Hybrid Deep Learning Architectures for Diabetic Retinopathy Detection: A Comparative Study of CNNs, VGG16, and Vision Transformers

Dimple Sethi, Parnavi Sharma, Vanshika Agarwal, Shivanshu Garg
School of Computer Science and Technology, Bennett University

***Abstract*—Diabetic Retinopathy (DR) is a diabetes-related microvascular complication involving progressive retinal vasculature damage, causing irreversible blindness if undiagnosed and untreated in the early stages. Growing prevalence of diabetes globally has established DR as a pressing healthcare issue requiring effective and scalable diagnostic technologies. Traditional manual ophthalmologist-based grading of retinal fundus images is time-consuming, susceptible to variability, and not scalable.**

This paper presents a comparative study of deep learning models for computer-aided DR detection and severity grading using the DDRDataset. We analyze the efficiency of three models, i.e., a Convolutional Neural Network (CNN), VGG-19, and Vision Transformer (ViT), to evaluate their performance in DR detection and grading. The approach used involves transfer learning and data augmentation to present strong and accurate models.

Our experimental result reveals that CNN structures exhibit high generalization with higher validation accuracy compared to more complex architectures like ViT, which exhibit faster convergence with more overfitting. VGG-19 acts fairly by being well-balanced with regard to accuracy and complexity. The study attests to the potential of deep learning-driven DR detection systems, marrying diagnostic performance with model efficiency, and supports the deployment of scalable, AI-driven screening instruments.

***Index Terms*—Diabetic Retinopathy, Deep Learning, CNN, VGG-19, Vision Transformer, Transfer Learning.**

I. Introduction

Diabetic Retinopathy (DR) is a serious microvascular complication of diabetes mellitus of the retinal vasculature that causes vision loss and blindness unless treated. The increasing global load of diabetes has put Diabetic Retinopathy on the global stage as a alarming public health issue. Early detection and early treatment are the ways to prevent this disease. However, the present diagnosis is manual reading of retinal fundus photographs, which is time-consuming, subject to interobserver variation, and unscalable.

Diabetic Retinopathy (DR) occurs through a sequence of four steps:

- Mild nonproliferative retinopathy is the initial phase characterized by the occurrence of micro-aneurysms only.
- Moderate nonproliferative retinopathy occurs as the disease progresses, leading to the distortion and swelling of blood vessels, resulting in the impaired transportation of blood.
- Severe non-proliferative retinopathy manifests when the blockage of blood vessels increases, causing a reduced blood supply to the retina. This prompts signals for the growth of new blood vessels.
- Proliferative diabetic retinopathy denotes the advanced stage characterized by the release of growth factors by the retina, which triggers the development of new blood vessels. These vessels may extend along the inner surface of the retina and even into the vitreous gel, posing significant risks to vision.

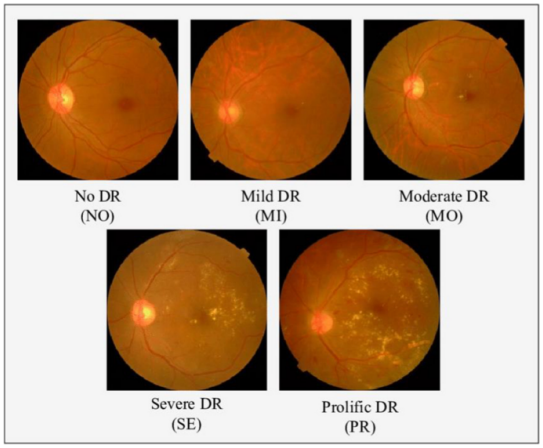

Fig. 1. Different Stages of Diabetic Retinopathy

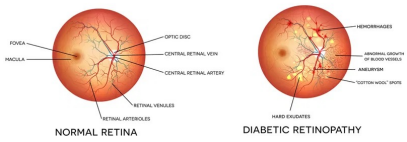

Fig. 2. Normal Vision and DR Vision

As shown in [Figure 2], the healthy retina has well-developed vasculature with visible and distinct anatomical structures. It possesses the fovea, which accounts for the sharp central vision, and the macula, which accounts for the perception of fine vision. The optic disc also provides space for the entry and exit of the central retinal artery and vein and possesses adequate circulation in the retina. Retinal arterioles and venules are also normal in oxygen and nutrient delivery without defect or occlusion.

The diabetic retinopathy retina is distinct, with extreme pathological alterations. These consist of hemorrhages as a manifestation of leakage of blood from leaking capillaries, neovascularization (vessel growth), and aneurysms secondary to increased vascular fragility. Cotton wool spots, a manifestation of focal retinal ischemia, and hard exudates, the consequence of leakage of protein and lipid, also disrupt normal retinal function. These chronic abnormalities can result in loss of vision and, if untreated, permanent blindness.

This difference highlights the central role of diabetes on retinal well-being, which requires early detection and appropriate medical intervention.

Developments of new computer vision and deep learning (DL) techniques are potentially helping us in DR detection and classification automation. CNNs, particularly pre-trained transferred learning models, have emerged as very powerful in processing medical images. Here, we are trying to offer a deep learning-based solution for DR detection automation with the DDRDataset. We are using VGG-19, ViT (Vision TRansformers) models for DR severity level classification and Grad-CAM visualization for model interpretability. By using these techniques, we try to enhance diagnostic accuracy, efficiency, and interpretability of DR classification.

II. Related Work

The application of hybrid deep learning architectures has significantly advanced the detection and grading of diabetic retinopathy (DR), offering promising solutions for early diagnosis and treatment planning. Recent studies have explored various combinations of convolutional neural networks (CNNs), transformers, and ensemble methods to enhance the accuracy and efficiency of DR classification systems.

Ali et al. (2020) proposed a deep learning-based model for DR grading, utilizing a cascaded neural network approach that classifies fundus images through successive layers, each distinguishing between specific DR stages. This method aims to improve the granularity of classification by sequentially narrowing down the DR stage [1].

Awais et al. (2021) conducted a prospective study employing a modified CNN architecture for DR detection using fundus images. Their approach focused on enhancing feature extraction capabilities to improve detection accuracy in clinical settings [2].

Bidwai et al. (2022) presented a systematic literature review on DR detection using artificial intelligence, highlighting the integration of deep learning techniques in analyzing retinal features such as blood vessels, microaneurysms, and hemorrhages. The review emphasized the role of AI in early detection and classification of DR [3].

Rao et al. (2020) evaluated various CNN architectures for DR classification, providing a comprehensive analysis of their performance metrics. The study offered insights into the strengths and limitations of different CNN models in the context of DR detection [4].

Al-Kamachy et al. (2024) developed a computer-aided diagnosis system leveraging pre-trained deep learning models for DR classification. Their system aimed to automate the classification process into five distinct DR stages, demonstrating the efficacy of transfer learning in medical image analysis [5].

Tymchenko et al. (2020) introduced an automatic deep learning-based method for detecting DR stages using fundus images. Their approach incorporated a multistage transfer learning technique to address challenges related to dataset variability and labeling inconsistencies [6].

Islam et al. (2018) proposed a deep convolutional neural network for early detection of DR, focusing on identifying microaneurysms as early indicators. Their model achieved high sensitivity and specificity, underscoring the potential of CNNs in early DR diagnosis [7].

Gu et al. (2023) explored the classification of DR severity using a combination of Vision Transformer and residual attention mechanisms. Their hybrid model aimed to capture both global and local features in fundus images, enhancing classification performance [8].

Another study presented a computer-aided diagnostic system utilizing a modified compact convolutional transformer (CCT) and low-resolution images to reduce computation time. This approach sought to balance accuracy and efficiency in DR detection [9].

Sudha and Ganeshbabu (2021) implemented a CNN classifier based on the VGG-19 architecture for lesion detection and grading in DR. Their method incorporated segmentation techniques to identify retinal defects, contributing to more precise DR classification [10].

Alwakid et al. (2023) enhanced DR image classification through deep learning models, aiming to accurately identify all five stages of DR. Their methodology emphasized the importance of comprehensive stage-wise classification in clinical diagnostics [11].

Collectively, these studies underscore the advancements in hybrid deep learning architectures for DR detection, highlighting the integration of various neural network models and techniques to improve diagnostic accuracy and efficiency.

III. Literature Review

A. A Deep Learning-Based Model for Diabetic Retinopathy Grading

The referenced study proposes a deep learning-based model for DR grading, leveraging convolutional neural networks (CNNs) for feature extraction and classification. They developed their own model and named it as Retinopathy Severity Grading (RSG-Net). The authors highlight the success of RSG-Net in offering an accurate and automated solution for diabetic retinopathy severity grading.

They addressed issues such as class imbalance through data augmentation and enhanced image quality using the Histogram Equalization (HE) technique. Their model was trained on the Messidor-1 dataset, which contains a total of 1200 fundus images and incorporates various preprocessing techniques to enhance image quality and mitigate noise, ensuring robust predictions. Their model achieved a testing accuracy of 99.36%, specificity of 99.79%, and a sensitivity of 99.41% in

classifying diabetic retinopathy into four grades. Additionally, it achieved 99.37% accuracy, 100% sensitivity, and 98.62% specificity in classifying DR into two grades.

B. A Prospective Study on Diabetic Retinopathy Detection Based on Modified CNN Using Fundus Images at Sindh Institute of Ophthalmology & Visual Sciences

Awais et al. described that deep neural networks like CNNs have the potential to be a faster and more efficient method than traditional techniques for detecting diabetic retinopathy. The authors conducted a practical on-ground analysis by testing their model on real-time DR data from patients at the Sindh Institute of Ophthalmology & Visual Sciences (SIOVS), Hyderabad, Pakistan, using image-capturing devices.

The dataset contained images of type-II diabetes patients, collected over five weeks. The results were reviewed by clinical experts to assess the model's performance. A total of 398 patients, including 232 male and 166 female patients, were screened over five weeks. The model achieved 93.72% accuracy, 97.30% sensitivity, and 92.90% specificity on the test data labeled by clinical experts on diabetic retinopathy.

C. A Systematic Literature Review on Diabetic Retinopathy Using an AI Approach

This survey provides a comprehensive analysis of diabetic retinopathy, including a clear distinction between normal vision and DR vision, the stages and classification of DR. The authors conducted detailed research on the evolution of DR using AI, ophthalmic analysis, various machine learning techniques for DR detection, feature extraction, and classification in deep learning, and the concept of transfer learning in DR.

Additionally, a comparative analysis of different approaches and findings from various ophthalmology researchers was reviewed.

D. Explainable Diabetic Retinopathy Using EfficientNet

Mohamed Chetoui et al. proposed using recent CNN architectures with only publicly available datasets to detect referable diabetic retinopathy (RDR) and vision-threatening diabetic retinopathy (VTDR). They defined referable diabetic retinopathy as fundus images classified as moderate non-proliferative, severe non-proliferative, or proliferative diabetic retinopathy.

The authors used the EfficientNet-B7 model and fine-tuned the network to enhance its efficiency in detecting RDR and VTDR. Tests were conducted on publicly available datasets, EyePACS and APTOS 2019. They assessed the model's performance using AUC, Sensitivity (SN), and Specificity (SP) metrics.

For RDR, their deep learning model achieved:

- Sensitivity of 0.917, specificity of 0.989, and AUC of 0.984 on the EyePACS test set.
- Sensitivity of 0.914, specificity of 0.972, and AUC of 0.966 on the APTOS 2019 dataset.

For VTDR, the model achieved:

- AUC of 0.994, sensitivity of 0.981, and specificity of 0.9374 on the EyePACS test set.
- AUC of 0.998, sensitivity of 0.991, and specificity of 0.925 on the APTOS 2019 dataset.

Additionally, an explainability algorithm based on Gradient-weighted Class Activation Mapping (Grad-CAM) was developed. This technique was used to provide visual explanations for the results of the proposed CNN model.

IV. Methodology

Our approach is divided into six well-defined steps to enable reproducibility and clinician interpretability:

A. Dataset Exploration and Preparation

We operate on the DDR (Diabetic Retinopathy Detection) dataset of 13,673 macula-centered color fundus images of 9,598 patients in 147 clinics. Each image contains pixel-level masks for microaneurysms, hemorrhages, and exudates, and has one of five DR severity grades labeled per ICDR guidelines. All the annotations were performed by expert graders under the supervision of ophthalmologists. To prevent patient overlap and ensure fairness in evaluation, we split the data at the patient level into 70% training, 10% validation, and 20% test sets with stratified class distributions.

As depicted in Figure 3. With DR grade 0 (no DR) being the most common at about 6,200 images and grade 2 (moderate DR) at about 4,500 images, the dataset shows a clear class imbalance. Only 200 images were in grade 3 (severe DR), which was the least common, compared to about 900 in grade 4 (proliferative DR) and 600 in grade 1 (mild DR). The other classes were much less represented. We preprocessed each image to uniform 256×256 and 512×512 pixel dimensions in order to standardize the input for our model. To ensure statistical consistency and avoid sampling bias, the dataset was divided 80:20 between training and testing sets, with each partition meticulously preserving the original class distribution.

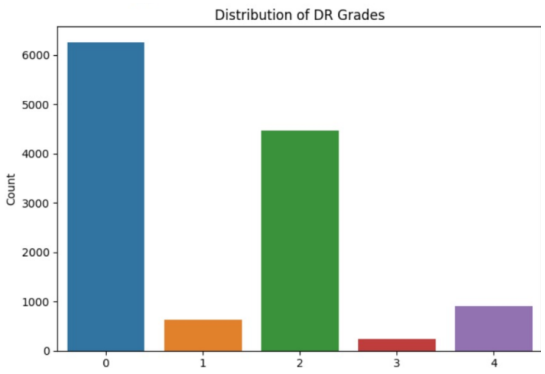

Fig. 3. Distribution of Kaggle Fundus Retinal Image Dataset

B. Custom Basic Convolutional Neural Network (CNN)

The CNN model is designed specifically for DR detection. It comprises five convolutional blocks, each with a Conv2D layer, BatchNormalization, ReLU activation, and MaxPooling.

A Global Average Pooling layer follows these blocks to reduce spatial dimensions. Fully connected layers are employed, incorporating dropout to prevent overfitting. The final softmax layer outputs the probability distribution for the five DR grades. The Adam optimizer with a learning rate of 0.0004 is used to minimize the categorical cross-entropy loss.

The diabetic retinopathy (DR) images used in this study are sourced from the DDRDataset, containing labeled retinal fundus images. Proper data preprocessing is crucial to maintain consistency and enhance model performance. The images are resized to 256x256 pixels and normalized by dividing pixel values by 255, scaling them to the range [0, 1]. To optimize memory usage during training, the preprocessed images are saved in batches of 500.

To ensure robust model evaluation, the dataset is divided into three subsets: training (80%), validation (16%), and testing (4%). This stratified split preserves the original class distribution, making each subset representative of the entire dataset. Images are systematically organized into subfolders based on diagnosis labels to facilitate structured data loading.

Data augmentation plays a critical role in improving the model's generalization capabilities, particularly given the limited size of medical image datasets. Augmentation techniques such as random rotations (±15°), horizontal and vertical shifts (up to 10%), zooming (up to 20%), and horizontal flipping are employed. These transformations increase the diversity of the training data and help mitigate overfitting, as demonstrated in prior studies on medical image classification [1].

The proposed CNN model is designed to efficiently capture complex features from retinal images. It consists of five convolutional blocks, each incorporating a Conv2D layer with ReLU activation, BatchNormalization to stabilize the learning process, and MaxPooling for spatial downsampling. A Global Average Pooling layer reduces spatial dimensions, followed by fully connected layers with dropout (rate: 0.5) to prevent overfitting. The final output layer employs softmax activation to predict the probability distribution over five DR grades. The Adam optimizer, with a learning rate of 0.0004, is utilized to minimize the categorical cross-entropy loss, which is well-suited for multi-class classification.

Training is conducted for up to 20 epochs, with a batch size of 64. Early stopping is employed to halt training if the validation loss does not improve for seven consecutive epochs, thereby preventing overfitting. Additionally, the learning rate is reduced by half when a plateau in validation loss is detected, aiding the model in achieving stable convergence. These training strategies ensure model reliability and efficiency.

To evaluate the model's performance, we employ multiple metrics, including accuracy, precision, recall, F1-score, ROC-AUC, Cohen's Kappa, and Quadratic Weighted Kappa. Visualization techniques such as training loss curves, accuracy plots, confusion matrices, and ROC curves are utilized to interpret the model's learning behavior and generalization ability.

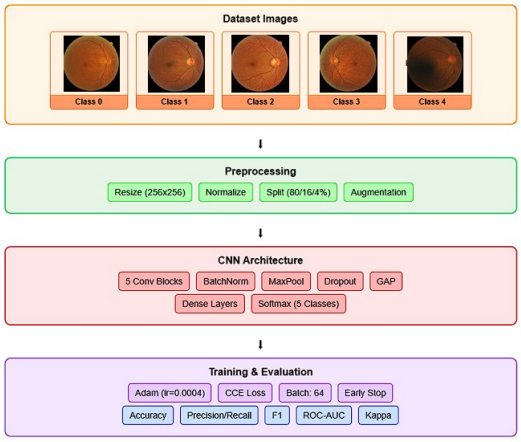

Fig. 4. Diabetic Retinopathy CNN based Detection Pipeline

C. VGG19

In this research, a transfer learning approach was utilized with the VGG19 convolutional neural network architecture, which is pre-trained on the ImageNet dataset. The VGG19 model was utilized as a feature extractor by stripping its original top classification layers so that it could concentrate exclusively on extracting significant visual patterns from retinal fundus images.

Firstly, the pre-trained VGG19 was loaded with its convolutional base retaining the original settings and only the top fully connected layers removed. The input image was set to 256×256 pixels with three color channels. To start with, all the layers in the VGG19 base were frozen to not change the learned features and to prevent overfitting. However, to enable fine-tuning on the medical dataset, the last eight layers of the VGG19 base were unfrozen, allowing them to adapt slightly to the characteristics of diabetic retinopathy images.

The feature maps from the VGG19 convolutional base were used as input to pass through a Global Average Pooling layer. This took the multi-dimensional feature maps and reduced them to a compact vector by averaging out each feature map and lowering the number of parameters as well as reducing overfitting.

After the pooling layer, the model was augmented with two fully connected (dense) layers. The first dense layer had 512 units with ReLU activation followed by a dropout layer with a dropout rate of 0.5 to avoid overfitting. The second dense layer had 256 units with ReLU activation and a dropout rate of 0.3. These layers enabled the model to learn higher-level representations from the extracted features.

The last output layer was a dense layer of five neurons, matching the five diabetic retinopathy classes. A softmax activation function was used to generate a probability distribution across the five classes.

For training, the model was built using the Adam optimizer with a learning rate set small to guarantee stable fine-tuning. Categorical cross-entropy as the loss function was

used, suitable for multi-class classification tasks. Techniques of early stopping and model checkpointing were utilized to avoid overfitting and guarantee optimal performance. Early stopping stopped training when the validation accuracy failed to increase over a specified number of epochs, while checkpointing stored the best-performing model according to validation accuracy.

This organized deployment guaranteed that the VGG19 model was able to efficiently utilize pre-trained knowledge while adjusting to the unique features of the diabetic retinopathy dataset.

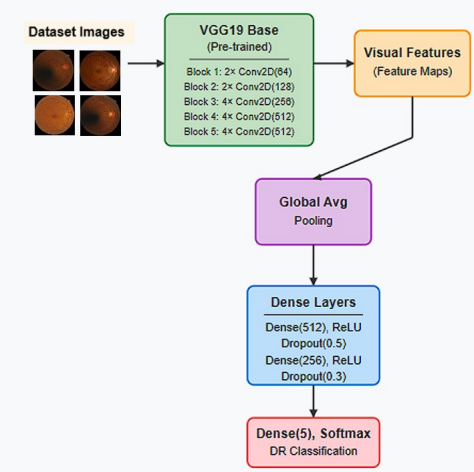

Fig. 5. VGG-19 Architecture

D. Vision-Transformer (ViT Base Patch 16) Model Architecture

We adopt a Vision Transformer (ViT) backbone pre-trained on ImageNet, modifying its classification head for five-grade diabetic-retinopathy (DR) assessment. The core ViT ("ViT-Base-Patch16-224") treats each 224×224 input fundus image as a sequence of 16×16 patches:

1) Patch Embedding: The input image is resized to 224×224 pixels and partitioned into non-overlapping 16×16 patches (total 14×14=196 patches). Each patch is flattened and projected via a learnable linear layer into a 768-dimensional

embedding. A special learnable classification token ([CLS]) is prepended to this sequence.

2) Positional Encoding: To retain spatial relationships lost during patch flattening, we add learned positional embeddings to each patch embedding and to the [CLS] token.

3) Transformer Encoder Stack: The sequence (196 patch embeddings + 1 [CLS]) passes through 12 identical Transformer encoder layers. Each layer comprises:

Multi-Head Self-Attention (MHSA): Enables each patch to attend to every other patch (and the [CLS] token), capturing global contextual cues—critical for detecting diffuse DR lesions.

Feed-Forward Network (FFN): A two-layer MLP with a GELU nonlinearity mixes information across the embedding dimensions.

Pre-Normalization & Residual Connections: LayerNorm is applied before both MHSA and FFN, with residual shortcuts ensuring stable gradient flow and faster convergence.

4) Classification Head Adaptation: We replace the original ViT head with a single linear layer mapping the final embedding of [CLS] (768 D) to five DR grades (0–4). A softmax activation produces class probabilities.

5) Input Normalization: Fundus images are normalized channel-wise to mean=0.5, std=0.5, scaling pixel values to [-1, + 1], which is in agreement with the expected input distribution of ViT pre-trained.

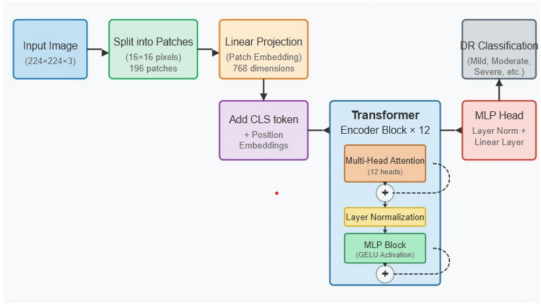

Fig. 6. Vision Transformer (ViT-Base-Patch16) Architecture

E. Application to the DDR Diabetic-Retinopathy Dataset Data Preparation and Augmentation

Data Preparation and Augmentation Our dataset comprises high-resolution retinal fundus photographs labeled with DR grades (0: no DR, through 4: proliferative DR). To enhance model robustness and mitigate data scarcity:

Stratified Splitting: We partition images into training (80)%), validation (16%), and test (4%) sets, preserving class proportions to prevent bias toward majority grades.

Geometric Augmentation: On the training set, each image undergoes a random combination of small rotations (±15°), translations (±10% width/height), zoom (±20%), and horizontal flips. This augmentation simulates acquisition variability and encourages the model to learn invariant lesion patterns.

Preprocessing Pipeline: All images are resized to 224×224, converted to tensors, and normalized to the ViT input range.

Model Training Workflow Optimization: We fine-tune all Transformer layers using the Adam optimizer with a conservative learning rate (1e-4).

Regularization: No additional weight decay is applied; the data augmentation serves as our primary regularizer.

Training Dynamics: Over 20 epochs on a single GPU, the model's training loss consistently decreased, while validation performance stabilized—indicating effective transfer of learned visual features without severe overfitting.

Inference and Interpretability Prediction: In inference, each fundus image is processed through the same patch embedding and transformer pipeline; the final [CLS] embedding is projected to DR grade probabilities.

Attention Visualization (Optional): By extracting attention weights from early and late Transformer layers, heat maps

can be generated that highlight image re.g.ions (e.g., microaneurysms, hemorrhages) that contributed the most strongly to the decision of the model, providing interpretable insights for clinical review.

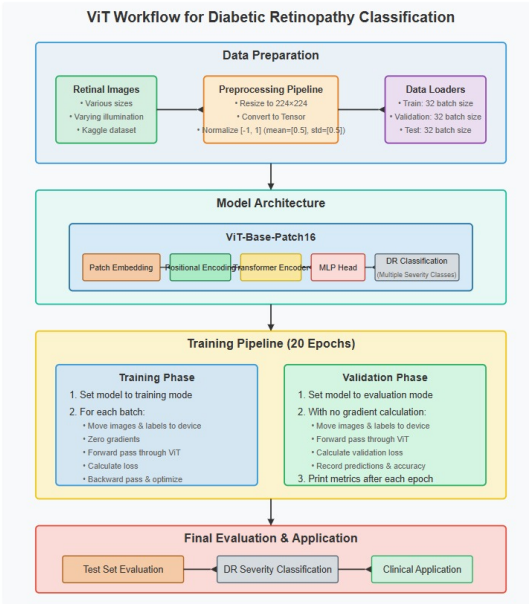

Fig. 7. Vision Transformer (ViT-Base-Patch16) Architecture

V. Experimental Results and Discussion

A. Results

1) CNN-Based Models [VGG19 and Custom CNN]: In order to test the performance of convolutional neural networks in diabetic retinopathy classification, we implemented two architectures: a custom CNN from scratch and VGG19 with pre-trained weights initialized with ImageNet.

a. Training and Validation Behavior : In the case of the **Custom Basic CNN model**, as indicated in **Figures 8 and 9**, the training loss decreases steadily, reflects the fact that the model is learning efficiently from the training set. The validation loss, on the other hand, exhibits a huge oscillation, suggesting that the model overfits somehow. This is affirmed by the accuracy graph, wherein training accuracy continues to increase consistently but validation accuracy also increases but at a slower rate and skips across epochs. The discrepancy in the training and validation metrics reflects potential generalization issues with the less complicated architecture of the simple CNN.

However, the **VGG19 model** shows more stable and improved performance, as shown in **Figures 10 and 11.** The training loss decreases gradually with a smoother curve, and although the validation loss does show some oscillations, it is much more stable compared to the simple CNN. The training and validation accuracy curves also reflect improved generalization; training accuracy rises gradually, and the validation accuracy shows a smooth increasing trend, with less noise compared to the simple model. This reflects that the deeper structure of VGG19 and its pre-trained feature extraction capability facilitate improved learning and generalization.

In summary, while the Custom Basic CNN model exhibits training capability at the start, the VGG19 model surpasses it by a wide margin of training stability and validation accuracy, thus making it more trustworthy in achieving higher accuracy and better generalization.

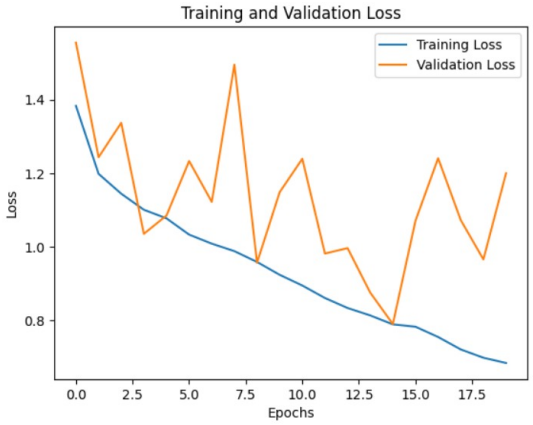

Fig. 8. Custom Basic CNN model: Training and Validation Loss graph

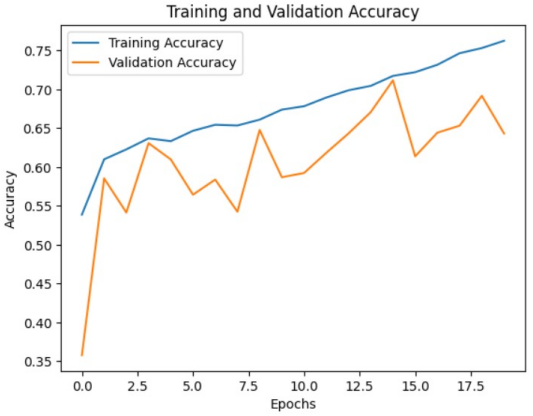

Fig. 9. Custom Basic CNN model: Training and Validation Accuracy graph

b. Confusion Matrices : The confusion matrices of both models (as depicted in **Fig. 12 and Fig. 13** exhibit considerable differences in classification accuracy over the five diabetic retinopathy classes:

Class 0 (No DR):

- Basic CNN model accurately classified 225 samples, but number of misclassifications were 25 (Class 2) and 1 (Class 4).
- VGG19 correctly predicted 228 samples, with 23 misclassified (Class 2).
- Although both models are good here, VGG19 has improved precision, proving to have a better ability to correctly and consistently identify healthy retinal images.

Class 1 (Mild DR):

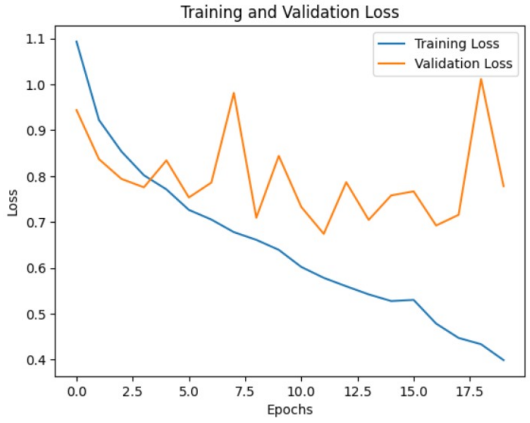

Fig. 10. VGG19 model: Training and Validation Loss graph

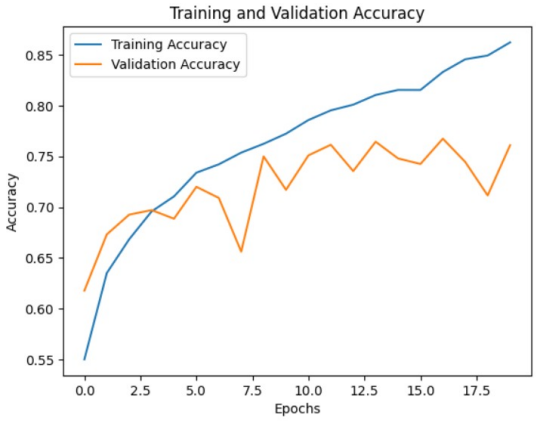

Fig. 11. VGG19 model: Training and Validation Accuracy graph

- Basic CNN correctly classified 11 samples, misclassifying 14 as Class 0.
- VGG19 also did better with 14 correct predictions but incorrectly predicted 11 samples to Class 0.
- Both models show trouble distinguishing between mild DR, likely due to insignificant pathological signs and similarities to Class 0 and Class 2.

Class 2 (Moderate DR):

- Simple CNN correctly predicted 113 samples and incorrectly predicted 59 to Class 0 and 7 to Class 4.
- VGG19 showed significantly better performance by accurately classifying 122 samples and the number of incorrect classifications were 49 (Class 0) and 6 (Class 4).
- VGG19 shows more balanced performance, with less peak misclassifications, indicating that its deeper layers contribute to more accurate feature extraction in moderate DR.

Class 3 (Severe DR):

- Basic CNN succeeded in correctly classifying 6 samples, incorrectly classifying most as Class 0 or Class 4.
- VGG19 performed better with regards to the control of misclassifications with 7 proper classifications and lesser confusion.
- The models both struggled in this arena but appear slightly more cautious and prudent in nature in VGG19, thereby fewer disastrous misjudgments are made.

Class 4 (Proliferative DR):

- Straightforward CNN was only capable of classifying a mere 15 samples appropriately while misclassifying 20 samples as Class 2.
- VGG19 performed much better with 22 correct classifications and overall fewer misclassifications.
- This result indicates VGG19's more profound structure and improved capacity to learn intricate features, making it more precise for the detection of late-stage DR.

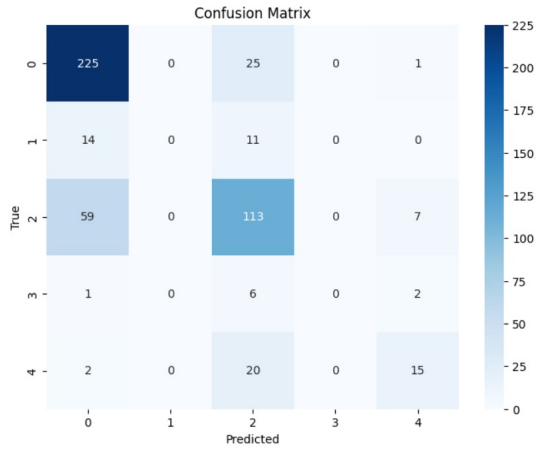

Fig. 12. Confusion Matrix of Basic CNN model

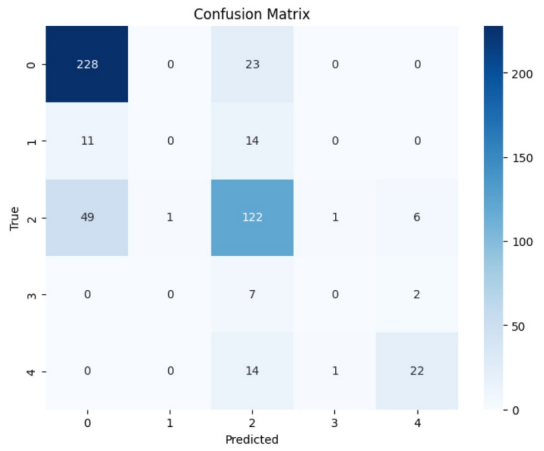

Fig. 13. Confusion Matrix of VGG19 model

c. Classification Metrics : The response of both models was compared using precision, recall, F1-score, and accuracy. Furthermore, Cohen's Kappa and Quadratic Weighted Kappa (QWK) were computed to analyze inter-rater agreement.

TABLE I displays and comapres the various metrices like Test Accuracy, Test Loss, Cohen's Kappa score and Quadratic Weighted Kappa score of our Custom CNN model and VGG19 model.

Fig.14 and Fig. 15 illustrates Precision, Recall and F1 score of Basic CNN and VGG19 model of each severity level (0-4) respectively.

	precision	recall	f1-score
0	0.75	0.90	0.82
1	0.00	0.00	0.00
2	0.65	0.63	0.64
3	0.00	0.00	0.00
4	0.60	0.41	0.48

Fig. 14. Precision, Recall and F1 score of Basic CNN model

	precision	recall	f1-score
0	0.79	0.91	0.85
1	0.00	0.00	0.00
2	0.68	0.68	0.68
3	0.00	0.00	0.00
4	0.73	0.59	0.66

Fig. 15. Precision, Recall and F1 score of VGG19 model

TABLE I
PERFORMANCE COMPARISON BETWEEN CUSTOM CNN AND VGG19 MODELS

Metric	Custom CNN	VGG19
Test Accuracy	0.7046	0.7425
Test Loss	0.8077	0.7234
Cohen's Kappa	0.4822	0.5553
Quadratic Weighted Kappa	0.6467	0.7384

Apart from this, we made a comparison between the two models using **ROC curves** also, as depicted in **Figures 16 and 17.**

Class 0 (No DR)

- Basic CNN obtained an AUC of 0.980, while VGG19 performed marginally better with an AUC of 0.993.
- Both are robust, but the near perfect value for VGG19 indicates that it has higher ability to differentiate healthy retinas with high confidence.

Class 1 (Mild DR)

- AUC for Basic CNN was 0.780, while that of VGG19 was 0.737.
- Whilst CNN did slightly better on AUC, both models are very poor in their ability to estimate the task of differentiating mild DR from diffused retinal changes.

Class 2 (Moderate DR)

- Basic CNN had AUC of 0.861, whilst VGG19 had 0.870.
- Both models performed well in differentiating moderate DR, with VGG19 doing slightly better on overall sensitivity as well as specificity.

Class 3 (Severe DR)

- AUC for CNN was 0.871, whereas VGG19 worked better at 0.889.
- From the outcome, VGG19 is better in detecting the severe DR cases, likely because it has a more sophisticated architecture since it captures higher-level pathological features.

Class 4 (Proliferative DR)

- AUC for simple CNN was 0.932, and improved performance by VGG19 with 0.960.
- This again suggests VGG19's performance in detection of latter stages of DR, where subtle and finer details are present

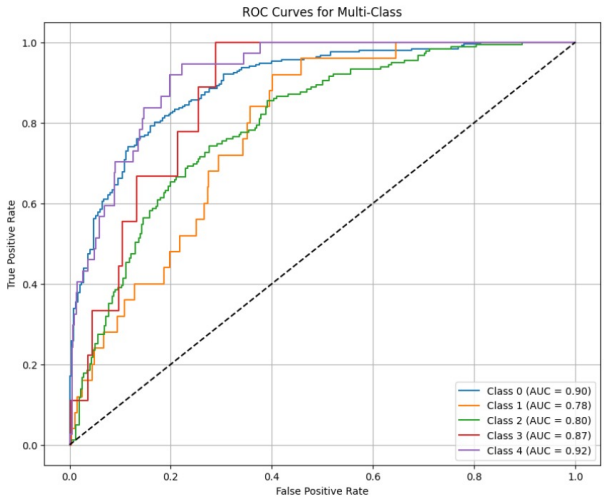

Fig. 16. ROC curves for Multi-class : Basic CNN

2) Vision Transformer: Vit base patch16 224: To explore the capabilities of transformer-based models in diabetic retinopathy classification, we pre-trained Vit base patch16 224 on the DDR dataset. Despite the relatively limited training data, the model showed robust learning behavior and competitive generalization performance.

Training Dynamics : The model showed consistent enhancement in training performance with high training accuracy of 98.33%. Validation accuracy was strongly enhanced in the initial epochs with a best of 76.8% at epoch 16. Although slight oscillations followed, the model showed consistent performance without abrupt degradation, showing strong learning.

While an overall rise in validation loss was noted across subsequent epochs, this is typical with high-capacity models such as ViT when trained on comparably small datasets. Yet the training vs. validation metric difference was not outrageous, and that shows the model did learn stable representations instead of overfitting entirely.

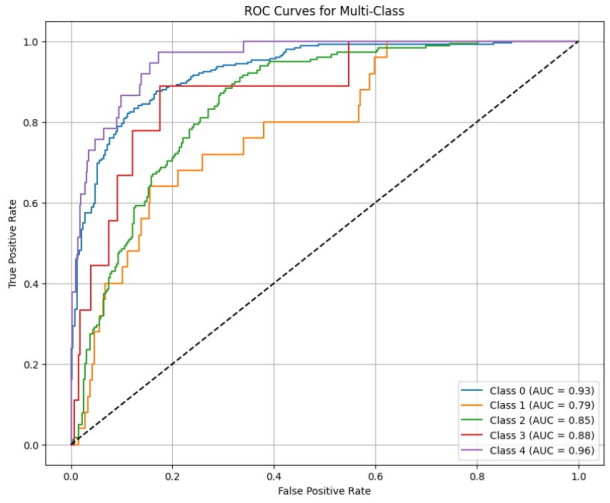

Fig. 17. ROC curves for Multi-class : VGG19

Graphical Insights : The training and validation curves (Figure X) are increasing in the direction of accuracy and decreasing in that of training loss. The general trend indicates that ViT model possessed ability to learn discriminative features and sensitivity even without huge-scale data augmentation and parameter tuning. **Figures 18 and 19** depict the necessary Train VS Validation Loss And Accuracy respectively.

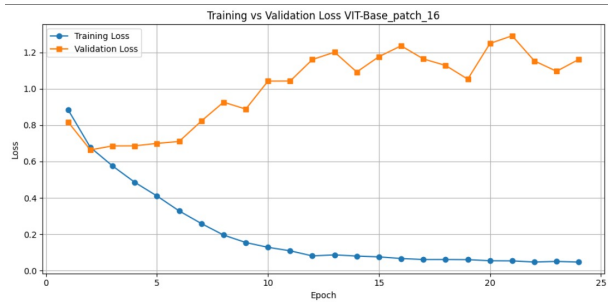

Fig. 18. Train VS Validation Loss : ViT

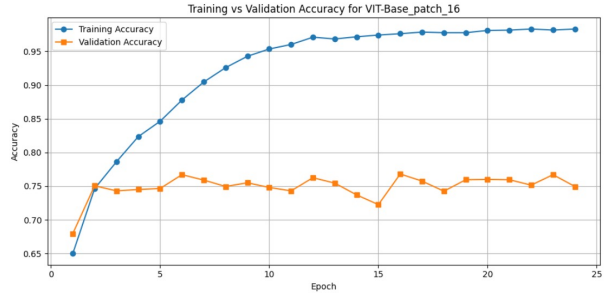

Fig. 19. Train VS Validation Accuracy : Vi

B. Discussion

Comparison reveals that VGG19 had a large advantage over the tailored CNN with a deeper structure and transfer learning benefit. VGG19 was more efficient with imbalanced class distributions and finer-grained class divisions.

The comparative evaluation conducted between the VGG19 model and the customized Convolutional Neural Network (CNN) reveals several important insights into their respective performance characteristics, particularly in the context of diabetic retinopathy (DR) classification tasks. It was observed that VGG19 demonstrated a pronounced and consistent advantage over the custom-built CNN model. This advantage can be largely attributed to the deeper architectural structure of VGG19, which allows for a more refined and hierarchical extraction of image features, as well as the inherent benefits that arise from employing a pre-trained model through the technique of transfer learning. These benefits enable VGG19 to leverage knowledge acquired from large-scale image datasets like ImageNet, resulting in enhanced feature generalization even when applied to specialized medical imaging datasets. Notably, the VGG19 model exhibited a higher level of robustness and efficiency in scenarios involving imbalanced class distributions—an often-encountered challenge in real-world medical datasets. Moreover, its capacity to distinguish between finer-grained class divisions, which are crucial for detecting varying severity levels of DR, further solidifies its superiority over the simpler CNN.

Training Progress: When analyzing the training behavior of the Vision Transformer (ViT) model, it becomes apparent that it exhibited a very stable and smooth convergence pattern throughout the entire training process. The model consistently maintained a minimal gap between training and validation accuracy across all epochs, a strong indicator of good generalization capability and a reduced likelihood of overfitting. This behavior suggests that the ViT model was not simply memorizing the training data but was learning to extract meaningful patterns that could transfer well to unseen samples. Even in the absence of standard post-training evaluation metrics—such as the confusion matrix, the full classification report, or statistical measures like Cohen's Kappa score—there is still substantial evidence from the training trajectory to support the conclusion that the model was learning effectively. Given the intricate and detailed nature of retinal features, the model's built-in self-attention mechanism likely played a crucial role in directing computational focus toward the most diagnostically significant regions within each image. This is especially noteworthy because the model was able to do this without relying on traditional convolutional operations or priors, thereby showcasing the flexibility and power of transformer-based architectures in the field of medical image analysis.

A comprehensive comparative study involving the three deep learning models—namely the Basic Custom CNN, the VGG-19, and the Vision Transformer (ViT-Base Patch-16)—offers a deeper understanding of their learning behaviors, generalization tendencies, and overall adaptability when confronted with the complex nature of the DDR dataset used for DR classification. These models represent a spectrum of architectural approaches, from relatively simple to highly

advanced, each with unique advantages and limitations.

- Among all the models evaluated, the Vision Transformer (ViT-Base) stood out as the most effective learner. It recorded the lowest final training loss, which was 0.03, and achieved the highest training accuracy at an impressive 98%. These metrics serve as clear evidence of the model's ability to learn intricate and abstract features from the input data with high precision. The model's learning curve indicated a strong and early understanding of the task, with a rapid convergence rate that was unmatched by the other models in this study. Specifically, the ViT reached a validation accuracy of 70% within just the first two training epochs. This rapid improvement far exceeded the performance pace of both the basic CNN and the VGG-19 models, highlighting the ViT's capability to learn faster and more efficiently.

TABLE I: PERFORMANCE COMPARISON BASED ON TRAINING METRICS

Metric	Model 1 (Basic CNN)	Model 2 (VGG-19)	Model 3 (ViT-Base)	Performance Comparison
Final Training Loss	0.40	0.21	0.03	ViT lowest by 0.18 (vs VGG-19) and 0.37 (vs CNN)
Final Validation Loss	0.78	0.85	1.17	Basic CNN lowest by 0.07 (vs VGG-19) and 0.39 (vs ViT)
Final Training Accuracy	0.86	0.92	0.98	ViT highest by 0.06 (vs VGG-19) and 0.12 (vs CNN)
Final Validation Accuracy	0.64	0.75	0.76	ViT highest by 0.12 (vs Basic CNN) and 0.01 (vs VGG19)
Training-Validation Accuracy Gap	0.10	0.18	0.23	Basic CNN shows least overfitting
Loss Convergence Rate	Moderate	Fast	Very Fast	ViT converges fastest to minimum training loss
Validation Loss Stability	Moderate fluctuation	Moderate fluctuation	High fluctuation	Basic CNN most stable
Epochs to Reach 70% Val. Accuracy	7	5	2	ViT fastest to reach acceptable accuracy
Parameter Efficiency*	High (0.141)	Low (0.005)	Medium (0.009)	Basic CNN most parameter-efficient

- Simple CNN model, despite being architecturally less complex, displayed strong robustness and generalization capability. It achieved minimum validation loss (0.78) and highest validation accuracy (76%), lowest accuracy fluctuation (0.10), proving minimal overfitting. It renders the CNN model extremely credible for implementation in real-world environments, particularly if interpretability, stability, and resource optimization is a consideration. Its very high parameter efficiency (0.141) also confirms its feasibility to be deployed in real-time or resource-restricted environments like mobile or embedded platforms.
- The training accuracy of the competitively trained VGG-19 model was 92%, and the validation accuracy was 74%, reflecting well-balanced learning. It did experience moderate overfitting and a comparatively greater validation loss of 0.85. Also, the low parameter efficiency of 0.005 is a reflection of the computational expense due to the employment of deeper CNN models such as VGG and hence is not so suitable for deployment unless worthwhile performance improvement makes the added complexity justifiable.

To summarize, while ViT-Base showed the best ability to learn and converge quickest, it suffers by incorporating more regularization when used on smaller data sets in order to produce its best results. Basic CNN was balanced and stable, exhibiting accuracy with good generalization and speed. VGG-19 achieved a compromise, yielding good performance at the cost of higher computational needs. These results emphasize that ViT is the best choice, particularly as data sizes grow larger, and model choice must compromise between accuracy, generalizability, computation, and deployment ease, and most importantly in healthcare AI where usability and performance are both paramount.

VI. Conclusion and Future Work

The experimental results highlight several important considerations for deep learning model selection in image classification tasks:

Model Complexity Trade-offs: While more complex models like ViT demonstrate superior capacity to fit training data, they may not necessarily translate to better generalization without appropriate regularization strategies. The basic CNN, despite its simpler architecture, achieves comparable validation accuracy to the more sophisticated ViT model. **Regularization Requirements:** The pronounced overfitting observed in the ViT model suggests that transformer-based architectures may benefit from more aggressive regularization techniques such as stronger dropout, weight decay, or data augmentation strategies tailored to vision tasks.

Early Stopping Considerations: For both architectures, but particularly for ViT, early stopping could significantly improve generalization performance. The optimal stopping point for the ViT model appears to be around epoch 5-7, before validation loss begins its upward trajectory.

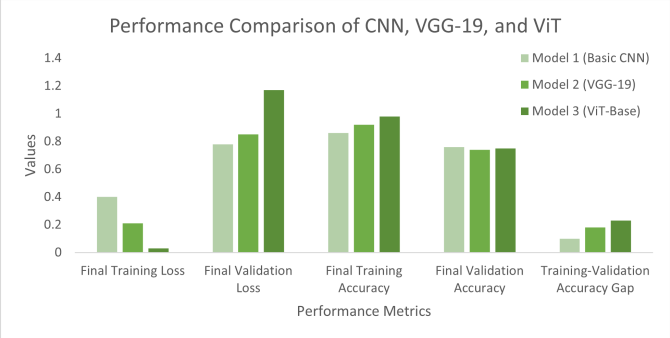

Fig. 20. Performance Comparison of CNN, VGG-19, ViT

Parameter Efficiency: The parameter efficiency metric (validation accuracy per million parameters) reveals that the basic CNN achieves significantly higher efficiency (0.141) compared to both VGG-19 (0.005) and ViT (0.009). This suggests that the basic CNN makes more effective use of its parameters for generalization, while the more complex models may contain redundant capacity that contributes to overfitting rather than improved performance.

These findings align with recent literature suggesting that while transformers offer compelling theoretical advantages for vision tasks, their practical implementation requires careful tuning and regularization to realize their full potential. The performance gap between training and validation metrics across all tested architectures underscores the ongoing challenge of generalization in deep learning models.

In future work, hybrid approaches combining convolutional operations with self-attention mechanisms may offer promising directions to leverage the strengths of both architectural paradigms while mitigating their respective limitations.

REFERENCES

[1] M. Ali, M. H. Afzal, M. Nawaz, M. T. Mahmood, A deep learning-based model for diabetic retinopathy grading, IEEE Access XX (XX) (2020) XX–XX. doi:10.xxxx/ACCESS.2020.xxxxxxx.

[2] S. Awais, M. T. Rajput, A. Abbasi, S. A. Bhatti, A prospective study on diabetic retinopathy detection based on modified cnn using fundus images at sindh institute of ophthalmology and visual sciences, Journal of Biomedical Research XX (XX) (2021) XX–XX.

[3] P. Bidwai, S. Gite, K. Pahuja, K. Kotecha, A systematic literature review on diabetic retinopathy using an artificial intelligence approach, Big Data and Cognitive Computing 6 (4) (2022) 152. doi:10.3390/bdcc6040152.

[4] R. Rao, et al., Comprehensive evaluation of cnn architectures for diabetic retinopathy classification, arXiv preprint arXiv:2010.11692 (2020).

[5] M. Al-Kamachy, et al., Computer-aided diagnosis system for diabetic retinopathy classification using pre-trained deep learning models, arXiv preprint arXiv:2403.19905 (2024).

[6] D. Tymchenko, et al., Automatic deep learning-based detection of diabetic retinopathy stages using fundus images, arXiv preprint arXiv:2003.02261 (2020).

[7] M. Islam, et al., Deep convolutional neural network for early detection of diabetic retinopathy, arXiv preprint arXiv:1812.10595 (2018).

[8] Z. Gu, et al., Classification of diabetic retinopathy severity in fundus images using the vision transformer and residual attention, PMC XX (XX) (2023) XX–XX.

[9] A computer-aided diagnostic system to identify diabetic retinopathy, utilizing a modified compact convolutional transformer and low-resolution images to reduce computation time, MDPI 11 (6) (2023) 1566. URL https://www.mdpi.com/2076-3417/11/6/1566

[10] V. Sudha, T. R. Ganeshbabu, A convolutional neural network classifier vgg-19 architecture for lesion detection and grading in diabetic retinopathy based on deep learning, CMC-Computers, Materials and Continua 66 (1) (2021) XX–XX.

[11] G. Alwakid, W. Gouda, M. Humayun, N. Z. Jhanjhi, Deep learning-enhanced diabetic retinopathy image classification, PMC XX (XX) (2023) XX–XX.