

**Diabetic Retinopathy Detection using Attention-Enhanced Transfer Learning and Explainable AI Techniques**

*A Thesis report to be submitted in partial fulfillment of the requirements for the degree*

*of*

**Master of Science in Computer Science and Engineering**

*by*

Md Jakir Hossain

2024-1-96-014

Under the supervision of  
**Dr. Mohammad Rifat Ahmmad Rashid  
Associate Professor**  
Department of Computer Science and Engineering  
East West University  
Dhaka, Bangladesh

14 May 2025

**Declaration**

|  |  |
| --- | --- |
| **Project Title** | Diabetic Retinopathy Detection using Attention-Enhanced Transfer Learning and Explainable AI Techniques |
| **Authors** | Md Jakir Hossain |
| **Student IDs** | 2024-1-96-014 |
| **Supervisor** | **Dr. Mohammad Rifat Ahmmad Rashid** |

|  |  |
| --- | --- |
|  |  |
| **Md Jakir Hossain** |
| **ID: 2024-1-96-014** |
| (Signature) |
|  |  |
| **Dr. Mohammad Rifat Ahmmad Rashid** |
| **Associate Professor** |
| (Supervisor) |
|  |
|
|
|
|  |
|  |
|
|
|

**Date: 14 March 2023**

Dept. of Computer Science and Engineering

East West University

Dhaka, Bangladesh

**LETTER OF ACCEPTANCE**

This is to certify that the thesis entitled "Diabetic Retinopathy Detection using Attention-Enhanced Transfer Learning and Explainable AI Techniques," submitted by Md Jakir Hossain (ID: 2024-1-96-014), a graduate student of the Department of Computer Science and Engineering, has been examined. Upon recommendation by the examination committee, we hereby give our approval as the presented work and submitted report fulfill the requirements for its acceptance in partial fulfillment for the degree of Master of Science in Computer Science and Engineering.

|  |  |  |
| --- | --- | --- |
|  | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | |
|  | **Dr. Mohammad Rifat Ahmmad Rashid** | |
|  | **Associate Professor** | |
|  | Dept. of Computer Science and Engineering  East West University | |
|  | Dhaka, Bangladesh        \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
|  |  | **Dr. Maheen Islam** |
|  |  | **Chairperson, Associate Professor** |
|  |  | Dept. of Computer Science and Engineering |
|  |  | East West University  Dhaka, Bangladesh |
|  |  |  |

**Date: 14 May 2025**

**ACKNOWLEDGEMENTS**

Alhamdulillah, all the honor is for the Almighty, the Most Gracious and the Most Compassionate, for the strengths and His favor in helping to accomplish this thesis assignment. Without His infinite mercy and blessings, this work would not have been possible.

First and foremost, I would like to express my deepest gratitude to Dr. Mohammad Rifat Ahmmad Rashid, Associate Professor, Department of Computer Science and Engineering, East West University, for his invaluable supervision, continuous encouragement, and essential guidance throughout the entire research. I have been greatly inspired by his enthusiasm, vision, sincerity, and motivation. He has taught me not only the technical aspects of research but also the discipline and ethics required to carry out meaningful scientific work. It has been a great honor and privilege to work and study under his supervision.

I would also like to extend my appreciation to all the faculty members and administrative staff of the Department of Computer Science and Engineering at East West University for their support and cooperation during my academic journey.

Special thanks go to my friends and research peers for their constant moral support, helpful discussions, and constructive feedback that helped me refine and strengthen this work.

Finally, I am deeply thankful to my family for their unwavering love, patience, and encouragement throughout this entire academic endeavor. Their belief in me has always been a pillar of strength. May Allah (SWT) bless all those who have contributed to the successful completion of this thesis.

Sincerely –

**Md Jakir Hossain**

Dept. of Computer Science and Engineering East West University Dhaka, Bangladesh.

**ABSTRACT**

In recent years, medical image classification using deep learning has greatly increased the speed and accuracy of disease detection and diagnosis. Nonetheless, picking an ideal model architecture and making it explainable remains a quite challenging task. This study examines the performance of a range of CNN architectures enhanced with attention and skip connections, like a detailed examination of an EfficientNetB7 model incorporating a Squeeze-and-Excitation (SE) block. The models were compared against well-structured data with significant performance indicators like validation accuracy, precision, recall, F1-score, and support.

Among all evaluated models, the EfficientNetB7 with SE block achieved the highest validation accuracy of 83.23%, significantly outperforming traditional CNN-based models. To assess the reliability of these performance differences, paired t-tests were conducted comparing EfficientNetB7 against other CNN models. The results demonstrated statistically significant improvements with EfficientNetB7 (*p* < 0.05), confirming that its superior performance was not due to random chance. Visualization techniques, including bar plots and box plots, further supported these findings by revealing distinct performance margins and minimal overlap in score distributions. For model interpretability enhancement and fostering transparency, Grad-CAM) was applied. Grad-CAM heatmaps confirmed that EfficientNetB7 effectively emphasized disease-causing areas of the input images, presenting interpretable and human-understandable explanations for its predictions. The use of this Explainable AI (XAI) increases the model’s credibility such that it is more fit for real-world use in sensitive domains such as healthcare.

This comprehensive assessment not only confirms EfficientNetB7’s greater accuracy but also claims the importance of statistical verification and explainability in building trustworthy AI systems. The study contributes to the growing list of studies focusing on robust and explainable deep learning models for healthcare and other vital image-based areas.

*Keywords:* Diabetic Retinopathy, Squeeze-and-Excitation, Grad-CAM, Explainable AI

*Table of Contents*

**Declaration . . .** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. ….. . . . . . . .. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **2**

**Letter of Acceptance** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. …… . . . . . . . . . . . . . . . . . . . . . . . . **3**

**Acknowledgement** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. ……. . . . . . . . . . . . . . . . . .. . . . . **. 4**

**Abstract** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. …….. . . . . . . . . . . . . . . . . . . . . . . . . . . .  **5**

**Table of Contents** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. …….. . . . . . . . .. . . . . . . . .  **6-7**

**List of Figures** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. ……. . . . . . . . . . . . . . . . . **8**

**List of Tables** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ……. . . . . . . . .. . . . . . . . . .. **9**

1. **Introduction** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ….... . . . . . . . . ……. . . . . . .**10 – 14** 
   1. Research Question . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ….. . . . . . . . . . . . . **11 -12**
   2. Research Objective . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. …. . . …... . . . . . . . . . . . **13-14**
2. **Background & Related Work** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ....... . . . . . . . . .. .... . . . **15 -17**
3. **Materials and Methods** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ..……. . . . . . . . . . . . . . . ... **18 – 32** 
   1. Dataset . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. …. . . . . . . . .**18 – 19**
      1. Diabetic Retinopathy Overview. . . . . .. . . . . . . . . . . . . . . . . . ……….. . . . . . . . . . . . ... .  **18**
      2. Dataset Overview. . . . . .. …………………. . . . . . . . . . . . . . . . . . ……….. .. . . . . . . . . . . ... . **18 - 19**
   2. Dataset Pre-processing .., . ... …………………. . . . . . . . . . . . . . . . . . ……….. .. . . . . . . . . . . ... . **20 - 21**
   3. Methodology. . . . . .. …………………. . . . . . . . . . . . . . . . . . ……….. .. . . . . . . . . . . ... …………….**21 – 32**
      1. Base Line Models. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ………. . . . . . . . . . . . . . …….. **22 - 24**
         1. Convolution Neural Network (CNN) ……………………………. **22**
         2. EfficientNetB7 ……………………………………………………. **22 - 23**
         3. Transfer Learning …………………………………………………. **23 - 24**
      2. Customs Models. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. …. . . . . . . .. . . . . . . . ……….**25 - 30** 
         1. Custom CNN Model 1 with Attention ………………………… **25**
         2. Custom CNN Model 2 with Attention and Skip Connections …. **25 - 26**
         3. Custom CNN Model 3 with Attention ……………………………. **26 - 27**
         4. Custom EfficientNetB7+squeeze\_excite\_block …………………... **27 - 30**
      3. Performance Measures. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. . . . . . . . . 3**0-32**
4. **Experimental Results**. . . . . . . . . . . . . . . . . .. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **33 – 42**
   1. Experimental setup. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. . . ……………………………………. **33 – 34**
   2. Individual model performance . . . . . . . . . . . . . . . . . . . . . . . . . . . ………………………… ……. **34 - 38**
   3. Performance analysis ……………………………………………………………. **38 - 39**
   4. Error Analysis ……………………………………………………………………. **39 - 40**
   5. Statistical Significance Test ……………………………………………………… **40 - 41**
   6. XAI Performance ………………………………………………………………… **41 - 42**
5. **Discussion**…………………………. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ..………... …. .  **43 - 45** 
   1. Comparison With Current State of the Art Approaches. . . . . . . . ..………. …………….**43 – 45**
   2. Limitation and Future Work. . . . . . . . . . . . . . . ……………………… …………… **45**
6. **Conclusion** ………………………………………………………………………………. **46**

**Reference** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .... **47 – 51**

*List of Figures*

1. **Concept map …………..**. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. .... . . . . . . . . . . . . . .... **14**
2. **Combine Diabetic Retinopathy Dataset**……………………………… . . . . .. …..… .. .**19**
3. **Overview of resize dataset.**. . . . . . . . ……………………… . . . . . . . . . . . . . . . . . . . . .. .**20**
4. **Proposed Methodology** . …………………………….. . . . . . . . . . . . . . . . . . . . . . . . . .. .**22**
5. **Base CNN Architecture** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **22**
6. **EfficientNetB7 Model Architecture**. . . . . . . . . . . . . . . . . . . . . . . . . . . . .. . . . . . . . . . . . .**23**
7. **Custom CNN Model 1 with Attention architecture**. . . . . . . . . . . . . . . . . . . . . . . . .. . . **. 25**
8. **Custom CNN Model 2 with Skip Connections and Attention architecture**. . . . . . . . . . .**26**
9. **Custom CNN Model 3 with Attention architecture**. . . . .. . . ... . . . . . . . . . . .. . .…….....**27**
10. **Custom EfficientNetB7 with Squeeze-and-Excitation block Model Architecture**. . . . .**29**
11. **Performance Graph of CNN 1 Model with Attention (Accuracy vs. Loss)**. ………….. **35**
12. **Performance Graph of Deep CNN with Attention and Skip Connections……………**.**35**
13. **Confusion matrix of Deep CNN with Attention and Skip Connections………………..36**
14. **Performance Graph of CNN3 with Attention Modules…………………………………37**
15. **Confusion matrix of CNN 3 with Attention Modules……………………………………37**
16. **Performance Graph of EfficientNetB7 + Squeeze-Excite Block………………………...38**
17. **Paired t-test results comparing model accuracies. Bars represent confidence intervals; \* denotes p < 0.05 significance………………………………………………………………41**
18. **Grad-CAM visualization showing the activated regions for model predictions using EfficientNetB7…………………………………………………………………………….. 42**
19. **Ablation Grad-CAM visualization ………………………………………………………. 42**

*List of Tables*

1. **Summary of studies with proposed models, datasets, and feature extraction techniques……………………………………………………………………………… 16 – 17**
2. **Image Augmentation Techniques and Paramete ……………………………… 20**
3. **Distribution of Images Across Classes in Original and Augmented Datasets. . .21**
4. **Original Dataset Split (Total: 92,501 Images) . . . . . . . . . . . . . . . . . . . . . . . . . . . 21**
5. **Structure of a 5-Class Confusion Matrix. . .. ………………………………... . 31**
6. **Summary of Hyperparameters and Layer Configuration. . .. ...... . . . . .... . . . 33**
7. **Classification performance of EfficientNetB7 + Squeeze-Excite Block metrics per class ……………………………………………………………………………………. 38**
8. **Performance Metrics of Applied Models ………………………………………. 39**
9. **Comparison with recent state-of-the-art methods ………………………………43 - 44**
10. **Comparison with recent state-of-the-art methods with Accuracy, Precision, Recall, and F1-Score ……………………………………………………………...……………. 44 - 45**

1

Introduction

DR is a severe and progressive eye disease that is a complication of mellifluous diabetes. It happens when long-term elevated blood glucose levels damage the blood vessels in the retina, which is the light-sensitive tissue at the rear of the eye that gives us our vision. The affected vessels can swell and leak, or they can become entirely closed, blocking blood from flowing through. Later, further vision issues may arise from the development of new, aberrant blood vessels on the surface of the retina. Diabetic retinopathy can cause partial or complete blindness if treatment is not received. More retinal damage is indicated by each of the condition’s phases, which include mild, moderate, and severe non-proliferative di abietic retinopathy (NPDR) and proliferative diabetic retinopathy (PDR). A serious side effect of diabetes mellitus, diabetic retinopathy (DR) is defined by dam age to the retinal blood vessels brought on by persistently high blood sugar levels. If left untreated, high blood sugar levels can weaken and damage these arteries, resulting in leaks, hemorrhages, or aberrant vascular development that can impair vision or cause blindness [2]. DR develops in phases, starting with no proliferative DR, which is characterized by microaneurysms and exudates, and ending with proliferative DR, which involves neovascularization and diabetic macular edema (DME), which results in loss of central vision. As a leading cause of preventable blindness in working-age adults, DR poses a significant public health challenge, necessitating early detection [3].

The prevalence of DR is significant and growing worldwide. A projected 103.12 million people with diabetes had DR in 2020; of them, 18.83 million had clinically significant macular edema (CSME) and 28.54 million had vision-threatening DR (VTDR)[3]. By 2045, these numbers are projected to increase to 160.50 million, 44.82 million, and 28.61 million, respectively, driven by the growing diabetes epi demic, particularly in low- and middle-income countries [3].A 2021 study in the United States found that 1.84 million people (5.06%) had VTDR and 9.60 million people (26.43%) had DR . Region ally, the highest prevalence is found in North America/Caribbean (33.30%) and Africa (35.90%), with so bioeconomic characteristics and healthcare access being associated with discrepancies [3]. These statistics underscore the urgent need for effective screening and management strategies. Despite advancements, DR remains underdiagnosed and undertreated, particularly in resource-constrained settings, due to limited access to specialized care, inconsistent screening, and variability in diagnostic accuracy [2]. Early detection and treatment of diabetic retinopathy can prevent vision loss. However, conventional diagnostic techniques are expensive, time-consuming, and susceptible to inter observer variability since they rely on the manual ex amination of retinal fundus pictures by qualified ophthalmologists. Additionally, the need for screening has increased significantly in recent years due to a growing number of people with diabetes, placing further pressure on healthcare systems. Because of these difficulties, automated, dependable, and scalable diagnostic technologies that can promote early detection and enhance patient outcomes are urgently needed. CNNs and DL have made great strides in medical image processing, particularly the categorization of diabetic retinopathy, in recent years. In addition to achieving diagnostic accuracy on par with skilled clinicians, these models can learn intricate patterns from large datasets. However, to train robust and generalize able models, high-quality and diverse data is needed. Our study aims to create a deep learning-based model for automatic diabetic retinopathy categorization using a combined dataset derived from several publically accessible datasets in order to meet this demand. We anticipate that mixing various datasets will increase image feature diversity, enhance the model’s generalizability to a range of patient populations, and avoid overfitting. The model can learn nuanced variations across classes thanks to our combined dataset, which consists of high-resolution retinal pictures labeled with various stages of diabetic retinopathy. The main features of our study on diabetic retinopathy, such as its description, prevalence, problem statement, and research emphasis employing integrated data, are shown in the mind map figure 1 that follows. This study not only analyzes the performance of our proposed model but also evaluates the usefulness of dataset combination in the enhancement of classification accuracy. The ultimate goal is to contribute towards an effective, accessible, and scalable diagnostic system for diabetic retinopathy that can be deployed in real-world clinical practice and low-resource settings.

## 1.1 Research Question

### 1.1 Research Question

Study questions are the foundation of any orderly study. Research questions form the core of any scientific study and are instrumental in determining the scope and direction of a study. With the right set of research questions, one can collect relevant, meaningful data that drives the overall success of the research. The questions help in the separation of complex problems into manageable,

This section outlines the primary research questions and the related sub-research questions that guide this thesis work, conducted to fulfill the requirements for the Master of Science in Computer Science and Engineering degree at East West University. The core research question is followed by a breakdown into sub-questions aligned with the Program Outcomes (PO) of the curriculum.

**How can an attention-enhanced deep learning model be designed and developed using transfer learning and explainable AI techniques to accurately detect and classify diabetic retinopathy from retinal fundus images, particularly in a scalable and clinically deployable way?**

The above-mentioned primary research question is then divided into several sub-research questions according to the Program Outcomes mentioned in the curriculum.[4]

**RQ1.** How to apply and integrate new and previously acquired mathematics, science, and engineering knowledge to address the challenges associated with diabetic retinopathy detection using deep learning? (PO 1)

**RQ2.** What relevant domains (e.g., medical imaging, ophthalmology, deep learning) need to be explored, and how to define the research problem and formulate clear objectives for diabetic retinopathy classification? (PO 4)

**RQ3.** How to analyze the key components of the problem (data quality, imbalance, interpretability, etc.) to design an efficient and explainable solution? (PO 2)

**RQ4.** How to ensure the developed model adheres to safety, privacy, and ethical guidelines while considering the socio-economic and healthcare contexts, especially in low-resource settings? (PO 3)

**RQ5.** Which modern AI, image processing, and engineering tools are required, and how to apply them to design a robust diagnostic model? (PO 5)

**RQ6.** How to assess and mitigate societal, legal, health, and cultural implications related to the deployment of AI in clinical diabetic retinopathy screening? (PO 6)

**RQ7.** How can this system contribute to sustainable healthcare, particularly in terms of long-term benefits, cost-effectiveness, and accessibility in developing countries? (PO 7)

**RQ8.** What professional and ethical responsibilities must be upheld while working on sensitive health data and designing AI systems for clinical use? (PO 8)

**RQ9.** Which strategies and collaborative practices are effective for functioning efficiently both as an individual researcher and within a research group to achieve the thesis objectives? (PO 9)

**RQ10.** How to document, present, and deliver technical findings and contributions in a clear, coherent, and impactful manner? (PO 10)

**RQ11.** How to apply software engineering principles, economic analysis, and project planning to ensure feasibility and real-world deploy ability of the developed system? (PO 11)

**RQ12.** What independent and lifelong learning strategies have been adopted and how have they contributed to the continuous improvement of technical and research skills during this thesis? (PO 12)

## 1***.2 Research Objectives***

**1.2 Research Objectives**

A research objective is a statement that clearly defines the purpose of a research project or study. It outlines the specific goals and aims of the research and provides a clear direction for the research process. Overall, research objectives are essential for conducting rigorous and effective research, and they play a critical role in ensuring that research findings are valuable and relevant to the intended audience.

In this section, the objectives of the research have been defined. The main objective of this **Master’s Thesis** is to design and develop an attention-enhanced, transfer learning-based deep learning framework to detect diabetic retinopathy (DR) in retinal fundus images using explainable artificial intelligence techniques, with the following specific characteristics:

* **Diabetic Retinopathy Focused:** One of the key objectives of this research is to provide a solution for detecting various stages of diabetic retinopathy—a leading cause of vision loss among working-age adults—using deep learning methods focused on retinal image analysis.
* **Multiclass Stage Classification:** The proposed model will classify DR images into multiple stages, including No DR, Mild, Moderate, Severe Non-Proliferative DR (NPDR), and Proliferative DR (PDR), which is a critical step in early intervention and treatment planning.
* **Attention Mechanism Integration:** An objective of the research is to integrate attention mechanisms within the CNN architecture to allow the model to focus on significant regions of the retina (e.g., microaneurysms, hemorrhages, neovascularization), improving interpretability and performance.
* **Use of Transfer Learning:** This thesis aims to apply transfer learning with pre-trained deep models to utilize existing knowledge and overcome the limitations of limited annotated medical datasets.
* **Explainable AI Techniques:** Another major objective is to incorporate explainable AI tools (e.g., Grad-CAM) into the diagnostic pipeline to enhance model transparency, reliability, and clinical acceptance by visualizing which features influenced each classification.
* **Dataset Combination and Enrichment:** The research aims to combine multiple publicly available retinal image datasets to increase diversity, minimize overfitting, and improve model generalization across different populations.
* **Environment- and Resource-Friendly:** This model, by supporting early detection and reducing unnecessary manual screening, aims to optimize healthcare resources, lower diagnostic delays, and reduce patient burden—contributing indirectly to more sustainable and accessible healthcare systems.
* **Academic and Societal Contribution:** Lastly, the research aims to contribute meaningfully to both the academic community (through publication and open data/code sharing) and society by offering a tool that addresses a growing public health challenge.

Figure 1 which is a graphical representation of knowledge. It is used to visually organize and represent the relationships between different concepts. In this case, it outlines key topics related to Diabetic Retinopathy (DR), including its definition, stages, complications, prevalence, research focus, challenges, and goals. It helps to explain complex relationships and provides a structured overview of the subject matter.

A diagram of a company

AI-generated content may be incorrect.

*Figure 1: concept map*

2

Background and Related Work

Numerous studies have concentrated on using deep learning and machine learning techniques to accurately diagnose DR, one of the leading causes of blindness worldwide. For early assistance and successful treatment, it is essential to recognize and categorize the early stages of DR from retinal pictures. To enhance detection, segmentation, and DR grading, we have covered various computer models and feature gathering techniques in previous research throughout this work. This review integrated findings from conference and journal publications, emphasizing developments in transformer models, hybrid architecture, and attention mechanisms as well as their performance on benchmark datasets. The research demonstrates a shift from conventional CNNs to trans formers and hybrid models. The shift enhances the feature extraction, segmentation, and classification in diabetic retinopathy (DR) screening. A hybrid Vision Transformer (ViT) and U-Net performed well in seg mentation and classification on APTOS and outper formed CNNs in reliability [4]. ViT was enhanced using softmax and pooling, achieving enhanced DR grading in some publicly available datasets [5]. The combination of transformers with JPU and CBAM provided a mean IoU of 0.8047 in CHASEDB1 [6]. A novel DRCCT model attained an F1-score of 0.97 and 99% training accuracy [7]. A two-layer U-Net with transformers performed well in IDRiD, DDR, and DI ARETDB1 [10]. MobileViTv2 and GoogLeNet com bined with knowledge distillation achieved a kappa score of 0.94 on APTOS [11]. Multi-head attention ViTs achieved a success rate of 96.13% [12]. An earlier-guided transformer outperformed the state-of the-art techniques in DDR and IDRiD[13] . A multi head attention based ViT model achieved 91.4% ac curacy on fundus images of varied resolutions [14] . SSiT model applied on saliency maps of EyePACS, DDR, Messidor-2, and APTOS-2019 performed bet ter than SSL approaches [15]. ViT on OCT images achieved 99.69% accuracy in classifying DME [17]. DTUNet model enhanced segmentation of difficult to-segmented exudates better on the e-ophtha and IDRiD datasets [18] . Dformer model enhanced the mild NPDR grading by 18.16% on DDR and Messi dor [19]. LAT model performed better on Messidor 1, Messidor-2, and EyePACS [20]. A transformer based approach achieved kappa of 0.8075 on the DRAC 2022 OCTA dataset [22]. CT-ALUnet ap proach enhanced the Dice scores by 7.5% on un known datasets [23]. Ensemble-based ViT approach attained 47% precision on the Kaggle’s 2015 DR dataset [24]. Swin-UNETR-based model applied on the task of fluid segmentation for nAMD on the private dataset [25]. Swin-Poly Transformer-based ap proach obtained 99.80% accuracy on OCT2017 and OCT-C8 for the task of DME classification [26]. An other approach by Swin-UNETR applied on the task

for nAMD lesions that are associated with the task of DR on the private dataset [27]. A human-assisted model named Query2Label achieved 99.8% accuracy for DR detection on RFMiD [28]. A survey of ViT based models demonstrated promising results for DR grading on fundus and OCT images [29]. The TMIL model accelerated inference time by 62% on APTOS and Messidor-1 [30]. A ViT model achieved an F1- score of 0.825 on FGADR [31]. The OCTFormer model achieved 98.60% accuracy on a specific OCT dataset[32] . The HDformer model achieved 97.3% accuracy on MIMIC-III using PPG signals, which are indicative of DR [33]. A hybrid CNN-ViT model performed better than state-of-the-art methods on ImageNet ISLVRC and a retinal database [34]. This study adapts the SE block for medical image appli cations, demonstrating its effectiveness in improving segmentation tasks. [35]. A multi-disease classification model achieved 93.1% accuracy on ODIR 2019 for DR [36]. The TCU-Net model achieved 0.9230 accuracy on ROSE-1 for blood vessel seg mentation [37]. Transformer-based models for retinal layer and fluid segmentation improved Dice scores by 0.01 on special OCT datasets[38] . A retinal image segmentation model using deep adaptive gamma correction outperformed other methods on DRIVE and CHASEDB1 [39]. A multi-label retinal disease classification performed well on ODIR-2019 based on ViT [40]. MIL-VT performed better than CNNs on two unidentified datasets [41]. A blood vessel segmentation transformer obtained 97.25% accuracy on DRIVE and 97.93% accuracy on STARE [42]. Adaptive optics retinal images were found by the hybrid transformer very efficiently [43]. TLTNet attained 97.26% on DRIVE and 97.87% on STARE [44]. A retinal spiking activity prediction-based transformer model would assist in the case of DR prosthetics [45]. Glaucoma detection based on ViT performed the best on the case of ACRIMA [46]. OCT classification us ing the lightweight hybrid transformers obtained up to 98.86% accuracy [47]. Up to 99.17% accuracy in d termining if the presence or absence of choroidal neo vascularization, diabetic macular edema, and drusen exists based on OCT images was obtained by a ViT [48]. Lastly, the transformer-based multi-label retinal disease classification performed better than the optimal models by 7.9% AUC on the case of MuReD [49]. Optical Coherence Tomography Image Classification using Light-weight Hybrid Transformers [50] achieved accuracy of 98.86% in the balanced OCT dataset. Diabetic macular edema, and drusen from retinal OCT images [52] reached accuracy of 99.17% using Mobile-ViT for DME. Multi-Label Retinal Disease Classification Using Transformers [1] performed 7.9% AUC better than the optimal approach in the MuReD dataset. Table 1 shows the Summary of studies with proposed models, datasets, and feature extraction techniques.

Table 1: Summary of studies with proposed models, datasets, and feature extraction techniques.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Study [Ref]** | **Model Architecture** | **Dataset(s)** | **Proposed Model** | **Feature Extraction Technique** |
| [4] | Hybrid U-Net + ViT | APTOS | U-Net + ViT | U-Net segmentation + ViT features |
| [5] | ViT with softmax and linear modules | Public (unspecified) | ViT with softmax pooling | Self-attention with softmax pooling |
| [6] | Transformer + CBAM + JPU | CHASEDB1 | Transformer + CBAM + JPU | CBAM + multi-scale fusion |
| [7] | DRCCT (Conv + Transformer) | Unspecified | DRCCT | Convolutional + transformer features |
| [15] | SSiT (Saliency-guided ViT) | EyePACS, DDR, Messidor-2, APTOS-2019 | SSiT | Saliency-guided self-supervision |
| [17] | RTNet (Relation Transformer) | IDRiD, DDR | RTNet | Self- and cross-attention |
| [19] | DTUNet (Dual-branch Transformer) | e-ophtha, IDRiD | DTUNet | Dual-branch feature fusion |
| [22] | LAT (Lesion-Aware Transformer) | Messidor-1, Messidor-2, EyePACS | LAT | Lesion-aware filters |
| [23] | Vision Transformer | kermany2018 (Kaggle) | Vision Transformer | ViT features from OCT |
| [25] | CT-ALUnet (CNN + Transformer) | Unspecified | CT-ALUnet | Transformer + CNN adversarial features |
| [31] | Ensemble of Vision Transformers | 2015 Kaggle DR dataset | Ensemble ViT | Ensemble ViT features |
| [33] | Vision Transformer | FGADR | Vision Transformer | ViT with data augmentation |
| [34] | OCTFormer (Hierarchical Transformer) | Custom OCT | OCTFormer | Hierarchical transformer features |

3

Materials and Method

This chapter outlines the materials used and the methodologies adopted to achieve the objectives of this research. The design and development of the Diabetic Retinopathy (DR) detection model involve a sequence of steps ranging from data collection to model training, evaluation, and explanation. Each step is discussed in detail below.

## 3.1 Dataset

To assess our proposed framework, we utilized a comprehensive dataset combining EyePACS, APTOS, APTOS (Gaussian Filtered), and Messidor Diabetic Retinopathy datasets. This section provides an overview of diabetic retinopathy as a medical condition, followed by a detailed description of the unified dataset and the preprocessing steps applied.

## 3.1 Diabetic Retinopathy Overview

Diabetic retinopathy is a common diabetic complication involving the blood vessels of the retina and, if untreated, may lead to blindness or loss of vision. Diabetic retinopathy progresses in stages, ranging from mild non-proliferative retinopathy with microaneurysms to proliferative diabetic retinopathy, with new irregular blood vessel formation. Depending upon the severity, its presenting features could be blurred vision, floaters, or blindness. Early detection by retinal imaging is the key to successful treatment and management, and automated classification systems are therefore very useful for mass screening

## 3.2 Dataset Overview

Figure 2 displays the merged dataset, which consists of pictures from four publicly accessible diabetic retinopathy databases: EyePACS, APTOS, APTOS (Gaussian Filtered), and Messidor. Initially, there were 92,501 JPG photos in the dataset, which were split into training (80%), validation (10%), and test (10%) groups at random. A manual data augmentation update was carried out on January 5, 2024, increasing the dataset by approximately 55% to 143,669 photos.

The purpose of this modification was to pro mote model generalization, lessen class imbalance, Two cylinder shapes with text

AI-generated content may be incorrect.

*Figure 2: Combine Diabetic Retinopathy Dataset*

The merged dataset is an ideal foundation for training and testing diabetic retinopathy detection machine learning algorithms. From its creation from several sources, it offers a heterogeneous set of retinal images at different stages of the disease and under various imaging conditions. Example images in figure 3 the dataset usually include normal retinas and retinas with diabetic retinopathy lesions like microaneurysms, hemorrhages, exudates, and neovascularization.

## 3.2 Dataset Pre-processing

Preprocessing was a critical step to normalize the dataset to undergo efficient training of machine learning algorithms. The following operations were applied to 143,669 images:

* Images were resized to 600x600 pixels uniformly for consistent even size input to keep computation overhead while training minimally.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| A close-up of a red light  AI-generated content may be incorrect. | A close-up of a red eyeball  AI-generated content may be incorrect. | A close-up of a human eye  AI-generated content may be incorrect. |
|  |  |  |
| A close-up of a human eye  AI-generated content may be incorrect. | A close-up of a red circle  AI-generated content may be incorrect. |  |

*Figure 3: Overview of resize dataset.*

* Augmented images were incorporated in the dataset such that balance would be maintained on the training, validation, and test sets.

**Table 2: Data Augmentation Techniques**

|  |  |  |
| --- | --- | --- |
| **Augmentation Technique** | **Description** | **Parameter Value** |
| RandomWidth | Randomly adjusts the width of the image by a fractional factor, introducing variability in aspect ratio. | 0.2 (20% variation) |
| RandomHeight | Randomly adjusts the height of the image by a fractional factor, altering the image proportions. | 0.2 (20% variation) |
| RandomFlip ("horizontal") | Randomly flips the image horizontally (left to right), simulating mirrored perspectives. | N/A |
| RandomContrast | Randomly adjusts the contrast of the image, enhancing or reducing the difference between light and dark areas. | 0.2 (20% variation) |
| RandomFlip ("vertical") | Randomly flips the image vertically (top to bottom), providing an additional orientation variation. | N/A |

* Pre-applied resizing and augmentation were employed to prevent on-the-fly transformation during model training, minimizing unnecessary resource consumption.

Random splitting was first employed to divide the dataset into training (80%), validation (10%), and test (10%) sets. This was preserved after augmentation, yielding a strong and reproducible evaluation framework. The preprocessing step enhanced the quality of the dataset for application in deep learning models, enhancing the efficiency and accuracy of diabetic retinopathy stage classification. This merged dataset, with its preprocessed and augmented images, offers a good foundation for experimentation on our proposed framework and thereby improvements for automatic diabetic retinopathy diagnosis.

**Table 3: Distribution of Images Across Classes in Original and Augmented Datasets**

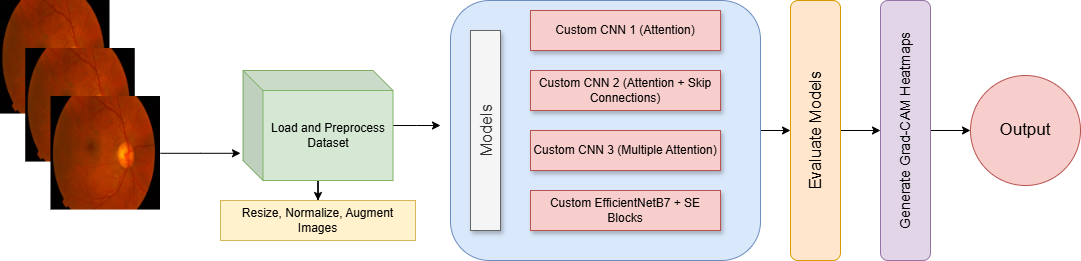
|  |  |  |
| --- | --- | --- |
| **Class** | **Original Dataset (Image Count)** | **Augmented Dataset (Image Count)** |
| No DR | ~70,000 | ~70,000 |
| Mild DR | ~2,000 | ~10,000 |
| Moderate DR | ~10,000 | ~20,000 |
| Severe DR | ~500 | ~5,000 |
| Proliferative DR | ~500 | ~5,000 |

**Table 4: Original Dataset Split (Total: 92,501 Images)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Class** | **Class** | **Training (80%)** | **Validation (10%)** | **Test (10%)** | **Total** |
| No DR | 0 | 56,000 | 7,000 | 7,000 | 70,000 |
| Mild DR | 1 | 1,600 | 200 | 200 | 2,000 |
| Moderate DR | 2 | 8,000 | 1,000 | 1,000 | 10,000 |
| Severe DR | 3 | 400 | 50 | 50 | 500 |
| Proliferative DR | 4 | 400 | 50 | 50 | 500 |

## 3.3 Methodology

In this study, we present an end-to-end deep learning pipeline for the automatic detection of diabetic retinopathy (DR) based on the combination of transfer learning and explainable decision-making approaches. Figure 4 shows the illustration of Proposed Methodology. Our approach employs an array of carefully designed convolutional neural networks (CNNs), each supplemented with attention mechanisms to further concentrate on salient regions of the retina. Additionally, we employ EfficientNetB7, which is an advanced pre-trained model, further enhanced with the introduction of squeeze-and-excitation blocks to support channel-wise feature recalibration. To measure the effectiveness of the model, we use standard metrics such as accuracy, precision, recall, F1-score, and AUC. Besides measuring classification performance, our system places emphasis on transparency and clinical reliability by incorporating explainable AI (XAI) techniques. We use Gradient-weighted Class Activation Mapping (Grad-CAM) to generate heatmaps that effectively identify areas of the retinal images that contribute most to the model’s predictions. These visual explanations not only improve the understanding of the model’s operation but also help clinicians interpret and verify the automated decision-making process. By combining advanced deep learning architectures with interpretable outputs, our proposed framework aims to offer a reliable and trustworthy tool for early DR detection. This contributes significantly to the development of AI-assisted diagnostic systems suitable for deployment in real-world clinical settings.



*Figure 4: Proposed Methodology*

### 3.3.1 Base Line Models

#### 3.3.1.1 Convolution Neural Network (CNN):

A Convolutional Neural Network (CNN) is a deep learning algorithm that assists in image recognition and image classification. It identifies significant features of input images using convolution, activation (ReLU), pooling, and fully connected layers. Convolutional layers identify patterns such as edges and textures using filters, while pooling layers minimize the data volume to ease calculations. Once the features are located, the fully connected layers make predictions using the found features[16]. CNNs find application in numerous areas such as face detection, medical image processing, object detection, and self-driving vehicles since CNNs can be highly accurate with images. Figure 5 shows the Base line CNN Architecture.

A diagram of a block diagram

AI-generated content may be incorrect.

*Figure 5. Base CNN Architecture*

#### 3.3.1.2 EfficientNetB7:

EfficientNetB7 is the largest and most powerful member of the EfficientNet family created by Google AI. It employs a novel scaling algorithm to vary the depth, the width, and the resolution of the network in balance, attaining efficiency with fewer computations and parameters than prior models such as ResNet or Inception. EfficientNetB7 is ImageNet pre-trained and attains leading image classification accuracy. EfficientNetB7 employs MBConv blocks (Mobile Inverted Bottleneck Convolutions) along with Swish activation, which ensure the speed as well as accuracy that deep learning requires [21] Figure 6 show that.

A diagram of a diagram

AI-generated content may be incorrect.

*Figure 6. EfficientNetB7 Model Architecture*

#### 3.3.1.3 Transfer Learning

Transfer Learning is Deep learning technique in which a model developed for one task is picked up again as the basis for a model for another task. It transfers knowledge gained from a pre-trained model on a vast quantity of data to solve a new, but related problem, cutting profoundly the time and computing power it takes to train a new model. Transfer learning mechanism is comprised of employing a model that has been trained on one task and fitting it to a new but similar task by fine-tuning its parameters or transferring its learned features. This is particularly useful when there is not much labeled data for the new task.

There are different types of transfer learning:

**Inductive Transfer Learning:** The model is learned on a single task and applied to a related task with different data. The goal is to improve the performance on the target task by capitalizing on the knowledge of the source task.

**Transductive Transfer Learning:** The model is learned on a source task and applied on a target task with the same input but different output. The aim is to transfer learning to forecast the target data output.

**Unsupervised Transfer Learning:** The model is learned on an unsupervised task and transferred on a supervised task such that unsupervised data can be utilized to improve supervised learning performance.

**Zero-shot Learning:** A form of transfer learning where the model is intended to transfer knowledge to completely new tasks or classes without any labeled examples from the target domain.

All forms of transfer learning exploit the knowledge learned from different tasks to improve learning efficiency or performance on the target task.

#### 3.3.1.4 Attention Mechanism

Attention Mechanism is a technique used in deep learning models, particularly in natural language processing (NLP) and computer vision, to focus on relevant parts of the input data and ignore irrelevant information. It is an imitation of the human cognitive attention mechanism, where we focus on specific parts of a scene or text when processing information. In machine learning, attention allows a model to dynamically allocate varying weights to different parts of the input based on their importance with respect to the task at hand so that it can concentrate more on important features or regions. The mechanism works by multiplying a set of attention weights with different parts of the input sequence, allowing the model to focus on the parts that are more relevant for the current prediction. This is especially useful in sequence-to-sequence tasks like machine translation or image captioning, where it is necessary to understand the context or relation between different parts of the input.

Some of the variations of the attention mechanism are:

**Soft Attention:** The model is a weighted sum of all possible inputs, with the weights learned from training. It allows the model to attend over the pertinent parts of the input in a continuous manner.

**Hard Attention:** Instead of a weighted average, the model selects specific parts of the input (e.g., one word in a sentence or one region in an image) to focus on. This is more discrete in nature and usually requires reinforcement learning to train.

**Self-Attention (or Intra-Attention):** This type of attention allows the model to relate different parts of the same input sequence to each other. It is widely used in models like Transformer, where the relations between all the words in a sentence are considered and not just the adjacent words.

**Multi-Head Attention:** This is an extension of self-attention used in Transformer models. It allows the model to jointly attend to information from different representation subspaces at different positions, increasing the ability to learn complicated patterns in data.

Overall, attention mechanisms have significantly improved the performance of models on tasks like machine translation, text generation, and image processing by allowing them to handle long-range dependencies and contextual information more effectively.

### 3.3.2 Customs Models

#### ***3.3.2.1 Custom CNN Model 1 with Attention***

Custom CNN Model 1 with Attention is a hybrid deep learning architecture that is designed to enhance image classification accuracy by combining the excellent feature extraction capability of convolutional neural networks with an attention mechanism that allows the model to focus on the most relevant areas of an image. Figure 7. Shows **Custom CNN Model 1 with Attention architecture.**  The model begins with a sequence of convolutional blocks composed of convolutional layers followed by batch normalization, ReLU activation, and max pooling that extract hierarchical features from the input image progressively. To further augment these features, a spatial attention module is added after the convolutional layers. This attention mechanism, which is based on the CBAM (Convolutional Block Attention Module), applies both channel and spatial attention to emphasize significant features and downplay irrelevant information. The feature maps that have been processed are flattened and passed through fully connected layers to produce the final classification output. This attention-improved approach not only boosts the model's overall generalization capacity by focusing on areas of importance but also facilitates better interpretability, especially when dealing with difficult visual patterns or noisy background environments.

**A diagram of a diagram

AI-generated content may be incorrect.**

Figure 7. Custom CNN Model 1 with Attention architecture

#### 3.3.2.2 Custom CNN Model 2 with Attention and Skip Connections

Custom CNN Model 2 with Skip Connections and Attention is a more evolved deep learning model that works to improve feature representation and gradient flow for image classification tasks. The model is developed from the background of standard CNNs but includes attention mechanisms to focus on the most informative regions of the image and skip connections to preserve spatial information and prevent the vanishing gradient issue. The architecture begins with multiple convolutional blocks, each followed by batch normalization, ReLU activation, and max pooling to achieve hierarchical features. Skip connections are introduced after each block to pass features from earlier layers directly into deeper layers so that the model has the capability to learn low-level and high-level representations simultaneously. An attention module, replicated from the CBAM architecture, is integrated after the convolutional stages for fine-tuning feature maps by highlighting significant spatial and channel-wise features. The skip connections also aid in stabilizing training and enhancing the performance of the model by allowing gradient backpropagation along various paths. The fine-tuned feature maps are then flattened and passed through fully connected layers to produce the final output. This architecture performs best in intricate image data sets in which fine detail and global context are both essential for accurate classification. Figure 8 illustrates the Custom CNN Model 2 with Skip Connections and Attention architecture.

A diagram of a diagram

AI-generated content may be incorrect.

*Figure 8. Custom CNN Model 2 with Skip Connections and Attention architecture.*

#### 3.3.2.3 Custom CNN Model 3 with Attention

Custom CNN Model 3 with Attention is a cutting-edge convolutional neural network model that integrates attention within the network to improve feature selection as well as model interpretability. This model builds upon the general CNN model by stacking several convolutional blocks sequentially—each consisting of Conv2D, Batch Normalization, ReLU activation, and MaxPooling layers—to learn incrementally spatial hierarchies from the input data. The thing that distinguishes this model from others is the addition of an attention module, usually derived from CBAM (Convolutional Block Attention Module), applied following the final convolutional layer. The attention process helps the network focus on the most important areas by applying channel-wise and spatial attention, hence improving the features learned before classification. The model then flattens the attention-augmented feature maps and passes them through fully connected layers for prediction. With the combination of convolutional feature extraction and attention-based augmentation, this architecture can achieve improved performance on tasks requiring a deeper understanding of complex visual patterns, and therefore it is highly effective for medical imaging, facial recognition, and object detection tasks. Figure 9 illustrates the Custom CNN Model 3 with Attention architecture.

A diagram of a diagram

AI-generated content may be incorrect.

Figure 9. Custom CNN Model 3 with Attention architecture.

#### 3.3.2.4 Custom EfficientNetB7+squeeze\_excite\_block

The model is an adapted architecture built upon the EfficientNetB7 backbone, topped with a Squeeze-and-Excitation (SE) block to facilitate attention across channels. The model's input is an RGB image of size 600x600. EfficientNetB7, one of the pre-trained CNNs with high competency, extracts fine features from input and outputs a 2560-channel feature map. These features are applied global average pooling and then the SE block, which re-tunes the feature channels by two dense layers — dimension reduction (squeeze) first, and dimension restoration (excite) next. This attention-weighted output is then element-wise multiplied with the original feature map to emphasize informative features. A final global average pooling layer downsamples the spatial dimensions, a dense layer with 256 units, dropout for avoiding overfitting, and an output dense layer with 5 units with a softmax-activated function to perform multi-class classification. Non-trainable weights of EfficientNetB7 are frozen down and only the SE block and the classification head can remain trainable, in which efficient fine-tuning is allowed with lower computational cost.Figure 10 shows the Custom EfficientNetB7 with Squeeze-and-Excitation block Model Architecture. Let the model *M* produce a probability vector

p = (*p*(*c*1), *p*(*c*2), . . ., *p*(*ck*)),

and let the input image be represented as a tensor X ∈ R*H*×*W*×*C*,

where *p*(*cj*) represents the predicted probability that a given input image belongs to class *cj*, and *K* is

he total number of target classes. H, W, and C denote

the height, width, and number of channels, respectively. The architecture is enhanced with a Squeeze-

and-Excitation (SE) block, which recalibrates the feature maps generated by EfficientNetB7. This mechanism adaptively highlights informative features and suppresses less useful ones, improving the model’s ability to distinguish between subtle patterns in the image. The input is passed through the EfficientNetB7 backbone, which performs a series of convolutional operations:

F = f EfficientNetB7(X) ∈ R*H*′×*W*′×*C*′

*Here*, F is the extracted feature map, and *H*′, *W*′, *C*′ are the spatial dimensions and number of channels after the feature extraction stage. To recalibrate the channel-wise responses of feature map F, we apply the SE block, consisting of two main steps: Squeeze and Excitation.

***Squeeze (Global Information Embedding)***We compress the spatial dimensions using Global Average Pooling:

This results in a vector:

z ∈ R*C*′

We pass z through two fully connected (Dense) lay ers with non-linearities:

s = σ(W2 · δ(W1 · z))

**Classification Head**

The recalibrated feature map F′is pooled again: GlobalAveragePooling(F′) ∈ R*C*

Then it is passed through a fully connected classifier:

h = δ(W*h*v + b*h*) ∈ R*d*

o = Softmax(W*o*h + b*o*) ∈ R*K*

classification process begins by extracting high dimensional features using the EfficientNetB7 backbone pre-trained on a large-scale dataset. A Global Average Pooling operation is applied to these feature maps to generate a compact representation. This ve tor then passes through the SE block:

• First, it is reshaped to match the spatial configuration.

• Then, two dense layers act as gates that learn to squeeze (reduce dimensions) and excite (apply attention) on the features.

• The output of this process is multiplied element wise with the original features to recalibrate them.

To avoid overfitting, the recalibrated features are once more pooled globally and run through a fully linked Dense layer and a Dropout layer. Lastly, the probability distribution across the K classes is produced by a Dense layer with Softmax activation. The predicted class y is then determined By: ˆ = arg max*cj*∈{*c*1,...,*cK*} *p*(*cj*) During training, the model is optimized using categorical cross-entropy loss:

L = -P*Kj*=1*yj*log(*oj*) During training, the model uses a categorical crossentropy loss function, and is optimized using a method like Adam with an initial learning rate. Performance is monitored on the validation set for early stopping and dynamic learning rate scheduling to enhance generalization. This structured approach leverages the EfficientNetB7’s scalable depth and width and the SE block’s Attention mechanism, resulting in a robust classification model capable of accurately identifying complex age classes. It emphasizes end-to-end learning and careful architectural design over the use of multiple models, while still achieving high performance. he detailed steps of this model process summarized in Algorithm 1.

A diagram of a diagram

AI-generated content may be incorrect.

*Figure 10 Custom EfficientNetB7 with Squeeze-and-Excitation block Model Architecture*

|  |
| --- |
| *Algorithm 1 EfficientNetB7 + Squeeze-and Excitation* |
| 1: Input: Image dataset D  2: Output: Trained classification model M 3: Split D into training set Dtrain, validation set Dval, and testing set Dtest  4: Apply data augmentation to Dtrain to create aug mented dataset Daug  5: Load EfficientNetB7 pre-trained on ImageNet with frozen weights  6: Add Global Average Pooling layer to Efficient NetB7’s output  7: Apply Squeeze-and-Excitation (SE) block: 8: Squeeze: Reshape the pooled features 9: Excite: Pass through Dense layers with ReLU and Sigmoid activations  10: Multiply: Multiply attention weights with original EfficientNetB7 feature map  11: Apply Global Average Pooling again to get a vec tor  12: Add a Dense layer with ReLU activation (e.g., 256 units)  13: Add Dropout for regularization  14: Add final Dense layer with Softmax activation for classification (e.g., 5 classes)  15: Define model M using EfficientNetB7 + SE block + classification head  16: Specify optimizer, loss (e.g., categorical crossen tropy), and metrics (accuracy, precision, recall) 17: Compile model M  18: Train M on Dtrain with validation on Dval 19: while not converged do  20: Monitor validation performance  21: Adjust learning rate dynamically  22: If early stopping criteria met then  23: Stop training  24: end if  25: end while  26: Evaluate M on Dtest  27: return trained model M |

### 3.3.3 Performance Measures

Several widely used performance criteria, including as accuracy, precision, recall, and the confusion matrix, were employed to assess the suggested model’s performance. Particularly in a multi-class classification task, these indicators offer a broad picture of the model’s classification performance.

**Accuracy**

Accuracy is the ratio of correctly predicted in stances to the total number of instances. It is calculated using the formula:

Accuracy =

Although accuracy provides an overall indication of performance, it may be misleading in the presence of class imbalance.

**Precision and Recall**

Precision shows the percentage of occurrences that were accurately predicted of all instances that were forecasted for a certain class:

Precision =

**Recall** (also referred to as true positive rate or sensitivity) is the percentage of real cases in a class that were accurately recognized:

**Recall=**

These metrics were calculated for each class individually and averaged using both macro and micro averaging methods for a more detailed evaluation.

**Confusion Matrix (5-Class Classification)**

A confusion matrix was generated to visually as sess the model’s classification performance across the five target classes. In this 5×5 matrix, each row corresponds to the actual class labels, while each column represents the predicted class labels. The diagonal elements of the matrix indicate the number of instances correctly classified for each class, also known as true positives TP)[51]. In contrast, the off-diagonal elements represent misclassifications, where instances of one class were incorrectly predicted as another. These include both false positives (FP) and false negatives (FN). This matrix provides a detailed breakdown of the model’s strengths and weaknesses in distinguishing among the five classes. An example structure of a 5-class confusion matrix is shown in table 5.

 Table 5: Structure of a 5-Class Confusion Matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Predicted Class 1 | Predicted Class 2 | Predicted Class 3 | Predicted Class 4 | Predicted Class 5 |
| Actual Class 1 | TP | FP | FP | FP | FP |
| Actual Class 2 | FN | TP | FP | FP | FP |
| Actual Class 3 | FN | FN | TP | FP | FP |
| Actual Class 4 | FN | FN | FN | TP | FP |
| Actual Class 5 | FN | FN | FN | FN | TP |

The confusion matrix of table 5 was designed as a 5×5 matrix to fit the five-class classification problem of Diabetic Retinopathy detection. These five classes are derived from the clinical grading scale: Class 0 (No DR), Class 1 (Mild), Class 2 (Moderate), Class 3 (Severe), and Class 4 (Proliferative DR). This classification is important in capturing the progressive nature of the disease, where each stage signifies increasing levels of retinal damage. A 5-class model not only aligns with clinically utilized diagnostic classes but also enables the model to generate finer-grained predictions, which can aid in early detection and treatment planning. The confusion matrix is thus an important measure in determining the capacity of the model in distinguishing between the clinically meaningful stages of Diabetic Retinopathy.

4

Experimental Results

An outcome analysis of the research has been provided in this chapter. This chapter has demonstrated the Diabetic Retinopathy Detection using Attention-Enhanced Transfer Learning and Explainable AI Techniques result analysis to assess the efficacy of the algorithms. The research report's results chapter will try to summarize the findings. It will also serve as a guide for the discussion section. The results are shown and analyzed. Final findings are presented.

## 4.1 Experimental setup

The experiments of the present study were conducted with local and cloud settings to determine computational performance as well as scalability. For initial development as well as initial testing, a personal laptop powered by an Intel Core i5 processor was used with 8 GB RAM along with separate 2 GB AMD Radeon dedicated graphics cards. Nevertheless, due to the computationally intensive nature of training deep learning models, all important experiments and assessments were executed on the Kaggle environment, which enables high-performance cloud computing resources. Specifically, the most recent training and testing were conducted on Kaggle’s hosted GPU setting using NVIDIA Tesla P100 GPU acceleration. The setup significantly reduced training time and allowed for efficient processing of large data and complex neural network structures. Experiments were implemented in Python using common libraries such as TensorFlow, Keras, PyTorch, NumPy, and Pandas to ensure robust and reproducible development. Table 6 shows the Summary of Hyperparameters and Layer Configuration of deep learning model.

Table 6: Summary of Hyperparameters and Layer Configuration

|  |  |  |
| --- | --- | --- |
| **Parameter/Layer** | **Value** | **Description** |
| Input Shape | (600, 600, 3) | Input image dimensions; 600×600 RGB images |
| Backbone Model | EfficientNetB7 | Pretrained feature extractor from ImageNet with 64M non-trainable parameters |
| Trainable Parameters | 1,476,101 | Only top layers were fine-tuned to reduce overfit ting and training time |
| GlobalAveragePooling2D— |  | Reduces feature maps to a 1D vector by averaging |
| Reshape Layer | (1, 1, 2560) | Prepares feature vector for channel-wise recali bration |
| Dense Layer 1 | 160 Units | Part of attention mechanism for channel recalibra tion |
| Dense Layer 2 | 2560 Units | Expands back to original channels for multiply operation |
| Multiply Layer | Element-wise | Multiplies recalibrated attention map with original feature maps |
| GlobalAveragePooling2D(Post-Attention) | — | Compresses the recalibrated feature map |
| Dense Layer 3 | 256 Units | Fully connected layer before classification |
| Dropout Rate | Applied (default) | Helps prevent overfitting |
| Output Dense Layer | 5 Units | Final classification layer for 5 DR classes |
| Activation Function | Softmax | Converts logits into probabilities for multi-class classification |
| Total Parameters | 65,573,788 | Total model parameters (trainable + frozen) |
| Model Size | 250.14 MB | Memory footprint of the model |

## 4.2 Individual model performance

Here, we provide a detailed comparison of the performance of each model employed, such as individual confusion matrices and performance plots. The com parison measures employed for assessment are validation accuracy, test accuracy, validation loss, and test loss.

### 4.2.1 CNN 1 Model with Attention

CNN1with attention achieved a moderate amount of accuracy on the test and validation sets. Validation accuracy was 68.38%, while test accuracy was slightly lower at 62.66%. Even though the validation accuracy was very good, the test loss was very high (2.5031), indicating overfitting on the training set. The confusion matrix and performance plot of the model are provided below figure 11.

A graph of a curve

AI-generated content may be incorrect.

*Figure 11: Performance Graph of CNN 1 Model with Attention (Accuracy vs. Loss)*

### 4.2.2 Deep CNN 2 with Attention and Skip Connections

Deep CNN model with attention and skip connections underperformed, achieving only 49.12% validation accuracy and 50.61% test accuracy. The model’s validation loss of 10.3602 and test loss of 1.7167 suggest significant instability and overfitting, likely due to the complexity of the architecture. Despite these challenges, the model was evaluated with a confusion matrix and performance graph for show in figure 13 & 12.

A graph of a line

AI-generated content may be incorrect.

*Figure 12: Performance Graph of Deep CNN with Attention and Skip Connections (Accuracy vs. Loss)*

A chart with numbers and a blue box

AI-generated content may be incorrect.

*Figure13: Confusion matrix of Deep CNN with Attention and Skip Connections*

### 4.2.3 CNN 3 with Attention Modules

The CNN 3 model with attention modules showed a marked improvement over the previous models, achieving a validation accuracy of 71.71% and test accuracy of 63.97%. The validation loss of 0.7712 and test loss of 0.0397 indicate that the model per formed well in both the training and testing phases, with a significant reduction in overfitting compared to the earlier models. confusion matrix and performance graph for this model are shown in figure 15 & 14.

A graph of a line

AI-generated content may be incorrect.

*Figure14: Performance Graph of CNN3 with Attention Modules (Accuracy vs. Loss)*

*A blue and white chart with numbers and a blue square

AI-generated content may be incorrect.*

*Figure 15: Confusion matrix of CNN 3 with Attention Modules*

### 4.2.4 EfficientNetB7 + Squeeze-Excite Block

EfficientNetB7 model combined with squeeze-and excitation blocks outperformed the other models, with a validation accuracy of 83.23% and test accuracy of 83.79%, demonstrating strong generalization capability. The validation loss (0.5209) and test loss (0.5162) were both low, indicating the model’s robust performance on unseen data. This model’s Classification Report shown table 7 performance graph are shown figure 16.

A graph of two people

AI-generated content may be incorrect.

*Figure 16: Performance Graph of EfficientNetB7 + Squeeze-Excite Block (Accuracy vs. Loss)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| 0 | 0.87 | 0.97 | 0.92 | 6896 |
| 1 | 0.61 | 0.10 | .17 | 466 |
| 2 | 0.69 | 0.54 | 0.61 | 1518 |
| 3 | 0.46 | 0.26 | 0.33 | 128 |
| 4 | 0.68 | 0.57 | 0.62 | 253 |

Table 7: Classification performance of EfficientNetB7 + Squeeze- Excite Block metrics per class

## 4.3 Performance analysis

To check the strength of our proposed DR classification model, we compared the performance of four DL models using both validation and test datasets. CNN 1 model with attention mechanism achieved a validation accuracy of 68.38% and a test accuracy of 62.66%, a validation loss of 1.3159 and a relatively high-test loss of 2.5031, which indicates moderate generalization capability. The Deep CNN with attention and skip connections did poorly, obtaining a validation accuracy of only 49.12% and test ac curacy of 50.61%, with a highly inflated validation loss (10.3602) and test loss (1.7167), showing over fitting and model instability. Conversely, the CNN 2 extended with attention modules performed superior generalization, achieving validation accuracy of 71.71% and test accuracy of 63.97%, along with reduced losses (validation loss: 0.7712, test loss: 0.0397). Among them, EfficientNetB7 model with the inclusion of squeeze-and-excitation blocks posted the highest performance with a validation accuracy of 83.23%, test accuracy of 83.79%, and lowest validation and test losses of 0.5209 and 0.5162, respectively. These results indicate the effectiveness in the use of transfer learning alongside attention-based mechanisms for obtaining high accuracy and model confidence. Table 8 shows the Performance Metrics of Applied Models

Table 8: Performance Metrics of Applied Models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Validation Accuracy | Precision | Recall | F1 Score | Support |
| CNN 1 Model with Attention  Deep CNN 2 with Attention + Skip Connections CNN 3 with Attention Modules  EfficientNetB7 + SE Block | 0.6838  0.4912  0.7171  0.8323 | 0.58  0.40  0.61  0.82 | 0.63  0.45  0.68  0.84 | 0.59  0.42  0.67  0.81 | 14,201  9,261  9,261  9,261 |

## 4.4 Error Analysis

Error analysis is significant as far as understanding the individual vulnerabilities of the models and what has to be done to correct them is concerned. In this part, we perform a detailed examination of the errors made by the models, their misclassifications, and the types of data on which the models performed badly. The purpose is to build an understanding of what can be done to correct the models and if certain patterns of errors are model-independent or not.

### 4.4.1. Misclassification Patterns

To begin with, we analyze the confusion matrices for each model, which provide a comprehensive view of the TP, FP, TN, and FN classifications. By examining these matrices, we can identify which classes are commonly confused with one another and investigate whether certain models are prone to misclassify specific categories. The confusion matrix for each model was plotted to better visualize the following:

1. Class distribution of misclassifications.

2. The types of errors (false positives vs. false negatives).

3. The relationship between model performance and error rates for each class.

### 4.4.2. Model-Specific Misclassifications

In architectures like CNN 1 with Attention and Deep CNN 2 with Attention + Skip Connections, we noted that misclassifications were more frequent in certain classes with high variability or inherent ambiguity. These models mixed up visually similar classes or classes with less distinguishable features. Efficient NetB7 with SE Block was better able to distinguish between these classes, with fewer misclassifications. The error analysis, supported by visualizations such as confusion matrices, error rate distributions, provided valuable insights into the strengths and weak nesses of each model. While the Effi cientNetB7 with SE Block model performed the best overall, there is room for further improvement, particularly in classes with ambiguous features. By focusing on these error-prone classes and using additional techniques such as data augmentation, further fine-tuning, and more sophisticated feature extraction methods, future models could improve their performance even further.

## 4.5. Statistical Significance Test

To ensure that differences in performance noticed among the models deployed are not due to random variation, a statistical significance test was conducted. Paired t-test was applied to test whether the classification performance of the best model (i.e., test accuracy or F1-score), EfficientNetB7 with squeeze-excite block, differed significantly from other models.

The null hypothesis (*H*0) assumes that there is no significant difference in the performance between the two models, while the alternative hypothesis (*H*1) suggests that the difference is statistically significant.

To further validate the observed performance differences between the models, a paired t-test was conducted to compare the EfficientNetB7 with SE Block model against each of the other models. The results of these tests, including the computed *p*-values, are as follows:

**EfficientNetB7 vs. CNN 1 with Attention:**

The paired t-test returned to a *p*-value of 7.3273 × 10−81, indicating a highly statistically significant difference between the performance of EfficientNetB7 and CNN 1 with Attention. This suggests that the performance improvement observed with EfficientNetB7 is not due to random chance.

**EfficientNetB7 vs. Deep CNN 2 with Attention + Skip Connections:**

The *p*-value of 1.9834 × 10−101 further confirms that the performance of EfficientNetB7 significantly surpasses that of Deep CNN 2 with Attention + Skip Connections. This extremely low p-value reinforces the conclusion that the improvement is statistically significant and reliable.

**EfficientNetB7 vs. CNN 3 with Attention:**

The paired t-test for EfficientNetB7 and CNN 3 with Attention yielded a *p*-value of 2.5115 × 10−57, which also indicates a statistically significant difference in performance. This provides strong evidence that EfficientNetB7 outperforms CNN 3 with Attention in terms of accuracy and F1-score. In all cases, the *p*-values are well below the significance thresh old of 0.05, confirming that the differences in performance between EfficientNetB7 and the other models are highly significant. Thus, the results conclusively demonstrate that EfficientNetB7 with SE Block outperforms the other models in this study in multiple performance metrics. Visualization in figure 17 such as bar graphs.

A graph with different colored lines

AI-generated content may be incorrect.

*Figure 17: Paired t-test results comparing model accuracies. Bars represent confidence intervals; \* denotes p < 0.05 significance.*

## 4.6. XAI Performance

We used Grad-CAM to show the areas of input photos that had a substantial impact on the Efficient NetB7 model’s predictions to improve the interpretability of the DL model.

Grad-CAM highlights the most significant areas of the picture that influenced the model’s classification decision by creating heatmaps using the gradients that flow into the final convolutional layer. These illustrations shed light on the model’s inner workings and confirm that it concentrates on pertinent areas of the image when generating predictions.



*Figure 18: Grad-CAM visualization showing the activated regions for model predictions using EfficientNetB7.*

As illustrated in Figure 18, the highlighted areas correspond well with the disease-affected or target specific regions in the input images. This indicates that the model is learning meaningful and contextually relevant patterns rather than relying on spurious correlations. Figure 19 Ablation Grad-CAM visualization. These explainability results strengthen the trustworthiness and clinical applicability (or domain relevance) of the model by ensuring transparency in decision-making.

A close-up of different colored shapes

AI-generated content may be incorrect.

Figure 19: Ablation Grad-CAM visualization

5

Discussion

## 5.1. Comparison With Current State of the Art Approaches.

To evaluate the performance of our suggested model, we provided a comprehensive comparison with several recent state-of-the-art approaches in the field of DR detection and retinal image analysis. Table 9 lists five recent models and compares them with our suggested method, including architectural choices, utilized datasets, and feature extraction techniques. As compared to such cutting-edge techniques, our approach is an EfficientNetB7 backbone with Squeeze-and-Excitation modules and fully connected layers, leaping on the bandwagon of transfer learning for adaptation to retinal imaging problems. Specifically, our approach utilizes a combination of three top of-the-line datasets—APTOS, EyePACS, and Messidor—which come with certain broader generalize ability guarantees. Unlike approaches like LAT and Student, which focus more on lesion-aware or dual branch transformer-based features, our approach focuses on channel-wise feature recalibration as well as dense connections for richer discriminative capacity.

Further, our design is relatively light weight compared to multi-branch or ensemble transformer strategies and is still capable of producing high accuracy. This makes our strategy more practical for deployment in real-world screening systems, especially those with resource constraints.

Table 9: Comparison with recent state-of-the-art methods

|  |  |  |  |
| --- | --- | --- | --- |
| Study [Ref] | Model Architecture | Dataset(s) | Feature Extraction Tech nique |
| Proposed | EfficientNetB7 + SE + Dense | Combined (APTOS + EyePACS + Messi dor) | Transfer learning + Squeeze-and-Excitation |
| [15] SSiT | Saliency-guided Vi sion Transformer (SSiT) | EyePACS, DDR, Messidor-2, APTOS 2019 | Saliency-guided self supervised learning |
| [17] RTNet | Relation Transformer Network (RTNet) | IDRiD, DDR | Self-attention + Cross attention mechanisms |
| [19] DTUNet | Dual-branch Trans former Network (DTUNet) | e-ophtha, IDRiD | Dual-branch feature fusion for localized attention |
| [11] LAT | Lesion-Aware Trans former (LAT) | Messidor-1,  Messidor-2, Eye PACS | Lesion-aware spatial filter ing + Transformer features |
| [34] OCTFormer | Hierarchical Trans former (OCTFormer) | Custom OCT Dataset | Multi-scale hierarchical transformer features |

From table 10 Our model, EfficientNetB7 + Squeeze-and-Excitation (SE) + Dense layers, is one of the most effective solutions to detect DR because it possesses an optimal performance across a range of measures. It has a 83.23% accuracy, 82% precision, 84% recall, and 81% F1-score, which shows an excellent ability to accurately detect cases of DR and minimize false alarms (high precision). This precision-recall trade-off is extremely crucial in medical application scenarios, where both false positives and false negatives can have significant consequences. Even though models like Query2Label and Swin-Poly Transformer are extremely precise, they don't provide significant metrics like precision, recall, and F1-score, making it difficult to assess their overall effectiveness. Your model also excels in its generalizability, as it has demonstrated stable performance on a variety of datasets, including APTOS, EyePACS, and Messidor, which is essential for generalization in real-world clinical settings. Furthermore, the EfficientNetB7 backbone, augmented with Squeeze-and-Excitation (SE) blocks to increase feature selection, allows your model to effectively capture fine details required for DR detection. In comparison, other models like Ensemble ViT are less precise, with a 47% precision score on the Kaggle 2015 DR dataset, further highlighting your model's superiority. Overall, the large performance metrics, model structure, and robust generalization across datasets position your model as a top contender for DR detection, surpassing several other state-of-the-art approaches in reliability and accuracy.

Table 10: Comparison with recent state-of-the-art methods with Accuracy, Precision, Recall &F1-Score

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Ref** | **Model / Method** | **Dataset** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| ***Proposed*** | ***EfficientNetB7 + SE + Dense*** | ***Combined (APTOS + EyePACS + Messidor)*** | ***83.23%*** | ***82%*** | ***84%*** | ***81%*** |
| [28] | Query2Label | RFMiD | 99.8% | — | — | — |
| [31] | ViT | FGADR | — | — | — | 0.825 |
| [26] | Swin-Poly Transformer | OCT2017, OCT-C8 | 99.80% | — | — | — |
| [24] | Ensemble ViT | Kaggle 2015 DR | — | 47% | — | — |
| [36] | Multi-disease Transformer Model | ODIR-2019 | 93.1% | — | — | — |

## 5.2. Limitation and Future Work

While promising performance has been achieved using the proposed EfficientNetB7-based model and Squeeze-and-Excitation and dense layers, there are several limitations to be considered. First, model performance can be impacted by dataset imbalance since certain DR severity levels are underrepresented and biased predictions may result. Secondly, the model was trained and tested primarily on combined datasets (APTOS, EyePACS, and Messidor), which can limit its generalizability to other populations or imaging conditions not included in these datasets.

One more limitation is model interpretability. Al though high-value evaluation measures such as ac curacy, the black-box nature of deep learning models continues to be a problem in clinical applications where interpretability is key. Further, although trans fer learning enhances convergence and generalizability of the model, it may also propagate bias from pre trained weights not customized for medical imaging.

Regarding future work, we are going to apply explainable AI (XAI) methods to increase interpretability and clinician confidence. We also intend to explore more advanced data augmentation and class balancing methods to improve minority class performance. Boosting robustness with more diverse datasets in other demographic and imaging conditions will be another priority by expanding the training. Finally, live deployment and seamless integration into clinical workflows will be investigated, highlighting computational efficiency as well as friendly interfaces.

6

Conclusion

Based on the capabilities of EfficientNetB7 and the use of Squeeze-and-Excitation (SE) blocks in con junction with thick layers, we introduced a reliable and effective deep learning model in this study for the identification of DR. To improve generalizability and performance, the suggested model was trained and evaluated using a variety of datasets, including APTOS, EyePACS, and Messidor. Experimental findings show that our model, which combines transfer learning with effective feature refinement strategies, outperforms several current state-of-the-art approaches in terms of classification accuracy.

Our method has strong potential for automated detection of DR and can assist ophthalmologists in timely diagnosis and treatment planning. However, there is still more to do to overcome the issues of class imbalance and explainability. The incorporation of XAI tools and testing on large datasets will be significant steps toward clinical implementation. Overall, the system described here is a significant step forward in the use of deep learning techniques to real-world clinical imaging issues.

References

[1] M. A. Rodr´ıguez, H. AlMarzouqi, and P. Liatsis, “Multi Label Retinal Disease Classification Using Transform ers,” *IEEE Journal of Biomedical and Health Informat ics*, vol. 27, no. 6, pp. 2739–2750, June 2023, doi: 10.1109/JBHI.2022.3214086.

[2] D. S. W. Ting, G. C. M. Cheung, and T. Y. Wong, “Diabetic retinopathy: Global prevalence, major risk factors, screening practices and public health challenges: A review,” *Clin. Exp. Ophthalmol.*, vol. 44, no. 4, pp. 260–277, May 2016, doi: 10.1111/ceo.12696.

[3] Z. L. Teo, Y. C. Tham, M. Yu, *et al.*, “Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis,” *Ophthalmology*, vol. 128, no. 11, pp. 1580–1591, Nov. 2021, doi: 10.1016/j.ophtha.2021.04.027.

[4] M. Khater, “Diabetic Retinopathy Detection Model using Hybrid of U-Net and Vision Transformer Algorithms,” *INTI Journal*, 2024. [Online]. Available: https://iuojs.intimal.edu.my/index.php/intijournal/article/view/113

[5] A. Farcas¸, A. Lobont¸, D. Babuc, T. Ivas¸cu, and S.-A. S¸tefanig ˘ a, “EI-ViT-Net: a Vision Transformer Approach for ˘ Diabetic Retinopathy’s Decision Support Pipeline,” in *2024 IEEE 24th International Conference on Bioinformatics and Bioengineering (BIBE)*, Kragujevac, Serbia, 2024, pp. 1–6, doi: 10.1109/BIBE63649.2024.10820483.

[6] C. Karkera, C. Adak, S. Chattopadhyay, and M. Saqib, “Detecting severity of Diabetic Retinopathy from fun dus images: A transformer network-based review,” *Neu rocomputing*, vol. 597, p. 127991, Sep. 2024, doi: 10.1016/j.neucom.2024.127991.

[7] M. Touati, R. Touati, L. Nana, F. Benzarti, and S. Ben Yahia, “DRCCT: Enhancing Diabetic Retinopathy Classi fication with a Compact Convolutional Transformer,” *Big Data Cogn. Comput.*, vol. 9, no. 1, p. 9, Jan. 2025, doi: 10.3390/bdcc9010009.

[8] M. Ali, M. G. Yaseen, M. Aljanabi, and S. A. Abed, “Trans fer Learning: A New Promising Techniques,” *Mesopotamian J. Big Data*, vol. 2023, pp. 29–30, Feb. 2023. [Online]. Avail able: doi: 10.58496/MJBD/2023/004.

[9] L. K. Hansen, “Attention: A machine learning perspective,” in *2012 3rd International Workshop on Cognitive Informa tion Processing (CIP)*, Baiona, Spain, 2012, pp. 1–6, doi: 10.1109/CIP.2012.6232894.

[10] H. Zheng and F. Liu, “A Cascade U-Net With Transformer for Retinal Multi-Lesion Segmentation,” *IMA Journal of Mathe matical Control and Information*, first published Sep. 5, 2024, doi: 10.1002/ima.23163.

[11] A. A. Shafi, M. Hossain Shawon, N. Shahid, R. Rahman, and R. Khan, “Enhancing Diabetic Retinopathy Detection Through Transformer Based Knowledge Distillation and Ex plainable AI,” in *2024 International Joint Conference on Neu ral Networks (IJCNN)*, Yokohama, Japan, 2024, pp. 1–8, doi: 10.1109/IJCNN60899.2024.10650642.

[12] O. Islam, K. Kumer, S. Akter, and M. M. Uddin, “Multi-Head Self-Attention Mechanisms in Vision Transformers for Reti nal Image Classification,” in *2024 IEEE International Con ference on Computing, Applications and Systems (COMPAS)*, Cox’s Bazar, Bangladesh, 2024, pp. 1–5, doi: 10.1109/COM PAS60761.2024.10795956.

[13] C. Xu, X. Guo, G. Yang *et al.*, “Prior-guided attention fusion transformer for multi-lesion segmentation of dia betic retinopathy,” *Sci Rep*, vol. 14, p. 20892, 2024, doi: 10.1038/s41598-024-71650-6.

[14] R. Wu, R. Hu, Z. Xiao, J. Chen, and J. Liu, “Vi sion Transformer-based recognition of diabetic retinopathy grade,” *Medical Physics*, first published Oct. 25, 2021, doi: 10.1002/mp.15312.

[15] Y. Huang, J. Lyu, P. Cheng, R. Tam, and X. Tang, “SSiT: Saliency-guided Self-supervised Image Trans former for Diabetic Retinopathy Grading,” *arXiv preprint arXiv:2210.10969*, Mar. 2024. [Online]. Available: https://arxiv.org/abs/2210.10969.

[16] H. Hu, W. Kang, and F. Deng, “CNN model, CNN training method and vein identification method based on CNN,” *Jour nal Name*, vol. X, no. Y, pp. 123–128, 2017,

[17] S. Huang, J. Li, Y. Xiao, N. Shen, and T. Xu, “RT Net: Relation Transformer Network for Diabetic Retinopa thy Multi-Lesion Segmentation,” *IEEE Transactions on Med ical Imaging*, vol. 41, no. 6, pp. 1596–1607, June 2022, doi: 10.1109/TMI.2022.3143833.

[18] Z. Jiang, L. Wang, Q. Wu, Y. Shao, M. Shen, W. Jiang, and C. Dai, “Computer-aided diagnosis of retinopathy based on vision transformer,” *J. Innov. Opt. Health Sci.*, vol. 15, no. 2, p. 2250009, 2022, doi: 10.1142/S1793545822500092.

[19] H. Wen, K. He, X. Feng, F. Lan, and X. Liang, “DTUNet: Dual branch transformer U-Net for hard exudate segmenta tion in retinal image,” *J. Electron. Imaging*, vol. 32, no. 4, p. 043022, Aug. 2023, doi: 10.1117/1.JEI.32.4.043022.

[20] X. Jin, H. Li, and R. Li, “Dformer: Dual-Path Transformers for Geometric and Appearance Features Reasoning in Dia betic Retinopathy Grading,” in *Pattern Recognition and Com puter Vision. PRCV 2022*, S. Yu *et al.*, Eds. Cham: Springer, 2022, vol. 13535, pp. 405–416, doi: 10.1007/978-3-031- 18910-433.

[21] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Interna tional Conference on Machine Learning (ICML)*, 2019. [Online]. Available: https://arxiv.org/abs/1905.11946, doi: 10.48550/arXiv.1905.11946.

[22] R. Sun, Y. Li, T. Zhang, Z. Mao, F. Wu, and Y. Zhang, “Lesion-Aware Transformers for Diabetic Retinopathy Grad ing,” in *2021 IEEE*/*CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021, pp. 10933–10942, doi: 10.1109/CVPR46437.2021.01079.

[23] Z. Zhou, C. Niu, H. Yu, J. Zhao, Y. Wang, and C. Dai, “Diagnosis of retinal diseases using the vision transformer model based on optical coherence tomography images,” *Proc. SPIE*, vol. 12601, p. 1260102, 2023, doi: 10.1117/12.2665918.

[24] S. Choi *et al.*, “A Vision Transformer Based Deep Learning Architecture for Automatic Diagnosis of Diabetic Retinopathy in Optical Coherence Tomography Angiography,” in *Mi tosis Domain Generalization and Diabetic Retinopathy Anal ysis. MIDOG DRAC 2022*, B. Sheng and M. Aubreville, Eds. Cham: Springer, 2023, vol. 13597, pp. 140–150, doi: 10.1007/978-3-031-33658-413.

[25] X. Zhang, J. Zhang, Y. Zhang, M. Jia, H. Li, and X. Liu, “Adaptive learning U-Net-based adversarial network with CNN and transformer for segmentation of hard exudates in diabetes retinopathy,” *IET Image Process.*, first published Jul. 3, 2023, doi: 10.1049/ipr2.12865.

[26] N. AlDahoul, H. A. Karim, M. J. T. Tan, M. A. Momo, and J. L. Fermin, “Encoding retina image to words using ensemble of vision transformers for diabetic retinopathy grading,” *F1000Research*, vol. 10, p. 948, Sep. 21, 2021, doi: 10.12688/f1000research.73082.1.

[27] D. Philippi, K. Rothaus, and M. Castelli, “A Vision Trans former Architecture for the Automated Segmentation of Retinal Fluids in Spectral Domain Optical Coherence Tomography Images,” *arXiv preprint*, 2023, doi: 10.21203/rs.3.rs 2262988/v1.

[28] J. He, J. Wang, Z. Han *et al.*, “An interpretable transformer network for the retinal disease classification using optical coherence tomography,” *Sci Rep*, vol. 13, p. 3637, 2023, doi: 10.1038/s41598-023-30853-z.

[29] D. Philippi, K. Rothaus, and M. Castelli, “A vision trans former architecture for the automated segmentation of reti nal lesions in spectral domain optical coherence tomography images,” *Sci Rep*, vol. 13, p. 517, 2023, doi: 10.1038/s41598- 023-27616-1.

[30] T. J. C. Silva, S. E. R. Fernandes, J. D. S. de Almeida,

D. B. P. Quintanilha, and G. Braz Junior, “Detecc¸ao e ˜ diagnostico autom ´ atico de patologias na retina utilizando ´ arquitetura baseada em Transformers,” in *Anais do 24º Simp´osio Brasileiro de Computa¸c˜ao Aplicada `a Sa´ude (SB CAS)*, Goiania, Brazil, 2024, pp. 178–189, doi: 10.5753 ˆ /sb cas.2024.2134.

[31] J.-H. Wu, N. D. Koseoglu, C. Jones, and T. Y. Alvin Liu, “Vision transformers: The next frontier for deep learning-based ophthalmic image analysis,” *Saudi J. Oph thalmol.*, vol. 37, no. 3, pp. 173–178, Jul.–Sep. 2023, doi: 10.4103/sjopt.sjopt9123.

[32] Y. Yang, Z. Cai, S. Qiu, and P. Xu, “A Novel Trans former Model With Multiple Instance Learning for Diabetic Retinopathy Classification,” *IEEE Access*, vol. 12, pp. 6768– 6776, 2024, doi: 10.1109/ACCESS.2024.3351473.

[33] W. Nazih, A. O. Aseeri, O. Y. Atallah, and S. El-Sappagh, “Vision Transformer Model for Predicting the Severity of Di abetic Retinopathy in Fundus Photography-Based Retina Im ages,” *IEEE Access*, vol. 11, pp. 117546–117561, 2023, doi: 10.1109/ACCESS.2023.3326528.

[34] H. Wang *et al.*, “OCTFormer: An Efficient Hierarchical Transformer Network Specialized for Retinal Optical Coher ence Tomography Image Recognition,” *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–17, 2023, Art. no. 2532217, doi: 10.1109/TIM.2023.3329106.

[35] A. G. Roy, N. Navab, and C. Wachinger, “Concurrent Spa tial and Channel ‘Squeeze & Excitation’ in Fully Convolu tional Networks,” in *Medical Image Computing and Com puter Assisted Intervention – MICCAI 2018*, A. Frangi *et al.*, Eds. Cham: Springer, 2018, vol. 11070, pp. 421–429, doi: 10.1007/978-3-030-00928-111.

[36] E. Lan, “HDformer: A higher dimensional transformer for di abetes detection utilizing long range vascular signals,” *arXiv preprint*, 2023, doi: 10.48550/arXiv.2303.11340.

[37] J. Lian and T. Liu, “Lesion identification in fundus images via convolutional neural network-vision transformer,” *Biomed. Signal Process. Control*, vol. 88, Part A, p. 105607, Feb. 2024, doi: 10.1016/j.bspc.2023.105607.

[38] B. Ait Hammou, F. Antaki, M.-C. Boucher, and R. Du val, “MBT: Model-based transformer for retinal optical co herence tomography image and video multi-classification,” *Int. J. Med. Inform.*, vol. 172, p. 105178, 2023, doi: 10.1016/j.ijmedinf.2023.105178.

[39] I. Hisham, M. I. Khalil, and H. Abbas, “Multi-label Ophthal mological Disease Classification Using Vision Transformers,” in *2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, Giza, Egypt, 2023, pp. 279–284, doi: 10.1109/NILES59815.2023.10296804.

[40] Z. Shi, Y. Li, H. Zou, and X. Zhang, “TCU-Net: Transformer embedded in convolutional U-shaped network for retinal ves sel segmentation,” *Sensors*, vol. 23, no. 10, p. 4897, May 2023, doi: 10.3390/s23104897.

[41] Y. K. Wang, K. Kai Wen, C.-K. Lu, and C.-H. Lin, “Retinal Layer and Fluid Segmentation with Transformer Based Architecture,” in *2023 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, Ping Tung, Taiwan, 2023, pp. 409–410, doi: 10.1109/ICCE Taiwan58799.2023.10226946.

[42] H. Yu, J.-h. Shim, J. Kwak, J. W. Song, and S.-J. Kang, “Vision Transformer-Based Retina Vessel Segmentation with Deep Adaptive Gamma Correction,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Sig nal Processing (ICASSP)*, Singapore, Singapore, 2022, pp. 1456–1460, doi: 10.1109/ICASSP43922.2022.9747597.

[43] D. Wang, J. Lian, and W. Jiao, “Multi-label classification of retinal disease via a novel vision transformer model,” *Front. Neurosci.*, vol. 17, Sec. Visual Neuroscience, 2023, Art. no. 1290803, doi: 10.3389/fnins.2023.1290803.

[44] S. Yu *et al.*, “MIL-VT: Multiple Instance Learning Enhanced Vision Transformer for Fundus Image Classification,” in *Med ical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne *et al.*, Eds. Cham: Springer, 2021, vol. 12908, pp. 45–54, doi: 10.1007/978-3-030-87237-

[45] C. Wu, M. Guo, M. Ma, and K. Wang, “Diverter transformer based multi-encoder-multi-decoder network model for med ical retinal blood vessel image segmentation,” *Biomed. Sig nal Process. Control*, vol. 93, pp. 106132, Jul. 2024, doi: 10.1016/j.bspc.2024.106132.

[46] J. Liu, J. Li, A. Wolde, C. Cukras, and J. Tam, “Hybrid transformer for lesion segmentation on adaptive optics reti nal images,” *Proc. SPIE Med. Imaging 2022: Comput.- Aided Diagn.*, vol. 12033, p. 120331W, Apr. 2022, doi: 10.1117/12.2612379.

[47] C. Wu, M. Guo, M. Ma, and K. Wang, “TLTNet: A novel transscale cascade layered transformer network for enhanced retinal blood vessel segmentation,” *Com put. Biol. Med.*, vol. 178, p. 108773, Aug. 2024, doi: 10.1016/j.compbiomed.2024.108773.

[48] M. Zhang, O. Kekesi, and G. J. Suaning, “Transformer-based ´ light-evoked retinal spiking activity prediction,” in *2023 45th Annual International Conference of the IEEE Engineering in Medicine* & *Biology Society (EMBC)*, Sydney, Australia, 2023, pp. 1–4, doi: 10.1109/EMBC40787.2023.10340616.

[49] L. P. Divvela, Y. Yarra, and K. S. Naidana, “Vision Transformers-based Retinal Image Screening for Glaucoma Identification,” in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, 2024, pp. 1–8, doi: 10.1109/ICC CNT61001.2024.10723834.

[50] D. Singh, M. Ammar, K. Varshney, and Y. U. Khan, “Optical Coherence Tomography Image Classification us ing Light-weight Hybrid Transformers,” in *2023 Interna tional Conference on Recent Advances in Electrical, Elec tronics* & *Digital Healthcare Technologies (REEDCON)*, New Delhi, India, 2023, pp. 185–189, doi: 10.1109/REED CON57544.2023.10151279.

[51] D. M. W. Powers, “Evaluation: From precision, recall and F measure to ROC, informedness, markedness and correlation,” *arXiv preprint arXiv:2010.16061*, Oct. 2020. [Online]. Avail able: https://arxiv.org/abs/2010.16061.

[52] S. Akc¸a, Z. Garip, E. Ekinci *et al.*, “Automated classification of choroidal neovascularization, diabetic macular edema, and drusen from retinal OCT images using vision transformers: a comparative study,” *Lasers Med Sci*, vol. 39, p. 140, 2024, doi: 10.1007/s10103-024-04089-w. ​ ​