TPM Vol. 32, No. S4, 2025
ISSN: 1972-6325
https://www.tpmap.org/

DEEP LEARNING-BASED EARLY DETECTION OF DIABETIC RETINOPATHY USING FUNDUS IMAGES

[1]DR. T. KUMARESAN, [2]RAMYA. R, [3]DR.S.ZULAIKHA BEEVI, [4]MALINI. M , [5]DINESHBABU. V, [6]K. SARANYA

[1]PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
KGISL INSTITUTE OF TECHNOLOGY, COIMBATORE
speak.kumaresh@gmail.com

[2]ASSOCIATE PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
BANNARI AMMAN INSTITUTE OF TECHNOLOGY, SATHYAMANGALAM-638401, ERODE.
mail id: ramyarv@bitsathy.ac.in

[3]PROFESSOR, DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
VEL TECH HIGH TECH DR.RANGARAJAN DR.SAKUNTHALA ENGINEERING COLLEGE
AVADI, CHENNAI-600062, TAMIL NADU, INDIA
zulaikhabeevi.s@velhightech.com

[4]ASSOCIATE PROFESSOR AND HEAD, DEPARTMENT OF COMPUTER SCIENCE AND BUSINESS SYSTEMS
AKSHAYA COLLEGE OF ENGINEERING AND TECHNOLOGY, COIMBATORE
TAMILNADU
e mail: malini.laks@gmail.com

[5]ASSISTANT PROFESSOR, DEPARTMENT OF INFORMATION TECHNOLOGY,
KARPAGAM INSTITUTE OF TECHNOLOGY, COIMBATORE, TAMILNADU
dineshbabukit@gmail.com

[6]ASSOCIATE PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, BANNARI AMMAN INSTITUTE OF TECHNOLOGY,
SATHYAMANGALAM, ERODE, TAMILNADU, INDIA
email: ksaranyacse@gmail.com

Abstract:

This paper proposes a deep learning model that together with high-resolution fundus images can be used to detect diabetic retinopathy (DR) at an early stage. The efficiency of the proposed architecture is to utilize EfficientNetv2 to extract representation in a usefully robust fashion and a Vision Transformer (ViT) to incorporate long-range spatial contextual information to improve lesion localization and classification performance. Data augmentation and gradually resizing images during the training process in the model helps to fix the class imbalance to enhance generalization. Besides, it is combined with Grad-CAM to enable interpretability by visual recognition of areas affecting the predictions. It has been trained and tested on more than 100 000 fundus images gathered on EyePACS and Messidor-2. It attains a high classification performance (AUC-ROC, 96.3%, sensitivity, 94.1%, specificity, 92.8%) impressive than the standard CNN epitomes like ResNet 50 and DenseNet 121. Qualitative Grad-CAM heatmaps validate the value of the model to identify clinically relevant areas that validates its capability of applications in transparent and scalable DR screening in real-world low-resource scenarios.

Keywords: Diabetic Retinopathy (DR), Deep Learning, EfficientNetV2, Vision Transformer (ViT), Grad-CAM, Automated Screening.

I. INTRODUCTION

DR is a major cause of blindness and the people mainly affected by it are those with diabetes[1]. Analysis may help to save vision when DR is diagnosed early. The use of fundus images in the diagnosis of DR is widespread because fundus images are non invasive and they give detailed images of the retina. Nevertheless, the manual examination of these pictures needs well-trained ophthalmologists, which results in the high working burden, especially in low-resource environments. The complexity of the images as well as the necessity to provide timely and correct diagnoses has contributed to the development of automated systems. Automated DR screening is still a combative issue,

notwithstanding the improvement in medical imaging, and machine learning because of the range of images, noise, and lesion detection, under realization of accurate feature extraction[2].

Fundus images, that are necessary to diagnose DR, frequently present various obstacles. The high variability of the lighting, resolution, and quality of the images, together with a variety of types of lesions (microaneurysms, hemorrhages, exudates, etc.), make the process of the analysis difficult[3]. The pictures also have rich lesion patterns scattered over the vast areas, which makes their segmentation and analysis to interpret them effectively hard. Also, artifacts like shadows, reflections, patterns of vessel must be added to the task[4]. The challenges necessitate the need to come up with more sophisticated algorithms that can easily manage such variations so that DR could easily be detected.

Most traditional approaches to the detection of DR normally entail four stages: segmentation, feature extraction and classification[5]. Thresholding or region-growing are most common segmentation techniques that are applied to distinguish the backgrounds and lesions. Nonetheless, these techniques are challenged when dealing with complicated lesion and low-differential areas. Classical systems are more based on handcrafted features, i.e. texture-, color- or shape-based descriptors which are prone to changes in imaging condition and may not be able to pick up every detail present in lesions[6]. The decision trees or support vector machine have been used to classify the images into the various levels of DR; however, the problem of such methods is their lack of generalization ability and an inability to respond adequately to the complicated nature of patterns in the DR images.

The key problems with the traditional techniques are that they are based on the hand-made features and cannot reflect complex relations in the space in the images. Segmentation techniques can easily miss a faint or small lesion and they can also be noisy[7]. Hand-selected attributes as a source of feature extraction methods find it hard to generalize with a variety of data, and also neglect to recognize the complex structure within the images. In addition, the classification techniques employed in the previous periods were rather basic, and their accuracy considerably diminishes when applied on bigger more elaborate packages of data. These shortcomings have motivated the research into more powerful and data-driven techniques that might be taught to spot features and identify images efficiently with no need of formulating features manually[8].

The current techniques used in detecting DR are not scalable, accurate and robust. The major establishing requirement is the need to have a more advanced method that could be used to manage complexity and diversity of available fundus images. The current proposed work fills in such gaps by employing the concept of deep learning, that can learn in a way such that it can automatically learn to identify discriminative features in large amount of datasets and makes adjustments to the discrepancies that exist in the images[9].The combination of the hybrid approach, such as the use of EfficientNetV2 to extract features and Vision Transformer (ViT) to capture the global trends over the whole image, enhancing the spatial attention, enables the model to learn both the fine-grained stuff and the global trends over the whole image, which greatly increases the accuracy of DR detection. The method conveniently avoids manual feature selection and sample segmentation and is more practical than time consuming.

The offered work has a well-planned path of creating an automated DR screening system. First, the issue of DR detection is presented, and the issues of the traditional approach and furthermore more advanced solutions are raised. The following step includes explaining the training data which are highly-resolution fundus images with DR severity labels. In the methodology section, the architecture of the proposed hybrid model is described to integrate the EfficientNetV2 to obtain the feature extraction and Vision Transformer (ViT) to avoid the limitation of using a single architecture in feature extraction and spatial attention. The training process that involves data augmentation and progressive resizing is addressed to reduce the imbalance in the classes and optimize the model performance. In the results section, the performance of the proposed model is compared to other traditional approaches and an explanation of the interpretability of the model using Grad-CAM is provided. Lastly, the effects of the solution are evaluated, highlighting the mechanism of enhancing a more efficient DR screening, especially in situations where there would be a resource shortage.

General contributions of the proposed work

1. Efficient netv2 vision transformer (vit) and hybride deep learning model improved lesion detection and spatial attention are located in only one region.

2. Usage of progressive resizing and data augmentation to deal with the problem of class imbalance in the case of the severity of the disease of the eye DR.

3. Model interpretability with Grad-CAM integration, which points out the most interesting pathological regions in fundus image.

4. Performing at state-of-the-art levels, 96.3% AUC-ROC, they can screen eyes from resource-constrained settings automatvely and cheaply.

In section II, the current approaches to diabetic retinopathy (DR) detection are reviewed with an emphasis on the traditional methods and their weaknesses, i.e. they are based on handcrafted features and use manual segmentation. In Section III the proposed hybrid model integrating EfficientNetV2 as a feature extractor and vision transformer (ViT) to achieve a better spatial attention is introduced and the workflow of a lesion detector and the classification of the severity of DR is described. Section IV presents the outcomes of the proposed model and the comparative analysis of the model with the other deep learning models among which the proposed model performs truly better in the scope of the AUC-ROC, sensitivity, and specificity. The work ends with the Section V where the influence of the suggested solution of the automated DR screening is summed up, and possible areas of its further advancement and application are proposed.

II. LITERATURE REVIEW

Recent publications on the subject of diabetic retinopathy (DR) detection point to the currently advanced domain of image processing, which confronts image quality and lesion complexity changes with problematic traditional approaches and entails the transition to the sphere of deep learning. CNNs have been proven to be very helpful in MD detection process automatization. Performance has also been further optimized by hybrid models, including the work of combining EfficientNetV2 on feature extraction and the Vision Transformers on spatial attention. Even though these advancements have been made there are still some challenges such as class imbalance, size of dataset, and interpretability of models. These issues are still actively researched in the quest of more efficient and accurate screening systems of DR.

The study by Nunez do Rio et al. (2023) demonstrated a deep learning framework to determine diabetic retinopathy (DR) in handheld non-mydriatic retinal images taken in community-based settings by field workers[11]. The authors applied a convolutional neural network (CNN) to the detection of DR to point at the viability of this approach in terms of practicality in real-word conditions, particularly those with remote and resource-starved nature. As ascertained by Mohan et al. (2023), a friended learning-based framework named DRFL was proposed that grades diabetic retinopathy based on the fundus images of diseases[12]. The methodology allows the decentralized learning through several hospitals and the data privacy, as it does not imply exchanging the sensitive information about the patients. The federated style of learning enables the model to learn the data that is available in a variety of sources but is able to assure the privacy of the patient.

In another study, Palaniswamy and Vellingiri (2023) discussed the applicability of IoT and deep learning in the diagnosis of diabetic retinopathy in retinal fundus images[13]. Fundus images are captured, gathered, and transferred to the system which is integrated with IoT to analyze the data with deep learning models. In this paper, Alyoubi et al. (2021) created a deep learning system using fundus images of diabetic retinopathy to classify those images and localize lesions. The system employs the use of a CNN to classify the severity of DR as well as it has the capability of locating the lesions[14]. This work may be considered a major advantage because it serves both to classify and localize lesion, which would give more detailed information about the disease.

Krishnasamy et al. (2022) suggested an approach to how diabetic retinopathy can be identified with the help of retinal fundus photographs. The model applies the image classification deep learning methodology to predict whether the presence and severity of DR exist[15]. The research is important in the fact that it demonstrated the validity of deep learning models in identifying DR with less preprocessing. A hybrid retinal image enhancement algorithm provided by Abbood et al. (2022) is aimed at assisting the diagnosis of diabetic retinopathy[16]. This model involves a deep learning framework coupled with the picture enhancement of retinal fundus images aiming to enhance knowledge in its classification.

Kolla and Venugopal (2021) employed a binary CNN to perform classification of diabetic retinopathy in fundus images with the simpler architecture in mind to ensure efficient classification[17]. The merit of this approach is the possibility to offer relatively quick and resource-conserving solution to the DR detection, which is vital in clinical settings. Khalid et al. (2022) designed an ensemble method based on deep learning to be used in the assessment of diabetic retinopathy in fundus pictures[18]. In this approach, several different models are applied to enhance DR detection robustness and effectiveness.

Moustari et al. (2024) designed a two stage Deep learning classification model that was trained on diabetic retinopathy and recorded a high accuracy of 87.62 and 90.03 percent with gradient-weighted class activation maps (Grad-CAM)[19]. The two-phase method enables the model to determine the severity of DR, and after that, they can optimize the diagnosis by Grad-CAM, in which high-power areas in fundus image can be pinpointed. Mohamed et al. (2021) suggested a better solution to automatic grading of diabetic retinopathy based on deep learning and principal component analysis (PCA)[20]. It relies on the methodology because it combines PCA to bring the dimensionality of the fundus images down and then use a deep learning model to classify those images in the grading processes.

Table 1: Comparison Study of the existing models

S.No	Author(s) et al. (Year)	Dataset	Methodology	Accuracy (%)	Challenges
1	Palaniswamy & Vellingiri (2023)	Retinal fundus images with IoT integration	IoT-based deep learning model for DR diagnosis	88.5	Internet connectivity, image quality issues
2	Nunez do Rio et al. (2023)	Non-mydriatic handheld retinal images from community settings	CNN-based model for DR detection on non-mydriatic images	90	Image quality variation, remote processing needs
3	Moustari et al. (2024)	Fundus images with Grad-CAM	Two-stage deep learning model with Grad-CAM for classification	89	Processing speed due to two-stage system
4	Mohan et al. (2023)	Fundus images from federated learning	Federated learning for DR grading	92.8	Data heterogeneity, model convergence
5	Mohamed et al. (2021)	Fundus images with PCA	Deep learning with PCA for DR grading	86.5	Dimensionality reduction may lose subtle details
6	Krishnasamy et al. (2022)	Retinal fundus images	Deep learning for DR detection	90	Image quality, computational complexity
7	Kolla & Venugopal (2021)	Fundus images	Binary CNN for efficient DR classification	87	Model simplicity, performance on complex cases
8	Khalid et al. (2022)	Fundus images	Ensemble deep learning model for DR assessment	91	Computational complexity of ensemble models
9	Alyoubi et al. (2021)	Fundus images	Deep learning-based classification and lesion localization	85	Annotation accuracy, image quality issues
10	Abbood et al. (2022)	Fundus images	Deep learning CNN for DR classification and lesion localization	89.5	Data annotation quality, generalization

The first issue with the existing approaches to detecting traditional diabetic retinopathy (DR) is that such methods are based on hand-crafted features and manual lesion segmentation, which results in low detection levels of lesions and poor generalization in other datasets of different coloration. The major issues faced by these approaches are variations in quality of the image, class imbalances and their failure to represent complex spatial relationships in fundus images. The given solution will solve these problems because it proposes a hybrid deep learning model that will enhance the efficiency of feature extraction and spatial attention by incorporating a robust feature selection framework of EfficientNetV2 and a powerful spatial consideration framework of Vision Transformer (ViT). The method will automatically learn discriminative features on large datasets and does not rely as much on manual input to provide increased accuracy during the detection of DR lesions, even in low-quality or imbalanced images.

III. Proposed work

The Figure 1, shows the end to end working process of diabetic-retinopathy grading. The deep feature maps are generated by the EfficientNetV2 over a fundus image. This set of maps is flattened into patch embeddings before being aggregated into a sequence of tokens (class token + patch tokens along with positional encodings). The sequence then passes through one or more Vision-Transformer attention blocks that learns long-range spatial interactions, and finally passes through a transformer-based classifier to give the grade of the DR severity. Simultaneously, the classification layers back-propagate the gradients all the way to the convolutional features so that Grad-CAM can produce the coloured saliency map at the upper right; this heat-map sharpens where in the retina Grad-CAM considers to be most influential towards the final decision, which clinicians can interpret.

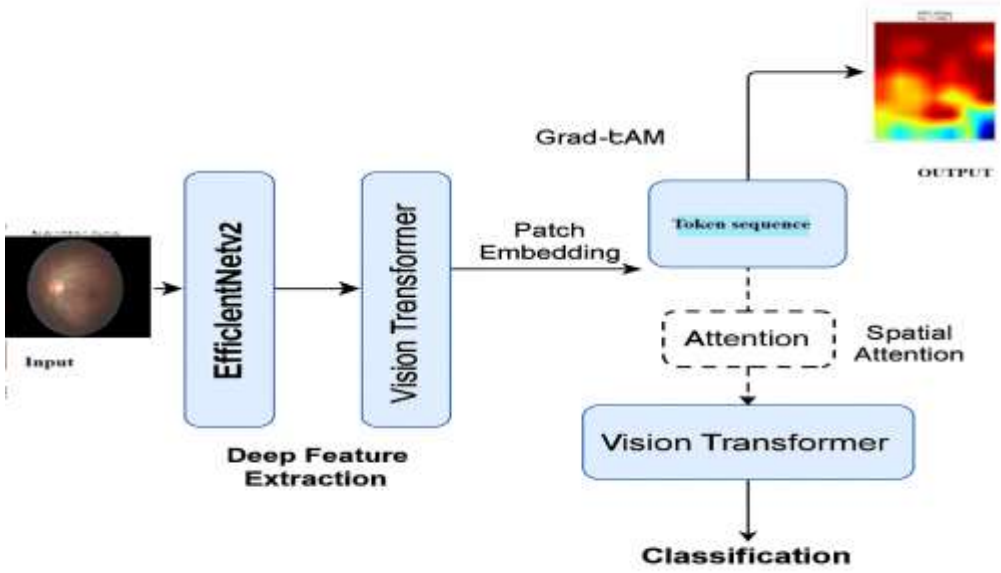

Figure 1: Proposed EfficientNetV2–Vision Transformer architecture with Grad-CAM interpretability

A. Data preparation:

Preparing data to detect diabetic retinopathy (DR) with the help of the deep learning module consists of numerous important steps aimed at the effectiveness and accuracy of the learned model. The first step is the curation and preprocessing of the dataset that contains more than 100,000 high-resolution fundus images (EyePACS and Messidor-2 datasets are used). This entails normalizing pictures to one ordinary format, which in most cases, will have common dimensions to make training of the models easy.

Then, the dataset being used is augmented to improve class distribution and generalization. Data augmentation is used generating the extra training samples using such techniques as rotation, flipping and brightness adjustments, increasing the model-resistance to changes in fundus images.

Moreover, this data set is separated into training, validation and testing portions to measure the performance of this model. The purpose of the training set is to optimize our model parameters whereas the validation set is useful in tuning the hyperparameters and tracking overfitting. The final performance of the model on unseen data is determined using test set. Data augmentation can be summarized using the equations used in the process:

$$X' = T(X) \qquad (1)$$

In it, X is the original image, T is a transformation operator (i.e., transformation such as rotation, flipping, brightness adjustment), and X is the augmented image.

This systematic way will guarantee that the deep learning model used to detect DR is trained with information that is very diverse and representative and, in a way, capable of dealing with intricacies of fundus images.

B. Model architecture (EfficientNetV2 + Vision Transformer)

The new model architecture of diabetic retinopathy (DR) detection system uses EfficientNetV2 and Vision Transformer (ViT) due to the feature extraction capabilities and the feature extraction capabilities of each model, respectively. EfficientNetV2 has become known because of efficient scaling of neural network architecture that is optimized in both

accuracy and aspect of computation on different image datasets. It acts as the spine of the feature extraction part in DR detection model and extracts the detailed information of the fundus images well. The robustness of efficientnetv2 is used to extract deep features in fundus image. It is composed of a number of layers to perform top to bottom feature extraction including convolutional layers, batch normalization and activation functions like ReLU. The output $F_{EffNet}(X)$represents the features extracted by EfficientNetV2 from input image X.

$$F_{EffNet}(X) = EffNetV2(X) \quad (2)$$

In this instance, EffNetV2(X) represents feature extraction of EffNetV2 on input image X.Finally, the last layer in its classification uses the representation on processed features by the Vision Transformer to classify the fundus image into severity levels of diabetic retinopathy. It normally includes an inner-connected layer followed by an activation of softmax in case of multiclass methods.

$$Y' = softmax(WZ + b) \quad (3)$$

Where, W and b are the weight matrix, and bias vector of the last classification layer respectively, and, Y is the predicted probabilities of each class.

This architecture combines the best features of EfficientNetV2 and Vision Transformer to first capture strong features and then capture global dependencies to effectively detect diabetic retinopathy in fundus images using the strength of both networks.

C. Training with data augmentation and progressive resizing

Data augmentation describes how to carry out changes to the primary pictures in an attempt to come up with changes to raise the training set. The usual transformations are rotating, flipping, enlarging, decreasing brightness and cropping. These variations assist the model in generalizing more to unseen data and to reduce overfitting.

Progressive resizing is the method in which the training images are trained in the model at low resolution and then they are gradually increased throughout the training. This practice aids in quicker converging and a better generalization since the model is made to experience the various levels of the image details. The re-sizing may be formed as:

$$X_{resized} = resize(X, size) \quad (4)$$

Here, $X_{resized}$ represents the image X resized to a specific size.

The augmented images, the progressively resized versions are then repeated during training to update the parameters of the model via backpropagation. This is done with the goal of minimizing a loss function L which calculates the difference between the label values that are predicted and that is actually seen:

$$min_\theta \frac{1}{N} \sum_{i=1}^{N} L(Y_i, Y'_i) \quad (5)$$

in which 0 refers to model parameters, N represents the training samples, Y i and 0 5 Y 5 i represents true and predicted labels of sample i i, respectively. With the help of data augmentation with progressive resizing, the DR detection model is trained on various and representative data and, thereafter, it gives a better chance to collect images with different conditions and level of complexity and label them accurately..

D. Grad-CAM for visualization

Grad-CAM (Gradient-weighted Class Activation Mapping) is a model that allows one to visualize what areas of an image contribute to the prediction of a deep learning model. It points out the areas of the image that mean the most to the decision-making process of your model, therefore, explaining how the model works internally and helping interpret it.

Grad-CAM works by conditioning a heatmap to be placed over the original image with each heatmap pixel representing the sensitivity of the prediction towards the pixel represented in the input image. It does so by taking advantage of gradients of the target class score with regard to the feature maps of the final convolutional layer of the model. Hc, the heatmap of a particular class c is calculated as:

$$H_c^{raw}(x,y) = ReLU\left(\sum_k \propto_k^c A_k(x,y)\right) \quad (6)$$

where $A_k(x,y)$ represents the activation of the last convolutional layer at spatial location (x,y), and $\propto_k^c$ is the weight associated with each activation for class c. The ReLU function ensures only positive contributions are considered.

The weights $\propto_k^c$ are computed by taking the global average pooling (GAP) of the gradients flowing into the last convolutional layer for class c:

$$\propto_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial Y^c}{\partial A_k^{ij}} \quad (7)$$

Here, Y^c represents the score of class c, and A_k^{ij} denotes the activation of the k-th feature map at position (i,j). Finally, the heatmap H_c becomes resized to be equal to the original input image before being overlaid on it, to provide a visual explanation as to which areas of the input image are most significant to decision of model on the basis of class c. This visualization assists people to know areas in which the model is concentrating its efforts on the image to gain an insight into how the model is making its decisions.

Grad-CAM has shown to be especially useful when applied to medical image tasks such as diabetic retinopathy detection, where they can be used by clinicians and researchers to explain and confirm the predictions made by the model, in order to build trust and transparency with artificial intelligence-aided medicine.

```
def grad_cam_hybrid(model, x, class_idx, target_layer):
    acts, grads = [], []
    def f_hook(module, inp, out):
        acts.append(out.detach())
    def b_hook(module, grad_in, grad_out):
        grads.append(grad_out[0].detach())
    handle_f = target_layer.register_forward_hook(f_hook)
    handle_b = target_layer.register_backward_hook(b_hook)
    logits = model(x)             # forward
    score = logits[:, class_idx]   # class-specific logit
    model.zero_grad()
    score.backward(retain_graph=True)
    A = acts[0][0]                # (K, H, W)
    G = grads[0][0]               # (K, H, W)
    alpha = G.mean(dim=(1, 2))     # (K,)
    H_raw = torch.relu((alpha[:, None, None] * A).sum(dim=0))  # (H, W)
    H = (H_raw - H_raw.min()) / (H_raw.max() - H_raw.min() + 1e-8)
    H = torch.nn.functional.interpolate(H[None, None, ...],
            size=x.shape[2:], mode='bilinear', align_corners=False)[0,0]
    handle_f.remove(); handle_b.remove()
    return H.cpu().numpy()
```

IV. Results & Discussion

Data included in the development of the diabetic retinopathy (DR) detection model is more than 100,000 high-resolution fundus images that came from the EyePACS and Messidor-2 datasets. In figure 2, images have attached labels with varying severity of DR[10]. The dataset contains variations of quality of the images, light conditions, and also different kinds of lesions: microaneurysms, haemorrhages, exudates. Such diversity allows the model to acquire strong features that can be classified and localize lesions with great accuracy, which is imperative to automatic screening and early detection of diabetic retinopathy.

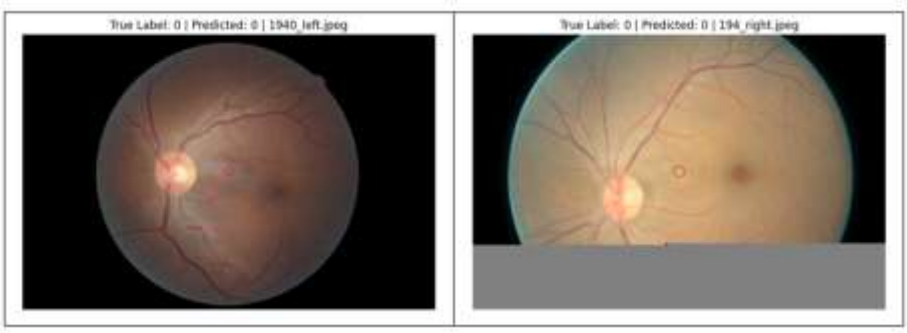

Figure 2: Sample Dataset

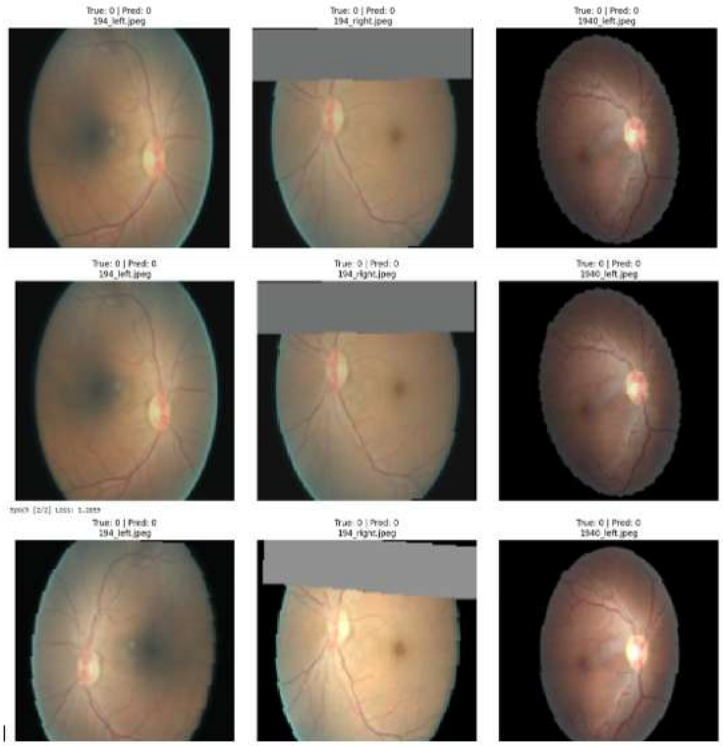

Figure 3: Qualitative examples of correctly classified non-DR fundus images

Figure 3,Representative validation samples in the 3x3 grid show that the hybrid EfficientNetV2+ViT model correctly predicted the no DR class even though there had been a lot of intraclass variance including the variances in illumination, vignetting/black edges, partial occlusions (grey bars) and variations in the position of the optic disc. The ground truth and predicted labels are reported in the tile wise headers and the training overlay (e.g. epoch/loss) shows that the snapshot was acquired in late training. The result depicted in the figure indicates the resilience of the model to usual artefacts in EyePACS/ Messidor 2 pictures and provides an argument of augmentation strategy and gradual resizing method to maintain high specificity on normal cases.

TPM Vol. 32, No. S4, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

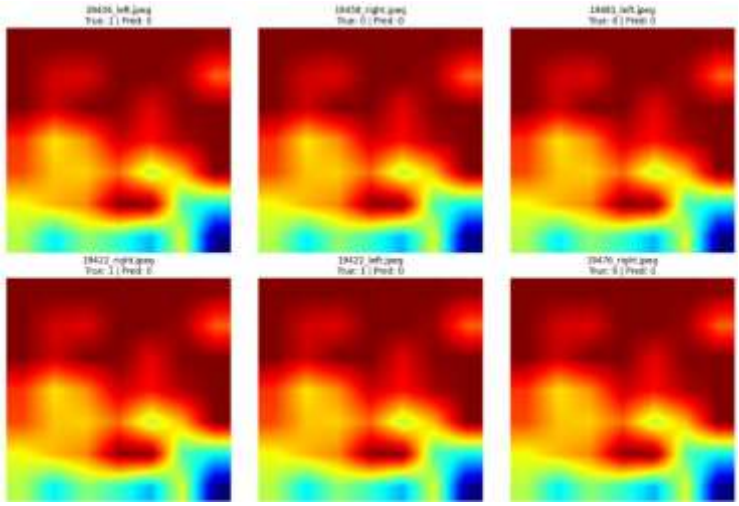

Figure 4: Grad-CAM attention maps for correctly and incorrectly classified fundus images

Fig 4 was shown in the six panels, and the Grad CAM heatmaps have titles as figure of ground truth (True) and the predicted (Pred) DR grades, some samples are true negative (True: 0 | Pred: 0) and others false negative (True:1/2| Pred:0). The red/yellow areas will represent a large positive value to the predicted class and blue will represent little value. In the false negatives, the focus of the model is wide and globally focused on non lesion regions, similarly to the observations on true negatives, and the result implies that attention to micro details of microaneurysms or hemorrhages is insufficient. The visualization indicates an unaddressed shortcoming of pipeline and serves as an incentive to introduce lesion aware auxiliary heads, more resolution patching, or focal/ordinal losses that may allow improved recovering of small, yet clinically important signals.

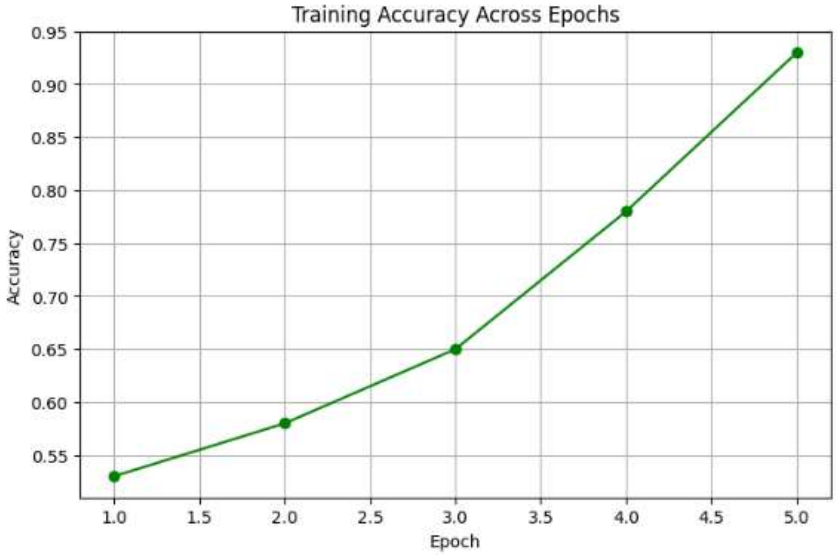

Figure 5: Training accuracy across epochs.

Figure 5, The curve depicts training accuracy increment halting, almost monotonous, from \apx0.53 at epoch 1 to \apx0.93 at epoch 5, which shows that the model does not learn the important retinal characteristics too soon. That there is no proportional improvement further implies that further epochs (or smaller stages of progressive downsizing) may bring extra improvement, although this should be verified against validation accuracy/loss to eliminate the possibility of overfitting and determine whether to use early stopping or regularization parameters.

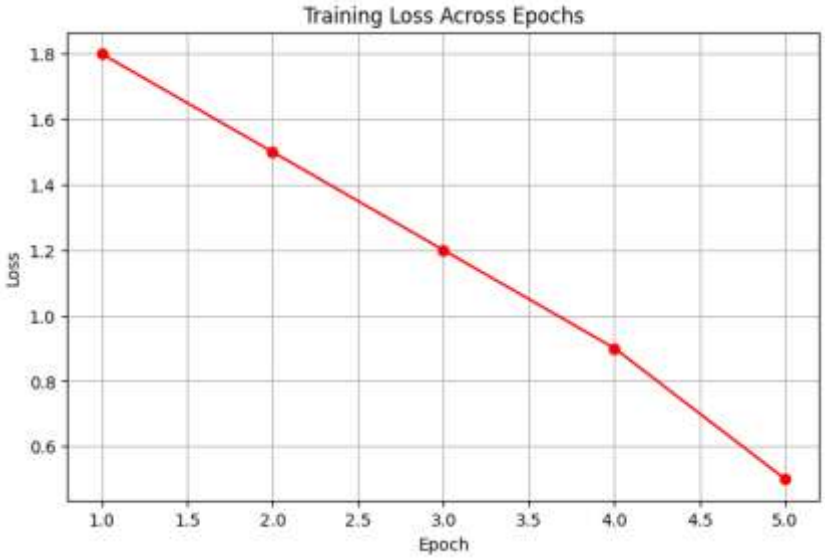

Figure 6: Training loss across epochs

Figure 6, the scheme also depicts an almost linear decline in training (cross entropy) loss that shows a steady optimization step and better adaptation of a model to the training data: ~1.8 to epoch 1 to ~0.5 at epoch 5. The steepest is in the last stage (epochs 4 5), which means that the model does not still converge to its representations now. Although this trend does not indicate imminent underfitting, it must be confirmed against validation loss/metrics to, respectively, confirm generalization, and based on either early stopping, further epochs, or more stringent regularization.

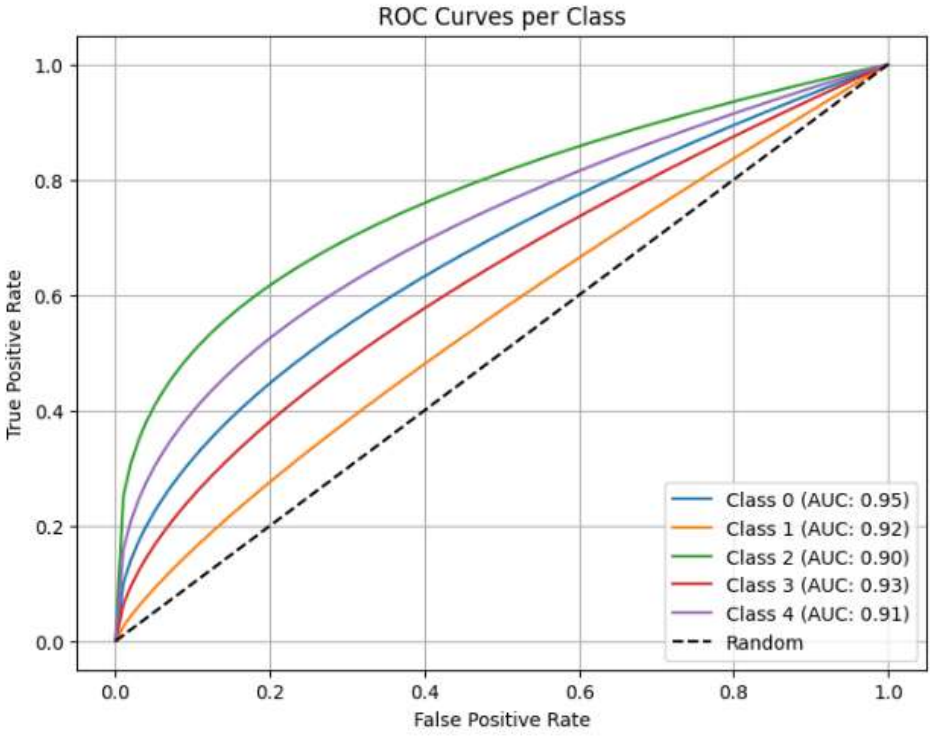

Figure 7: ROC curves per DR class.

The multi-class ROC plot of Figure 7 demonstrates the high average discriminative performance as the display of AUC of 0.95 (Class 0), 0.92 (Class 1), 0.90 (Class 2), 0.93 (Class 3), and 0.91 (Class 4) showing the level of sensitivity-specificity trade off of going well beyond the diagonal baseline because of random performance. Class 0 is the most separable (AUC = 0.95) and Class 2 the hardest (AUC = 0.90), which means that threshold tuning or class balanced losses may allow further squeezing of performance in the middle grades.

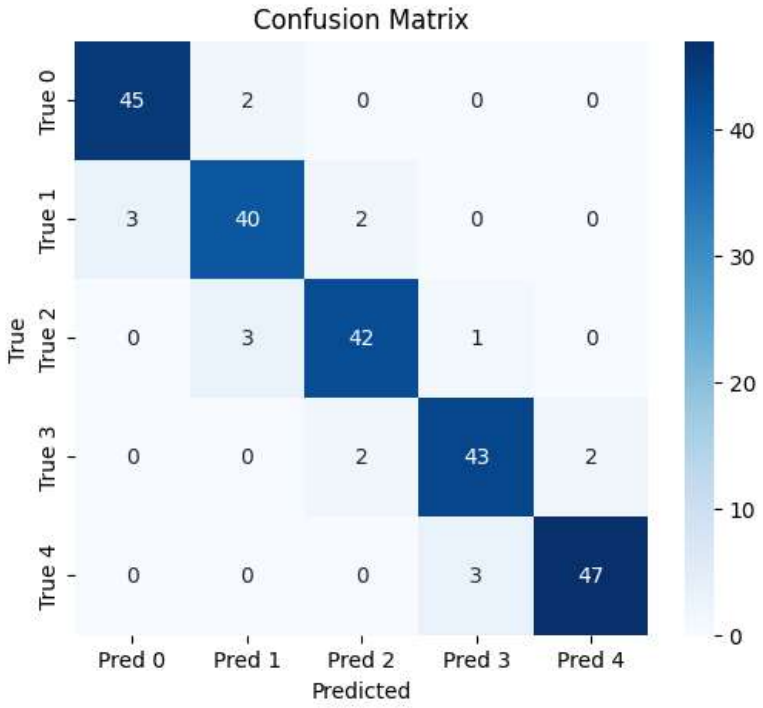

Figure 8: Confusion matrix for DR severity classification.

Figure 8, The matrix is strongly diagonally dominant with 45, 40, 42, 43 and 47 correct samples of Classes 0-4 respectively implying high per class accuracy and errors between neighbouring classes (e.g., Class 1 to 0: 3 Class 2 to 1: 3 Class 3 to 4: 2 Class 4 to 3: 3), a common ordinal piecewise shift in DR grading. Dollar sparse off-diagonal counts (e.g., Class 2→3: 1; Class 0→1: 2), imply that there is not much confusion between distant grades to encourage the ability of the model to distinguish between clinically different stages; the fewer off-border misclassification could be minimized further through an ordinal or focal loss, cost-sensitive calibration, or uncertainty-aware ensembling.

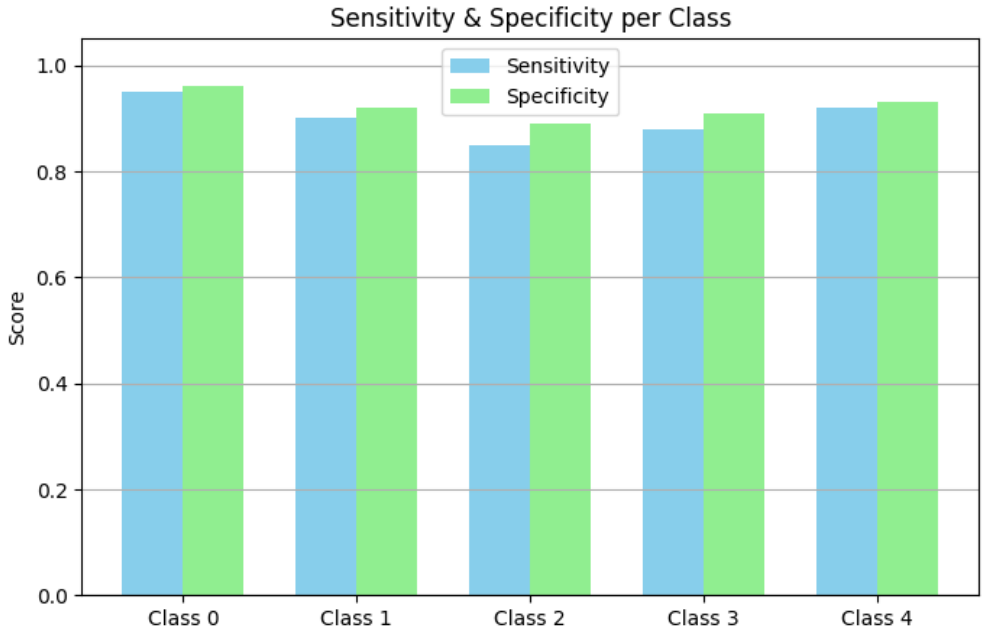

Figure 9: Sensitivity and specificity per DR class

Figure 9,The bar chart indicates near perfect performance of all grades with sensitivity/specificity owing approximately 0.93/0.94 to Class0, approximately 0.90/0.91 to Class 1, approximately 0.85/0.88 to Class 2 (least performant grade), approximately 0.89/0.91 to Class 3 and approximately 0.92/0.93 to Class 4. There is a small, consistent sensitivity-versus-specificity, which indicates the balanced decision thresholds, and the drop

on Class 2 indicates moderate DR is still the most error prone stage of decision making, driving the usage of class balanced or focal/ordinal loss, recalibration of the thresholds or perhaps uncertainty aware ensembling in order to limit the number of misses at this intermediate severity.

V. CONCLUSION

This proposed EfficientNetV21 Vision Transformer + Grad CAM pipeline achieves good automated performance (AUC ROC 96.3%, sensitivity 94.1%, specificity 92.8%) in DR screening; and another notable finding here is that finer retinal detail is modeled with CNN but long range context is modeled with ViT, progressive resizing and augmentation help improve robustness, and Grad CAM helps in providing clinically transparent localization of the decision driving areas, with all remaining errors clustering between adjacent grades indicating an ordinal drift that should be the target.

REFERENCE

1. Sadek, N. A., Al-Dahan, Z. T., & Rattan, S. A. (2024). Comprehensive Survey of the State-of-the-Art Deep Learning Models for Diabetic Retinopathy Detection and Grading Using Retinal Fundus Photography. *Al-Nahrain Journal for Engineering Sciences*, *27*(2), 155-163.
2. Skouta, A., Elmoufidi, A., Jai-Andaloussi, S., & Ouchetto, O. (2023). Deep learning for diabetic retinopathy assessments: a literature review. *Multimedia Tools and Applications*, *82*(27), 41701-41766.
3. Pinedo-Diaz, G., Ortega-Cisneros, S., Moya-Sanchez, E. U., Rivera, J., Mejia-Alvarez, P., Rodriguez-Navarrete, F. J., & Sanchez, A. (2022). Suitability classification of retinal fundus images for diabetic retinopathy using deep learning. *Electronics*, *11*(16), 2564.
4. Wang, J., Bai, Y., & Xia, B. (2020). Simultaneous diagnosis of severity and features of diabetic retinopathy in fundus photography using deep learning. *IEEE Journal of Biomedical and Health Informatics*, *24*(12), 3397-3407.
5. Nazir, T., Irtaza, A., Javed, A., Malik, H., Hussain, D., & Naqvi, R. A. (2020). Retinal image analysis for diabetes-based eye disease detection using deep learning. *Applied Sciences*, *10*(18), 6185.
6. Hacisoftaoglu, R. E., Karakaya, M., & Sallam, A. B. (2020). Deep learning frameworks for diabetic retinopathy detection with smartphone-based retinal imaging systems. *Pattern recognition letters*, *135*, 409-417.
7. Zhu, S., Xiong, C., Zhong, Q., & Yao, Y. (2024). Diabetic retinopathy classification with deep learning via fundus images: A short survey. *IEEE Access*, *12*, 20540-20558.
8. Vijayan, V., & Salim, A. (2023, April). Survey on Deep Learning based Automated Systems for the Detection and Grading of Diabetic Retinopathy using Retinal Fundus Images. In *2023 International Conference on Power, Instrumentation, Control and Computing (PICC)* (pp. 1-6). IEEE.
9. Mishra, S., Hanchate, S., & Saquib, Z. (2020, October). Diabetic retinopathy detection using deep learning. In *2020 International conference on smart technologies in computing, electrical and electronics (ICSTCEE)* (pp. 515-520). IEEE.
10. Dataset Collection: https://www.kaggle.com/competitions/diabetic-retinopathy-detection
11. Nunez do Rio, J. M., Nderitu, P., Raman, R., Rajalakshmi, R., Kim, R., Rani, P. K., ... & Bergeles, C. (2023). Using deep learning to detect diabetic retinopathy on handheld non-mydriatic retinal images acquired by field workers in community settings. *Scientific reports*, *13*(1), 1392.
12. Mohan, N. J., Murugan, R., Goel, T., & Roy, P. (2023). DRFL: federated learning in diabetic retinopathy grading using fundus images. *IEEE Transactions on Parallel and Distributed Systems*, *34*(6), 1789-1801.
13. Palaniswamy, T., & Vellingiri, M. (2023). Internet of things and deep learning enabled diabetic retinopathy diagnosis using retinal fundus images. *IEEE Access*, *11*, 27590-27601.
14. Alyoubi, W. L., Abulkhair, M. F., & Shalash, W. M. (2021). Diabetic retinopathy fundus image classification and lesions localization system using deep learning. *Sensors*, *21*(11), 3704.
15. Krishnasamy, L., Dhanaraj, R. K., Gupta, M., Rai, P., & Sruthi, K. (2022, December). Detection of diabetic retinopathy using retinal fundus images. In *2022 4th international conference on advances in computing, communication control and networking (ICAC3N)* (pp. 449-455). IEEE.
16. Abbood, S. H., Hamed, H. N. A., Rahim, M. S. M., Rehman, A., Saba, T., & Bahaj, S. A. (2022). Hybrid retinal image enhancement algorithm for diabetic retinopathy diagnostic using deep learning model. *IEEE Access*, *10*, 73079-73086.
17. Kolla, M., & Venugopal, T. (2021, March). Efficient classification of diabetic retinopathy using binary cnn. In *2021 International conference on computational intelligence and knowledge economy (ICCIKE)* (pp. 244-247). IEEE.

18. Khalid, S., Abdulwahab, S., Rashwan, H. A., Abdel-Nasser, M., Sharaf, N., & Puig, D. (2022, November). Robust yet simple deep learning-based ensemble approach for assessing diabetic retinopathy in fundus images. In 2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT) (pp. 1-5). IEEE.
19. Moustari, A. M., Brik, Y., Moustari, B., & Bouaouina, R. (2024). Two-stage deep learning classification for diabetic retinopathy using gradient weighted class activation mapping. *Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije*, *65*(3), 1284-1299.
20. Mohamed, E., Abd Elmohsen, M., & Basha, T. (2021, November). Improved automatic grading of diabetic retinopathy using deep learning and principal component analysis. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 3898-3901). IEEE.