based Feature Fusion model in grading

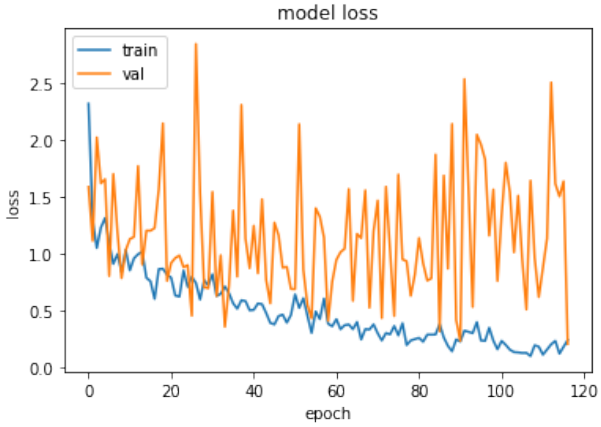


Figure 11: Line plots of the training and validation accuracy for the suggested CNN based Feature Fusion system in grading of diabetic retinopathy

6 Discussions

In this section, we conduct a detailed performance analysis of the developed classification model for diabetic retinopathy detection and grading. We compared the obtained results with those of other existing solutions to verify the effectiveness of the feature fusion model in diagnosis. Additionally, we compared the fused CNN model to the individual pretrained models utilized in its creation, using the evaluation measures for analysis. This comparative analysis indicates the significance of feature map fusion and elucidates how the fused model enhances the classification performance more effectively than the individual models used in the fusion.

Experiments were conducted to assess the performance of ResNet50 and EfficientNetV2L models in diabetic retinopathy detection. The EfficientNetV2L and Resnet50 models were trained and tested using the IDRiD dataset and EfficientNetV2L achieves higher traning accuracy. Specifically, EfficientNetV2L demonstrated a training accuracy of 94.29% and a validation accuracy of 80.95%, while ResNet50 attained a training accuracy of 91.43% and a validation accuracy of 80.95%. The comparison results between the feature fusion CNN model and the individual models in diabetic retinopathy detection are summarized in Table **??**. The findings indicate that the fused CNN model exhibited a 10% performance enhancement over both EfficientNetV2L and ResNet50.

Table 1: Performance comparison of the feature fusion CNN model and individual models in Diabetic Retinopa thy detection.

Method	Performance Evaluation Metrics				
	Accuracy	***Sensitivity***	***Specificity***	***AUC***	***F1 Score***
Resnet50	80.95	80.95	80.95	87.76	80.95
EfficientNetV2L	80.95	80.95	80.95	85.03	80.95
Feature Fusion Model	95.24	95	95	97.05	95

Table 2: Performance comparison of the feature map fusion model and individual models in Diabetic Retinopathy Grading.

Method	Performance Evaluation Metrics				
	Accuracy	***Sensitivity***	***Specificity***	***AUC***	***F1 Score***
Resnet50	63.89	62.5	-	87.86	94.05
EfficientNetV2L	90.25	91.42	89.58	-	94.05
Feature Fusion Model	92	88	98	94.05	94.05

The Fig.**??** displays the training and validation accuracy across epochs for ResNet50 in grading of diabteic retinopathy. The training and validation loss across epochs for Resnet50 is shown in Fig.**??**.

The effectiveness of ResNet50 and EfficientNetV2L models was assessed for diabetic retinopathy grading using the IDRiD dataset. The EfficientNetV2L model achieved a validation accuracy of 90.25%, with a sensi tivity of 91.42% and a specificity of 89.58%. On the other hand, the Resnet50 model achieved a lower validation accuracy of 63.89% and a sensitivity of 62.5% in grading. The comparison of the proposed feature fusion CNN model with individual models in diabetic retinopathy grading is outlined in Table **??**. The analysis indicates that the performance of the feature fusion model is better than ResNet50 and EfficientNetV2L models.

The model was also tested on the DeepDRiD dataset to confirm its generalizability. The results obtained from DeepDRiD dataset for diabetic retinopathy detection and grading are summarized in Table 3. The model achieved an accuracy of 92.86% in grading and 90.48% in diabetic retinopathy detection. These performance metrics demonstrate the effectiveness of the model in both detection and grading tasks. The model demonstrated consistent performance in the detection and grading irrespective of the datasets used.

Table 3: Performance of the feature fusion CNN model in DeepDRiD dataset.

Evaluation	Performance Evaluation Metrics					
	Accuracy	***Sensitivity***	***Specificity***	***ROCAUC***	***PRCAUC***	***F1 Score***
Diabetic Retinopathy detection	90.48	90.48	90.48	90.48	90.48	90.48
Grading	92.86	78.57	91.67	97.05	84.62	85

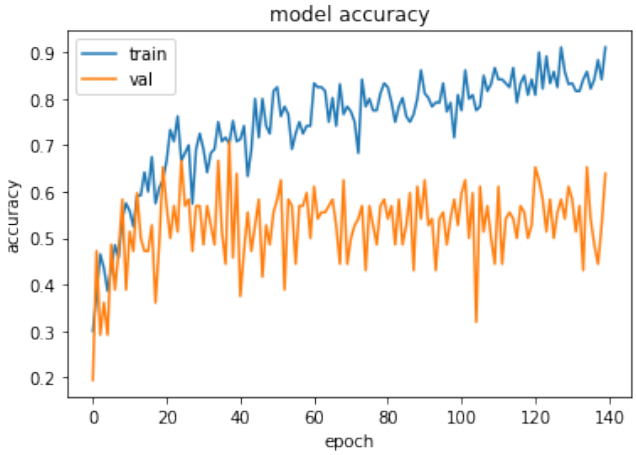


Figure 12: Line plots of the training and validation accuracy for Resnet50 in Diabetic Retinopathy Grading

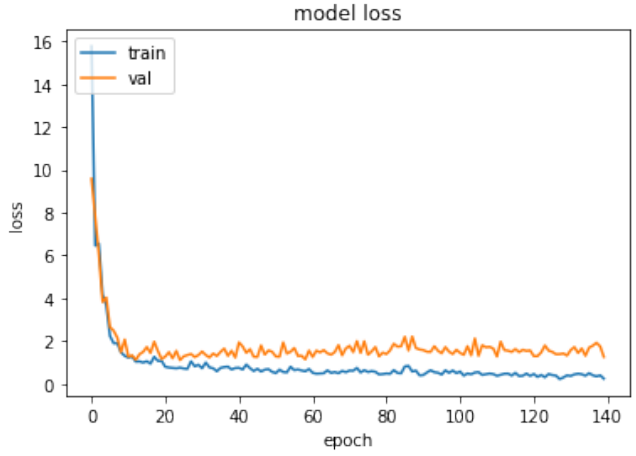


Figure 13: Line plots of training and validation loss for Resnet50 in grading

The Kappa Score and Geometric Mean are calculated for both datasets and presented in Table 4. These measures are commonly evaluated to verify the robustness of the model in managing class imbalance within the dataset. A high kappa score value, greater than 0.8 is desirable for the models used in medical diagnosis and score values, greater than 0.85, denote almost perfect agreement and ensure the efficiency of the fused model. Similarly, the high geometric mean values demonstrate the ability of the model to accurately identify all five categories while effectively handling class imbalance.

The graph in Fig.**??** illustrates the accuracy of training and validation for the grading of diabetic retinopathy using EfficientNetV2L, plotted against the epochs. Fig.**??** demonstrates the loss of training and validation over epochs for EfficientNetV2L.

From the analysis, it is evident that the efficiency of the feature map fusion model is better than the individual CNN models, both in diabetic retinopathy detection and grading. The F1 score findings reveal that the fused CNN model exhibits greater precision and recall values in both detection and grading compared to Efficient NetV2L and ResNet50. Additionally, in classification, the feature fusion model demonstrates higher sensitivity compared to the individual models.

The Table **??** presents a summary of the performance comparison among various methods for diabetic retinopa thy detection alongside our developed model. The comparison demonstrates that our CNN based feature fusion inference model performs better in diabetic retinopathy detection than other works. Through

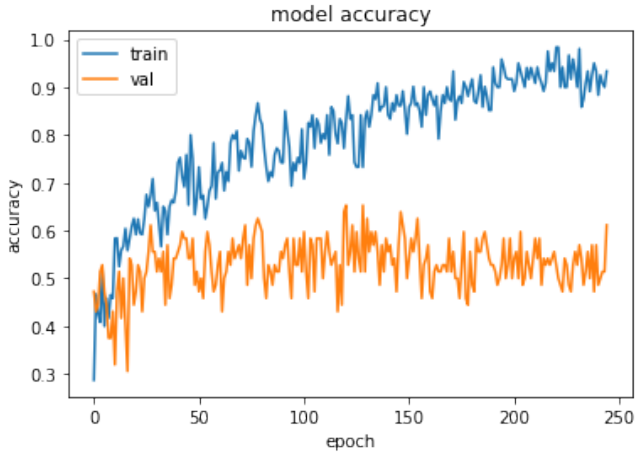


Figure 14: Line plots of training and validation accuracy values over epochs for EfficientNetV2L in Diabetic Retinopathy Grading

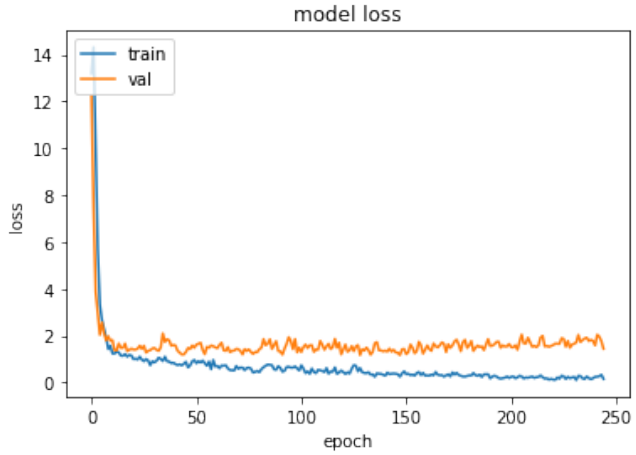


Figure 15: Line plots of training and validation loss over epochs for EfficientNetV2L in Grading

performance analysis, it is evident that our proposed model achieved the highest accuracy of 95.24% for diabetic retinopathy detection. Moreover, our model exhibited a higher sensitivity value of 95% compared to the

Table 4: kappa score and Geometric mean for Diabetic Retinopathy detection and grading on IDRiD and DeepDRiD datasets.

Evaluation	***Dataset***	Performance Evaluation Metrics	
		Kappa score	***Geometric Mean***
Diabetic Retinopathy Detection	DeepDRiD	0.90	0.86
Grading	DeepDRiD	0.89	0.84
Diabetic Retinopathy Detection	IDRiD	0.93	0.95
Grading	IDRiD	0.88	0.92

Table 5: Performance comparison of existing Diabetic retinopathy detection methods with the suggested CNN model using IDRiD dataset.

Reference	**Method**	Performance Evaluation Metrics			
		Accuracy	***Sensitivity***	***Specificity***	***AUC***
Bhardwaj et al. [**?**]	QEIRV-2 model	92.38	-	-	-
Elwin et al.[**?**]	Ar-HGSO-based deep learning	90.25	91.42	89.58	-
Das et al. [**?**]	EfficientNetB4	79.11	-	-	-
Ours	Feature map fusion model	95.24	95	95	0.9705

CNN model uti lizing the Ar-HGSO optimization algorithm, which demonstrated a sensitivity of 91.42%. The high sensitivity value indicates that it is capable of performing effectively on new data that it has not previously encountered, thereby ensuring the generalizability of the model.

The performance comparison of various existing diabetic retinopathy grading methods with the suggested feature fusion model using the IDRiD dataset is summarized in Table **??**. Feature fusion model have superior performance over traditional machine learning methods like neural networks, SVM, and decision trees in terms of accuracy and sensitivity for grading. Specifically neural network method used by Kirange et al.[**?**] achieved a sensitivity of 69.11% and SVM method achieved a sensitivity of 81%. In contrast, the feature fusion model achieved a sensitivity of 88%, significantly higher than the traditional methods and indicates that the model can minimize the number of false negatives. A high value of sensitivity is a demanding criteria in medical diagnosis, where a missing positive instance could have serious consequences. The proposed feature fusion model also significantly improves accuracy compared to other deep learning based methods for grading. The

Table 6: Performance comparison of existing Diabetic retinopathy grading methods with the suggested fused CNN model using IDRiD dataset.

Reference	**Method**	Performance Evaluation Metrics		
		Accuracy	***Sensitivity***	***Specificity***
Kirange et al.[**?**]	Neural Networks	65.95	69.11	-
Elwin et al.[**?**]	Ar-HGSO-based deep learning	91.42	92.54	91.42
Kirange et al. [**?**]	SVM	74.46	81	-
Zhan et al. [**?**]	CF-DRNet	56.19	64.21	87.39
Acharya et al.[**?**]	SVM with morphological features	85.99	82	86
Sakaguchi et al.[**?**]	Graph Neural Network	79.3	-	-
Kirange et al.[**?**]	Decision Trees	65.95	70.72	-
Ours	Feature map fusion model	92	88	98

main reason for performance improvement of the feature fusion model is the ability to capture a robust feature representation. The comparison shows that our suggested feature fusion model performs better in the grading of diabetic retinopathy compared to other approaches.

Performance analysis ensures that the developed method can be utilized in a beneficial manner to assist ophthalmologists in screening of diabetic retinopathy. The suggested CNN based feature fusion model demon strates remarkable efficiency in the challenging IDRiD dataset.

The proposed selective feature fusion model reduce the computational overhead introduced during feature fusion, thereby supporting the real-time deployment of the model. A comparative summary of memory requirements between the conventional feature fusion model and the selective feature fusion model is presented in Table 8 for diabetic retinopathy detection and in Table 9 for grading of diabetic retinopathy. The tables demonstrate that the selective feature fusion model reduces computational overhead compared to the baseline fusion model.

The selective fusion model needs only fewer trainable parameters (922.68 KB or 923.86 KB) compared to the baseline fusion model (1.37 MB or 1.371 MB) for both IDRiD and DeepDRiD datasets. This indicates that the selective fusion model has a smaller memory footprint, which directly reduces computational overhead. Selective fusion consistently uses less memory (132.02 MB or 132.48 MB) for both datasets and this ensures that the model can be run more easily on resource-limited hardware. The slective fusion model also achieves lower latency compared to the baseline model. The improvements of the selective feature fusion model in terms of trainable parameters, memory requirement and inference latency lead to reduced computational overhead for model deployment, supporting real-time and resource-efficient screening in clinical settings.

To ensure the real-time applicability of the model, average inference latency per image is computed on different hardware platforms: a high-performance GPU (NVIDIA A100), a resource-constrained environment (NVIDIA T4), and an Intel Xeon CPU-only system. The results are summarized in Table 8 for diabetic retinopathy detection and in Table 9 for diabetic retinopathy grading. The average inference time per image for both the IDRiD and DeepDRiD datasets is computed, and the values demonstrates that the models are suitable for real-time diagnosis.

Table 7: Comparison of memory utilization and latency between the baseline fusion model and the proposed selective feature fusion approach for Diabetic Retinopathy detection

Dataset	***Model***	**Trainable Parameters**	**Memory (MB)**	**Latency (s)**		
				Tesla T4	**Intel Xeon CPU**	**NVIDIA A100**
IDRiD	Baseline Fusion	1.37	132.48	0.19	2.05	0.15
	Selective Fusion	922.68 KB	132.02	0.17	1.38	0.11
DeepDRiD	Baseline Fusion	1.371 MB	320.48	0.45	2.03	0.31
	Selective Fusion	922.68	132.08	0.38	1.43	0.27

The clinical real-time is usually considered less than 1 second per image for interactive screening and both GPUs are within real-time range(0.1–0.4s per image) for diabetic retinopathy detection. For diabetic retinopathy grading, NVIDIA A100 GPU achieves the lowest latency (0.25–0.57s), and Tesla T4 GPU is a cost-effective option, still within the real-time range.

The analysis shows that the machine with GPU support provide real-time deployment capabilities for both diabetic retinopathy detection and grading. CPU-only deployment is acceptable for near real time diagnosis and it can be further optimised by applying techniques such as model compression, channel pruning, quantization and architecture optimization. While the model compression technique optimising the model by pruning and

Table 8: Comparison of memory utilization and latency between the baseline fusion model and the proposed selective feature fusion approach for Diabetic Retinopathy Grading

Dataset	***Model***	**Trainable Parameters**	**Memory (MB)**	**Latency (s)**		
				Tesla T4	**Intel Xeon CPU**	**NVIDIA A100**
IDRiD	Baseline Fusion	1.37 MB	132.49	0.99	2.82	0.64
	Selective Fusion	923.86 KB	132.02	0.91	1.95	0.57
DeepDRiD	Baseline Fusion	1.37 MB	132.49	0.62	2.43	0.38
	Selective Fusion	923.86 KB	132.02	0.48	1.79	0.25

quamitisation, channel pruining method removes or compresses less effective convolutional filters to reduce model complexity and preserving accuracy.

7 Conclusion and Future Scope

The application of deep learning-based computer-aided diagnosis systems presents a promising approach for the early detection and grading of diabetic retinopathy using fundus images. In this study, a CNN-based selective feature fusion architecture leveraging ResNet50 and EfficientNetV2L models is suggested to enhance the diagnosis of diabetic retinopathy. The framework combines the feature maps generated by Resnet50 and EfficientNetV2L and additional layers are added after the fusion process to improve the effeciency of the model. By fusing fea tures extracted from different models, the feature fusion process can capture more diverse and complementary information. This allows the model to learn a richer and more discriminative representation of the data, leading to improved performance. The suggested feature fusion architecture produces reliable results when applied to the publicly available IDRiD dataset. The model achieved an accuracy of 95.24% in diabetic retinopathy detection and 91.74% accuracy in grading. The performance of the developed model in the DeepDRiD dataset ensure the generalizability of the model and the results ensured that the model demonstartes consistency in detection and grading irrespective of the datasets used. Notably, the fused model performs better than alternative methods in both detection and grading within the IDRiD dataset. Moreover, comparative analysis reveals supe rior performance of the feature fusion model compared to individual models used for feature map generation. Although the proposed selective feature fusion model exhibits exceptional performance and reduces the computational overhead, there is still scope for improvements in feature selection and reduction of memory requirements. Furthermore, efforts should focus on optimizing the proposed method for real-time application to retinal images while maintaining high performance levels observed in standard dataset evaluations. Different lightweight alternatives, such as attention-based fusion and selective channel pruning, can also be used instead of selective feature fusion. These improvements will enhance the utility and effectiveness of the feature fusion model in clinical settings, ultimately contributing to improved detection of diabetic retinopathy.

8 Acknowledgments

The authors would like to thank Centre for Engineering Research and Development (CERD), Government of Kerala, India for granting the research fellowship.

9 Data Availability

The data that support the findings of this study are openly available in IDRiD at https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid and the DeepDRiD dataset at https://github.com/deepdrdoc/DeepDRiD.

References

[1] Diabetic retinopathy screening: a short guide. Increase effectiveness, maximize benefits and minimize harm. Copenhagen: WHO Regional Office for Europe; 2020

[2] Bourne, Rupert, Adelson, Trends in Prevalence of Blindness and Distance and Near Vision Impairment Over 30 Years and Contribution to the Global Burden of Disease in 2020 ,Lancet Glob Health,2020.

[3] Tsiknakis N, Theodoropoulos D, Manikis G, Ktistakis E, Boutsora O, Berto A, Scarpa F, Scarpa A, Fotiadis DI, Marias K. Deep learning for diabetic retinopathy detection and classification based on fundus images: A review. Comput Biol Med. 2021 Aug;135:104599. doi: 10.1016/j.compbiomed.2021.104599. Epub 2021 Jun 25. PMID: 34247130.

[4] Padhy S.K, Takkar B, Chawla R, Kumar A, Artificial intelligence in diabetic retinopathy: A natural step to the future. Indian J Ophthalmol. 2019 Jul;67(7):1004-1009. doi: 10.4103/ijo.IJO_1989_18. PMID: 31238395; PMCID: PMC6611318.

[5] Prasanna P, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, Fabrice Meriaudeau, April 24, 2018, "Indian Diabetic Retinopathy Image Dataset (IDRiD)", IEEE Dataport, doi: https://dx.doi.org/10.21227/H25W98.

[6] Wilkinson CP, Ferris FL 3rd, Klein RE, Lee PP, Agardh CD, Davis M, Dills D, Kampik A, Pararajasegaram R, Verdaguer JT; Global Diabetic Retinopathy Project Group. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology. 2003 Sep;110(9):1677-82. doi: 10.1016/S0161-6420(03)00475-5. PMID: 13129861.

[7] Gazala Mushtaq and Farheen Siddiqui,"Detection of diabetic retinopathy using deep learning methodology", IOP Conference Series: Materials Science and Engineering, 1070 (2021) 012049. https://doi.org/10.1088/1757-899X/1070/1/012049.

[8] Bilal A, Zhu L, Deng A, Lu H, Wu N. AI-Based Automatic Detection and Classification of Diabetic Retinopathy Using U-Net and Deep Learning. Symmetry. 2022; 14(7):1427. https://doi.org/10.3390/sym14071427

[9] S. H. Abbood, H. N. A. Hamed, M. S. M. Rahim, A. Rehman, T. Saba and S. A. Bahaj, "Hybrid Retinal Image Enhancement Algorithm for Diabetic Retinopathy Diagnostic Using Deep Learning Model," in IEEE Access, vol. 10, pp. 73079-73086, 2022, doi: 10.1109/ACCESS.2022.3189374.

[10] Momeni Pour, A., Seyedarabi, H., Abbasi Jahromi, S.H., & Javadzadeh, A. (2020). Automatic Detection and Monitoring of Diabetic Retinopathy Using Efficient Convolutional Neural Networks and Contrast Limited Adaptive Histogram Equalization. IEEE Access, 8, 136668-136673. https://doi.org/10.1109/ACCESS.2020.3005044.

[11] Li X, Jiang Y, Zhang J, Li M, Luo H, Yin S. Lesion-attention pyramid network for diabetic retinopathy grading. Artif Intell Med. 2022. Apr 126:102259. doi: 10.1016/j.artmed.2022.102259. Epub 2022 Feb 25. PMID: 35346445. https://doi.org/10.1016/j.artmed.2022.102259.

[12] AbdelMaksoud, E., Barakat, S.I.,& Elmogy, M.M. Diabetic Retinopathy Grading System Based on Transfer Learning. 2020. ArXiv, abs/2012.12515. https://doi.org/10.1109/ICDABI51230.2020.9325672.

[13] Elswah, D.K., Elnakib, A., & El-Din Moustafa, H. Automated Diabetic Retinopathy Grading using Resnet. 2020 37th National Radio Science Conference (NRSC), 248-254. https://doi.org/10.1109/NRSC49500.2020.9235098.

[14] Luo L, Xue D, Feng X. Automatic Diabetic Retinopathy Grading via Self-Knowledge Distillation. Electronics. 2020; 9(9):1337. https://doi.org/10.3390/electronics9091337

[15] Bhardwaj C, Jain S, Sood M. Deep Learning-Based Diabetic Retinopathy Severity Grading System Employing Quadrant Ensemble Model. J Digit Imaging. 2021 Apr;34(2):440-457. doi: 10.1007/s10278-021-00418-5. Epub 2021 Mar 8. PMID: 33686525; PMCID: PMC8289963.

[16] Das D, Biswas SK, Bandyopadhyay S. Detection of Diabetic Retinopathy using Convolutional Neural Networks for Feature Extraction and Classification (DRFEC). Multimed Tools Appl. 2022 Nov 29:1-59. doi: 10.1007/s11042-022-14165-4. Epub ahead of print. PMID: 36467440; PMCID: PMC9708148.

[17] Lin CL, Wu KC. Development of revised ResNet-50 for diabetic retinopathy detection. BMC Bioinformatics. 2023 Apr 19;24(1):157. doi: 10.1186/s12859-023-05293-1. PMID: 37076790; PMCID: PMC10114328.

[18] Elwin JGR, Mandala J, Maram B, Kumar RR Ar-HGSO: Autoregressive-Henry gas sailfish optimization enabled deep learning model for diabetic retinopathy detection and severity level classification. Biomed Signal Process Control 2022 77:103712. https://doi.org/10.1016/j.bspc.2022.103712

[19] Liu R, Wang X, Wu Q, Dai L, Fang X, Yan T, Son J, Tang S, Li J, Gao Z, Galdran A, Poorneshwaran JM, Liu H, Wang J, Chen Y, Porwal P, Wei Tan GS, Yang X, Dai C, Song H, Chen M, Li H, Jia W, Shen D, Sheng B, Zhang P. DeepDRiD: Diabetic Retinopathy-Grading and Image Quality Estimation Challenge. Patterns (N Y). 2022 May 20;3(6):100512. doi: 10.1016/j.patter.2022.100512. PMID: 35755875; PMCID: PMC9214346.

[20] Organization W.H. Report of a WHO consultation. WHO; Geneva, Switzerland: 2005. Prevention of blindness from diabetes mellitus; pp. 1–48.

[21] Kukkonen H, Rovamo J, Tiippana K, Näsänen R. Michelson contrast, RMS contrast and energy of various spatial stimuli at threshold. Vision Res. 1993 Jul;33(10):1431-6. doi: 10.1016/0042-6989(93)90049-3. PMID: 8333164.

[22] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[23] He, Kaiming et al. "Deep Residual Learning for Image Recognition." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 770-778. https://doi.org/10.1109/CVPR.2016.90.

[24] Tan, Mingxing and Quoc V. Le. EfficientNetV2: Smaller Models and Faster Training. ArXiv abs/2104.00298 (2021): n. pag.

[25] Kirange, D.K., Chaudhari, J.P., Rane, K.P., Bhagat, K.S.& Chaudhri, N. Diabetic retinopathy detectionand grading using machine learning. Int. J. of Adv.Trends in Comput. Sci. & Engine., 2019. 8(6), 3570-3576.https://doi.org/10.30534/ijatcse/2019/139862019

[26] Sakaguchi, A., Wu, R., & Kamata, S. I. (2019). Fundus image classification for diabetic retinopathy using disease severity grading. In Proceedings of the 2019 9th International Conference on Biomedical Engineering and Technology, ICBET 2019 (pp. 190-196). (ACM International Conference Proceeding Series). Association for Computing Machinery. https://doi.org/10.1145/3326172.3326198

[27] Wu Z, Shi G, Chen Y, Shi F, Chen X, Coatrieux G, Yang J, Luo L, Li S. Coarse-to-fine classification for diabetic retinopathy grading using convolutional neural network. Artif Intell Med. 2020 Aug;108:101936. doi: 10.1016/j.artmed.2020.101936. Epub 2020 Jul 24. PMID: 32972665.

[28] U.R. Acharya, C.M. Lim, E.Y.K. Ng, C. Chee, T. Tamura, Computer-based detection of diabetes retinopathy stages using digital fundus images, Proc. Inst. Mech. Eng. H J. Eng. Med. 223 (5) (2009) 545–553. https://doi.org/10.1243/09544119JEIM486.