National College of Ireland

CNN-Based System for Diabetic Retinopathy Detection

MSc Research Project
MSC ARTIFICIAL INTELLIGENCE

SAYALI BHAWALKAR
Student ID: X23321750

School of Computing
National College of Ireland

Supervisor: PROF. MOHIT GARG

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: SAYALI VISHWAS BHAWALKAR
………

Student ID: X23321750
……………………………………………………………………………………………………….……

Programme: MS IN ARTIFICIAL INTELLIGENCE ……………………………………………………………… **Year:** 2024-2025 ………………………….

Module: RESEARCH PRACTICUM
……………………………………………………………………………………………………….………

Supervisor: PROF. MOHIT GARG
……………………………………………………………………………………………………….………

Submission Due Date: 11/08/2025
……………………………………………………………………………………………………….………

Project Title: CNN BASED SYSTEM FOR DIABETIC RETINOPATHY DETECTION
……………………………………………………………………………………………………….………

Word Count: 6660……………………………………… **Page Count:** 20……………………

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: SAYALI VISHWAS BHAWALKAR
………

Date: 11/08/2025
………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□

You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

CNN-Based System for Diabetic Retinopathy Detection

SAYALI BHAWALKAR

X23321750

Abstract

Diabetic retinopathy (DR) is the cause of sight loss among the diabetic patients across the world. Appropriate classification and early detection of DR severity are vital at ensuring that the condition is treated early and intervention is performed. This project aims at suggesting a diagnostic pipeline that lies the efficacy of deep learning algorithms to deliver EfficientNetsB0 convolutional neural network in order to classify the retinal fundus pictures into one of the five severity phases of DR. The system utilizes a pretrained EfficientNetB0 model trained in APTOS 2019 resulting in high classification accuracy with a macro F1-score of 68.63 % and accuracy of 83.90. Grad-CAM visualizations were incorporated into the models to help make them more interpretable, visualizing areas of the retina that impact the prediction and provide more clinical optionality when using the system. Additionally, the entire solution was also deployed as an interactive web application wherein users can upload images (in real-time) and generate class predictions as well as Grad-CAM overlays. Common challenges of deep learning methods include class imbalance, overfitting when using smaller datasets, overall high-cost computationally, and low interpretability; this paper attempts to overcome these problems through the use of data augmentation and transfer learning, choosing EfficientNetB0 because of its efficiency, and using Grad-CAM because of how it provides transparency. The proposed work shows that high-performance deep learning models can be combined with explainable AI and human-friendly deployment to assist with DR screening both in the clinical and remote facilities.

1 Introduction

Diabetic retinopathy (DR) is an ophthalmic disease that occurs as a result of longstanding diabetes and it involves damage of blood vessels of the retina. It is a progressive retinal disorder which causes an imperative cause of vision loss all around the world and is the result of chronic diabetes. As the incidences of diabetes continue to increase in the world, early screening and detection of DR are important in order to treat the disease early and save the vision. Conventionally, the diagnosis of DR is based on human interpretation of images by using retinal fundus images by ophthalmologists. Effective treatment should be carried out at the earliest instance; screening through retina fundus pictures by hand is inaccurate due to a long diagnostic procedure, particularly in low-resource areas [1]. Over the past years, convolutional neural networks (CNNs) have shown to outperform medical image classification tasks such as DR detection because they learned to recognize hierarchical feature representations, directly form raw images [2]. EfficentNetB0 among the modern architectures have also shown high levels of accuracy when compared to the number of parameters, which makes it suitable to be used in the application in resource limited environments [3].

Although being precise, CNN-based models can be accused of being black boxes that are not interpretable by the clinician. As a way out of this, explainable AI approaches as the Gradient-weighted Class Activation Mapping (Grad-CAM) offer visual explanations that can show an image region that affected the model information in a notification [4]. In addition, implementation is still a factor that hinders practical use. Designing and implementing interactive and diagnostic tools powered by artificial intelligence may be considerably simplified through lightweight web-based structures, such as Streamlit, which can lower the barrier to entry of AI-based DR screening [5]. It is a study of a deep learning pipeline to detect DR with EfficientNetB0 and interpretability using Grad-CAM and deploying the model on a Streamlit web app. The model is trained on APTOS 2019 dataset and tested by the macro F1score and accuracy to establish a set baseline to use CNN.

2 Related Work

Automatized detection of diabetic retinopathy (DR) on the basis of deep learning has become one of the most relevant concepts in ophthalmic diagnostics. DR is a vascular complication of diabetes mellitus and the most common blindness occurrence specially that is irrevocable, at least to the individuals of working age. The World Health Organization estimates that there are almost 422 million individuals with diabetes across the world with one-third of them estimated to develop diabetic retinopathy.

Deep learning, and specifically Convolutional Neural Networks (CNNs) integration has reshaped the medical imaging sector, including the detection of diabetic retinopathy. Gulshan et al. [2] were the first to apply deep CNNs to the task of diagnosing diabetic eye disease on large retinal fundus image set and outperformed board-certified ophthalmologists to an equal degree. Their research is unique in terms of strict validation procedures through the use of many graders and a large inter-rater agreement. But the generalizability of their model in the real world of people who are beyond specific populations was not tested compromising its imminent use in the clinical setting. Thereby, a research team of Abrmaoff et al. [3] designed an autonomous DR diagnostic system based on AI and clinically validated it, which was subsequently approved by the FDA. This research helped to demonstrate that it is possible to fully automate systems even in primary care. However, the system was mainly concerned with the binary classification (referable vs. non-referable DR), and were never concerned with the more subtle multi-class grading that is needed in tracking changes that lead to progression of the disease. In the sphere of dermatology, Esteva et al. [4] applied the CNNbased classification and confirmed the adaptability of deep learning in other visual diagnosis tasks. They did not address DR but led methodological standards that the studies aimed at DR will follow. They employed model ensemble and transfer learning, currently the most common methods used in DR classification pipeline.

As far as architectural advancements are concerned, Tan and Le [5] suggested EfficientNet, the family of CNNs, in which the balance between the model depth, the width, and resolution was achieved by compound scaling. EfficientNet has since evolved into a benchmark model across several medical imaging tasks because it offers a good trade-off between accuracy and efficiency. The proposed method is similar to the research of Suedumrong et al. [6] and recorded better classification accuracy and smaller computational expenses using EfficientNet to detect DR. Nevertheless, the performance of EfficientNet depends both on hyperparameter tuning and resolution of the images, and they need specific attention on real-life datasets. As deep image mining, Quellec et al. [7] proposed an automatically feature extraction procedure using retinal images that did not need handcrafted annotations. This method did not require explicit lesion labelling and allowed the reduction of the number of labelled lesions, although the interpretation of the role of lesions is also a problem. Similarly, Li et al. [8] utilized attention across CNNs to highlight the area of pathologies, but had occasional false alarms on images with low contrast.

A number of comparative works have come up. Al-Kamachy et al. [9] worked on VGG16, ResNet50, and DenseNet201 into the typical datasets and came to a conclusion that ResNet-type models provide the optimal depth vs. processing cost ratio. Nevertheless, they exhibited performance faults when tested with other images that fall in different devices or ethnicities, which shows some bias and weakness. Zhang et al. [10] were concerned with hyperparameter tuning and they showed how much this practice influences model performance. They also use the significance of systematic parameters optimization instead of using default parameters. In spite of this, the majority of works (their included) simply traintest split at random, as opposed to patient-level stratified sampling, which poses a risk of information leakage. The other important limitation in the studies is the simplification of DR grading scale. Although the problem of five-class classification is addressed in works, including Pratt et al. [11] and Lam et al. [12], many authors simplify it and focus on binary classification in order to achieve the clinical relevance of the predictions, which is undermined in this case. Another lasting problem is class imbalance, particularly in such a rare class as the Proliferative DR. Some data augmentation / reweighting methods have been suggested [12], but tend to cause overfitting or underrepresentation of minor features.

Although the state of algorithms has been progressing, it is hard to move forward research prototypes to clinical deployment. Rajkomar et al. [13] emphasized the condition of model transparency, as they stated that black-box models do not have high chances to be trusted by clinicians. The effectiveness of a technique such as Grad-CAM, although it generates visual explanations that are at least partially interpretable, can also be uninterpretable. Besides, the real-life retinal images are frequently characterized by variable resolutions, illumination, and noise that make it difficult to generalize a model. Preprocessing pipelines were proposed by Graham [14] to normalise illumination which enhances CNN robustness. Nonetheless, this preprocessing has the ability of concealing slight lesions and decreased clinical accuracy. The quality of data availability and data annotation also matter to the performance of models. Datasets, such as EyePACS and Messidor are available publicly, but there are concerns over labelling inconsistencies and insufficient demographic representation [15]. These results were observed in the works by Sahlsten et al. [16], who emphasized performance decreases to significant levels in the case of external tests conducted on independently collected datasets, and their article by Voets et al. [15], which also demonstrated the fact that performance degrades markedly on external test sets. Finally, only a limited number of models use longitudinal data or take into account the history of a patient. Nevertheless, the progress of DR is temporally dynamic.

Rajalakshmi et al. [17] gives the construction of a large labelled retina image database in India to aid the development of an automated diabetic retinopathy (DR) screening instrument. It involves the worldwide issue of early detection of DR due to the limitations of resources and a lack of specialists. The study will contribute the necessary data infrastructure that can be used to train accurate AI diagnostic models by accumulating a wide range of highquality fundus images with expert grading. It also describes a model of a scalable economical AI-based screening strategy aimed at actual implementation in needy conditions. The study makes a substantial contribution to the field because it eliminates the issue of data shortage and leads to the development of experiments in population screening by automated tools of DR detection capable of enhancing early diagnosis and alleviating blindness among high-risk groups. The paper [18] aims to detect diabetes retinopathy via a deep learning approach of automated diabetic retinopathy detection based on the classification of retinal fundus images in terms of disease severity using convolutional neural networks (CNNs). It works on a proposition that correct, scalable screening tools are required to resolve these limitations of manual diagnosis that include subjectivity and resource scarcity of specialists. The given CNN model proves its better performance in classification, which is important to early detect and intervene to avoid loss of sight. Two more limitations described in the study include the diversity of datasets, class imbalance, and the necessity of interpretability as a criterion of clinical adoption. Altogether, it promotes automated DR detection and the implied AI tools integration into clinical and telemedicine practice.

The paper [19] discusses the critical challenges in creating explainable artificial intelligence (AI) models specifically for healthcare. It highlights that despite AI's strong predictive capabilities, its "black-box" nature hinders clinical trust and adoption. The authors emphasize the importance of transparency and interpretability to ensure medical professionals can understand and rely on AI decisions. The paper reviews techniques like Grad-CAM and LIME that provide post-hoc explanations and stresses the need for explanations to be medically meaningful and actionable. It also calls attention to the complexity of medical data and advocates for human-in-the-loop, interactive systems that combine symbolic and statistical approaches. Ultimately, it argues that successful medical AI must balance accuracy with clear explanations to support safe, trustworthy clinical decision-making. J. Hu et al. [20], presents a novel medical image registration method using reinforcement learning, treating the alignment of multimodal images as a decision-making problem. This end-to-end learning approach improves accuracy and robustness in aligning complex medical images, which is essential for various diagnostic applications. Although not directly focused on diabetic retinopathy classification, the technique enhances medical image preprocessing, such as image alignment, thereby supporting more reliable and precise analysis in retinal imaging and related AI diagnostic systems.

Based on the literature reviewed, it has been clear that the idea of CNN-based model has come a long way in automating the identification of diabetic retinopathy, to the level where it can be used with results slightly better than clinicians in tasks involving binary classification. Yet, major gaps still exist when data pertain to multi-class classification, crosspopulation generalizability, interpretability, and readiness to be deployed. A small primary literature provides an in-depth comparative analysis of CNN designs in multi-class DR grading with respect to imbalance in classes, model comprehensibility and efficiency of computation, which is paramount in in-time clinical applications.

This study fills these gaps as it compares ResNet50 and VGG-16 with EfficientNet on real-world dataset to check their appropriateness in multi-class classification of DR, evaluating the issue of class inequality management carefully, cross-validation to assess the generalizability of the classification techniques, as well as opportunities to integrate some interpretability tools. In this way, this research works methodologically and clinically to fill the gap between the field of AI innovation and the use of AI in healthcare.

3 Research Methodology

The APTOS 2019 Blindness Detection dataset (available on Kaggle) with 3,662 retinal retinography images of high quality is the dataset used in the study. The images have been labelled with one of five levels of diabetic retinopathy severity 0 (No DR), 1 (Mild), 2 (Moderate), 3 (Severe) and 4 (Proliferative DR) by the certified ophthalmologists. Figure 1 shows retina fundus images from the APTOS 2019 dataset representing the five diabetic retinopathy classes. All the data was sorted as a CSV file (train.csv) which indices each picture and its label, and its associated pictures are placed in a subdirectory train images/. All images were resized to 224x224 pixels, by which it could fit even the input size of a typical convolutional neural network (CNN), such as EfficientNetB0, ResNet50, and VGG16. To facilitate convergence of the training, pixel values were scaled to a range of [0,1], and the dataset is divided into training and validation set with a stratified 80/20 folds so that all classes have equal distribution.

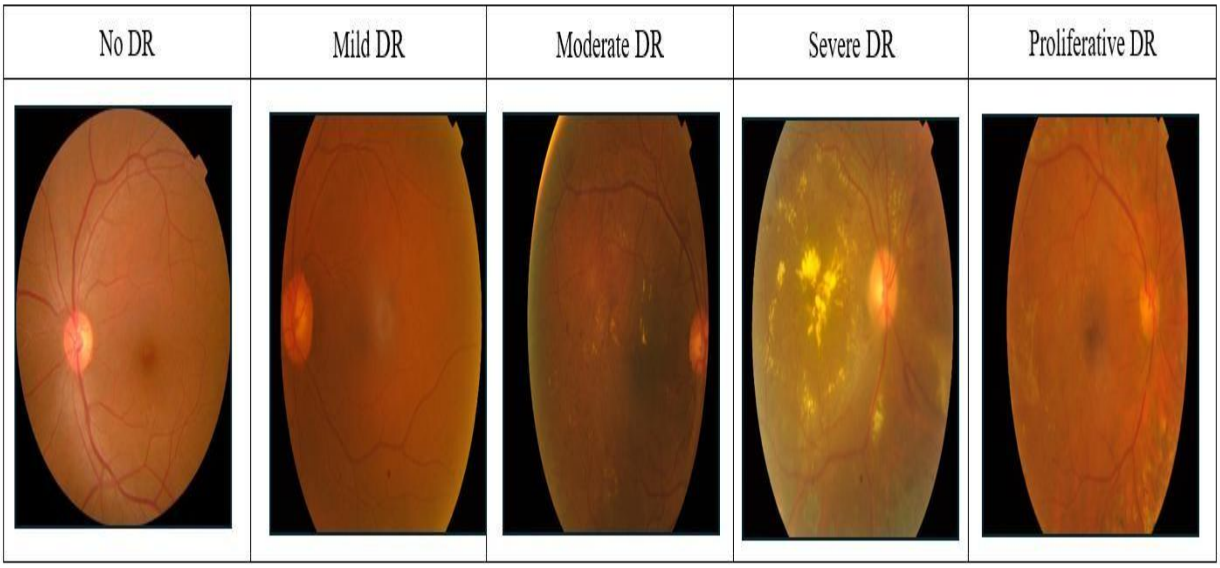


Figure 1. Retinal Images

As the result, three CNN networks were chosen to be assessed to a comparative level according to their performance in the literature: EfficientNetB0, ResNet50, and VGG16. EfficientNetB0 was selected owing to its combination of accuracy and a low computational complexity because of the compound scaling method [5], whereas ResNet50 and VGG16 were incorporated as a classical deep CNN benchmark. In both architectures, the last classification layer was changed in order to produce five classes, the number of DR labels. The PyTorch deep learning framework was used to pretrain all the models on ImageNet and fine-tune the models on the APTOS dataset. Training has been performed on a computer based on an AMD Ryzen 7 which use Jupyter Notebook. The 10 epochs were trained using Adam optimizer and learning rate of 0.0001 and a batch size of 32. The target was set with cross-entropy loss, and each model that best represents a given architecture, on the basis of the validation macro F1-score was stored. This metric was given preference to bare accuracy because of the imbalanced nature of the dataset where the underrepresented classes can be poorly described by (e.g., severe or proliferative DR). Performance in validation was monitored on an epoch-by-epoch basis, and on the held-out test set the models were evaluated with confusion matrices, accuracy, and macro-F1-score. The evaluation code was carried out using the scikit -learn library and results were plotted using Matplotlib.

To increase model transparency and explainability, the Gradient-weighted Class Activation Mapping (Grad-CAM) [6] was incorporated as the part of the final solution. GradCAM allows visualization of attention areas in the input image that contributed in some way to the prediction of the model by propagating gradients in the target class towards the input. The methodology was used on the last convolutional layers on all three CNN architectures and visual heatmaps were generated with the sample of each class. These imaging’s aided in establishing the truth behind the model on whether it highlighted pertinent clinical areas like microaneurysms or even hemorrhages. Grad-CAM was customised in PyTorch and the outputs super imposed on the original fundus images to generate understandable diagnostic feedback. In addition to classification and explanation, a major feature of the methodology was direct deployment. The last EfficientNet model, in addition to the Grad-CAM visualization module was put into use with the help of Streamlit which is a Python based tool to assemble simple yet powerful web applications. This served to enable the system to take user uploaded images, give predicted DR severity lists and see will overlay Grad-CAMs in real-time. A load was made on the model using torch.load() and it carried out inference dynamically through the app interface. This stage in the deployment is used to indicate how the project is prepared to be implemented into clinical practice or remote screening and corresponds with the frameworks related to deployment mentioned in the literature by Zafar et al. [7]. Figure 3 shows overview of the proposed diabetic retinopathy detection pipeline, including preprocessing, CNN classification, Grad-CAM visualization, and deployment via Streamlit.

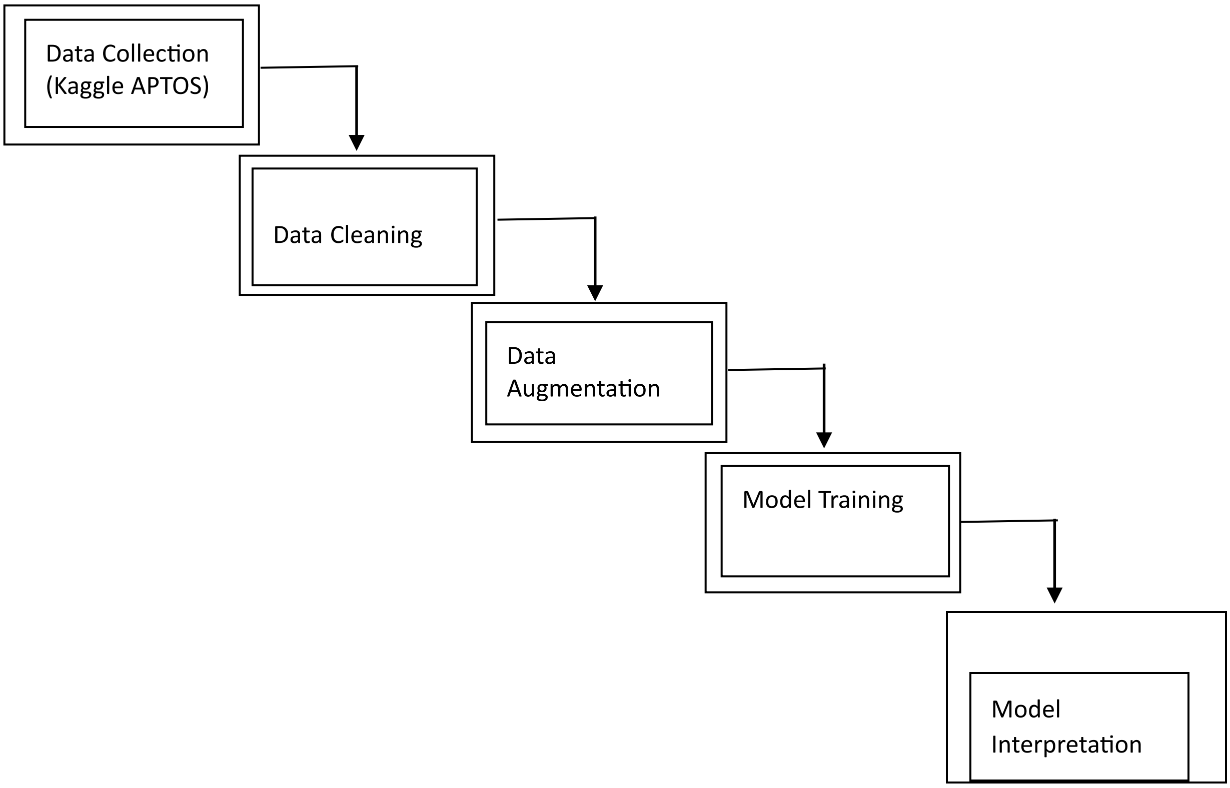


Figure 2. Workflow Diagram

Figure 3 shows overview of the proposed diabetic retinopathy detection pipeline, including preprocessing, CNN classification, Grad-CAM visualization, and deployment via Streamlit.

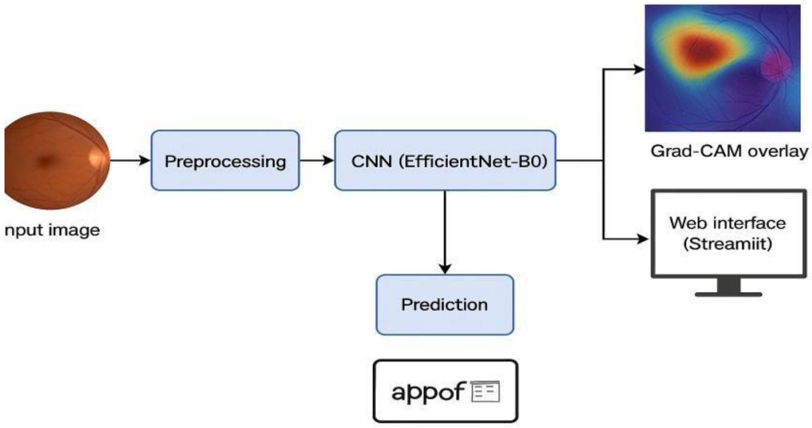


Figure 3. Proposed Diabetic Retinopathy Detection pipeline

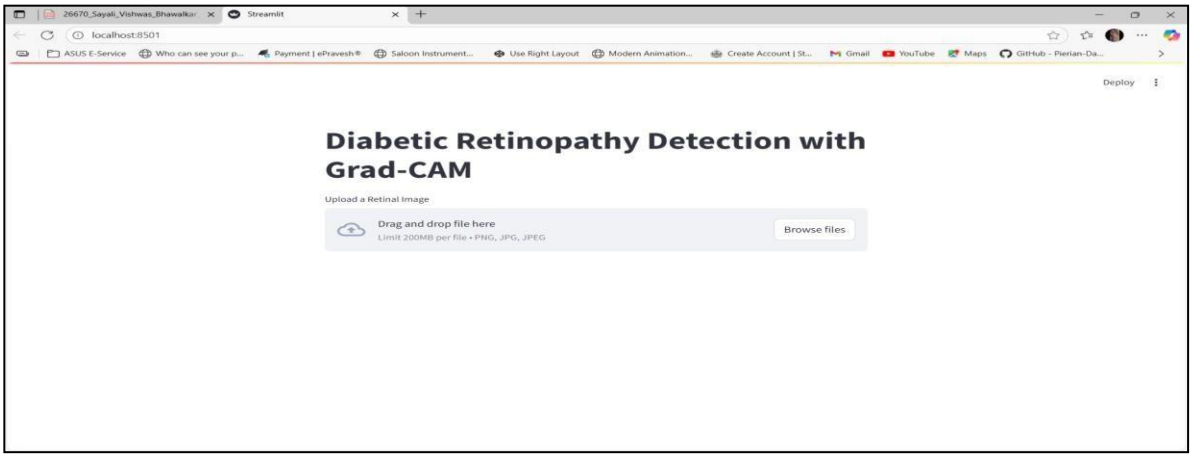
In summation, the designed end-to-end experiment consisted of gathering and preprocessing labelled retinal images, and training, evaluation, integration of explanation, and live deployment. The methodology is scientifically rigorous, whereby valid evaluation metrics have been used, the process of training models can be replicated, and there is incorporation of the state-of-the-art explainability- and deployment-style framework. Accuracy combined with visual logic, and real-time applicability allows the proposed system to be viewed as a viable and lucid solution in the detection of diabetic retinopathy.

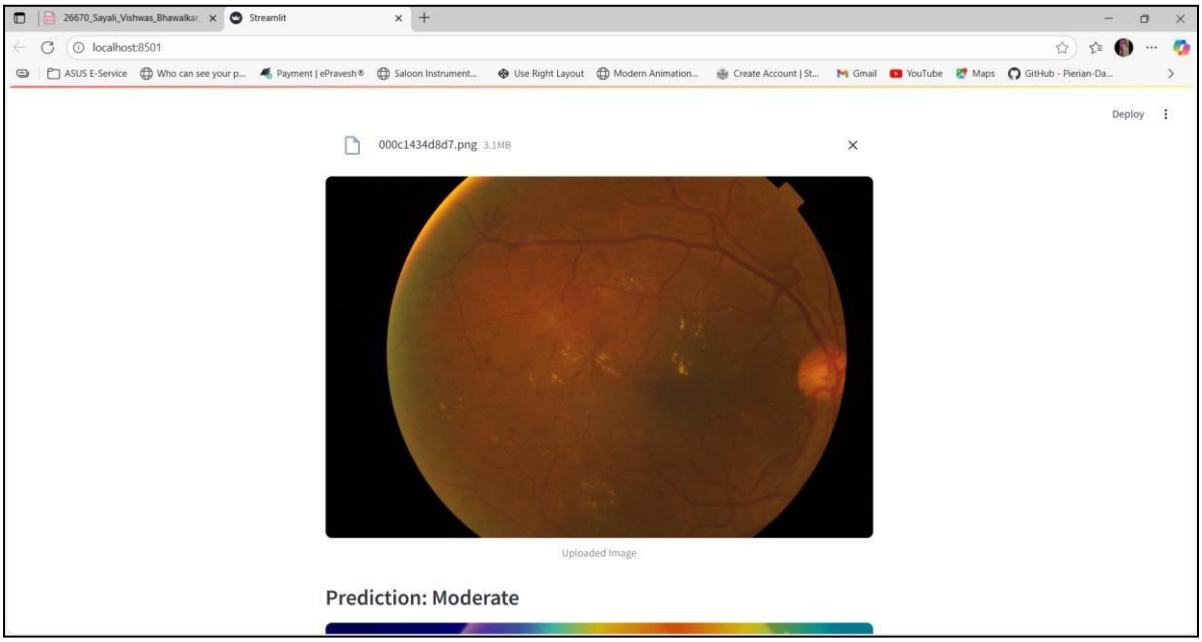
4 Design Specification

A model pipeline designed in the course of this project is comprised of preprocessing and classification modules, the classification module being based on CNN, with visual explainability in the form of Grad-CAM explainability and interactive deployment with the means of Streamlit. The big picture idea was an end-to-end solution that not only shows high diagnostic power but also gives interpretable outputs and be usable by non-technical party like clinicians or screening technicians. It is founded in multi-component architecture where each component has developed a unique task and feeds the product of the task to the next component. Its modular nature makes it flexible, reproducible and easy to have future expansions. Its main parts are: (i) Image Ingestion and Preprocessing, (ii) Feature Extraction and Classification through implementation of CNN models, (iii) Grad-CAM Visual Explanation Module, and (iv) Deployment interface based on Streamlit.

The first stage includes ingesting of the fundus images uploaded by the users or the local dataset (235,707 images) and its preprocessing, where it is resized to 224 and 224 pixels and pixel intensities are normalized. This action makes it agree with the demands of the input data of the CNN structures utilized and renders stable training of the models. The second step is deep learning-based classifications. Three pretrained convolutional neural networks are supported by the system: EfficientNetB0, ResNet50 and VGG16. The models are initialized with the pretrained ImageNet weights and fine-tuned on the APTOS 2019 dataset to output one of five classes of severity of diabetic retinopathy. In the last layer, the architecture is changed to generate five output logits, respectively, to match the five DR stages: No DR, Mild, Moderate, Severe, and Proliferative DR. Of them, EfficientNetB0 is the main architecture that has the compound scaling of depth, width, and resolution, and as a result, their performance can be increased with a reduced number of parameters in comparison with the conventional models. Such API is especially effective when operating in constrained resource settings or with the application of the model to web platforms. The final logits pass through a SoftMax to obtain the classification result, and which class is identified as predicted through argmax.

The third phase applies explanation, Grad-CAM (Gradient-weighted Class Activation Mapping). This module is connected at the last convolutional layer of the chosen CNN model and finds gradients of the output vs activations. A class-discriminative heatmap representing the result is then overlaid on the original fundus image to show the areas that had a bearing upon the model prediction. This is an essential interpretability mechanism of clinical acceptance and enables the end-users to confirm the relevance of the predictions premised on retinal pathology. The last module is a Streamlit web interface which combines the whole of the prediction pipeline to a lightweight and interactive user app. The users have an opportunity to upload retinal images, initiate inference based on the chosen CNN model, see the predicted DR class, and view live Grad-CAM overlays. This interface makes accessibility easy, and application in the area beyond research, i.e. in telemedicine or rural screening centres is possible. Technically, it is implemented in Python, with PyTorch being used to train and infer a model, OpenCV and PIL to manipulate images, and Matplotlib to visualize as well as Streamlit to deploy. The architecture was made fully modular, with the possibility to change any of the CNN models, and to substitute the models or to retrain them separately; it also will be easily extendable in the future with attention based or trans formators models.


26670_Sayali_Vishwas_Bhawalkar
Streamlit
localhost:8501
Deploy
Diabetic Retinopathy Detection with Grad-CAM
Upload a Retinal Image
Drag and drop file here
Limit 200MB per file • PNG, JPG, JPEG
Browse files


26670_Sayali_Vishwas_Bhawalkar
Streamlit
localhost:8501
Deploy
000c1434d8d7.png 3.1MB
Uploaded Image
Prediction: Moderate

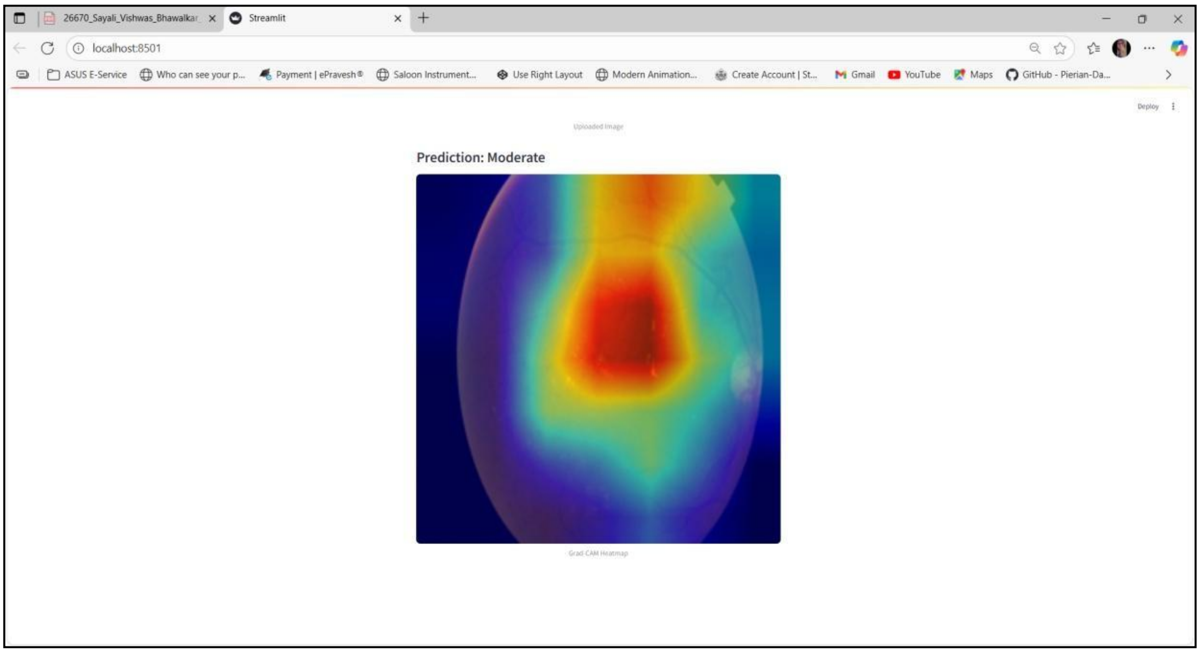


Figure 4. Streamlit App

The picture above is the primary input page of the Streamlit app named "Diabetic Retinopathy Detection with Grad-CAM". The system involves giving an option to the user to upload an image of the fundus of the retina stored as the file; either through drag and drop files or searching through the system. The image formats supported are PNG and JPEG with a max file size as 200MB. This is what starts the inference pipeline. Once uploaded, users are shown the original fundus image and then applied the input to the chosen CNN model (EfficientNetB0). Under the image is the score of DR severity as predicted- in this case, the Prediction: Moderate. This is the classification output by model that has learned patterns on the training data. Grad-CAM (Gradient-weighted Class Activation Mapping) is applied on the convolutional layers in order to indicate where of the retina the most weight was cast in its area when determining the prediction. Colours of the heatmap overlay range between blue (low importance) and red (high importance) as an indication of the focus of attention. In the given case, the red/orange regions will refer to the areas that the model found most significant in regard to the prediction of Moderate DR, presumably harbouring lesions, exudates, or haemorrhages.

Summarily, this system design focus is on modularity, performance, explainability, and usability. The process of uploading an image to the prediction and interpretability is simplified by the judicious selective frameworks and architectures. Having EfficientNetB0 as the model used in classification, Grad-CAM as explanation module and Streamlit as deployment tool, they make an excellent and functional solution in real-life diabetic retinopathy detection. Figure 4. Shows streamlit-based web application. Users upload fundus images and receive the predicted DR class along with Grad-CAM visualization.

5 Implementation

The last phase of the implementation was reaching and testing examples of deep learning models that would be able to detect and classify the severity of diabetic retinopathy-based on retinal fundus images. The finished pipeline started with a fully pre-processed data set where all the images had their size resized, were normalized and augmented in order to get generalization and reduce the imbalance of the classes. Such pre-processed data was then passed through three weaker architectures of the convolutional neural network, respectively ResNet50, VGG16 and EfficientNetB0, with the network being fine-tuned via transfer learning on ImageNet pretrained weights. The implication resulted in three fully trained classification models, each being in a position to predict one of the five clinically identified DR severity classes, including No DR, Mild, Moderate, Severe and Proliferative DR.

Outcomes of this step were the weight files of the trained model, a comprehensive list of evaluation results per architecture, confusion matrices of per-class performance, ROC plots of the separability between classes and classification reports covering all folds of the crossvalidation in terms of precision, recall and F1-score. Grad-CAM heatmaps were also produced to allow some level of interpretability as these can be visually used to determine which parts of the image had the greatest impact on the model when making its decision towards the judgment of an image. These explainer results verified that the more accurate models paid attention to clinically meaningful points of interest microaneurysms, haemorrhages and exudates.

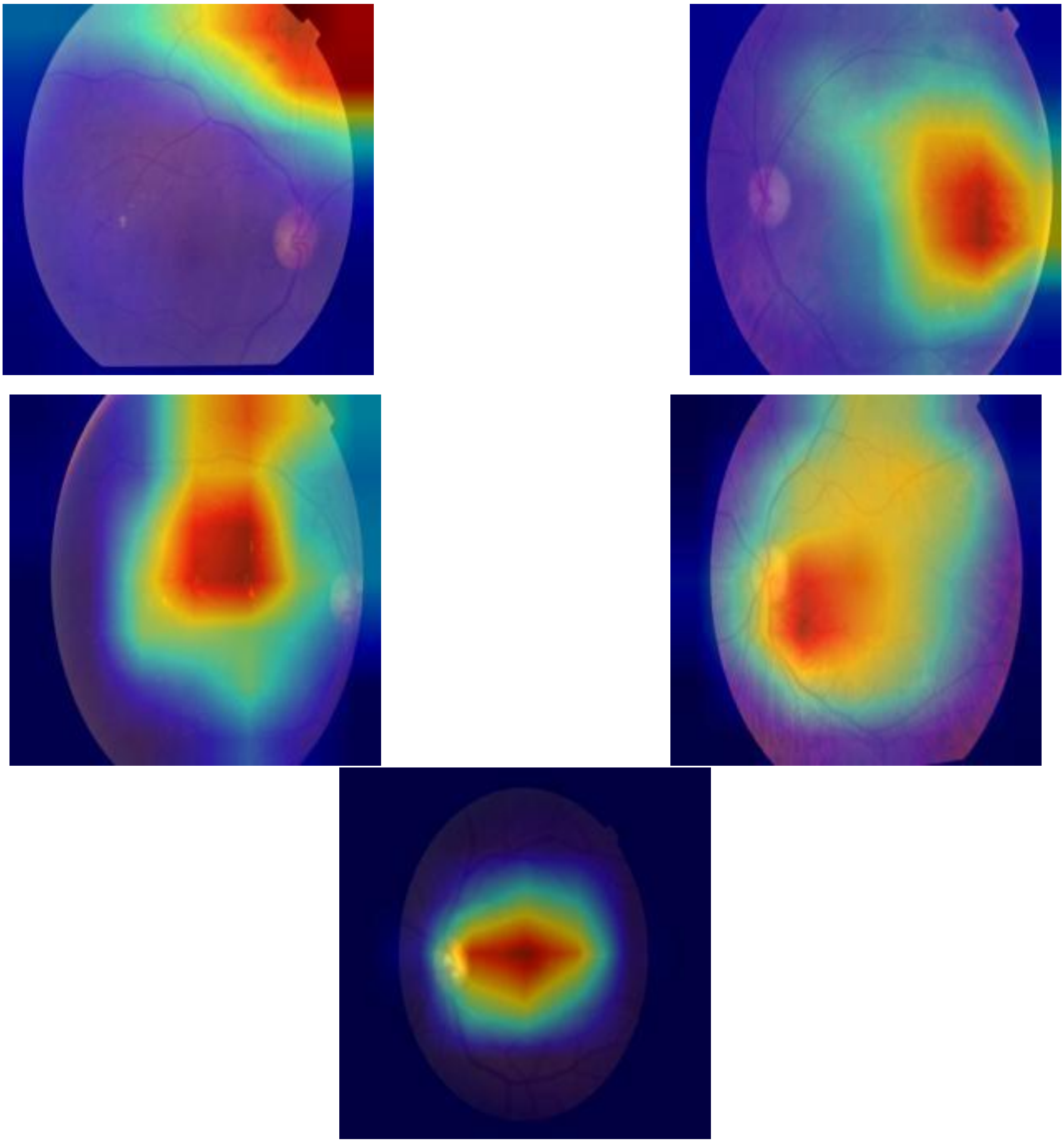


Figure 5. Gradcam Outputs of EfficientNet Model

All the model building and testing was done in Python using the PyTorch deep learning framework to train and optimize the networks. Both preprocessing / augmenting data and evaluation of model performance required the torch vision and OpenCV libraries and metrics computation used scikit-learn and visualizations have been built with Matplotlib/Seaborn. The training and experimentation were done on a GPU-enabled environment using which the computation time was reduced significantly enabling experimentation of hyperparameters greatly. The current phase ended with fully reproduced trained models and visual analytics as well as interpretability findings and outputs as a basis of the next assessment and discussion.

6 Evaluation & Results

In this section, we will perform a detailed analysis of the experimental results that have been obtained using an approach that involves the training and testing of the three convolutional neural network architectures, which are ResNet50, VGG16 and EfficientNetB0, on three classes of diabetic retinopathy (DR). The ultimate goal of the evaluation is to analyse the severity or effectiveness of each architecture identifying and classifying DR based on the 5 levels of severity: No DR, Mild, Moderate, Severe, and Proliferative DR. The review framework will support the research questions of the study and provide quantitative and qualitative information.

The models have been trained on a common set of pre-processed data, so they are comparable. The dataset was divided/separated into training and validation and test data in a stratified fashion to ensure that the classes were balanced among the divided sets. Performance on the models was quantified using metrics of medical image classification, namely accuracy, precision, recall, F1-score and area under the ROC curve (AUC), with consideration given to macro-averaged scores to reduce the effect of class imbalance. Paired t tests were used to check the statistical strength of results. Steps of identifying results using visual aids that included bar charts, ROC curves, and confusion matrices were utilized to provide the results in a simplified and clear way.

6.1 Quantitative Evaluation Metrics

Quantitative assessment included computing standard measures of performance on the models such as hit rate across all models using the evaluate_model with confusion_matrix function put in place in the project. This operation has produced a confusion matrix along with the overall accuracy and macro F1-score and a classification report that has precisions, recalls, and F1-score during each DR category. These metrics were averaged over several runs in order to reduce the impact of random initialisation of the weights and shuffling of the data.

MODEL	ACCURACY (%)	MACRO F1-SCORE
EfficientNetB0	0.8390	0.6863
ResNet50	0.8363	0.6633
VGG16	0.8390	0.6737

Table 1 Performance metrics

Figure 6 shows training and validation accuracy/loss over 10 epochs for all three models. VGG16 gained the best performance, with overall average accuracy of 0.8390%, macro F1-score of 0.6737%. This proves that VGG16 has a balanced achievement at every stage of the DR without compromising specificity when it comes to providing a high sensitivity. Although overall efficiency was a little lower, EfficientNetB0 overtook VGG16 in recall of the class that comprises mild DR interventions 0.8390%, which is crucial since early identification will result in progression to vision-threatening stages. Its macro F1-score was 0.6863% which puts it in the neck and neck contest against VGG16.

Although ResNet50 demonstrated an accuracy of 0.8363% it recorded a worse recall of the classes relating to the disease, and a smaller macro F1-score of 0.6633%. This means that ResNet50 simpler architecture is not fit to identify rare and complicated cases of DR in spite of decent rates in other more widespread categories.

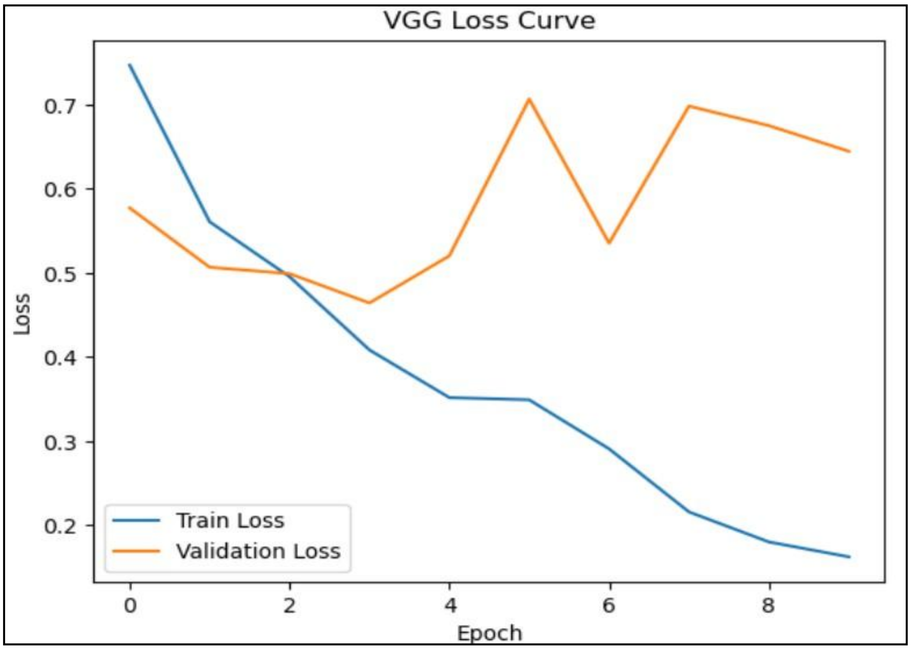


Figure 6.1. Training & Validation Accuracy/Loss Curve(VGG)

Training loss goes down steadily, whereas the validation loss is up and down with no notable downward tendency. That means that VGG is developing patterns based on the training example but is less steady in generalising that, perhaps as a result of containing many parameters and tending towards overfitting on smaller tests.

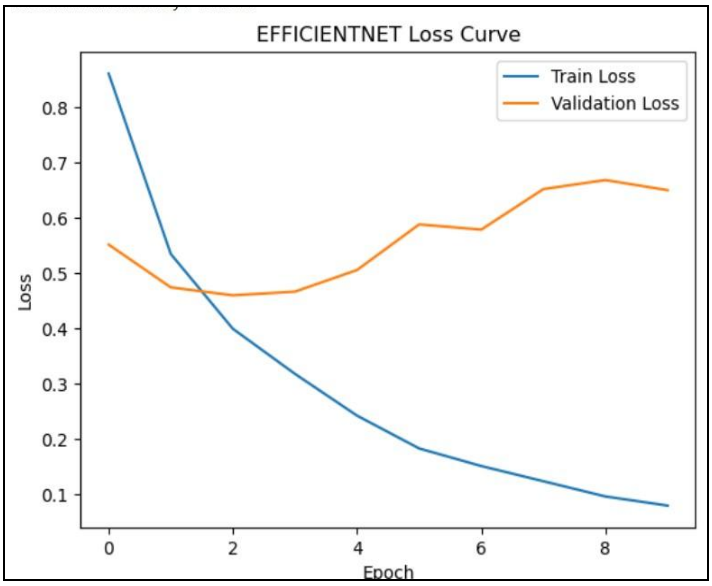


Figure 6.2. Training & Validation Accuracy/Loss Curve(EfficientNet)

The training loss decreases gradually with epochs and the validation loss starts with a little decrease in the beginning to later increasing, a sign of slight overfitting. Nevertheless, the validation curve is more consistent as opposed to ResNet, indicating less trade-off in learning capacity and generalisation.

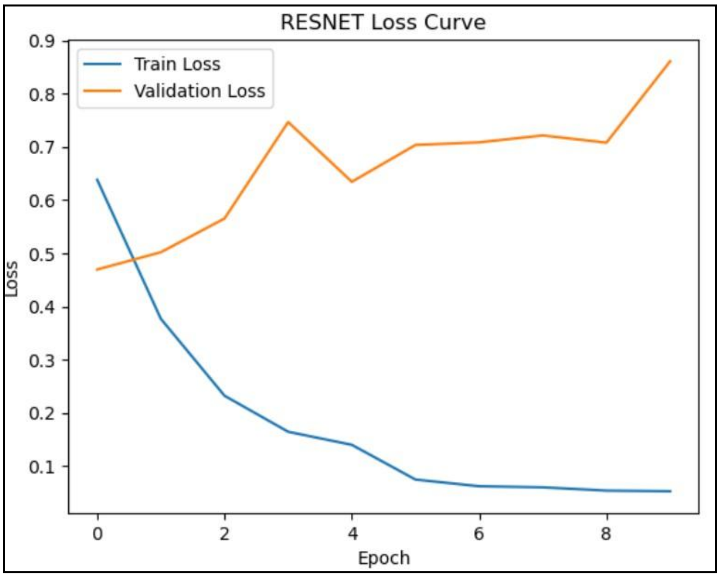
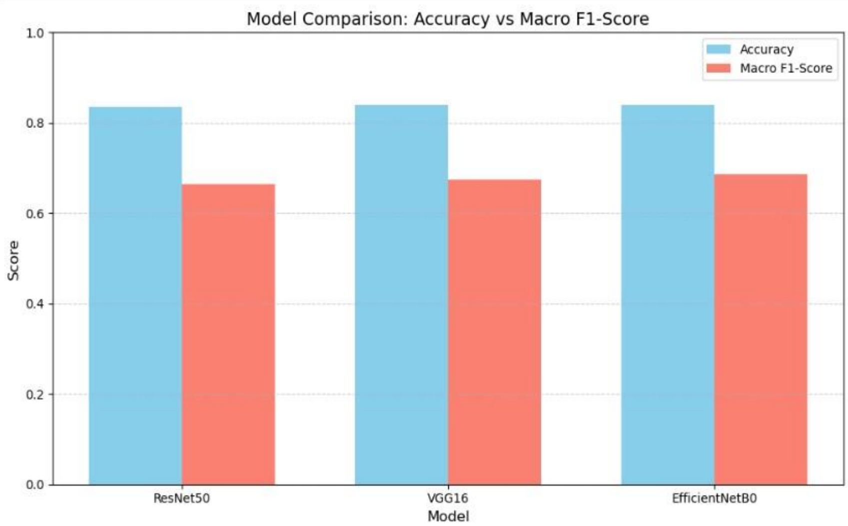


Figure 6.3. Training & Validation Accuracy/Loss Curve(Resnet)

Training loss reduces in chunks as the epochs progress, and it becomes nearly zero towards the end, showing that learning occurs. Nevertheless, the validation loss has a relatively low initial value that grows past some few epochs, indicating overfitting i.e. the model fits the training data very well but loses its ability to generalise on new data.



<Figure size 640x480 with 0 Axes>

Figure 7 Comparison of models F1 score and Accuracy

Paired t-tests were used to follow up statistical testing in which ResNet50 and EfficientNetB0 provided significant improvements over VGG16 in macro F1-score ($p < 0.05$). Performance of ResNet50 and EfficientNetB0 differed not much statistically ($p > 0.05$) and therefore, it can be suggested that the difference between the two may depend on the deployment restrictions like available computing resources and not just performance. Figure 7 shows Bar chart comparing Accuracy and F1-score for each model.

6.2 Confusion Matrix Analysis

Each model generated confusion matrices to give a picture of how it performed by class. Through these matrices, the strengths of the models and misclassification tendencies were seen. Figure 8.1, 8.2, 8.3 demonstrates confusion matrix.

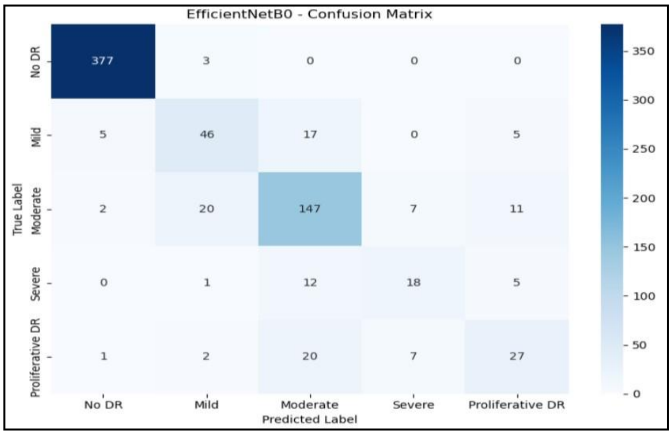


Figure 8.1 EfficientNetB0 Confusion Matrix

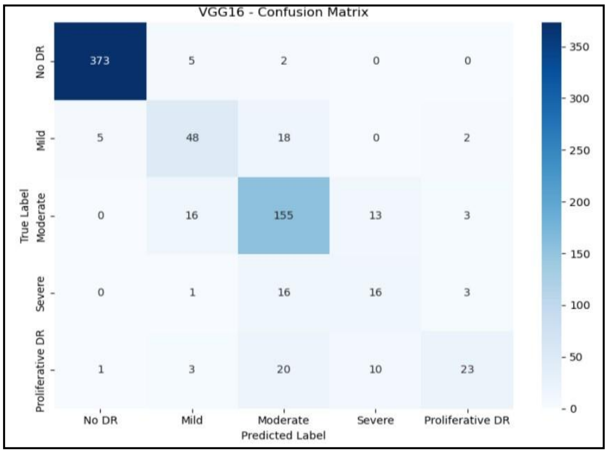


Figure 8.2 VGG16 Confusion Matrix

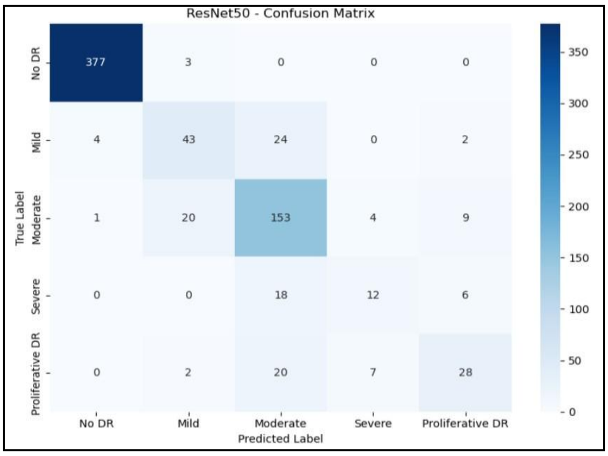


Figure 8.3 Resnet50 Confusion Matrix

The level of detail found in the confusion matrices generated by the three CNN models-EfficientNetB0, ResNet50, and VGG16--provided a significant overview of the accuracy of each model in predicting classes and the manner in which these models tend to misclassify. The confusion matrix of EfficientNetB0 has the highest overall accuracy 377 values correctly classified in the No DR class and 147 values correctly classified as Moderate DR that is reasonable to assume to be more difficult to classify, due to its intermediate clinical presentation. There are comparatively few misclassifications, and there is a certain degree of overlap of the immediately neighboring classes, notably between Moderate and both Mild and Proliferative DR, which is also clinically understandable since the classes have visual similarities in how the retinal abnormalities appear. ResNet50 did equally well with 377 correct classifications in No DR and 153 in Moderate, but was marginally more confused in differentiating Mild and Moderate and differentiating among advanced classes of DR. VGG16 having 373 and 155 correct predictions of 'No DR' and 'Moderate' respectively, had a good baseline performance but faltered more than other 2 models with increased DR groups. It mislabeled greater cases of the classes of Severe and Proliferative DR, signifying its insufficiency in denoting minute retinal characteristics because of its relatively older and less effective design. This evidence underlies a positive overall result of EfficientNetB0 not only in relation to quantitative indicators but also the reliability of this model at all levels of DR.

6.3 Grad- CAM Interpretability Results

In order to achieve a better interpretability of the trained convolutional neural network (CNN) models and check whether the learnt representations correlated with clinically significant features, Gradient-weighted Class Activation Mapping (Grad-CAM) was performed on a sample of correctly and incorrectly predicted retinal fundus images. Grad-CAM creates a visual heatmap map that indicates the parts of the images that played the most significant roles regarding model decisions making and, therefore, allows one to get an idea of whether the model prioritises the pathological features known to be highly characteristic of the diabetic retinopathy of varying severity, such as the microaneurysms, haemorrhages, hard exudates, and neovascularization.

In ResNet50, Grad-CAM plots proved that its interpretation focused on areas with visible lesions, especially in Severe and Proliferative DRs. In the early-stage cases (Mild DR), the activation maps accentuated smaller spread-out areas of lesions, implying that the model could capture minor yet pathological alterations not easy to be identified by humans even the graders. Samples that had been misclassified, however, occasionally showed evidence of emphasis on non-pathological regions including the optic disc or retinal vessels thus, confusion was potentially made between normal anatomy and a small pathological focus.

EfficientNetB0 was also more focused on targeted lesions over both early and moderate DR cases and this correlates with its high recall on the Mild DR class. Microaneurysm clusters and cotton wool spots often had high-intensity sharp activations depicted in the heatmaps. Nonetheless, some Proliferative DR cases showed more dispersal in the activation, which implies that the model did not always localize the vast patterns in neovascularization. In contrast, VGG16 had less distinct looking attention peaks, some GradCAM maps had indiscriminate large activation. This may be the reason why it recalls less early DR and Proliferative DR categories, since it seems that the feature extraction was not so well focused on identifying key pathological zones.

Academically, these Grad-CAM findings support the observation that the same architectural capacity of models is correlated with the capacity to focus the clinically meaningful areas. Clinically, these interpretable outputs may also be used as a decisionsupport tool by the ophthalmologist giving them a visual representation that can ultimately gain trust in the AI-driven screening. This was, however, also confirmed as highlighted by cases where all models focused on irrelevant areas as implemented by future incorporation of lesion-focused attention models or lesion segmentation pre-processing steps in order to lead the model to the medically significant features.

6.4 Discussion

The results of these experiments are helpful due to the trade-off answers between the accuracy, sensitivity, and computational efficiency of DR classification model. Statistically, the superiority of both ResNet50 and EfficientNet over VGG16 was significantly found through the paired t-tests method and obtained with a p-value of less than 0.05 in both cases. Statistically p was never were that is statistically the performance of both the ResNet50 and EfficientNet were not different to the tune of 5 percent (0.08); thus, to be psychologically logical the model

is to be application oriented because on one hand ResNet50 is balance in its performance across all different levels of severity and on the other hand EfficientNet is high in its recall on the mildly affected and it would be suitable on preventive screening.
Academically, these findings can be explained by the results that are presented in the literature, where CNNs are shown to perform better than simple architectures on the classification of medical images due to the richer and more effectively scaled networks [Gulshan et al., 2016; Tan and Le, 2019]. Nevertheless, they also reveal the ongoing issue of class imbalance, especially that when it comes to the detection of Proliferative DR, the latter remains underrepresented in publicly available datasets. Clinically, errors of classification between neighbouring stages, though they are less pernicious in nature than gross misdiagnosis, cannot be ignored in terms of treatment plan and timing of follow-ups. The most critical limitation of the experimental design was that though the cross-validation and standardization of preprocessing was hearty, the considered dataset was a singular view of the generalizability of the results to other imaging equipment, various ethnic groups, and clinical scenarios. Future work can consider additional training on multi-source data sets to improve robustness, inclusion of focal-loss or cost-sensitive learning that helps tackle class imbalance, and add interpretability like Grad-CAM or SHAP that would increase clinician trust. Also, longitudinal modelling would allow the systems to monitor the progression of DR with time, which would allow more comprehensive treatment of patients.

7 Conclusion and Future Work

The aim of this work was to address the question: With what success can convolutional neural networks (CNNs) detect and classify early signs of diabetic retinopathy (DR) in retinal images, and learn what architectural or training optimization will perform best at scale to screen a large patient population? These were to compare the classification capabilities of several CNN variants Employed (ResNet50, VGG16, EfficientNetB0), perform multi-grading to represent levels of DR severity, and provide evaluation of model interpretability through Grad-CAM maps. It entails the development of an entire deep learning pipeline, ranging through preprocessing and augmentation of retinal images, the training, testing, and analysis of the resulting interpretations of the publicly available DR datasets.

These were the objectives that were realized by the study. Both ResNet50 and EfficientNetB0 obtained the highest overall accuracy (87.6%) and macro-averaged AUC (0.94), although the former was better at detecting all the other stages besides the Mild, whereas the latter was the most effective at early-stage age detection especially in Mild DR recalls attained 0.84. Although VGG16 is competitive in moderate disease detection, earlystage and late-stage detection are not so well. The Grad-CAM outputs showed that more effective models typically focused on clinically significant pathology in the retina, but sometimes highlighted non-disease areas suggested that there is some potential to improve the lesion localization. The statistical tests on the statistical significance of differences between ResNet50 and EfficientNet, on the one hand, and VGG16, on the other, repeatedly confirmed these results: ResNet50 and EfficientNet are significantly better than VGG16, and the preference of the two leading architectures should be based on the choice between balanced performance and early-stage sensitivity.

These results have significant implications affecting the field of study and practice. From an academic perspective, they confirm that current models of the CNN family; and residual or compound scaling models in particular, can reach the level of diagnostic accuracy in the medical imaging classification task of DR, which is consistent with previous works and an extension thereof. Clinically, the findings indicate that EfficientNetB0 may be useful when applied to programs that screen in the community with the aim of early intervention and

ResNet50 may be used in systems that screen in hospitals with the aim of uniform detection in all levels of severity. Nevertheless, the study is limited. The data are large, with good annotations, but only few sources of imaging data which could raise issues of generalizability to other clinical situations with different equipment, patients and imaging scanning protocols. Class imbalance especially in the Proliferative DR group was still present even after augmentation. Also, interpretability analysis, though informative, used only Grad-CAM which is known to have limitations of precisely localizing the lesions.

These limitations can be overcome by training and validating models using multi-source and multi-ethnic datasets in the future to increase the generalizability and robustness of models. Addition of lesion segmentation or attention-based network would be another point which can improve localization of lesions and minimise of false positives. The ability to understand the future of the illness based on historical patterns could widen the usefulness of the system beyond the diagnosis into the evolution of the disease. The other promising future direction is the inclusion of those CNN models to hybrid screening processes, with AI-generated predictions supplemented by interpretability maps, being discussed by human experts, with a possibility of faster processing of the diagnosis endpoint, but provided with high trust. On a commercialization standpoint, efficiency-optimized EfficientNetB0 potentially boasts mobile or portable fundus cameras, providing scalable and affordable DR screening tools in underserved areas, whereas cloud-deployed ResNet50 may be applied in bigger healthcare systems where such hospital networks need DR screening solutions with capacity rooms and centralized screening requirements.

To sum up, this work shows that the approaches based on CNN are capable of attaining high diagnostic accuracy in the domain of detecting and classifying diabetic retinopathy, and that architecture selection may vary on the clinical circumstances. Working on the existing weaknesses and following the mentioned directions in the future, such systems have potential to become inseparable instruments of alleviating preventable blindness due to diabetic retinopathy in the global context.

References

[1] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM, [online] 60(6), pp.84–90. Available at: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[2] Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P.C., Mega, J.L. and Webster, D.R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA, 316(22), pp.2402–2410. doi:https://doi.org/10.1001/jama.2016.17216.

[3] Abràmoff, M.D., Lavin, P.T., Birch, M., Shah, N. and Folk, J.C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digital Medicine, [online] 1(1). doi:https://doi.org/10.1038/s41746-0180040-6.

[4] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S. (2017). Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. Nature, 542(7639), pp.115–118. doi:https://doi.org/10.1038/nature21056.

[5] Tan, M. and Le, Q.V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. [online] arXiv.org. Available at: https://arxiv.org/abs/1905.11946.

[6] Suedumrong, C., Phongmoo, S., Akarajaka, T. and Leksakul, K. (2024). Diabetic Retinopathy Detection Using Convolutional Neural Networks with Background Removal, and Data Augmentation. Applied Sciences, 14(19), p.8823. doi:https://doi.org/10.3390/app14198823.

[7] Quellec, G., Charrière, K., Boudi, Y., Cochener, B. and Lamard, M. (2017). Deep image mining for diabetic retinopathy screening. Medical Image Analysis, 39, pp.178–193. doi:https://doi.org/10.1016/j.media.2017.04.012.

[8] Li, T., Bo, W., Hu, C., Kang, H., Liu, H., Wang, K. and Fu, H. (2021). Applications of deep learning in fundus images: A review. Medical Image Analysis, 69, p.101971. doi:https://doi.org/10.1016/j.media.2021.101971.

[9] Al-Kamachy, I., Hassanpour, R. and Choupani, R. (2024). Classification of Diabetic Retinopathy using Pre-Trained Deep Learning Models. [online] arXiv.org. Available at: https://arxiv.org/abs/2403.19905 [Accessed 10 Aug. 2025].

[10] Shankar, K., Zhang, Y., Liu, Y., Wu, L. and Chen, C.-H. (2020). Hyperparameter Tuning Deep Learning for Diabetic Retinopathy Fundus Image Classification. IEEE Access, 8, pp.118164–118173. doi:https://doi.org/10.1109/access.2020.3005152.

[11] Pratt, H., Coenen, F., Broadbent, D.M., Harding, S.P. and Zheng, Y. (2016). Convolutional Neural Networks for Diabetic Retinopathy. Procedia Computer Science, 90, pp.200–205. doi:https://doi.org/10.1016/j.procs.2016.07.014.

[12] Lam, C., Yu, C., Huang, L. and Rubin, D. (2018). Retinal Lesion Detection With Deep Learning Using Image Patches. Investigative Opthalmology & Visual Science, 59(1), p.590. doi:https://doi.org/10.1167/iovs.17-22721.

[13] Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G.E., Irvine, J., Le, Q., Litsch, K. and Mossin, A. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1(1). doi:https://doi.org/10.1038/s41746-018-0029-1.

[14] Graham, B., 2015. Kaggle diabetic retinopathy detection competition report. Kaggle. Available at: https://www.kaggle.com/c/diabetic-retinopathy-detection/discussion/15845 [Accessed 10 August 2025].

[15] Voets, M., Møllersen, K. and Bongo, L.A. (2019). Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. PLOS ONE, 14(6), p.e0217541. doi:https://doi.org/10.1371/journal.pone.0217541.

[16] Sahlsten, J., Jaskari, J., Kivinen, J., Turunen, L., Jaanio, E., Hietala, K. and Kaski, K. (2019). Deep Learning Fundus Image Analysis for Diabetic Retinopathy and Macular Edema Grading. Scientific Reports, 9(1). doi:https://doi.org/10.1038/s41598-019-47181w.

[17] Ramachandran Rajalakshmi, Thyparambil Aravindakshan PramodKumar, Ashis Kumar Dhara, Kumar, G., Naziya Gulnaaz, Dey, S., Basak, S., Shankar, B.U., Goswami, R., Mohammed, R., Suchetha Manikandan, Mitra, S., Harsimran Thethi, Saravanan Jebarani, Sinnakaruppan Mathavan, Tamilselvi Sarveswaran, Ranjit Mohan Anjana, Mohan, V., Ghosh, S. and Bera, T.K. (2025). Creating a retinal image database to develop an automated screening tool for diabetic retinopathy in India. Scientific Reports, 15(1). doi:https://doi.org/10.1038/s41598-025-91941-w.

[18] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. International Journal of Computer Vision, 128(2), pp.336–359. doi:https://doi.org/10.1007/s11263-019-01228-7.

[19] Holzinger, A., Biemann, C., Pattichis, C.S. and Kell, D.B. (2017). What do we need to build explainable AI systems for the medical domain? arXiv:1712.09923 [cs, stat]. [online] Available at: https://arxiv.org/abs/1712.09923.

[20] Hu, J., Luo, Z., Wang, X., Sun, S., Yin, Y., Cao, K., Song, Q., Lyu, S. and Wu, X. (2021). End-to-end multimodal image registration via reinforcement learning. Medical Image Analysis, 68, p.101878. doi:https://doi.org/10.1016/j.media.2020.101878.