

University of Nottingham
UK | CHINA | MALAYSIA

Diabetic Retinopathy Image Classification with Quantitative Saliency-oriented Visual Explanation and SLIC-G Image Preprocessing Methods

Lim Wei Xiang

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

September 2023

To my beloved wife, parents, family members, and friends who supported me unconditionally.

List of Publications

1. W. X. Lim, Z. Chen, A. Ahmed, "The adoption of deep learning interpretability techniques on diabetic retinopathy analysis: a review," *Journal of Medical & Biological Engineering & Computing (MBEC)*, vol. 60, no. 3, pp. 633 - 642, 2022

2. W. X. Lim, Z. Chen, "A Quantitative Interpretable Convolutional Neural Network to Diagnose Diabetic Retinopathy," *Transactions on Computational Science and Computational Intelligence*, status "Accepted, To be Printed"

3. W. X. Lim, Z. Chen, "Enhancing Deep Learning Pre-Trained Networks on Diabetic Retinopathy Fundus Photographs with SLIC-G," *Transactions on Computational Science and Computational Intelligence*, submitted for publication, status "Accepted, To be Printed".

4. W. X. Lim, Z. Chen, A. Ahmed, T. Chandesa, I. Liao, "A Review of Machine Learning Techniques for Applied Eye Fundus and Tongue Digital Image Processing with Diabetes Management System," arXiv:2012.15025 [eess.IV], Dec. 2020.

5. E. F. Goh, Z. Chen, W. X. Lim, "Frequency Domain Convolutional Neural Network: Accelerated CNN for Large Diabetic Retinopathy Image Classification," arXiv:2106.12736 [cs.CV], June. 2021.

Declaration

I solemnly declare that, unless explicitly referenced to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification at this or any other university. This dissertation is a product of my own efforts and does not contain any outcomes from work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation comprises fewer than 65,000 words, including appendices, bibliography, footnotes, tables, and equations, and contains fewer than 150 figures.

Lim Wei Xiang

September 2023

Acknowledgements

I would like to express my deepest gratitude to my supervisors Dr. Zhiyuan Chen, Dr. Amr Ahmed, and Dr. Mohammad Aazam, for their unwavering guidance and support. The completion of this work would not have been possible without her relentless efforts and patience, which have been instrumental in making this work a reality.

I extend my heartfelt thanks to my parents, family members, and friends for their constant and unwavering love, care, and support, which have allowed me to surmount challenges throughout the years. Words cannot adequately convey my appreciation. This work is dedicated to all of you.

Abstract

The field of medical image analysis has seen remarkable progress, primarily propelled by advancements in Convolutional Neural Networks (CNNs). These sophisticated neural networks have demonstrated their value across various applications, including the detection of lung cancer and segmentation of brain tumors. Such instances highlight CNNs' potential to significantly enhance the accuracy of medical image analysis, suggesting their role in improving clinical outcomes. Despite these advancements, the inherently opaque nature of CNNs often categorizes them as "black boxes", raising concerns about their transparency.

Explainable Artificial Intelligence (XAI) represents a crucial approach to address the transparency concerns associated with Convolutional Neural Networks (CNNs) by elucidating the decision-making processes within these complex computational models. Gradient-based visual explanation methods have emerged as the most prevalent XAI techniques in medical image analysis. However, the validity of these gradient-based methods—where validity refers to the assessment of the explanation's accuracy provided by the method and its alignment with the expectations of end-users—remains in question. To address this gap, the first contribution of this research focused on quantitatively validating the heatmaps generated by gradient-based visual explanations, producing empirical scores that attest to their reliability. This validation is paramount for ensuring that the interpretations offered by XAI methods are not only technically accurate but also meaningful and trustworthy to end-users, including medical professionals.

Building upon the foundation of enhancing interpretability through valid explanations, the research recognizes that the performance of CNNs in medical image analysis is also contingent upon the quality of input data, necessitated by effective image preprocessing. Image characteristics and quality significantly impact algorithm performance and prediction accuracy. Traditional preprocessing methods, while widely used, fall short in addressing the pixel-related challenges inherent in digital image processing, such as the discretization of

image information and the computational burdens imposed by large images with high pixel counts. Bridging this gap, the second contribution introduces a novel superpixel-based image preprocessing method. This method not only mitigated computational demands but also optimized the image data for improved classification outcomes. In doing so, it complemented the first contribution by ensuring that the CNNs are operating on high-quality, efficiently processed images, thereby maximizing the utility and applicability of the validated explainable AI techniques. Together, these contributions represent a comprehensive approach to enhancing the transparency, interpretability, and efficiency of CNNs in medical image analysis, addressing both the 'why' and the 'how' of their decision-making processes.

In conclusion, the initial contribution demonstrated that the proposed quantitative evaluation method, focused on saliency evaluation, successfully quantitatively assess the created heatmaps by generating empirical scores that verify their validity. This approach allowed for the explicit validation of pixels that play a significant role in influencing the classification outcome. Additionally, the quality and features of images markedly affect the performance of algorithms and the precision of their predictions. Hence, the proposed Simple Linear Iterative Clustering (SLIC) augmented with a Gaussian smoothing filter — SLIC-G, served as an image preprocessing technique. It succeeded in segmenting images into regions that convey significant information with a substantially reduced dataset, rather than relying on all the pixels within an image. This method also improved image clarity by removing noise. Experimental outcomes indicated that employing SLIC-G for image preprocessing improved the results of classification.

Contents

List of Publications i

Declaration ii

Acknowledgements iii

Abstract iv

1 Introduction 1

1.1 Background 1

1.2 Problem Statement 5

1.3 Research Aims & Objectives 8

1.4 Scope & Limitation 9

1.5 Contributions 10

1.6 Organization of the Thesis 11

2 Literature Review 13

2.1 Introduction 13

2.2 Convolutional Neural Networks 14

2.2.1 Building Blocks of CNN 14

2.2.2 CNN Image Classification Pipeline 18

2.2.3 CNNs Adoption in Medical Image Classification 23

2.3 Gradient-based Visual Explanation 29

2.3.1 Sensitivity Analysis (SA) 30

2.3.2 Gradient x Input (GI) 30

2.3.3 Class Activation Map (CAM) 30

2.3.4 Gradient-based Class Activation Map (Grad-CAM) . . . 32

2.3.5 Guided Gradient-based Class Activation Map (Guided Grad-CAM) . . . 33
2.3.6 Integrated Gradient (IG) . . . 35
2.3.7 Visual Explanation Adoption in Diabetic Retinopathy Image Classification . . . 36
2.3.8 Strengths & Weaknesses . . . 38
2.4 Data-centric AI . . . 41
2.5 Diabetes & Diabetic Retinopathy . . . 43

3 Methodology 48

3.1 Introduction . . . 48
3.2 Machine Specifications and Software Frameworks . . . 49
3.3 DC-Check . . . 50
3.4 Data Acquisition . . . 52
3.5 Image Preprocessing . . . 53
3.6 Data Augmentation . . . 56
3.7 Model Building and Training . . . 56
3.8 Model Performance Evaluation . . . 58
3.8.1 F1 Score, Precision, and Recall . . . 58
3.8.2 Accuracy . . . 59
3.8.3 Confusion Matrix . . . 60
3.8.4 Receiver Operator Characteristics (ROC) Curve . . . 60
3.8.5 Geometric Mean . . . 61
3.8.6 Jaccard Similarity . . . 62
3.9 Gradient-based Visual Explanation . . . 62

4 Quantitative Saliency-Oriented Evaluation Metric 64

4.1 Introduction . . . 64
4.2 Methods & Materials . . . 65
4.2.1 Datasets . . . 65
4.2.2 Data Preprocessing . . . 67

4.2.3 Data Augmentation 69
4.2.4 Implementation Details 69
4.3 Result 73
4.3.1 Model Performance 73
4.3.2 Quantitative Saliency-oriented Evaluation Metric Performance 74
5 SLIC-G — An Optimised Image Preprocessing Method 80
5.1 Introduction 80
5.2 Methods & Materials 81
5.2.1 Datasets 81
5.2.2 Data Preprocessing 82
5.2.3 Data Augmentation 84
5.2.4 Implementation Details 86
5.3 Result 87
5.3.1 Model Performance 87
5.3.2 Quantitative Saliency-oriented Evaluation Metric Performance 91
6 General Discussion 94
6.1 Research Summary 94
6.2 Validity of Visual Explanation Methods 95
6.3 Superpixel-based Image Preprocessing 97
6.4 Limitations & Future Work 100
7 Conclusion 103
References 105
Appendices 121
A List of Abbreviations 121
B DC-Check Checklist A 122
C DC-Check Checklist B 129

List of Tables

2.1 Representative papers using CNN for classification. 28

2.2 Representative papers that provide visual explanation in Diabetic Retinopathy image analysis. 39

2.3 Strengths and weaknesses of gradient-based visual explanation methods. Strengths are depicted as +, weaknesses are -. 40

3.1 Machine Specifications 49

3.2 Software Frameworks 50

3.3 Overall data distribution of three different Diabetic Retinopathy (DR) datasets. 53

3.4 Distribution of classes across different tasks for the used datasets. 53

3.5 Formulation of gradient-based visual explanation methods. . . . 63

4.1 Data distribution of Kaggle Diabetic Retinopathy Dataset 2015 (KDRD2015) 66

4.2 The different used of image preprocessing methods for DR image classification. 68

4.3 The model hyperparameters that best fit the model classification. 71

4.4 Comparative analysis of classification results using various proposed image preprocessing methods. 73

5.1 Data distribution of the APTOS 2019 (APTOS2019). 82

5.2 The model hyperparameters that best fit the model classification. 87

5.3 The comparative analysis between the proposed method and other state-of-the-art pre-trained CNNs. 89

List of Figures

2.1 A representation of a CNN(LeNet5) [1]. 15

2.2 Convolution layer [2] 16

2.3 Max pooling illustration. 17

2.4 Global average pooling illustration. 17

2.5 A CNN image classification pipeline with different stages. 22

2.6 A fundus images with different types of diabetic retinopathy lesions. 46

3.1 Methodology Overview 49

3.2 DC-Check [3] 51

3.3 The architectural representation of DenseNet [4]. 58

4.1 Examples of pixel-level annotated ground truth images of diabetic lesions (From top left to bottom right: Microaneurysms (MA), Haemorrhages (HE), Hard Exudates (EX), Soft Exudates (SE)) 66

4.2 Before and after preprocessing fundus photgraphy images (Left to Right: Original, CLAHE, CLAHE-G Channel, CLAHE-L Channel, and Gaussian-filtered Weighted Contrast) 69

4.3 A workflow of the proposed quantitative saliency-oriented evaluation metric. 73

4.4 Training and testing loss of the fine-tuned DenseNet model using the Gaussian weight contrast images. 75

4.5 Training and testing accuracy of the fine-tuned DenseNet model using the Gaussian weight contrast images. 75
4.6 The confusion matrix of the fine-tuned DenseNet model using the Gaussian weight contrast images. 76
4.7 AUC ROC curve for non-referable and referable classes using the Gaussian weighted contrast images. 76
4.8 Example results of prediction score with heatmaps and quantitative saliency-oriented evaluation score. 78
4.8 Example results of prediction score with heatmaps and quantitative saliency-oriented evaluation score. 79

5.1 A workflow of the proposed image pre-processing module. 84
5.2 A sample retinal image that applied with the proposed hybrid image processing image method. (From top left to bottom right: Original retinal image, SLIC segmented retinal image, Gaussian-filtered CIELAB retinal image, SLIC-G retinal image). 85
5.3 The training and testing loss and accuracy curve on SLIC-G APTOS 2019 dataset. 89
5.4 Training and testing loss curve and testing accuracy curve on SLIC-G APTOS 2019 dataset. 89
5.5 Training and testing loss curve and testing accuracy curve on SLIC-G APTOS 2019 dataset. 90
5.6 Example results of prediction score with heatmaps and quantitative saliency-oriented evaluation score. 92
5.6 Example results of prediction score with heatmaps and quantitative saliency-oriented evaluation score. 93

B.1 Page 1 of DC-Check checklist for experiments in Chapter 4. . . 123
B.2 Page 2 of DC-Check checklist for experiments in Chapter 4. . . 124
B.3 Page 3 of DC-Check checklist for experiments in Chapter 4. . . 125
B.4 Page 4 of DC-Check checklist for experiments in Chapter 4. . . 126

B.5 Page 5 of DC-Check checklist for experiments in Chapter 4. . . 127
B.6 Page 6 of DC-Check checklist for experiments in Chapter 4. . . 128

C.1 Page 1 of DC-Check checklist for experiments in Chapter 5. . . 130
C.2 Page 2 of DC-Check checklist for experiments in Chapter 5. . . 131
C.3 Page 3 of DC-Check checklist for experiments in Chapter 5. . . 132
C.4 Page 4 of DC-Check checklist for experiments in Chapter 5. . . 133
C.5 Page 5 of DC-Check checklist for experiments in Chapter 5. . . 134
C.6 Page 6 of DC-Check checklist for experiments in Chapter 5. . . 135

Chapter 1

Introduction

1.1 Background

The advancement of Artificial Intelligence (AI) has been substantially influenced by the emergence of deep learning techniques. When contrasted with conventional machine learning algorithms like Decision Trees or Support Vector Machines, these deep learning approaches have demonstrated remarkable gains in an array of predictive tasks [5–7].

In the field of medical image analysis, deep learning algorithms, particularly Convolutional Neural Networks (CNNs), serve as successful machine learning tools for autonomously extracting and assimilating critical data, also known as featrues. The utility of applications can span a variety of domains, encompassing segmentation [8], detection of abnormalities [9], and classification of diseases [10].

Despite their efficacy in medical image analysis, CNNs often do not offer transparent insights into their decision-making procedures or final outcomes. This obscurity often results in both the engineers responsible for their creation and the end-users interpreting them as "black-box" entities [11–16]. These enigmatic models, which are data-driven, elude even the technical experts who cannot fully decode their internal mechanics or clarify the rationale for specific results [17, 18]. Such opacity can undermine trust, especially in medical

settings where the stakes are high, and may lead to the systems being disregarded [11–16]. Moreover, this lack of transparency has masked the emergence of algorithmic biases generated from flawed processes that unfairly disadvantage certain demographic groups. Such biases have been observed to result in widespread discrimination across various sectors, including employment, advertising, criminal justice, and healthcare [19].

Explainable Artificial Intelligence (XAI) constitutes a collection of techniques and processes designed to make the outcomes and results generated by machine learning algorithms understandable and trustworthy for human users. This interpretability extends to deep learning architectures like CNNs, enabling these models to articulate their operations in a manner comprehensible to humans.

Black-box models, like CNNs are being utilized in the medical settings for a range of prediction tasks. The opaque nature of these models can sometimes result in decisions that are not only unacceptable but could also give rise to legal complications, thus presenting a novel challenge for lawmakers. Regulations such as the European Union (EU)'s General Data Protection Regulation (GDPR) underscore the importance of XAI from a legal standpoint. Such regulations introduce the concept of a "right to explanation," allowing individuals to seek clarity on algorithmic decisions that have a significant impact on their lives [20].

From the standpoint of model development, various elements such as insufficient training data, data bias, outliers, adversarial examples, and model overfitting can contribute to the generation of erroneous outcomes in black-box models like CNNs. Understanding the internal workings and decision-making rationales of these black-box models becomes critical, especially when their decisions have far-reaching implications on human lives. The primary objectives of XAI are to dissect, troubleshoot, and refine these opaque models, thereby enhancing their robustness, bolstering safety and user trust, and mitigating or eliminating issues like incorrect behavior, bias, and discrimination [21]. Addi-

tionally, in scenarios where multiple models exhibit comparable performance metrics, XAI can offer insights into the features that drive the decision-making process, thereby aiding in model selection [22, 23]. XAI can also act as an auxiliary function in cases where the algorithm may be tuned towards a partial or incomplete objective, offering a more comprehensive understanding of the model's performance [24].

Additionally, from the perspective of human end-users, deep learning architectures such as CNNs are vulnerable to deception. Minor alterations to input data, such as the addition of noise, can lead CNNs astray, causing them to incorrectly identify irrelevant patterns as significant features, thereby yielding erroneous predictions [25, 26]. Such observations can undermine trust in these black-box models [16]. These models have the potential to make unjust decisions, particularly if they are trained on datasets that encapsulate human biases [16]. Consequently, enhancing the explainability or interpretability of these black-box models can bolster trust, as it allows for an understanding of the decision-making mechanisms and verifies whether the model functions as intended rather than simply executing based on its training [16, 27]. Additionally, the quest for fairness in decision-making, which cannot be fully captured by error metrics, often necessitates the development of models that are interpretable in nature [15].

Researchers in medical image analysis are increasingly using XAI to explain the results of their deep learning-based algorithms like CNNs [28–30]. In spite of recent advancement in XAI, several pressing challenges remain to be tackled. There is a urgent need for the establishment of formalized evaluation metrics [12, 13, 18, 31, 32]. The absence of established ground truth for explanations in many cases serves as the primary obstacle to executing comprehensive evaluations [31, 33, 34]. Other contributing factors, including the expenses associated with achieving predictive accuracy and computational efficiency, keep the question of what constitutes the best explanation unanswered [35]. To date, various evaluation criteria have been introduced, covering aspects

like reliability, credibility, utility, validity, completeness, compactness, comprehensibility, human-centricity, correctness or fidelity to the original model, complexity, and generalizability [33]. Recently, the work in [36] conducted a study to assess the validity of various visual explanation techniques in medical imaging, specifically focusing on chest X-rays. However, the study is limited to chest X-rays, indicating that additional research is needed to evaluate the efficacy of these visual explanation methods in other imaging modalities and anatomical regions.

Another significant obstacle in XAI is the dependency of the system's explanations on the quality of the data used for training [37–39]. Factors such as data bias, incompleteness, and inaccuracy can compromise the integrity of the data. The repercussions of utilizing subpar data are manifest not just in the model performance but also in the system's ability to provide reliable explanations [40]. For instance, a lung cancer risk prediction model trained on American patients data may yield imprecise estimates for a resident of Delhi, owing to differing environmental pollution levels. The quality of data, thus, is context-sensitive; what is deemed high-quality in one scenario may be insufficient in another [39]. Besides reducing accuracy, using flawed data could also result in unfair decision-making and impair the system's explainability. In light of this, it is recommended to exercise caution regarding the manner in which data is gathered, along with understanding any constraints inherent to the collected dataset [41]. There is an active research that prioritize data quality in the field of machine learning, known as data-centric AI [42–44]. The primary aim of this research is to focus on the collection of large volumes of high-quality data. A recent work [3] proposed a practical checklist, known as DC-Check, to guide data-centric approaches at different phases of the machine learning pipeline. Similarly, in [45], a dataset-specific checklist aimed at two major stakeholders: those who create datasets and those who utilize them. Both checklists aimed to increase transparency and accountability in the machine learning community, minimize the inadvertent societal biases in machine

learning algorithms, improve the replicability of machine learning studies, and assist researchers and practitioners in choosing datasets that are most appropriate for their specific tasks.

The landscape of AI, particularly in medical image analysis, has been profoundly shaped by deep learning algorithms like CNNs. While these algorithms offer exceptional predictive performance, their "black-box" nature raises serious concerns about transparency, trust, and fairness. Despite this, XAI is emerging as a viable solution to demystify these complex models, enhancing both their robustness and societal acceptance. A key dimension to the successful implementation of XAI is the quality of training data. Flawed or biased data not only compromises model performance but also inhibits the system's ability to produce reliable explanations. Regulatory frameworks like the EU's GDPR have started to underscore the necessity of explainability in decision-making algorithms, emphasizing its legal implications. However, the field still faces major hurdles such as the lack of formalized evaluation metrics and the inherent complexity of establishing what constitutes an 'optimal' explanation. Concurrently, the machine learning community is putting a spotlight on data-centric AI, focusing on the acquisition of high-quality data as a means to improve both model performance and explainability. As we navigate these challenges, the synthesis of high-quality data, advanced modeling techniques, and robust evaluation metrics will be essential in fulfilling the promises of both AI's predictive power and societal accountability.

1.2 Problem Statement

In medical image analysis, visual explanation serves as the predominant type of XAI [12, 13, 46]. This type of XAI is often referred to as post-hoc explanation, given that it offers insights (i.e. attribution maps or heatmaps) into the predictions subsequent to the training of the deep learning models like Convolutional Neural Networks (CNNs) [12, 13, 46]. Numerous studies have

investigated these heatmap-based methods in different medical anatomical locations such as bladder, brain, breasts, and cardiovascular [13].

However, these studies have only focused on the applicability and effectiveness of the visual explanation methods. In spite of the effectiveness of visual explanation methods in explaining the CNNs decision-making process by emphasizing relevant image areas for diagnosis, there is a lack of establishment of formalized evaluation metrics to evaluate the generated heatmaps [12, 13, 18, 31, 32]. In other words, there is a lack of validation of the heatmaps generated by the visual explanation methods. The validity, in this context, refers to evaluating whether the explanation provided by the method is accurate and aligns with end-user expectations. Consequently, the validity of visual explanations has not been rigorously investigated, and previous studies have focused solely on improving sharpness of the heatmaps [47, 48].

In summary, while visual explanation methods are increasingly recognized as a vital component in the field of medical image analysis to elucidate the decision-making processes of deep learning models like CNNs, a critical research gap persists in the literature concerning the evaluation of these visual tools [12, 13, 46]. Despite advances in applying these methods across various anatomical areas and their proven effectiveness in emphasizing diagnostically relevant regions, existing research has predominantly focused on the applicability and resolution of these visualizations [13, 47, 48]. A glaring omission lies in the lack of formalized metrics to assess the validity of generated heatmaps, leaving questions unanswered about the accuracy and alignment of these visual explanations with end-user expectations [12, 13, 18, 31, 32]. This underscores an urgent need for an examination of the validity of visual explanations to ensure both their scientific rigor and practical utility in clinical settings.

Additionally, models such as CNNs rely heavily on the quality of the training data for their operation and the explanations they generate [37–39]. Issues like biased data, incomplete datasets, and inaccurate information can undermine not only the model's performance.

In real-world scenarios, medical images display a significant degree of variability, noise, and temporal fluctuations [49]. These variations can stem from various factors, including disparate imaging protocols, patient cohorts, and environmental variables, among others. Prior research has indicated that image preprocessing plays a vital role in mitigating noisy artifacts and enhancing image classification performance [50–52]. Numerous image preprocessing methods, such as Contrast Limited Adaptive Histogram Equalization (CLAHE), Gaussian filtering, green channel extraction, and grayscale conversion, have been employed to improve image quality [53, 54]. However, these pixel-based image preprocessing methods suffer from pixel-related issues inherent in digital image preprocessing. Specifically, the two primary challenges associated with pixel preprocessing are the discretization of image information and the computational infeasibility of many algorithms in handling large image sizes with a high number of pixels [55, 56].

In addition, the machine learning community has primarily focused on optimizing algorithms for state-of-the-art CNNs on benchmark datasets while neglecting various factors relevant in real-world scenarios, such as data curation, data quality assessment and cleaning, and subgroup characterization [3]. There is a lack of research that focus on data-centric methodology [3, 45]. There is a need to standardize methodology for articulating the design aspects of data-centric machine learning pipelines.

In conclusion, the effectiveness and trustworthiness of CNNs heavily reliant on the quality of training data. Existing research has focused predominantly on optimizing algorithms, often neglecting critical considerations pertinent to real-world applications such as data quality, curation, and subgroup analysis [3, 37–39]. Furthermore, real-world medical imaging data introduce additional complexities, including variability, noise, and temporal fluctuations that can significantly impact model performance [49]. While image processing techniques have been explored to mitigate some of these issues, they come with their own set of challenges, including computational inefficiency and the dis-

cretization of image information [50–52, 55, 56]. Compounding these challenges is the absence of a standardized process for designing data-centric machine learning pipelines, a void that exacerbates the difficulty of engaging practically with data-centric AI systems. As a result, a more holistic approach that integrates algorithmic improvements with robust data quality measures and standardized design methodologies is imperative for the advancement and trustworthiness of machine learning applications in real-world scenarios.

1.3 Research Aims & Objectives

The overall aim of this research is to conduct medical image classification, assess the model performance, and the trustworthiness of its explanations, in a holistic approach that integrates algorithmic improvements with robust image quality measures and standardized design methodologies. The specific objectives include:

1. To develop a comprehensive end-to-end deep learning framework for the Diabetic Retinopathy (DR) image classification, encompassing both the model training and inference stages, in order to facilitate a robust and efficient diagnostic process.

   RQ1: What is the comparative performance of the proposed pre-trained DenseNet architecture in categorizing Diabetic Retinopathy fundus images relative to other leading Convolutional Neural Networks?

2. To develop an innovative image preprocessing method aimed at improving image quality and augmenting model classification accuracy within the context of binary and multiclass DR image classification.

   RQ2: How do the proposed image preprocessing techniques impact the performance of a pre-trained DenseNet model?

3. To develop an innovative quantitative saliency-oriented evaluation metric designed to evaluate the subjective heatmaps produced by gradient-based visual explanation method, thereby facilitating a comprehensive understanding of their effectiveness and reliability.

   RQ3: How well is the suggested quantitative evaluation method capable of quantifying the subjective saliency heatmaps?

   RQ4: What is the correlation between subjective heatmaps and numerical metrics?

4. To assess the efficacy of the proposed methodologies within the context of a data-driven paradigm, a rigorous and systematic evaluation will be conducted.

1.4 Scope & Limitation

The scope and limitations of this research includes the following:

1. This research predominantly centers on Diabetic Retinopathy (DR) as the case study of interest. However, the methodologies and findings demonstrated herein hold the potential to be generalized and applied to a broader array of biomedical contexts and case studies.

2. The datasets employed in this study were publicly accessible and widely recognized within the research community. These datasets can be obtained from open platforms like Kaggle [57] or Google's Datasets Search [58]. Importantly, these datasets exhibit a high level of class imbalance, aligning well with our research objective of improving dataset quality.

3. Pre-trained Convolutional Neural Networks (CNNs) were used in this research. Specifically, pre-trained DenseNet model was adopted due to the lack of utilization in DR image classification [59] and its unique architectural design [4]. The utilization of pre-trained CNNs was motivated by

their ready-to-use nature and the opportunity to apply transfer learning technique for efficient fine-tuning to suit our specific application requirements.

4. This study focused on evaluating the importance of each input feature to the CNN's predictions, a method often termed as gradient-based visual explanation. It does not encompass the investigation of how individual neurons within a specific layer contribute to the overall output of the CNN, termed as layer attribution. Additionally, it does not explore how each input feature affects the activation of specific hidden neurons, a concept known as neuron attribution.

1.5 Contributions

The contributions of this thesis based on the aims and objectives are as follows:

1. Diabetic Retinopathy (DR) appears in various stages, each characterized by unique lesions for diagnostic evaluation. Consequently, this research explored the applicability of DenseNet pre-trained Convolutional Neural Network (CNN) model for the task of multi-class DR image classification. The use of transfer learning was employed for effective adaptation to our specialized application needs. The experimental outcomes indicated that a fine-tuned DenseNet architecture performed efficiently in the classification of multi-class DR images, as compared to other pre-trained CNNs.

2. The performance and explanatory capability of CNNs are highly dependent on the quality of the training data. Previous research has employed a variety of image processing techniques to ameliorate image quality [53, 54]. Yet, these pixel-based methods present inherent challenges such as computational burden and image information discretization. In contrast, this research proposed a novel image pre-processing technique,

named SLIC-G. This technique integrates two key operations: Simple Linear Iterative Clustering (SLIC) for superpixel segmentation and Gaussian smoothing operation for image sharpening. The primary objective of SLIC-G partitions the DR images into meaningful regions, with significantly less data compared to using all pixels in an image and enhances the quality of fundus images. Experimental outcomes demonstrated that training datasets pre-processed with the proposed SLIC-G method lead to enhanced model accuracy.

3. A cost-effective and efficient quantitative evaluation metric was introduced to evaluate heatmaps produced by gradient-based visual explanation methods. Employing the Jaccard similarity coefficient—a recognized measure for segmentation evaluation—in conjunction with pixel-level ground truth images of DR lesions, our method provides a numerical score for heatmap validity. This score substantiates the heatmap's accuracy in highlighting regions of importance. Hence, this evaluation strategy affirms that heatmap analysis is not merely a subjective exercise, but is also backed by empirical evidence.

1.6 Organization of the Thesis

The structure of this thesis is organized as follows: Chapter 2 offers a review of the literature, encompassing cutting-edge gradient-based visual explanation techniques as well as core deep learning algorithms like Convolutional Neural Networks (CNNs) and SLIC superpixel segmentation. Chapter 3 details the methodologies used for formulating the new visual explanation evaluation metric and the SLIC-G image enhancement technique. Chapter 4 elucidates the implementation specifics and outcomes of the newly proposed visual explanation evaluation measure. In Chapter 5, the execution details and results of the innovative SLIC-G image processing method are presented. Chapter 6 synthesizes the thesis' key findings, while also providing a broader discussion,

identifying limitations, and suggesting avenues for future research. Finally, Chapter 7 concludes the thesis.

Chapter 2

Literature Review

2.1 Introduction

The present chapter represents a foundation, meticulously designed to achieve the research objectives and the goal of this thesis. Initially, we delve into the fundamentals of Convolutional Neural Networks (CNNs), encompassing a comprehensive understanding of convolution layers, pooling layers, and fully-connected layers. The chapter's scope also encompasses an extensive explanation of the necessary steps to architect any robust deep learning image classification pipeline. Furthermore, prior studies that have leveraged CNNs for medical image classification were also reviewed.

In the subsequent section pivot towards an in-depth exploration of gradient-based visual explanation methods. Every method has been elucidated at its core, laying bare its fundamental mechanisms. A careful review of past research using gradient-based visual explanation methods in Diabetic Retinopathy (DR), a field highly relevant to this research, was also presented, strengthening our position in the scholarly conversation.

Additionally, an introduction to a data-centric AI approach. This approach has substantial implications for the current research and has been introduced succinctly yet with clarity.

The chapter concluded with a synthesis of the critical literature concern-

ing diabetes and diabetic retinopathy, marking the end of this comprehensive compendium. By interweaving such diverse yet interconnected domains of knowledge, this chapter paves the way towards an academically rigorous, technologically informed, and clinically relevant exploration in the remaining sections of this thesis.

2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) represent a type of artificial neural network that exploits convolutional layers to extract informative features from input data, notably images. Traditional methods for object recognition and feature extraction in images were time-consuming. However, CNNs provide a scalable solution for image classification and object recognition, utilizing principles from linear algebra and matrix multiplication to identify image patterns.

In 1990, LeCun et al. [60] published the seminal paper establishing the modern framework of CNN, and later improved it in [61]. They developed a multi-layer artificial neural network called LeNet-5 which could classify handwritten digits. Like other neural networks, LeNet-5 has multiple layers and can be trained with the backpropagation algorithm [62]. It can obtain effective representations of the original image, which makes it possible to recognize visual patterns directly from raw pixels with little-to-none preprocessing.

This section highlights the essential components of CNNs that enable their utility in medical image analysis tasks.

2.2.1 Building Blocks of CNN

A CNN is comprised of various components, including convolutional layers, pooling layers, and fully connected layers. The CNN architecture typically involves a repetition of a sequence of convolutional layers and a pooling layer, which is subsequently followed by one or more fully connected layers. As the network grows in complexity with each layer, it gradually captures larger and

more sophisticated components of the image. Initially, the lower layers extract elementary features, such as edges and colors. Subsequently, higher layers of the CNN can detect increasingly larger and more complex elements or shapes of the object of interest, ultimately leading to its identification.

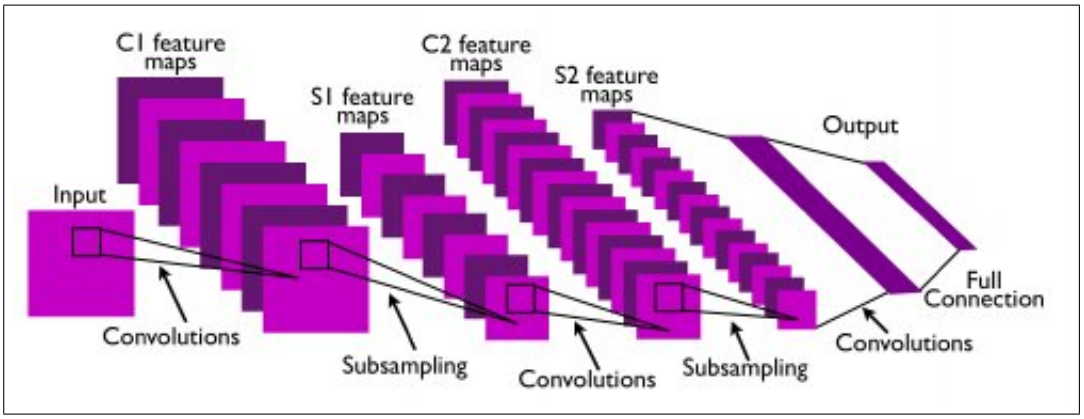


Figure 2.1: *A representation of a CNN(LeNet5) [1].*

Convolution Layer

The convolutional layer constitutes the fundamental unit of the CNN architecture. It is imperative to acknowledge that the layer's parameters and channels are constituted by a collection of learnable channels or neurons in the CNN. These neurons have a small receptive field. During the forward pass, each individual neuron traverses the dimensions of the input, subsequently computing the dot product of the filter pixels and the input pixels. The output of this computation is a two-dimensional (2D) feature map. Through this mechanism, the CNN learns channels, whereby it detects specific types of features at spatial locations within the feature map. Each neuron computes convolutions with small sections in line with LeCun et al. (2010), as shown in Eq. 2.1, where $y_i \, \epsilon \, Y, i = 1, 2, ...D$. The depth of the convolutional layer, denoted by D, is an important parameter. The index i refers to the index of the output feature map while index j refers to the index within a given filter or kernel and corresponds to the specific connection or weight applied to the input. Each filter, denoted by Wij, is a 3D matrix with dimensions [F $\times$ F $\times$ CX], where

F represents the receptive field size and CX represents the input feature-map depth. In the feed-forward process, each neuron in the next layer is connected to a specific receptive field in the previous layer, forming a weight vector. The receptive field extracts local features from the input image. For instance, if the receptive field size is three pixels and the feature-map input is an RGB image with dimensions [32 × 32 × 3], then the size of the filter is [3,3,3]. The receptive fields of neurons in the next layer are illustrated in Figure 2.2.

$$y_i = b_i + \sum_{x_i \epsilon x} W_{ij} \times x_i \tag{2.1}$$

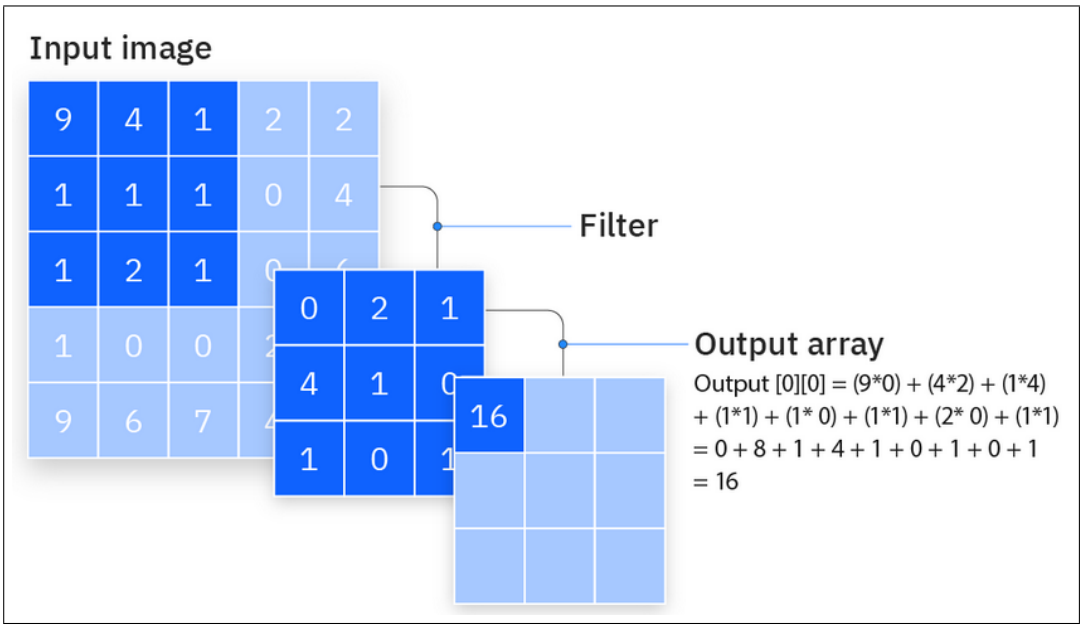


Figure 2.2: *Convolution layer [2]*

Pooling Layer

In the CNN architecture, the pooling layer plays a crucial role in downsampling, which reduces the input parameters. This process is achieved by sweeping a filter across the entire input, similar to the convolutional layer. However, in contrast to the convolutional layer, the pooling operation does not have any weights. Instead, the kernel applies an aggregation function to the values within the receptive field and generates the output array. While there are different variations of pooling, two primary types of pooling are:

- **Max Pooling** [63]: The filter traverses the input and picks the pixel with the highest value to transfer to the output array. It should be noted that this technique is frequently favored over average pooling.

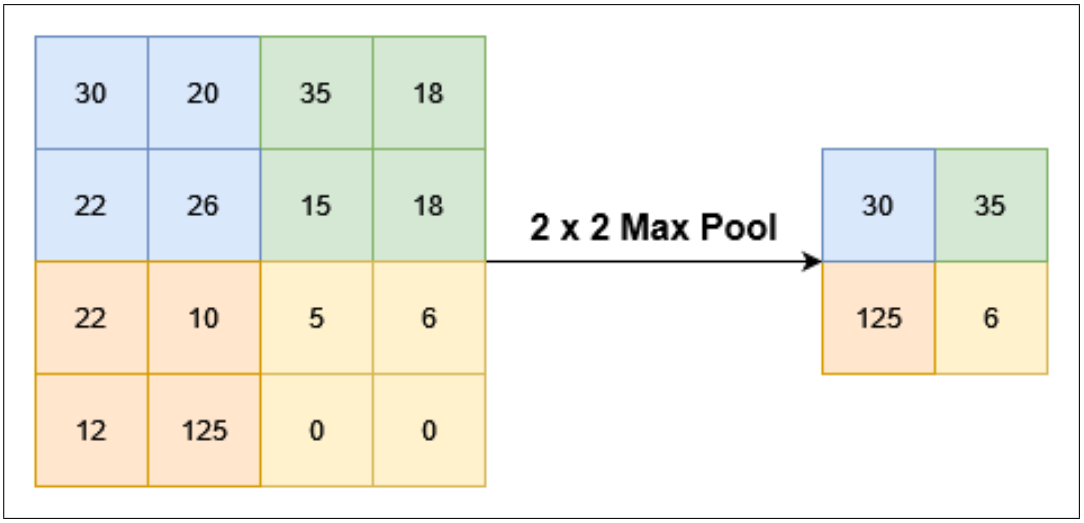


Figure 2.3: *Max pooling illustration.*

- **Global Average Pooling** [64]: The filter traverses the input, computing the mean value of the pixels within the receptive field, which is subsequently transferred to the output array.

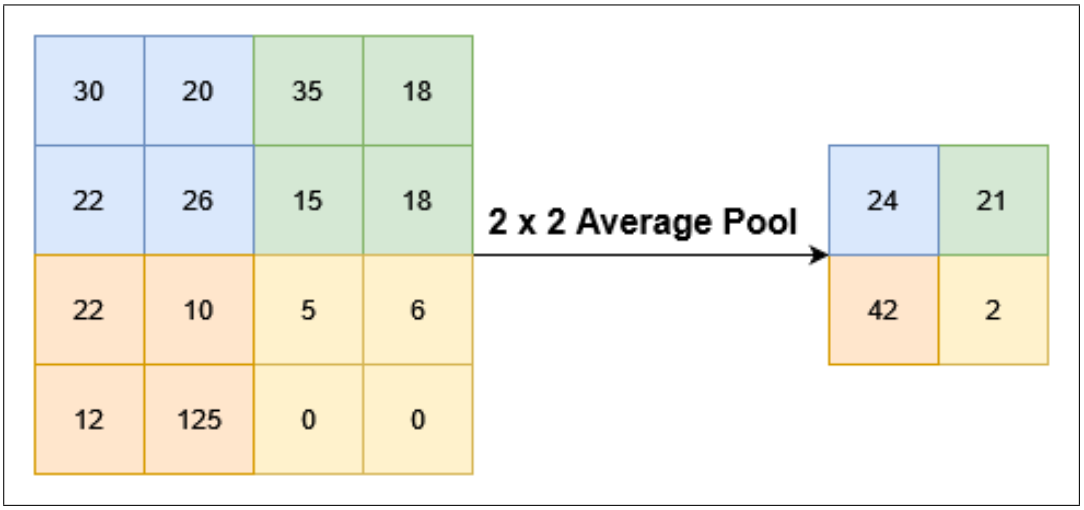


Figure 2.4: *Global average pooling illustration.*

Fully-Connected Layer

In this layer, the final convolution or pooling layer's output feature maps are transformed into a one-dimensional (1D) array of numbers and connected to

one or more fully connected layers, also known as dense layers. This dense layer aims to perform high-level reasoning [65, 66]. Each input is connected to every output by a learnable weight. Once the convolution and pooling layers have extracted and downsampled the features, a subset of fully connected layers maps them to the final outputs of the network, such as the probabilities for each class in classification tasks. The final fully connected layer usually has the same number of output nodes as the number of classes, and each fully connected layer is followed by a non-linear function, such as ReLU [67]. The activation function applied to the last fully connected layer is usually different from the others and needs to be selected according to each task. For a multi-class classification task, an appropriate activation function is the softmax function, which normalizes output real values from the last fully connected layer to target class probabilities. Each value ranges between 0 and 1, and all values sum to 1.

2.2.2 CNN Image Classification Pipeline

Achieving success in a deep learning endeavor demands more than mere expertise in constructing sophisticated neural networks such as Convolutional Neural Networks (CNNs). It is also crucial to comprehend and effectively execute various phases within the deep learning workflow for practical applications, such as medical image analysis. This section provides an overview of the CNN image classification process, as illustrated in Figure 2.5, encompassing stages like data acquisition, data preprocessing, data augmentation, model building and training, model performance evaluation, and classification prediction.

As illustrated in Figure 2.5, the upper section highlights the training phase, which incorporates stages like data acquisition, data preprocessing, data augmentation, model building and training, model performance evaluation. Hyperparameter tuning might be necessary contingent on the evaluation outcomes. Conversely, the lower section explicates the steps involved in assessing

the model's proficiency through the classification of new, formerly unseen test samples. Subsequent sections offer a detailed explanation of each component within this pipeline.

Data Acquisition

The process of image classification initiates with data acquisition. Within the scope of supervised image classification, the collected images are endowed with relevant class labels. These image datasets may be obtained from publicly accessible databases like Kaggle [57], IEEE DataPort [68], or Google Dataset Search [58]. The compiled image data are categorized as 'raw' since they contain undesirable elements such as noise, variance, and time-based fluctuations, which can adversely affect the performance of the model. These raw datasets are then subjected to further treatment in the ensuing phase, outlined in Section 2.2.2, through various image preprocessing techniques.

Data Preprocessing

The acquired image data may not be directly suitable for training machine learning models. These images often contain extraneous elements such as noise, contrast variations, and temporal inconsistencies that may degrade the performance of deep learning algorithms. The primary goal of the model is to classify images based on key characteristics present within them. Therefore, it is beneficial to alleviate the impact of such artifacts. This remediation is achievable through the application of image preprocessing techniques, designed to reduce noise and color disparities, thereby potentially boosting the model's classification capabilities [49].

Data Augmentation

Following the image preprocessing stage, the next step in the training phase involves data augmentation (see Figure 2.5). Data augmentation serves as a form of regularization, particularly significant in deep learning and Convolutional

Neural Networks (CNNs). The principal aim of regularization is to enhance a model's ability to generalize well on new, unseen data. This is accomplished by introducing specific constraints or penalties that discourage the model from over-adapting to the training dataset, thereby diminishing overfitting. In the context of data augmentation, the original dataset undergoes various transformations to produce new, but correlated, data points. This increases both the variability and size of the training dataset, equipping the model with a more expansive and resilient grasp of the data distribution [69, 70]. This is performed by applying basic image manipulations like rotation, mirroring, zooming, and horizontal or vertical translations, among other techniques. It should be emphasized that data augmentation is not applied during the model's validation phase. The evaluation must exclusively involve data that has been preprocessed, in order to provide an accurate measure of the trained model's true performance capabilities.

Model Building and Training

Once the gathered image data has been preprocessed and augmented, the next step entails the construction and training of deep learning models. A wide array of Convolutional Neural Networks (CNNs) is available, making it essential to approach this phase with a problem-specific mindset. Given the often limited availability of domain-specific image data, such as medical images, it is advisable to leverage pre-trained CNNs. Utilizing transfer learning techniques can efficiently tailor these pre-trained models to meet the unique demands of the specific application.

Model Performance Evaluation

During the training process, model performance is evaluated against a validation set, serving as an indicator of how well the model is likely to perform on new, unseen data. Choice of evaluation metric is critical, especially when dealing with imbalanced datasets. In such scenarios, accuracy may be mislead-

ing, and alternative metrics like precision, recall, F1-score, Receiver Operator Characteristics (ROC) Curve, and confusion matrices can provide a more nuanced understanding of model performance.

Hyperparameter Tuning

Hyperparameter tuning is an iterative process that requires experimentation with various learning rates, batch sizes, architectures, and regularization techniques. Monitoring the loss and performance metrics can offer insights into areas where the model may be faltering. Signs of unstable learning may necessitate reducing the learning rate or increasing the batch size. If there is a performance gap between the training and validation sets, overfitting is likely, warranting the introduction of regularization methods like dropout or early stopping. Conversely, poor performance on both sets may indicate underfitting, suggesting the need for a more complex model or adjustment of the learning rate.

In transfer learning, where neural network weights are initialized using pre-trained models instead of randomly, performance metrics tend to be more consistent across different training runs. To obtain robust evaluations, it is advisable to execute multiple training runs with different random seeds using the same hyperparameter settings.

Classification Prediction

In the testing phase (see Figure 2.5), the steps largely mirror those in the training phase, with the exception of data augmentation, model building and training, model evaluation, and hyperparameter tuning. The trained model employs its learned features to make predictions on new, unseen image data based on the highest prediction values.

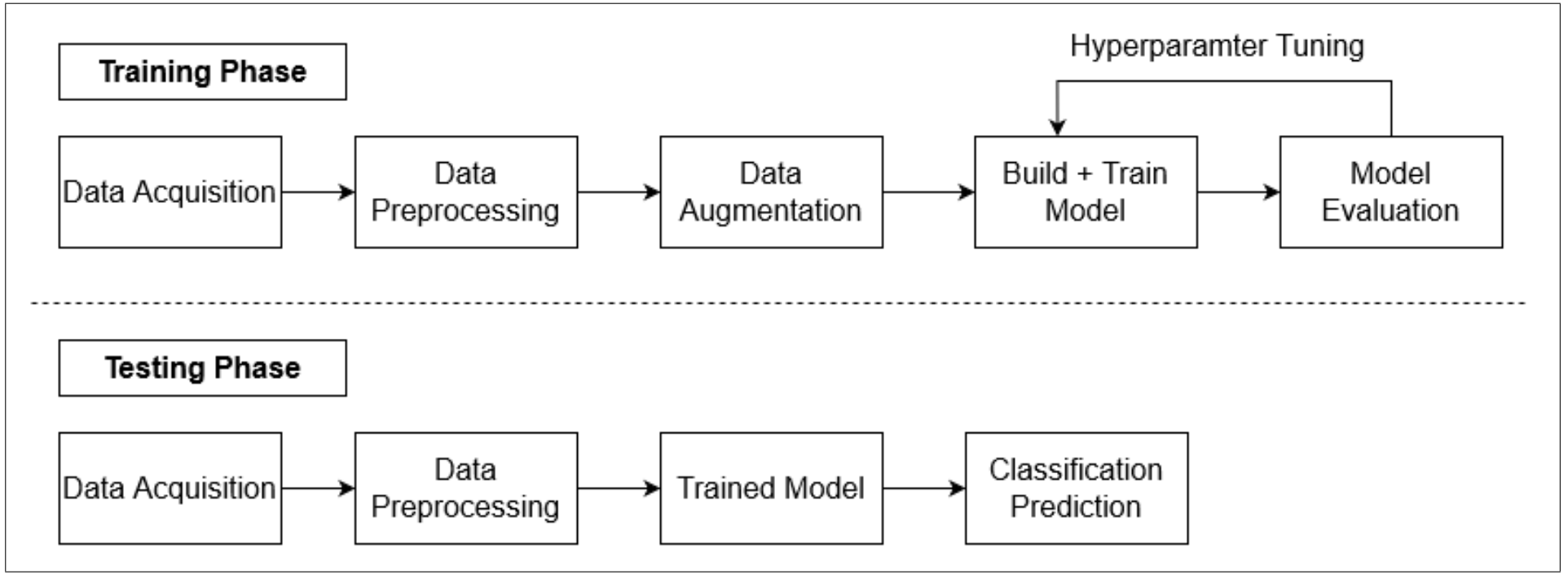


Figure 2.5: *A CNN image classification pipeline with different stages.*

2.2.3 CNNs Adoption in Medical Image Classification

The introduction of deep learning architectures like Convolutional Neural Networks (CNNs) in the field of medical image analysis was primarily focused on image classification tasks. The objective of medical image classification is to attribute one or more labels to specific target features, such as lesions, present in medical imagery, categorizing them into two or more groups. This section offers a systematic examination of how CNNs have been deployed for the task of medical image classification, spanning various types of medical imaging modalities. An overview of the adoption of CNNs in medical image analysis application doamin was carried out, organized according to the primary anatomical regions of the human body, namely the brain, chest, abdomen, and skin, in a descending sequence.

Brain

In [71], the authors introduced an innovative methodology for classifying brain tumors through a specialized form of CNN based on Complex Networks (CNNBCN). Unlike traditionally designed CNNs, the structure of CNNBCN is formulated through randomly generated graph algorithms rather than manual design and optimization. These graphs are subsequently converted into operational neural networks via a network generator. The modified version of the CNNBCN model boasts an accuracy of 95.49% in brain tumor classification, outperforming several models previously discussed in the literature. Moreover, in comparative tests, the modified CNNBCN also showed lower test loss in brain tumor classification than existing models like ResNet, DenseNet, and MobileNet. The findings suggest that computer-assisted diagnostic systems enhance the precision of tumor classification and that the modified CNNBCN model is beneficial not only for its high performance in image classification but also for advancing neural network design methodologies.

Similarly, another work [72] outlined a technique aimed at enhancing the

accuracy of practical applications, such as brain tumor image classification, by decreasing the computational complexity of traditional deep CNNs (DCNNs). The authors proposed modifications to the training algorithm to reduce the number of parameter adjustments. Specifically, their approach negated the need for weight adjustment in the fully connected layer and employed a straightforward assignment procedure to allocate the weights for this layer. The efficacy of this modified DCNN was assessed through its application on magnetic resonance images of brain tumors belonging to four different categories. The experimental results indicated that the new approach shows promise in its performance.

Chest

In 2019, the emergence of the COVID-19 virus posed a substantial global health risk [73]. Distinct from other respiratory diseases, COVID-19 is known to inflict lasting damage on the lungs, often leading to pneumonia characterized by inflammation and fluid accumulation in the lungs [74]. For severe pneumonia cases stemming from COVID-19, hospitalization and intensive respiratory support, such as oxygen therapy or mechanical ventilation, may be necessary.

Several research efforts have focused on leveraging Artificial Intelligence (AI) for diagnosing and managing this disease. In [75], the authors introduced a deep learning framework called CGNet for pneumonia detection. The framework was structured around a unique feature reconstruction technique and consists of three principal modules: feature extraction, graph-based feature reconstruction, and classification. This model was tested on publicly available X-ray and CT datasets related to pneumonia and COVID-19, achieving an impressive accuracy of 0.99 for COVID-19 detection.

Similarly, [76] developed an AI system designed to assist medical professionals in diagnosing COVID-19, particularly during high-demand periods in healthcare settings. Their system employed a multi-scale convolutional neural network for differentiating COVID-19 from other forms of pneumonia through

chest CT scans. Despite limited training data, the model demonstrated remarkable diagnostic performance, achieving a per-slice sensitivity of 99.5% and specificity of 95.6% for detecting COVID-19.

Furthermore, [77] introduced a specialized CNN model called OptCoNet aimed at automating COVID-19 screening. The architecture of OptCoNet included optimized components for feature extraction and classification. The Grey Wolf Optimizer (GWO) algorithm was applied for hyperparameter optimization. Evaluated against different classification strategies on a publicly available dataset, OptCoNet showed superior performance metrics, including an accuracy of 97.78% and an F1 score of 95.25%.

Abdomen

CNNs have demonstrated considerable potential in the use case of vital organs diagnosis, such as the stomach, liver, and prostate. In the domain of gastroenterological diseases, researchers have employed CNNs for image classification tasks involving the diagnosis of various gastric ailments, ranging from gastritis to gastric cancer.

In [78], the authors designed a CNN model specifically targeted at enhancing the diagnostic capabilities for Chronic Atrophic Gastritis (CAG). They trained their CNN-CAG model on 5,470 gastric antrum images from 1,699 patients, which were labeled based on pathological findings. This model exhibited a diagnostic accuracy, sensitivity, and specificity of 0.942, 0.945, and 0.940 respectively, outperforming even expert gastroenterologists.

In another study by [79], an ensemble of CNNs was developed for diagnosing five distinct gastric conditions. This model was trained on 21,217 gastroendoscopic images and evaluated on a test dataset comprising 1,091 images. Notably, the CNN ensemble displayed a higher diagnostic specificity and positive predictive value (PPV) than ten experienced endoscopists for early gastric cancer (EGC) and high-grade intraepithelial neoplasia (HGIN). The CNN's overall diagnostic accuracy was comparable to that of expert endo-

scopists across various gastric conditions, including lesion-free, EGC, HGIN, peptic ulcer (PU), advanced gastric cancer (AGC), and gastric submucosal tumors (SMTs), and it accomplished image recognition for the entire test set in 42 seconds.

[80] introduced a smart connected electronic gastroscopy system (SCEG) that utilizes an AdaBoost-based multi-column CNN (MCNN) for dynamic gastric cancer screening. This model significantly outstripped the performance of alternative methods.

Lastly, [81] presented a computer-aided diagnosis (CADx) system that employs a deep-learning architecture named Xception to differentiate between gastric cancer and pre-cancerous states. This system utilized Google's AutoAugment for data augmentation and incorporated the Simple Linear Iterative Clustering (SLIC) superpixel and the Fast and Robust Fuzzy C-Means (FRFCM) algorithm for image segmentation in its preprocessing stage. The model achieved an area under the receiver operating characteristic curve (Az) of 0.96, which is 0.06 higher than the Az obtained using the original data.

Skin

In the field of dermatology, particularly in the analysis of skin, the body's most extensive organ, CNNs have demonstrated notable efficacy. [82] created a hybrid classification model for dermoscopy image analysis that incorporated deep CNNs, Fisher vectors, and Support Vector Machines (SVM). This design was relatively simple, obviating the need for complex preprocessing steps. Additionally, the model's orderless representations were inherently more robust to geometric distortions. Tested on the ISBI 2016 Skin Lesion Challenge dataset, the model showed promising performance metrics without requiring intricate fine-tuning procedures.

In a focused effort on skin cancer detection, [83] employed Error-Correcting Output Codes (ECOC) SVM along with deep CNNs. Their dataset comprised RGB skin cancer images sourced from the internet, and was subject to cer-

tain noise elements like the presence of other organs and surgical tools. Image cropping was performed to minimize this noise. They employed a pre-trained AlexNet model for feature extraction and an ECOC SVM for the classification tasks. The algorithm was implemented on 3,753 images, covering four different skin cancer types. The results were impressive, yielding accuracy, sensitivity, and specificity scores of 95.1%, 98.9%, and 94.17%, respectively, for squamous cell carcinoma, actinic keratosis, and another instance of squamous cell carcinoma. The lowest average scores across these metrics were 91.8%, 96.9%, and 90.74%, which pertained to basal cell carcinoma, squamous cell carcinoma, and melanoma.

In [84], the authors developed a CNN trained on an extensive dataset of 11,444 dermoscopic images. The model aimed to categorize skin lesions into one of five diagnostic groups. The performance of this CNN was assessed against 112 dermatologists from 13 German university hospitals using a test dataset of 300 biopsy-verified images. The CNN showed statistically significant superiority ($p < 0.001$) in the classification of skin cancer images over the dermatologists, irrespective of the dermatologists' level of clinical experience.

The adoption of CNNs into the medical image analysis field marks a significant technological milestone. Across anatomical regions—including the brain, chest, abdomen, and skin—CNNs have been deployed to address an array of classification challenges with high degrees of accuracy and efficiency. The collective findings from these studies indicate that CNNs offer a transformative paradigm for medical image classification tasks across various anatomical landscapes. This merits further exploration and implementation, especially in real-world clinical settings. While CNNs have established their efficacy in laboratory and research contexts, the critical test will be their translational efficacy and interpretability in everyday medical practice. Future work should aim to address issues such as model explainability, data sensitivity, and real-world applicability to achieve a seamless integration of CNN-based diagnostic tools into healthcare systems.

Table 2.1: Representative papers using CNN for classification.

Reference	Organ	Data Modality	Application	Method
Huang et al. [71]	Brain	MRI	Brain tumor classification	CNN with a modified activation function
Hemmanth et al. [72]	Brain	MRI	Brain tumor classification	Conventional CNN with modified training algorithm.
Yu, Wang, and Zhang [75]	Lungs	- X-ray - CT scans	Pneumonia detection	Custom CNN with novel feature reconstruction method
Yan et al. [76]	Lungs	CT scans	COVID-19 diagnosis	CNN with optimized feature extraction and hyperparameter tuning.
Khan et al. [85]	Breast	Cytology images	Breast cancer classification	Ensembled CNNs
Nazeri, Aminpour, and Ebrahimi [86]	Breast	Microscopy images	Breast cancer classification	Two consecutive CNNs.
Levy and Jain [87]	Breast	Mammography images	Breast masses detection	AlexNet-based architecture with transfer learning and data augmentation.
Zhang et al. [78]	Gastric	Gastroendoscopic images	Chronic atrophic gastritis	Deep CNN.
Zhang et al. [79]	Gastric	- Gastroendoscopic images - High-grade intraepithelial neoplasia (HGIN) images	Gastric cancer detectio	Ensembled CNNs.
Wang et al. [80]	Gastric	Gastroendoscopic images	Gastric cancer screening	AdaBoost-based multi-column CNN.
Lee, Cho, Cho [81]	Gastric	Gastroendoscopic images	Gastric cancer classification	Xception-based CNN with novel image processing method.
Yu et al. [82]	Skin	Dermoscopy images	Skin lesion classification	Combination of CNN, Fisher vectors, and Support Vector Machine.
Dorj et al. [83]	Skin	Skin cancer images from Internet	Skin cancer classification	Combination of CNN and Support Vector Machine.
Maron et al. [84]	Skin	Dermoscopic images	Skin lesions classification	Deep CNN.

2.3 Gradient-based Visual Explanation

In medical image analysis, gradient-based visual explanation methods are the prevalent category of Explainable Artificial Intelligence (XAI) techniques [13, 32]. Such visualization-based explanations extend the functionality of pre-existing deep learning architectures, notably Convolutional Neural Networks (CNNs), by offering elucidations for otherwise uninterpretable outcomes [13]. These methods are tailored for local interpretability, providing insight for specific input instances and associated models [32]. Additionally, these techniques are categorized under post-hoc interpretability, delivering post-training interpretative resources like attribution maps or heatmaps [32].

From a mathematical standpoint, the objective of visual explanation algorithms is to generate attributions by ascribing a scalar "relevance" or "contribution" score to each individual input feature for a specific input data point. Given an output unit c, the aim is to quantify the influence $R^c = [R_1^c, ..., R_N^c] \in \mathbb{R}^N$ that each input variable x_i has on the output saliency $S_c(x)$. In the context of classification, one may select the output unit corresponding either to the correct class label (to identify the input features that activate that class) or to an incorrect class label (to discern reasons for misclassification). These attribution values are often visualized as heatmaps known as attribution maps. In these maps, red and blue hues signify the positive or negative influence, respectively, that each feature has on the target output, thereby spotlighting features that either promote or inhibit the target activation [32].

The conceptual foundations and technical specifications of multiple gradient-based visual explanation methods were examined. This examination was informed by existing scholarly work and exclusively centered on methods applicable to the image classification of Diabetic Retinopathy (DR). Additionally, a comparative study was undertaken to elucidate the merits and limitations of each gradient-based visual explanation method.

2.3.1 Sensitivity Analysis (SA)

Sensitivity Analysis (SA) stands as one of the pioneering techniques employed within the realm of deep learning for interpretability [32, 65]. The method (Equation 2.2) calculates attributions by determining the absolute value of the partial derivative of the target output S_c in relation to the input features x_i:

$$S_i^c(x) = \left| \frac{\partial S_x(x)}{x_i} \right| \tag{2.2}$$

Intuitively, the magnitude of the gradient signifies which input attributes, such as pixels in the context of image classification; require the minimum perturbation to elicit the maximum alteration in the target output, while omitting details concerning the directional aspects of this alteration. Sensitivity Analysis (SA) is generally characterized by a high level of noise [22], and the utilization of the absolute value obstructs the identification of both positive and negative evidential elements that could potentially be inherent in the input.

2.3.2 Gradient x Input (GI)

Gradient × Input is an extension of the Sensitivity Analysis (SA) method [88]. It was a technique proposed initially to enhance the sharpness of attribution maps generated by the method SA. The Grdient x Input method (Equation 2.3) computes the attribution by taking the signed partial derivatives of the output with respect to the input and multiplying them feature-wise by the input itself.

$$S_i^c(x) = \frac{\partial S_x(x)}{x_i} \cdot x_i \tag{2.3}$$

2.3.3 Class Activation Map (CAM)

The Class Activation Map (CAM) method visualizes a deep learning model like CNN by linearly combining activation maps at the penultimate layer with the last layer's fully-connected weights corresponding to a target class [89].

In other words, a class activation mapping of a particular class indicates the discriminative image regions used by the CNN to identify that class. However, there is an architectural contraint that require CNN's penultimate layer to be a global average pooling layer in order to work [89].

In the mathematical standpoint, let $I(x, y)$, be a given image. In the last convolution layer, the activation of a node k is $f_k(x, y)$. Then, for unit k , the result of performing global average pooling is:

$$F_k = \sum_{(x,y)} f_k(x, y) \tag{2.4}$$

Thus, the softmax for a given class c is:

$$S_c = \sum_k w_k^c F_k \tag{2.5}$$

where w_k^c is the weight corresponding to class c for unit k. Next,

$$P_c = \frac{exp^{S_c}}{\sum_{c=0} exp^{S_c}} \tag{2.6}$$

is then generated as the softmax output.

By connecting $F_k = \sum_{(x,y)} f_k(x, y)$ into the class score S_c, CAM can be defined as follows:

$$S_c = \sum_k w_c^k \sum_{(x,y)} f_k(x, y) \tag{2.7}$$

$$S_c = \sum_x^y \sum_k w_c^k f_k(x, y) \tag{2.8}$$

$$CAM_c = \sum_{(x,y)} w_c^k f_k(x, y) \tag{2.9}$$

2.3.4 Gradient-based Class Activation Map (Grad-CAM)

The Gradient-based Class Activation Map (Grad-CAM) serves as an interpretability tool for CNNs by furnishing visual clarifications for the model's decisions. This technique synthesizes a heatmap, commonly termed as a Class Activation Map, which illuminates those sections of the input image that are pivotal in steering the model toward a specific classification decision. Essentially, Grad-CAM leverages the gradients channeled into the terminal convolutional layer to ascribe significance scores to each neuron for a targeted decision [90].

In mathematical terms, to generate a class-specific discrimination map $L^c_{Grad-CAM} \in \mathbb{R}^{u \times v}$ with dimensions uxv, one must first calculate the gradient of the class score y^c with respect to the feature map activations A^k of a chosen convolutional layer. Mathematically represented as $\frac{\partial y^c}{\partial A^k}$, these backpropagated gradients are then subjected to global average pooling across the spatial dimensions (indexed by i and j) to derive the neuron importance weights α^c_k:

$$\alpha^c_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}} \tag{2.10}$$

Throughout the backpropagation of gradients for the computation of α^c_k, the operation essentially consists of successive multiplications of weight matrices and the gradient with regard to the activation functions until the gradient reaches the final convolutional layer. Consequently, this weight α^c_k embodies a partial linear approximation of the neural network downstream from A, and encapsulates the 'significance' of feature map k for a desired class c.

Subsequently, a heatmap is extracted through the application of the ReLU function, thereby eliminating negative contributions:

$$L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha^c_k A^k\right) \tag{2.11}$$

The authors of the study [90] elucidated how Grad-CAM serves as a gener-

alization of CAM. One fundamental distinction between the two lies in the architecture constraints: Grad-CAM dispenses with the necessity for the penultimate layer to be solely a global average pooling layer. Additionally, Grad-CAM employs the gradients that propagate back to the terminal convolutional layer to derive more intricate and encompassing measures of feature importance, as opposed to CAM, which directly calculates these weights from the output of the global average pooling layer to the classification scores.

2.3.5 Guided Gradient-based Class Activation Map (Guided Grad-CAM)

Despite the effectiveness of Grad-CAM in capturing the significant regions in an image that contribute to a particular classification, it has limitations in revealing fine-grained features, as indicated in existing literature [91]. To address this shortcoming, the authors extended their approach by integrating it with Guided Backpropagation, another explanation technique in the realm of XAI [90].

Guided Backpropagation, as an XAI strategy, emphasizes the specific input features, such as pixels in an image, that have a critical impact on a model's decision [91]. This method is an extension of the standard backpropagation algorithm traditionally utilized for training deep neural networks.

The standard backpropagation algorithm computes the gradient $\frac{\partial L}{\partial a}$ of the loss function L with respect to an activation a at layer l by the formula:

$$\frac{\partial L}{\partial a} = \sum_k \frac{\partial L}{\partial a_k} \frac{\partial a_k}{\partial a} \tag{2.12}$$

This equation indicates that to compute the gradient of the loss function L in relation to the activation a, one must aggregate the contributions along all paths by which a affects the loss. In this context, a_k denotes the activations in the layer that follows. The expression $\frac{\partial L}{\partial a_k}$ represents the gradient of the loss function with respect to a neuron's activation in the subsequent layer,

while $\frac{\partial a_k}{\partial a}$ is the rate of change of a neuron's activation in the following layer relative to a neuron's activation in the present layer. Essentially, this employs the chain rule to break down the gradient computation across the layers of the network.

In Guided Backpropagation, this calculation is tailored to include only the positive activations and gradients. Mathematically, this adjustment can be represented as:

$$\frac{\partial L}{\partial a} = \sum_k ReLU\left(\frac{\partial L}{\partial a_k}\right) \times ReLU\left(\frac{\partial a_k}{\partial a}\right) \tag{2.13}$$

In Eq. 2.13, the ReLU function is strategically utilized to modify the gradient before it undergoes backpropagation. Specifically, it's applied to the gradient of the loss in relation to the subsequent layer's activations $\left(\frac{\partial L}{\partial a_k}\right)$ as well as to the gradient of these subsequent activations relative to the current layer's activations $\left(\frac{\partial a_k}{\partial a}\right)$. The purpose of incorporating the ReLU function is to permit only the positive gradients to proceed backward through the network. That is, if a neuron's gradient is negative, it is reset to zero, thereby excluding it from the weight update process. Consequently, only features that positively impact the class score are taken into account, while those neurons that remain inactive are omitted. This approach is regarded as generating gradients that are more comprehensible visually, as it eliminates the influence of neurons that negatively affect the target class's score.

Subsequently, the Guided Grad-CAM visualization, denoted as V, can be computed via the equation:

$$V = G \odot L \tag{2.14}$$

Here, G represents the fine-grained visual explanations generated through Guided Backpropagation, and L signifies the coarser, class-specific regions highlighted by Grad-CAM. The operator $\odot$ symbolizes element-wise multiplication. The fusion of GG and LL culminates in a visualization that combines both high-resolution details and class-discriminative information.

2.3.6 Integrated Gradient (IG)

An advanced gradient-centric approach for visual explanation is Integrated Gradients (IG), which excels in accounting for the non-linear characteristics of deep learning models. IG computes gradients over a path that stretches from a predefined reference image x_0 to the input image in question [92].

Mathematically, IG attributes the prediction score to each input element relative to a chosen baseline x_0. The attribution is calculated by considering an interpolated series of inputs between the baseline and the actual input. The formal equation is given by:

$$s = (x - x_0) \odot \int_{\alpha=0}^{1} \frac{\partial f(x_0 + \alpha(x - x_0))}{\partial x} d\alpha \tag{2.15}$$

Due to the non-analytical nature of the integral, it's approximated through a finite summation over α in the interval $[0, 1]$:

$$s = (x - x_0) \odot \sum_{\alpha} \frac{\partial f(x_0 + \alpha(x - x_0))}{\partial x} \tag{2.16}$$

IG offers several notable features. First, it is implementation invariant, meaning that if two different network architectures yield functionally identical outputs, their attributions will also coincide [92]. Second, its generality allows it to be applied across various neural network architectures, as the method solely relies on gradients [92]. Lastly, IG adheres to the Sensitivity-N axiom, ensuring that features leading to differing outputs receive non-zero attributions [92].

Additionally, IG stands out for its capability to handle complex feature interactions effectively, as acknowledged in cited work [32]. The method is independent of architectural specifics, activation functions, or any other hyperparameters. The algorithm operates by accumulating gradients along a linear path between the baseline x_0 and the input x.

Despite its advantages, IG comes with computational drawbacks. Specifi-

cally, the method necessitates numerous evaluations of the model and its gradients for different inputs to approximate the integral numerically. This can be computationally taxing, particularly for large-scale models [32, 92].

2.3.7 Visual Explanation Adoption in Diabetic Retinopathy Image Classification

The gradient-based approaches for visual explanations have seen substantial application in the field of Diabetic Retinopathy (DR) image classification. This section provides a summary of existing literature specifically targeting this medical imaging domain and offers a comparative evaluation of the various methods in use.

As indicated in Table 2.2, fundus photography emerged as the principal type of image data for DR classification. Optical Coherence Tomography (OCT) was another data type utilized. Regarding the visual explanation techniques, Class Activation Map (CAM) stood as the predominant method, followed in popularity by Grad-CAM, Guided Grad-CAM, SA, and Integrated Gradients (IG).

Class Activation Map (CAM) (Refer to Section 2.3.3) has been extensively deployed in DR image classification, as cited in various studies [89, 93–106]. For instance, the work in [102] utilized CAM to offer a visual understanding of features and their spatial distributions that contributed to the Convolutional Neural Network (CNN) model's decisions, thereby linking it to manual grading techniques. These visualizations were directly correlated with the predicted class scores. Similarly, in [93], the CAM was employed to accentuate areas in images affected by Diabetic Retinopathy (DR). The authors further refined the CAM method to yield more detailed heatmaps. This was achieved by eliminating the strides in the first and third convolutional layers. Numerous other influential works [89, 94–101, 103–106] have similarly employed CAM for pinpointing and visually presenting important areas in images that are

critical for the classification decisions made by Convolutional Neural Networks (CNNs).

The Gradient-based Class Activation Map (Grad-CAM), as detailed in Section 2.3.4, serves as an extension to the Class Activation Map (CAM), and it too has found applications in the domain of Diabetic Retinopathy (DR) image classification [96, 107–111]. Specifically, the study in [96] leveraged Grad-CAM to identify and highlight target lesions in fundus images. Similarly, the Guided Grad-CAM technique, elaborated upon in Section 2.3.5, was employed by researchers in [112]. In this work, Guided Grad-CAM was utilized to spotlight key areas within the image and to evaluate decision uncertainty through pertinent analytical instruments. Notably, strong activations were observed at the optic disc location and the adjacent retinal blood vessels, which are areas clinicians typically concentrate on for determining laterality.

In Section 2.3.1, the Sensitivity Analysis (SA) technique is utilized by the author cited in [102] to investigate the distinctive features captured from input fundus photographs and ground truth. The study successfully identified pixels—or features—that played a role in diagnostic outcomes, revealing various lesion types like hemorrhages and microaneurysms within the vascular framework. Similarly, the research outlined in [113] employed a specialized version of sensitivity analysis for heatmap creation. To enhance the quality of these heatmaps, the authors introduced an optimization approach aimed at maximizing sparsity. This led to heatmaps characterized by fewer, yet more vividly emphasized regions, thus facilitating better interpretability and more precise identification of crucial image areas.

In [114], the authors explored the utility of the Integrated Gradients (IG) method (refer to Section 2.3.6) for classifying DR images within a simulated clinical setting. The authors found that heatmaps produced by the IG method were mainly effective for highlighting pathological features associated with positive DR predictions, but were less useful for cases with negative predictions. The study also revealed a potential risk of overdiagnosis when using such

heatmaps in the absence of any predicted DR, a finding substantiated by a significant negative effect. Overdiagnosis occurs when medical professionals identify and diagnose a disease or condition that would never have led to symptoms or issues throughout a person's life. It involves the discovery of an "abnormality" via a medical examination, which, despite being labeled as such, would not have caused any harm to the individual had it remained undetected. However, it was observed that clinician accuracy improved over the course of the study when heatmaps were used in conjunction with grading. By the study's conclusion, this level of accuracy was comparable to scenarios where only grading was employed for diagnosis.

In conclusion, gradient-based visual explanation methods have become indispensable tools in the medical imaging domain, particularly for the classification of Diabetic Retinopathy (DR) images. Among the techniques, Class Activation Map (CAM) is most prevalent, showing direct correlation with manual grading systems and aiding in the interpretability of Convolutional Neural Network (CNN) decisions. Grad-CAM extends these capabilities, offering even more nuanced insights. However, the utility of these methods is not without limitations. Specifically, heatmaps, including those generated by Integrated Gradients (IG), have shown the potential for overdiagnosis, especially when no DR is predicted in the image. Yet, there is an encouraging trend that clinicians can adapt to these advanced tools and improve their diagnostic accuracy over time.

2.3.8 Strengths & Weaknesses

While gradient-based techniques for visual explanations have seen extensive use in classifying Diabetic Retinopathy (DR) images, as documented in Table 2.2, these approaches come with both advantages and disadvantages. These have been systematically outlined and compared in Table 2.3.

Discussing the strengths first, there are two major benefits of using gradient-

Table 2.2: Representative papers that provide visual explanation in Diabetic Retinopathy image analysis.

Authors (Year)	**Image Modality**	**Main XAI Used**
Gondal et al. (2017) [93]	Fundus Photographs	CAM
Wang and Yang (2018) [94]	Fundus Photographs	CAM
Ahmad et al. (2019) [95]	Fundus Photographs	CAM
Jiang et al. (2019) [96]	Fundus Photographs	CAM
Kumar et al. (2019) [97]	Fundus Photographs	CAM
Li et al. (2019) [98]	Fundus Photographs	CAM
Liao et al. (2019) [99]	Fundus Photographs	CAM
Wang et al. (2019) [100]	Fundus Photographs	CAM
Perdomo et al (2019) [101]	OCT	CAM
Pratt et al. (2019) [102]	Fundus Photographs	CAM
Narayanan et al. (2020) [103]	Fundus Photographs	CAM
Shen et al. (2020) [104]	Fundus Photographs	CAM
Tu et al. (2020) [105]	Fundus Photographs	CAM
Wang et al. (2020) [106]	OCT	CAM
Zhou et al. (2020) [89]	Fundus Photographs	CAM
Kim et al. (2019) [115]	Fundus Photographs	Grad-CAM
Thakoor et al. (2019) [107]	OCT	Grad-CAM
Zhang et al. (2019) [108]	Fundus Photographs	Grad-CAM
Martins et al. (2020) [109]	Fundus Photographs	Grad-CAM
Meng et al. (2020) [110]	Fundus Photographs	Grad-CAM
Wang et al. (2020) [111]	OCT	Grad-CAM
Jiang et al. (2020) [96]	OCT	Grad-CAM
Jang et al. (2018) [112]	Fundus Photographs	Guided Grad-CAM
Pratt et al. (2019) [102]	Fundus Photographs	Sensitivity Analysis
Quellec et al. (2017) [113]	Fundus Photographs	Sensitivity Analysis
Sayres et al. (2019) [114]	Fundus Photographs	Integrated Gradient

Table 2.3: Strengths and weaknesses of gradient-based visual explanation methods. Strengths are depicted as +, weaknesses are -.

Gradient-based Visual Explanation Methods	**Heatmaps Quality**	**Ease of Use**	**Validity**	**Robustness**
Class Activation Map (CAM)	+/-	+	Not Tested	-
Gradient-based Class Activation Map (Grad-CAM)	+/-	+	+/-	-
Guided Gradient-based Class Activation Map (Guided Grad-CAM)	+/-	+	Not Tested	-
Sensitivity Analysis (SA)	-	+	Not Tested	+
Integrated Gradient (IG)	+	+	Not Tested	-

based visual explanation methods. Firstly, these methods are generally efficient in terms of computational resources. For example, Sensitivity Analysis (SA) and Gradient x Input (GI) only necessitate a single forward and backward pass through the neural network to generate the attribution maps. Integrated Gradients (IG), although relatively more resource-intensive, is still considered efficient when compared with other methods, requiring between 50 to 200 backward passes for numerical integration [32, 92]. Additionally, these methods scale well as the computational requirements do not hinge on the number of input features. It is worth noting that these methods integrate seamlessly with frameworks like TensorFlow [116] and PyTorch [117], leveraging GPUs for optimized gradient calculations [32].

The second benefit is the straightforwardness of implementing these methods. As stated by [92], these techniques can be readily applied to existing models like CNNs. When gradients are utilized, methods such as GI and IG can be adapted to any CNN architecture that supports gradient computations. Given that CNNs utilize backpropagation to efficiently calculate gradients, these explanation methods can be implemented without creating custom layers or operations [32, 92].

However, these methods are not without limitations. For one, the quality of the generated attribution maps can be inconsistent, being notably influenced by gradient noise [32]. In image classification tasks, these methods might identify image regions contributing to the classification, but the attributed values to adjacent pixels could show high variance, even contradicting each

other.

Secondly, these techniques tend to offer localized perspectives on feature importance, failing to encapsulate global feature interactions as noted by authors in [15]. This could limit their usefulness in contexts where decisions are based on complex feature relations. Moreover, the specific architecture of the CNN affects the performance and interpretation of these visual explanation methods [118]. Varied architectures could focus on different features, challenging the generalization of these methods.

Another shortcoming, as discussed in Section 1.2, is the lack of validated explanations these methods offer [12, 13, 32]. The term 'validity' here refers to how accurately the explanation matches the expectations of the end-user. Efforts to quantify this are limited but crucial, as shown in the study by Arun et al. [36], which focused exclusively on chest X-rays, implying the need for more research across other imaging modalities and anatomical sites.

Lastly, the resilience of these XAI methods can be assessed by altering parameters or data within the deep learning architecture and observing the impact on the generated explanations. Studies, such as in [119], have been conducted in this regard, revealing that some methods like guided backpropagation and guided Grad-CAM offer similar visual explanations even under these variations, often accentuating edges. Hence, careful consideration is recommended when adopting these methods for visualization purposes.

2.4 Data-centric AI

The rise of deep learning techniques has ignited significant breakthroughs in diverse areas ranging from personalized recommendations to medical diagnostics via computer vision. However, the broader machine learning community has traditionally focused on refining algorithms to excel on standard datasets, often neglecting vital factors relevant to practical deployments. Such overlooked factors include data quality and preprocessing, recognizing subpopu-

lations within data, ensuring model robustness, tracking performance, and accounting for shifts in the data landscape [3].

Despite monumental advancements in machine learning, there have been several prominent examples of system-level shortcomings. Google Health's diabetic retinopathy model serves as a notable example, falling short in its performance on lower-quality images compared to those used during its training phase [120]. Moreover, quality issues with data can trigger a series of downstream problems, known as data cascades, as highlighted by Sambasivan et al. [43]. The authors revealed that a large proportion of AI practitioners have faced data cascades, which can result in costly project adjustments, project cessation, and even societal harm. Studies also indicate that a large majority of industrial AI/ML models are susceptible to producing erroneous results due to data or algorithmic biases, and only about half successfully transition from prototype to production [121]. These case studies reiterate that having a highly predictive model is only one aspect of a reliable machine learning system in real-world scenarios. Central to these failures is often the quality of the data being used.

In light of these challenges, the focus within the machine learning community is gradually shifting towards a data-centric AI approach, emphasizing the importance of high-quality data [122]. Though this shift is in its early stages, several applications adopting data-centric AI methodologies exist. For instance, large language models like ChatGPT have minimal architectural modifications but rely heavily on high-quality, large-scale training data [123, 124]. As these models grow more powerful, they become increasingly capable of performing data-centric tasks like data labeling [124] and data augmentation [125]. Another example includes prompt engineering, which focuses solely on fine-tuning data inputs and has shown significant promise [126].

To guide this shift, Seedat et al. [3] have introduced DC-Check, a practical checklist advocating for a data-centric methodology at multiple stages of the machine learning workflow. Similarly, Gebru et al. [45] have presented a

checklist focused on dataset quality, aimed at dataset creators and users. This checklist seeks to improve transparency and accountability, minimize societal biases in machine learning models, enhance result reproducibility, and assist researchers and practitioners in making more informed dataset choices for their specific tasks.

2.5 Diabetes & Diabetic Retinopathy

Diabetes Mellitus (DM), commonly referred to as diabetes, is a chronic condition marked by high levels of blood sugar owing to either inadequate insulin production or ineffective utilization of the hormone by the body [127]. Elevated blood sugar levels, also known as hyperglycemia, can, if unmanaged over the long term, result in severe medical issues like cardiovascular disorders, nerve degradation (neuropathy), kidney dysfunction (nephropathy), and ocular complications (retinopathy) [128].

International organizations such as the International Diabetes Federation (IDF) [129], the Centers for Disease Control and Prevention (CDC) [130], and the World Health Organization (WHO) [131] are tracking the global incidence of diabetes. Recent data reveals that, in 2017, approximately 425 million adults between the ages of 20 and 79 were diagnosed with diabetes globally [132]. Projections suggest that this number could escalate to 693 million by the year 2045. Alarmingly, half of the adult population diagnosed with diabetes in this age range are not even aware of their condition [132], indicating an urgent requirement for robust screening, diagnosis, and treatment strategies on a worldwide scale.

The surging rates of diabetes are paralleled by an increased occurrence of Diabetic Retinopathy (DR), an eye condition caused by diabetes-induced damage to retinal blood vessels, resulting in blockages and leakages. The root cause of DR is hyperglycemia, and failure to control it can lead to vision impairment and even blindness in extreme cases [128]. The IDF Atlas in 2015

reported that over a third of the 425 million people with diabetes globally suffer from some form of DR [132]. According to [133], the International Association on the Prevention of Blindness (IAPB) disclosed that in 2015, 145 million individuals had some degree of DR, and 45 million had vision-threatening DR.

For effective disease diagnosis, periodic eye screening is imperative [129–131]. Medical experts rely on retinal imaging to identify different lesion types such as microaneurysms, hemorrhages, and both soft and hard exudates, which have unique attributes. Figure 2.6 showcases the variety of lesions detectable in a fundus photograph.

- **Microaneurysms**: The initial manifestation of DR is observed as microaneurysms. These are small circular reddish dots on the retina resulting from the weakened walls of the vessels, with a size less than 125 μm and well-defined margins [134].

- **Haemorrhages**: Haemorrhages manifest as larger, irregularly-shaped spots on the retina, with a size greater than 125 μm [135].

- **Exudates (Soft and hard)**: Hard exudates are yellowish spots that result from plasma leakage and are located in the outer layers of the retina. They have well-defined borders. On the other hand, soft exudates, also known as cotton wool spots, are white spots that occur due to the swelling of nerve fibers. They are usually oval or circular in shape [136].

Based on the extent of the disease, a variety of lesion combinations may appear. DR can be categorized into five separate stages as delineated in [137]. Figure 2.7 displays fundus images that correspond to these various DR severity levels. In this research, we have comprehensively examined all gradations of DR severity. Such an all-encompassing methodology is essential, as different severities frequently manifest through diverse lesion combinations. It is crucial for the Convolutional Neural Network (CNN) to accurately recognize and

extract these lesion combinations as notable features to enhance its predictive capabilities. Identifying these combinations is not a mere compilation of data; it serves as an essential prerequisite for elevating the efficacy of our predictive models. Therefore, by focusing on this aspect, we anticipate that our research will yield substantially improved diagnostic predictions for DR severity.

- **No DR (Stage 0)**: Absent or no sign of disease yet.
- **Mild DR (Stage 1)**: During the initial stages of the disease, the retina may exhibit small balloon-shaped swellings in the form of tiny blood vessels known as microaneurysms. These microaneurysms have the potential to release fluid into the retina, leading to capillary leakage.
- **Moderate DR (Stage 2)**: As the disease advances, the blood vessels that provide nutrients to the retina may undergo swelling and deformation, potentially leading to a loss of their blood transport functionality. These conditions can bring about distinct alterations in the retinal appearance, ultimately leading to macular edema, which can exacerbate the leakage of blood or exudate from the capillaries.
- **Severe DR (Stage 3)**: A significant number of blood vessels may become obstructed, causing a lack of blood supply to specific retinal regions, a phenomenon referred to as capillary non-perfusion. These areas may release growth factors that prompt the retina to initiate the development of novel blood vessels.
- **Proliferative DR (Stage 4)**: At this stage, the capillary non-perfusion phenomenon stimulates the growth of fresh blood vessels, which emerge along the internal surface of the retina and extend into the vitreous humor, the fluid that occupies the interior of the eye. These newly formed blood vessels are delicate and thus more susceptible to capillary leakage. Concurrently, the emergence of scar tissue can result in contraction, leading to retinal detachment, a situation where the retina separates from

the underlying tissue, akin to wallpaper peeling away from a wall. Such detachment may result in permanent loss of vision.

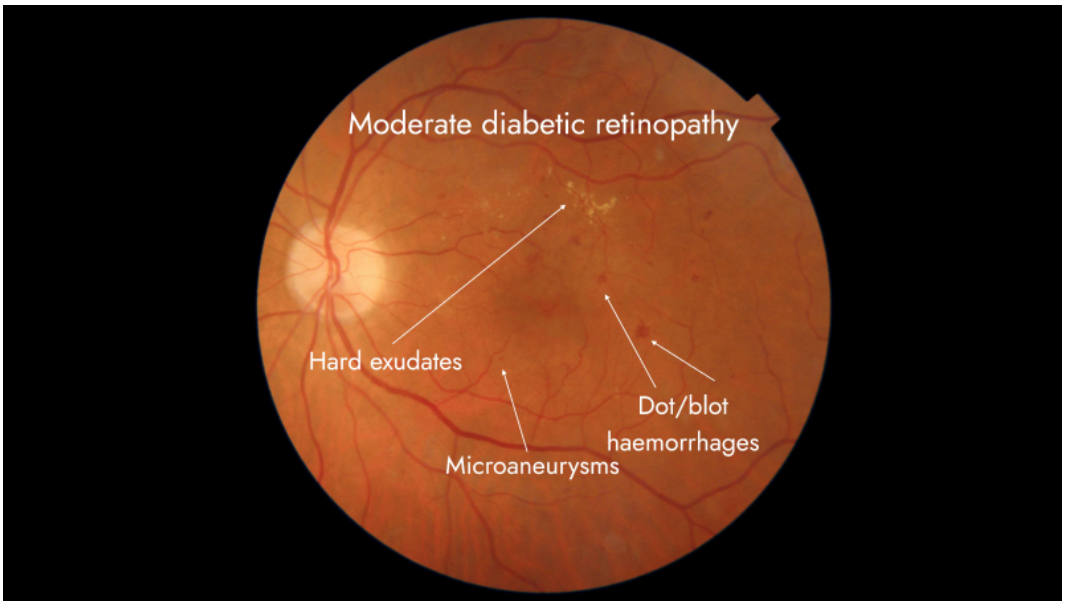
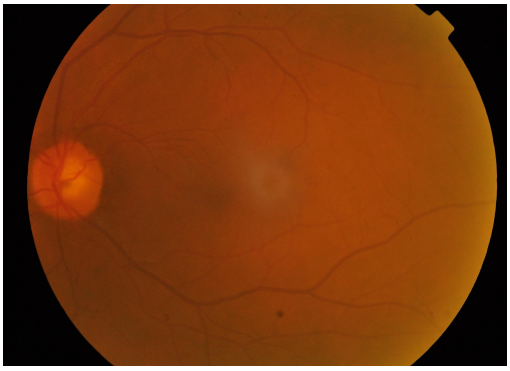
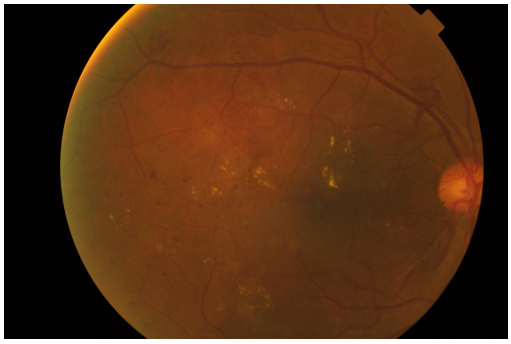


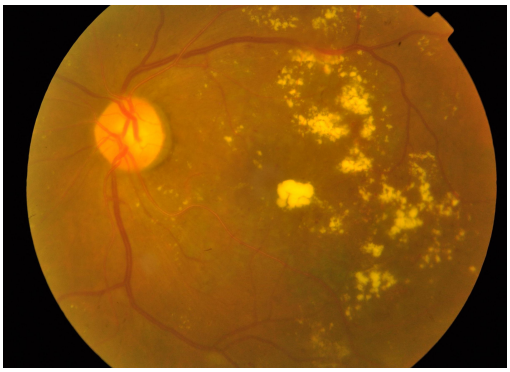
Figure 2.6: *A fundus images with different types of diabetic retinopathy lesions.*



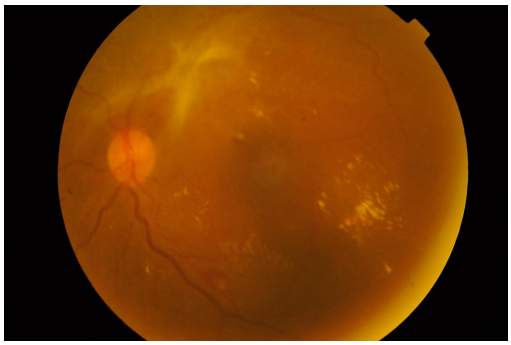
(a) *An example of fundus images with Mild DR (class 1)*



(b) *An example of fundus images with Moderate DR (class 2)*



(c) *An example of fundus images with Severe DR (class 3)*



(d) *An example of fundus images with Proliferative DR (class 4)*

Figure 2.7: *Examples of fundus images with respective diabetic retinopathy severity.*

Chapter 3

Methodology

3.1 Introduction

This chapter delineated the methodologies employed to achieve the research aim and objectives presented in Section 1.3.

The first of these pertained the development of image classification model, following an end-to-end deep learning pipeline. The pipeline started off with data acquisition, data preprocessing, data augmentation, model building and training, and model performance evaluation. Different methods were adopted in each phase and the rationale of using these methods were justified in this chapter.

Once the model has been trained and evaluated, a justifiable visual explanation method was used to generate the heatmaps that highlight the areas of importance in fundus photography images that contributed to the classification decision. Here, the next objective was to develop the proposed quantitative saliency-oriented evaluation metric to evaluate the validity of the generated heatmaps.

The next objective was to use a data-centric approach to produce a effective high-quality Diabetic Retinopathy (DR) fundus photography images. These images were then preprocessed by the proposed image preprocessing method to tackle the limitations of pixel-based image preprocessing methods. In addition,

the proposed quantitative saliency-oriented evaluation metric was used again to evaluate the validity of the generated heatmaps.

Figure 3.1 provides an overview of the workflow employed in this research, highlighting the manner in which the various objectives are pursued.

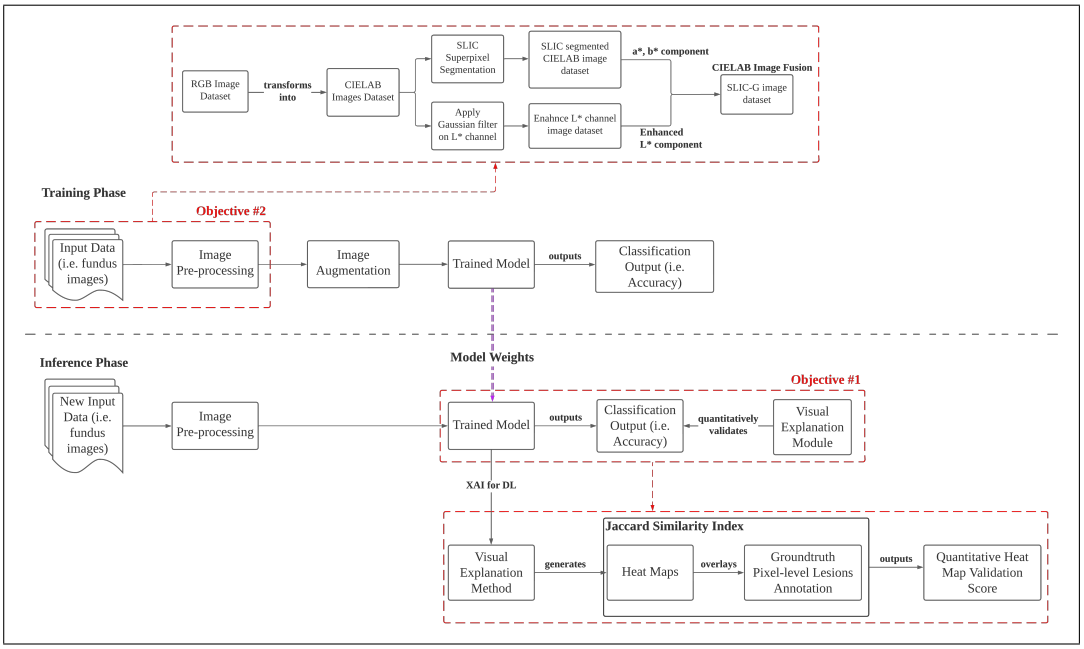


Figure 3.1: *Methodology Overview*

3.2 Machine Specifications and Software Frameworks

The computational environment employed for this research utilized the Ubuntu 20.04 operating system and was powered by an Intel(R) Core(TM) i7-10700F CPU @ 2.90GHz for its Central Processing Unit (CPU). The Graphics Processing Unit (GPU) in use was an NVIDIA GeForce RTX3070 with 8GiB of RAM, which is well-suited for executing Deep Learning (DL) tasks.

Table 3.1: Machine Specifications

Machine Specifications	
Operating System (OS)	Ubuntu 22.04
Central Processing Unit (CPU)	Intel(R) Core(TM) i7-10700F CPU @ 2.90GHz
Graphics Processing Unit (GPU)	NVIDIA GeForce RTX3070 @ 8GiB RAM

Regarding the software stack, the Open Computer Vision (OpenCV) Library and Pillow served as the libraries of choice for image preprocessing steps. For the construction and training of the machine learning model, the PyTorch framework in Python was employed. Additional libraries, including Numpy, MatPlotLib, and Scikit-learn, were utilized for data analytics and graphical representation of results. A summary of the hardware and software configurations used in this study can be found in Table 3.1 and Table 3.2, respectively.

Table 3.2: Software Frameworks

Software Frameworks	
Image Preprocessing	- Open Computer Vision (OpenCV) Library - Pillow (Python Imaging Library)
Deep Learning	PyTorch, TorchVision
Data Analysis and Visualization	Numpy, MatPlotLib, Scikit-learn

3.3 DC-Check

Prior to the commencement of any machine learning stages (as delineated in Figure 2.5), a methodical approach known as DC-Check was implemented. This approach was structured as a checklist intended to emphasize data-centric aspects across various phases of the deep learning workflow, ranging from the initial phase of data acquisition to the final stage of model deployment.

The rationale for the adoption of this framework lies in its potential to foster both transparency and conscientiousness during the developmental cycle, thereby contributing to the creation of machine learning systems that are both reliable and effective. The DC-Check provided a catalogue of queries; by systematically addressing these, key factors were recognized and addressed at each successive phase in the workflow. This checklist fostered a focus on data-centric issues throughout the deep learning pipeline's multiple stages:

- **Data**: Considerations to improve the quality of data used for model

training, such as data acquisition, data preprocessing, and data augmentation.

- **Training**: Considerations based on understanding the data to improve model training, such as data informed model design, domain adaptation, and robust training.

- **Testing**: Considerations around novel data-centric methods to test ML models, such as informed data splits, targeted metrics, classification evaluation.

- **Deployment**: Considerations based on data post-deployment, such as data and model monitoring, model adaptation and retraining, and uncertainty quantification.

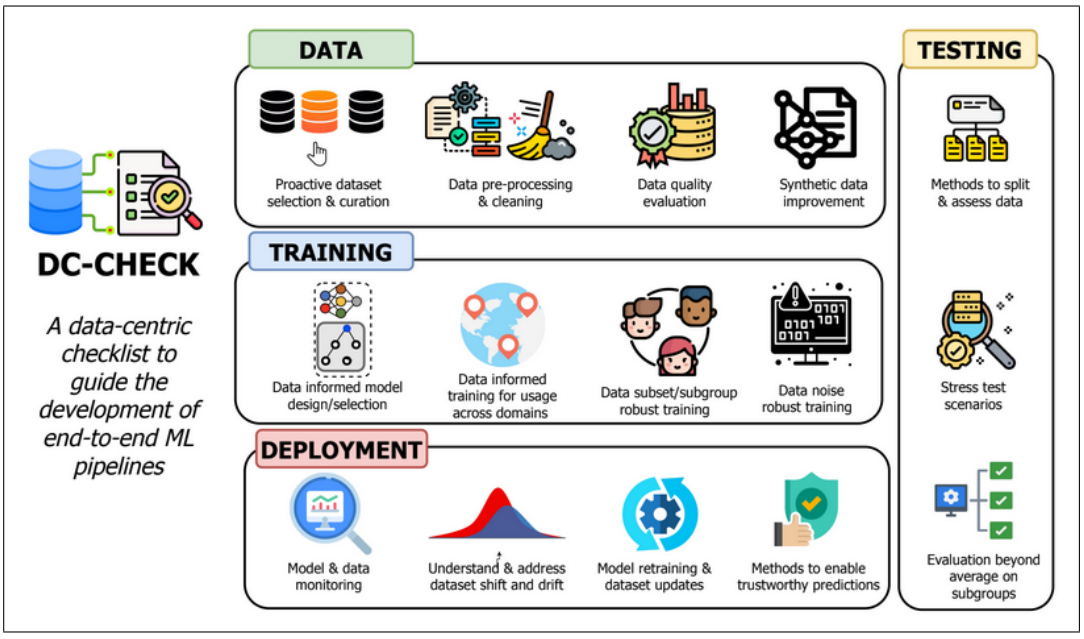


Figure 3.2: *DC-Check [3]*

It is crucial to acknowledge that the DC-Check framework was modified slightly to align with the work presented in this thesis. As the work in this thesis is primarily experimental and does not extend to real-world model deployment, modifications were made to the Deployment stage of the framework, specifically concerning Question 14 on the checklist. Question 14 pertains to

any tools utilized to foster trust in the model, hence the alteration to the framework. The whole checklist can be viewed in the Appendix B and C.

3.4 Data Acquisition

Three datasets were acquired and used in this research. The datasets were Kaggle Diabetic Retinopathy 2015 (KDRD2015) [57], APTOS 2019 (APTOS2019) [138], and the Indian Diabetic Retinopathy Image Dataset (IDRiD) [139]. All three datasets were publicly available and can be obtained from online repositories, namely Kaggle [57], IEEE Dataport [68], and Google Dataset Search [58]

The motivation for selecting these particular datasets rests on three factors: (1) the images were originated from real-world medical contexts, as the fundus images were gathered from different healthcare institutions [138–140]; (2) the datasets provides multiclasses which aligned with the aforementioned research objective; and (3) the datasets hold considerable recognition within the research community, particularly in data science competitions [57].

Table 3.3 illustrated the data distribution of each dataset. Based on the distribution, the KDRD2015 dataset provided the largest amount of images, followed by APTOS2019 and IDRID. It was observed that all three datasets were highly imbalanced as class 0 (i.e. No DR) has the highest amount of images whereas other classes were the least amount of images. The methodology adopted to tackled this imbalanced scenario was explained in Section 3.6.

All three datasets consisted of color fundus photography images aimed at classifying Diabetic Retinopathy (DR) at different stages of its progression. According to Table 3.4, all three datasets offered DR grading images for image classifcation. A noteworthy is the fact that the Indian Diabetic Retinopathy Image Dataset (IDRiD) offered DR images for classification and segmentation. The segmentation images, also known as pixel-level segmented ground-truth images of diabetic lesions, serving as a critical asset for scrutinizing the validity

of gradient-based visual explanations in Chapter 4 and 5.

Table 3.3: Overall data distribution of three different Diabetic Retinopathy (DR) datasets.

Dataset	Class	Category	Number of Images
Kaggle Diabetic Retinopathy Dataset 2015 (KDRD2015)	0	No DR	25,802
	1	Mild DR	2,438
	2	Moderate DR	5,288
	3	Severe DR	872
	4	Proliferative DR	708
		Total	**35,108**
Asia Pacific Tele-Ophthalmology Society 2019 (APTOS2019)	0	No DR	1,805
	1	Mild DR	370
	2	Moderate DR	999
	3	Severe DR	193
	4	Proliferative DR	295
		Total	**3,662**
Indian Diabetic Retinopathy Image Dataset (IDRID)	0	No DR	168
	1	Mild DR	25
	2	Moderate DR	167
	3	Severe DR	94
	4	Proliferative DR	62
		Total	**516**

Table 3.4: Distribution of classes across different tasks for the used datasets.

Dataset	Segmentation	Object Detection	Classification
Kaggle Diabetic Retinopathy Dataset 2015 (KDRD2015)	-	-	Diabetic Retinopathy Grading - Class 0: No DR - Class 1: Mild DR - Class 2: Moderate DR - Class 3: Severe DR - Class 4: Proliferative DR
APTOS 2019	-	-	Diabetic Retinopathy Grading - Class 0: No DR - Class 1: Mild DR - Class 2: Moderate DR - Class 3: Severe DR - Class 4: Proliferative DR
Indian Diabetic Retinopathy Image Dataset (DRID)	Lesion Segmentation - Hemorrhages (HA) - Microaneuryms (MA) - Soft Exudates (SE) - Hard Exudates (EX)	Detection of Organ Centroids - Optical Disc - Fovea Centrails	Diabetic Retinopathy Grading - Class 0: No DR - Class 1: Mild DR - Class 2: Moderate DR - Class 3: Severe DR - Class 4: Proliferative DR

3.5 Image Preprocessing

In the real-world, ophthalmologists examine fundus photography images to identify diabetic lesions such as exudates, hemorrhages, and microaneurysms for Diabetic Retinopathy (DR) diagnostic purposes. It requires years of practice and experience to accurately identify the affected area and make diagnosis [141].

Fundus photography images, however, may possess undesired artifacts, including noise and varying contrast levels, stemming from environmental conditions and the use of differing camera models [49]. Prior research has indicated that image preprocessing plays a vital role in mitigating noisy artifacts and enhancing image classification performance [50–52]. To allow automated deep learning system like Convolutional Neural Networks (CNNs) to perform image classification effectively, it is required that these artifacts must be removed so that the CNNs can learn to extract important features rather than noise, to make accurate and unbias prediction.

Different image preprocessing methods were used in this research. The Contrast Limited Adaptive Histogram Equalization (CLAHE) [142] was used to preprocess the fundus photgraphy images. The rationale of using CLAHE is that it prevents contrast over-amplification [142]. Furthermore, CLAHE improve contrast in both brighter and darker areas of an image, without saturating or washing out details [142, 143]. Prior literature used CLAHE as the go-to method of choice in DR image analysis [144]. Here, we extend the use of CLAHE by applying CLAHE in different images' color space such as CLAHE applied to green channel of images.

Furthermore, another image preprocessing method called Gaussian weighted contrast [145] had been adopted to preprocess the images. The rationale of adopting this method is that the author used this method and achieved great performance in DR image classification [145]. The Gaussian weighted contrast first blur the image and then take the difference with the original image, and add weights with a factor to increase contrast. In a nutshell, the Gaussian weighted contrast remove some of the variation between images due to different lighting conditions and camera resolution.

In this research, a novel image preprocessing method named SLIC-G was proposed to tackle the limitations of pixel-based image preprocessing methods as described in the Problem Statement section (Section 1.2). The SLIC-G took inspiration from Gaussian weighted contrast method and combined with Sim-

ple Linear Iterative Clustering (SLIC) superpixel segmentation method. The SLIC superpixel segmentation partitions the images into regions that provide meaningful descriptions with significantly less data, as opposed to utilizing all pixels in an image [146]. This method also curtails local redundancy within an image while preserving local structures and circumventing noisy, excessively segmented fragments [146]. Additionally, the researchers in [146] conducted a comparative analysis of the SLIC method against other superpixel segmentation techniques, specifically Turbopixels, Watershed, Felzenszwalb and Huttenlocher's method, and Mean Shift. The comparison was anchored on several factors, including the fidelity to image boundaries, computational efficiency, memory utilization, and the flexibility in managing the number of superpixels. The findings revealed that SLIC significantly excelled in terms of computational efficiency and memory consumption when compared to the other methods. It also either matched or surpassed the other techniques in regards to boundary adherence and the compactness of the superpixels. These evaluations were carried out using standard images from the Berkeley Segmentation Dataset. The SLIC-G image preprocessing method also incorporates the Gaussian smoothing operation. The adoption of the Gaussian smoothing operation is justified by its ability to enhance classification scores through image preprocessing, as indicated in [147]. This enhancement is primarily due to the Gaussian filter's capacity to improve image sharpness by reducing noise. Consequently, this presents an opportunity of developing a framework that utilizes SLIC to segment images into distinct regions based on perceptual attributes, further augmented by the image clarity provided by the Gaussian filter.

The technical implementation of each image preprocessing methods was described in Chapter 4 and 5, respectively.

3.6 Data Augmentation

Data augmentation, a regularization technique, can bolster the overall performance of deep learning models [148]. The use of data augmentation techniques is to artificially increase the size and diversity of a training dataset, with the aim of improving model performance. It works by applying various transformations to the original data samples to generate new instances. These transformations are carefully designed so as not to alter the semantic content of the data, ensuring that the augmented samples remain relevant to the task at hand. The rationale of using data augmentation lies in its capacity to introduce variability and thereby robustness into the training process. This serves to reduce overfitting by enlarging the dataset with diversified instances, the model is less likely to memorize the training data and more likely to generalize well to new, unseen data [70, 149]. Data augmentation also improve invariance. For example, by rotating an image, the model learns that the orientation of the object does not change its label. This is crucial for tasks where the object of interest can appear in varying orientations and scales [70, 149].

The RandAugment [150] data augmentation technique was adopted. The RandAugment holds several advantages over alternative augmentation methodologies, including the reduction or elimination of a separate search space, thus, significantly reducing training and computation. [150]. The method encompasses various data augmentation operations, such as random grid shuffle, rescale, horizontal and vertical flips, shear in x and y direction, translate in x and y direction, rotation, posterization, contrast, and sharpness.

3.7 Model Building and Training

In the scope of this research, the primary Convolutional Neural Network (CNN) selected for meeting the research objectives was a pre-trained DenseNet [4]. The justification for utilizing the pre-trained DenseNet lies in its underrep-

resentation in the domain of Diabetic Retinopathy (DR) image classification [59]. Additionally, the architecture of DenseNet distinguishes it from other pre-trained CNNs, contributing to its selection. Furthermore, the use of a pre-trained CNN enables the benefits of transfer learning—an optimization strategy that facilitates accelerated progress or enhanced performance in secondary tasks, such as medical image classification in this context [151].

The pre-trained DenseNet is a type of CNN that was introduced by [4]. One of the hallmark features of DenseNet is that it comprised of several "Dense Blocks" interconnected with the "Transition Layers". Within a dense block, each layer l is connected to every other layer in a feed-forward fashion. Mathematically, the output X_l of the $l-th$ layer in a dense block can be formulated by:

$$X_l = H_l[X_0, X_1, ..., X_{l-1}] \tag{3.1}$$

, where $[X_0, X_1, ..., X_{l-1}]$ represents the concatenation of the feature maps produced by all preceding layers, and H_l is a composite function of operations like Batch Normalization (BN), Rectified Linear Unit (ReLU), and Convolution (Conv).

There were multiple compelling advantages of employing DenseNet. First, the model is notably efficient in terms of parameter utilization [4]. Unlike conventional CNNs, where layers are connected solely to their immediate neighbors, DenseNet promotes feature map reuse, leading to fewer required parameters without compromising expressive capabilities. Secondly, the architecture supports improved gradient flow, obviating issues related to gradient vanishing, particularly in deep networks [152]. Thirdly, DenseNet enables feature reuse, which not only is effective for tasks requiring hierarchical and multi-scale features but also serves as a regularization technique, thereby lessening overfitting [4]. Moreover, the architecture is conducive for fine-tuning in domains like medical imaging, where labeled datasets are often scarce and costly to obtain [153]. Lastly, the design of DenseNet fosters significant parame-

ter sharing, thus mitigating computational and memory burdens during both training and inference [4].

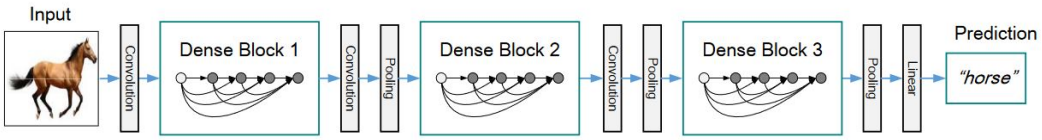


Figure 3.3: *The architectural representation of DenseNet [4].*

3.8 Model Performance Evaluation

Evaluation metrics are employed in CNN image classification to provide a quantitative assessment of the model's performance in terms of its accuracy in classifying images. This analysis aids in evaluating the effectiveness of the model and informs the development of improvement strategies.

In this research, several performance evaluation metrics have been utilized to assess the performance of CNN models. This segment elaborates on each of the evaluation metrics that are deemed suitable for diabetic retinopathy image classification and heat map generated by gradient-based visual explanation techniques.

3.8.1 F1 Score, Precision, and Recall

The F1 score, precision, and recall are commonly used metrics for evaluating the performance of classification algorithms, particularly in scenarios where the class distribution is imbalanced or the costs of different types of errors are not the same [154]. Below are formal definitions and explanations for each metric.

The Precision is the fraction of true positive predictions among the instances that the model classified as positive. Here, True Positive (TP) is the number of positive instances correctly classified by the model, and False Posi-

tive (FP) is the number of negative instances incorrectly classified as positive.

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

Recall (or Sensitivity or True Positive Rate), on the other hand, is the fraction of true positive predictions among the actual positive instances in the dataset. Here, False Negative (FN) is the number of positive instances incorrectly classified as negative.

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

Finally, F1-Score is the harmonic mean of precision and recall and is calculated as follows:

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.4}$$

3.8.2 Accuracy

Accuracy is a prevalent evaluation metric for image classification, which measures the ratio of correctly classified images to the total number of images. This metric is commonly used in traditional machine learning and deep learning approaches for image classification. In [155], the authors reported that accuracy is frequently used to assess the overall performance of image classification models, indicating their proficiency in distinguishing images across different classes.

Accuracy is a suitable metric for image classification tasks with balanced class distribution. However, in the case of imbalanced data sets, where some classes have a much lower number of images than others, accuracy may provide misleading results. In such scenarios, other metrics like precision, recall, F1-score, confusion matrix, and Receiver Operator Characteristics (ROC) Curve

may offer better insights into model performance [156].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.5}$$

3.8.3 Confusion Matrix

The confusion matrix serves as a crucial instrument for appraising the fidelity of predictions made by a machine learning model, particularly in the domain of image classification. This tabulated metric is prevalent for encapsulating both the predicted and actual categorizations rendered by a model. By enumerating True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), the matrix furnishes indispensable perspectives on the model's overall efficacy.

The utility of the confusion matrix in scrutinizing the performance of CNN models in image classification tasks has been corroborated by existing research [157]. The authors accentuate the matrix's capacity for exhaustive evaluation, which allows for the discernment of both the strengths and shortcomings inherent in a model. Moreover, the confusion matrix is instrumental for understanding the allocation of accurate and erroneous predictions, a facet crucial for the enhancement of the model's performance [158].

3.8.4 Receiver Operator Characteristics (ROC) Curve

In [159], the Receiver Operating Characteristic (ROC) curve is explicated as a graphical tool for evaluating the efficacy of a binary classification system. The curve visually delineates the trade-off between achieving high true positive rates (sensitivity) and minimizing false positive rates (1-specificity), across a spectrum of classification thresholds. The curve is constructed by plotting the true positive rate against the false positive rate for differing thresholds. An ideal classifier would have its curve converge towards the upper-left corner of the plot, denoting high discriminative power for both positive and negative

classes. Conversely, a curve that aligns with the diagonal line from the lower-left to the upper-right signifies that the classifier performs no better than random guessing.

In the context of multi-class image classification, ROC curves are instrumental for assessing classifier performance. Distinct ROC curves can be formulated for each class in the dataset as well as for aggregate classification efficacy. This involves configuring each class as the 'positive' class while designating the remaining classes as 'negative', followed by curve construction. The area under the ROC curve (AUC-ROC) can function as a unified metric for contrasting the performance of multiple classifiers.

3.8.5 Geometric Mean

Geometric Mean (G-mean) [160] is an assessment metric employed to gauge the overall classification effectiveness of a multi-class classifier. Differing from other performance metrics such as accuracy, G-mean factors in the sensitivity and specificity of each class and determines the geometric mean of these values by means of the following formula:

$$Geometric\,Mean = \left[\frac{TP}{TP+FN}\right] * \left[\frac{TN}{FP+TN}\right] \tag{3.6}$$

, where TPR represents the true positive rate (sensitivity) and TNR represents the true negative rate (specificity).

In the realm of multi-class image classification, G-mean is frequently utilized to appraise the performance of classifiers under imbalanced class distribution scenarios. This is owing to the fact that accuracy can be deceptive in the presence of imbalanced data sets, wherein a high accuracy rating can be achieved by assigning all instances to the majority class. Conversely, G-mean yields a more equitable evaluation of the classifier's performance across all classes by accounting for the sensitivity and specificity of each class.

3.8.6 Jaccard Similarity

Jaccard Similarity is an evaluation metric used to measure the similarity between two sets of data, and in the context of image classification, it is used to evaluate the performance of the classifier by measuring the intersection over the union (IoU) of the predicted and ground truth labels [161]. Jaccard Similarity is calculated using the formula:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.7}$$

where A is the predicted set and B is the ground truth set. The value of Jaccard Similarity ranges between 0 and 1, with 1 indicating perfect similarity between the predicted and ground truth labels.

Jaccard Similarity is frequently used in image classification to evaluate the performance of object detection models, such as semantic segmentation and instance segmentation models, as it provides a more accurate evaluation of the model's ability to detect and localize objects in the image [161]. Furthermore, Jaccard Similarity is a standard evaluation metric to assess the accuracy of the predicted segmentation [162]

3.9 Gradient-based Visual Explanation

Gradient-based visual explanation, also known as saliency mapping is the most common XAI in medical image analysis [13]. It refers to the process of generating interpretable and understandable representations (i.e. heat maps) of the underlying decision-making process in medical computer-aided diagnosis (CAD) systems. This can involve providing insights into why the system arrived at a particular diagnosis, highlighting key features that the system has considered. The goal of visual explanation is to increase the transparency, trust, and accountability of these systems, which are increasingly being used to support or even replace human decision-making in medical imaging.

Table 3.5: Formulation of gradient-based visual explanation methods.

Methods	**Attribution** $R_i^c(x)$
Class Activation Map (CAM)	$CAM_c = \sum_{(x,y)} w_c^k f_k(x,y)$
Gradient-based Class Activation Map (Grad-CAM)	$L^c_{Grad-CAM} = ReLU(\sum_k a_k^c A^k)$
Guided Gradient-based Class Activation Map (Guided Grad-CAM)	$\left\lvert \frac{\partial S_c(x)}{\partial x_i} \right\rvert \odot ReLU(\sum_k a_k^c A^k)$
Sensitivity Analysis (SA)	$\left\lvert \frac{\partial S_c(x)}{\partial x_i} \right\rvert$
Integrated Gradient (IG)	$(x_i - \overline{x}_i) \cdot \int_{\alpha=0}^{1} \frac{\partial S_c(\overline{x})}{\partial \overline{x}_i} \lvert \overline{x} = \overline{x} + \alpha(x - \overline{x}) da$

There are various gradient-based visual explanation as reviewed in Section 2.3. In Section 2.3.8, a comparative analysis of each methods was conducted. Based on Table 2.3, Sensitiviy Analysis (SA) was chosed as the primary method as it provides ease of implementation and its robustness. It is also because the SA is the predominant gradient-based visual explanation method. Furthermore, Table 3.5 illustrated that the formulation of gradient-based visual explanation methods. It can be seen that all variants of gradient-based visual explanation methods share a common formula, which is the partial derivative of the target out S_c with respect to the inputs x_i:

$$\frac{\partial S_c(x))}{\partial x_i} \tag{3.8}$$

With the facts and benefits from previous chapters, this research adopted SA as the baseline method, once successfuly, it can be applied easily to other gradient-based visual explanation method variants.

Chapter 4

Quantitative Saliency-Oriented Evaluation Metric

4.1 Introduction

Explainable Artificial Intelligence (XAI), particularly methods that employ gradient-based visual explanations, are increasingly incorporated into medical imaging analysis to clarify the outcomes yielded by deep learning architectures such as Convolutional Neural Networks (CNNs) [28–30]. However, as delineated in Chapter 1 Section 1.2, there is an imperative for formal metrics that assess the validity of heatmaps generated through gradient-based visual explanation techniques [12, 13, 18, 31, 32]. A principal challenge is the lack of established ground truth for these explanations, which hampers comprehensive evaluations [31, 33, 34].

This chapter addressed the above research problem by elucidating the technical aspects involved in creating a proposed metric for quantitative saliency evaluation. Initially, a pre-trained DenseNet model was employed for the predictive analysis of Diabetic Retinopathy (DR) using fundus photography images. Subsequently, Sensitivity Analysis (SA) was utilized to produce heatmaps that illuminate regions of interest within these fundus photography images that influenced the classification outcome. Lastly, the proposed metric quan-

titatively assessed these generated heatmaps against pixel-annotated ground truth images of diabetic lesions, thereby providing a numerical validity score for the heatmap.

This evaluation methodology established that heatmap analysis is both an empirical and objective exercise, underpinned by the tenets of post-hoc machine learning interpretability. This concept aimed to elucidate the mechanisms by which a pre-trained CNN generates predictions for any given set of inputs. The practical utility of this proposed metric has been substantiated within the realm of biomedical applications, specifically in classifying DR images. The method aimed to make meaningful advancements in improving both the accuracy and interpretability of machine learning applications in this crucial medical domain.

4.2 Methods & Materials

4.2.1 Datasets

Table 4.1 showed that the Kaggle Diabetic Retinopathy Dataset 2015 (KDRD2015) comprised of a total of 35,108 fundus photography images. These images were captured by various camera models, leading to artefacts such as contrast variations and noise. EyePACS [163] collected and annotated the fundus photography images on a 5-point scale, with the grades were divided into five imbalanced classes. Medical images are scarce, thus the rationale of using the KDRD2015 dataset is best as the dataset provided the largest DR data distribution. Moreover, prior literature in DR image classification also used KDRD2015 dataset as the benchmark of building, training, and evaluating CNN models [164–166].

The Indian Diabetic Retinopathy Image Dataset (IDRID) [139] was used as well in this study. Released in 2019, the IDRID dataset was part of the "Diabetic Retinopathy: Segmentation and Grading Challenge," held in association

Table 4.1: Data distribution of Kaggle Diabetic Retinopathy Dataset 2015 (KDRD2015)

Class	Category	Number of Images
0	No DR	25,802
1	Mild DR	2438
2	Moderate DR	5288
3	Severe DR	872
4	Proliferative DR	708
	Total	**35,108**

with the IEEE International Symposium on Biomedical Imaging. As outlined in Table 3.4, IDRID supported three key tasks in medical image analysis: DR grading, object detection, and image segmentation. The motivation for utilizing this dataset included: (1) the ability to use the IDRID set as a model inferencing dataset to assess the performance of the fine-tuned model, and (2) the provision of pixel-level annotated ground-truth images of DR lesions, useful for evaluating the proposed metric focused on quantitative saliency-oriented visual explanations. Figure 4.1 showcased the pixel-level annotated ground-truth lesion images from the IDRID dataset.

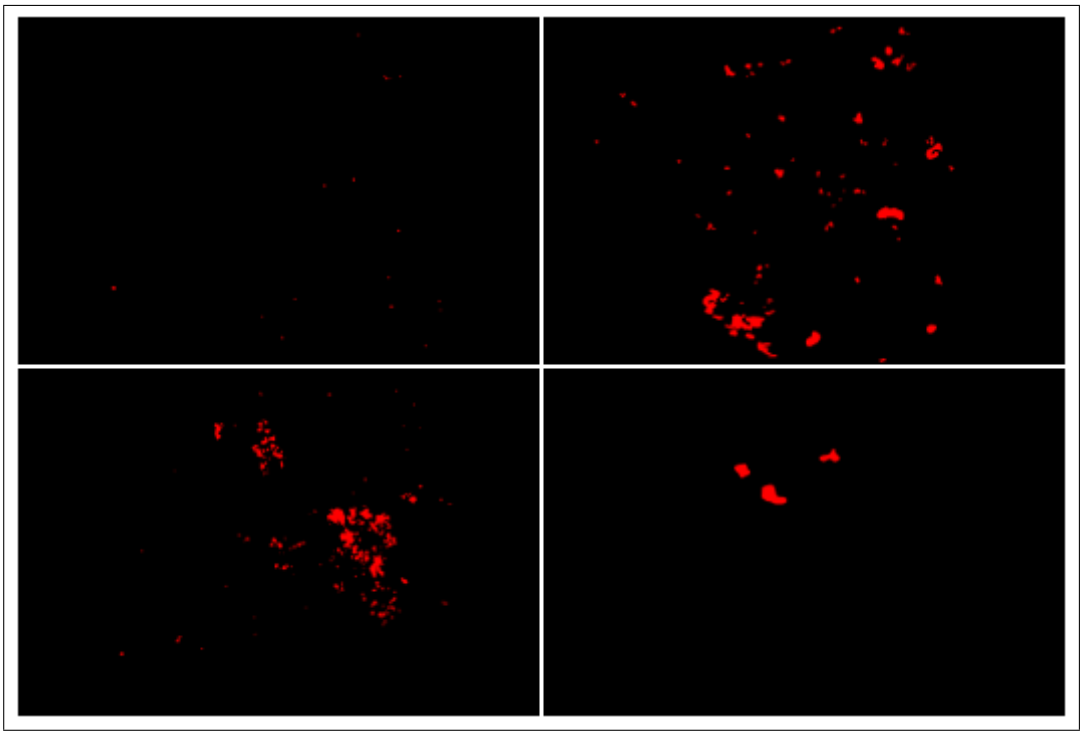


Figure 4.1: *Examples of pixel-level annotated ground truth images of diabetic lesions (From top left to bottom right: Microaneurysms (MA), Haemorrhages (HE), Hard Exudates (EX), Soft Exudates (SE))*

4.2.2 Data Preprocessing

Fundus photography images play a critical role in the diagnosis of Diabetic Retinopathy (DR). Recognizing DR-associated lesions in these images serves as a mandatory diagnostic approach for ophthalmologists. These lesions vary in terms of contrast, size, and texture within a single fundus photograph, as characterized in previous work [167]. The KDRD2015 consisted of color fundus photgraphy images that posed various challenges, such as inconsistent contrast levels and presence of noise [49]. To address these issues, image preprocessing techniques were employed in this study, aiming to enhance image quality and subsequently improve model performance.

In this study, the advanced image preprocessing technique known as Contrast Limited Adaptive Histogram Equalization (CLAHE) [142] was utilized. Preprocessing of the fundus photographs was executed based on a combination of CLAHE and additional image channels like the green and luminance channels. The green channel was specifically chosen because it offers enhanced visibility for lesions like microaneurysms, and provides high contrast between features such as blood vessels and the background [168, 169]. The luminance channel was selected as it aligns closely with human visual perception [170]. Consequently, the CIELAB color space was adopted to better simulate real-world conditions. Additionally, Gaussian-filtered Weighted Contrast, another preprocessing technique, was also employed in [171]. Therefore, the preprocessed fundus photography images included in Table 4.2.

Table 4.2: The different used of image preprocessing methods for DR image classification.

Proposed Image Preprocessing Methods	Description
Contrast Limited Adaptive Histogram Equalization (CLAHE)	1. The color space of the fundus photography images were first converted from RGB color to black and white color space. 2. CLAHE was then applied to the B/W fundus photography images.
CLAHE-Green Channel	1. The color space of the fundus photography images was first split into individual color channels of Red (R), Green (G), and Blue (B). The images with green channel was then extracted. 2. CLAHE was then applied to the green channel images.
CLAHE-Luminance Channel	1. The color space of the fundus photography images was first converted from RGB color space to CIELAB color space. 2. The LAB color space was then split into individual channels of L*, a*, and b*. 3. CLAHE was then applied to the L* channel and merged together with the a* and b* channels.
Gaussian Weighted Contrast [171]	A weighted Gaussian smoothing operation was applied to the fundus photography images that changed the contrast of the images.

Figure 4.2 depicted the image transformations corresponding to each of the previously mentioned image preprocessing techniques. In addition to these transformations, it is noteworthy that the black regions in the images were eliminated by locating the center and radius of the circle in the fundus photography images. Consequently, any fundus photography images that were nearly or completely black were manually removed from the dataset.

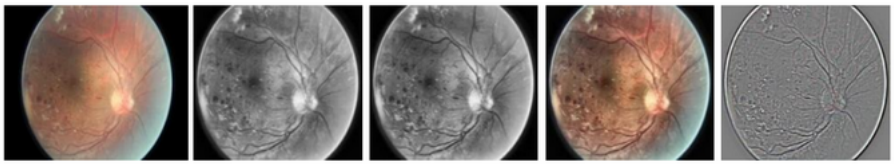


Figure 4.2: *Before and after preprocessing fundus photgraphy images (Left to Right: Original, CLAHE, CLAHE-G Channel, CLAHE-L Channel, and Gaussian-filtered Weighted Contrast)*

4.2.3 Data Augmentation

The Kaggle Diabetic Retinopathy Dataset 2015 (KDRD2015) displayed significant data imbalance. To counteract the risk of model overfitting, strategies involving data augmentation were implemented. In this study, the data augmentation methods employed consisted of image resizing, random image cropping, and random flipping horizontally, as well as random rotations ranging from 0 to 360 degrees. Furthermore, the RandAugment method described in Section 3.6 was used.

4.2.4 Implementation Details

Model Building, Training, and Testing

The Kaggle Diabetic Retinopathy Dataset 2015 (KDRD2015) was segmented into training and testing subsets, comprising 80% and 20% of the data, respectively. The KDRD2015 dataset went through image preprocessing phase using the proposed image preprocessing methods mentioned in Table 4.2. Moreover, the KDRD2015 dataset displayed significant data imbalance. To counteract

the risk of model overfitting, strategies involving data augmentation were implemented. In this study, the data augmentation methods mentioned in Section 4.2.3 were employed.

In training phase, transfer learning approach was employed. The majority of the pre-trained DenseNet model's parameters were preserved in their original state to maintain previously learned features, while only the terminal layer underwent training. In this study, the use case was treated as a binary classification problem. In other words, the fine-tuned DenseNet model classified fundus photography images into non-referable or referable classes. Referable Diabetic Retinopathy (DR) refers to any fundus images that the severity is more than mild are considered as referable [172]. Therefore, the KDRD2015 dataset was categorized into two classes: No DR (Class 0) as the non-referable class whereas Mild DR (Class 1), Moderate DR (Class 2), Severe DR (Class 3), and Proliferative DR (Class 4) as the referable class.

A weighted categorical cross-entropy loss was utilized as the preferred loss function. Aside from data augmentation to tackle data imbalance, class weights were incorporated into the CNN training as well. The weight of a class, W_c, was defined as follows:

$$Class\,Weights, W_c = \frac{Number\,of\,images\,of\,the\,largest\,class}{Number\,of\,images\,of\,class\,c} \tag{4.1}$$

The class weights were taken into consideration during the update of the models' weights through backpropagation. The minority classes were given a higher weight compared to the majority classes. Thus, the class weights reduced the risk of model overfitting. The Adaptive Momentum Optimization (Adam) was employed for model training. The learning rate was initially set to 0.003. The model was trained over a span of 200 epochs. Throughout the process, two optimization methodologies, namely ReduceLROnPlateau [173] and EarlyStopping [174] were employed. The ReduceLROnPlateau optimization method diminished the learning rate when a monitored metric ceases to

Table 4.3: The model hyperparameters that best fit the model classification.

Hyperparameters	**Values**
Batch Size	16
Loss Function	Cross Entropy Loss
Optimizer	- Adam - ReduceLROnPlateau - EarlyStopping
Initial Learning Rate	0.003

improve, whereas the EarlyStopping method halted the model training when a monitored metric no longer shows signs of improvement. Both these optimization methodologies monitor the validation loss metric. If the validation loss does not exhibit improvement over 10 epochs, the ReduceLROnPlateau method decreases the learning rate by a factor of 0.1. In case the validation loss fails to show progress with the new learning rate after 20 epochs, the model training is halted, implying the application of early stopping.

The performance of the fine-tuned DenseNet model was evaluated using the model performance metrics explained in Section 3.8. After the evaluation of the fine-tuned DenseNet model, the model was used as model inferencing using unseen images such as the IDRID fundus photography images to make predictions. The fine-tuned model tried to accurately classify the IDRID images. Then, the gradient-based visual explanation method called Sensitivity Analysis (SA) generated heatmaps that delineated the regions of interest in the image that contributed to the classification. Here, the proposed quantitative saliency-oriented evaluation metric from Chapter 4 was used as well to quantify the generated heatmaps to ensure the validity of the heatmaps.

Quantitative Saliency-oriented Evaluation Metric

As delineated in the Problem Statement (Section 1.2), the XAI field lacked a formalized evaluation metric for evaluating the heatmaps generated by gradient-based visual explanation techniques [12, 13, 18, 31, 32]. In other words, existing heatmaps suffered from validity concerns in terms of their alignment with accurate explanations and user expectations.

This research contributed by innovatively presenting a quantitative evaluation metric focused on saliency to assess the attribution maps (i.e. heatmaps). These maps are critical in spotlighting the salient regions within fundus photography images that influence the decision-making process of CNN. The Sensitivity Analysis (SA) method was employed as the visual explanation technique, substantiated by its merits outlined in Section 2.3.8. A notable advantage of the SA method is its computational efficiency—it necessitates only a single forward and backward pass through a Convolutional Neural Network (CNN) to construct the attribution maps. Moreover, its scalability is independent of the number of input features, and it easily integrates with established machine learning frameworks like TensorFlow [116] and PyTorch [117]. This ease of implementation comes from the method's capability to leverage existing GPU-accelerated gradient computations [32, 92].

The SA method operated by exploiting the backpropagation algorithm to identify the minimally mutable input components that most significantly impact the output prediction. This was achieved by computing the derivative of the class score concerning the input image, as formulated in Equation 2.2. Pixels with large-magnitude derivatives were indicative of having a substantial influence on the class score and correspondingly pinpoint the object's spatial location within the image. Consequently, the SA method served as a form of weakly-supervised object localization due to its ability to approximate object locations while solely relying on class labels for training.

To quantitatively assess the heatmaps generated via SA, the study introduced a metric rooted in the Jaccard Similarity, detailed in Section 3.8.6. This metric quantified the extent of overlap between the pixel-level annotated diabetic lesions in the IDRID dataset and the produced attribution maps. Specifically, the Jaccard Similarity score calculated the ratio of the intersecting pixels between the ground truth and the top-activated areas to the union

of pixels across both masks.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{4.2}$$

, where A is the pixel-level ground-truth mask and B is the attribution map. Figure 4.3 illustrated the operational sequence of the proposed quantitative saliency-oriented evaluation metric.

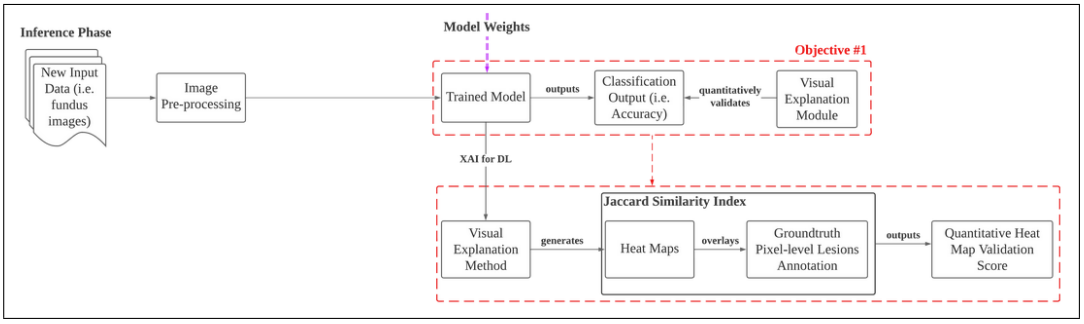


Figure 4.3: *A workflow of the proposed quantitative saliency-oriented evaluation metric.*

4.3 Result

4.3.1 Model Performance

The Kaggle Diabetic Retinopathy Dataset 2015 (KDRD2015) was preprocessed using different types of image preprocessing methods. The different types of preprocessed images were then used to fine-tuned the pre-trained DenseNet model. A comparative analysis of the classification results using different kind of image preprocessing methods was illustraated in Table 4.4.

Table 4.4: Comparative analysis of classification results using various proposed image preprocessing methods.

Proposed Image Preprocessing Methods	**Precision**		**Recall**		**F1 Score**		**Accuracy**
	c0	**c1**	**c0**	**c1**	**c0**	**c1**	
CLAHE	0.82	0.41	0.71	0.56	0.76	0.47	**0.67**
CLAHE-Green Channel	0.75	0.35	0.84	0.24	0.79	0.28	**0.68**
CLAHE-Luminance Channel	0.83	0.52	0.82	0.55	0.82	0.53	**0.75**
Gaussian Weighted Contrast	0.86	0.56	0.82	0.64	0.84	0.60	**0.77**

In Table 4.4, it can be observed that the pre-trained DenseNet model performed the best using the Gaussian weighted contrast images. The model achieved a classification accuracy of 77%, with a F1-score of 84% for the non-referable class (c0) and 60% for the referable class (c1). The second most effective was using the CLAHE-Luminance channel images, trailed by the CLAHE-Green channel images, and finally, the CLAHE images. the application of diverse image processing techniques to fundus images can bolster the classification performance of the classifier. Consequently, we infer that image quality is a critical factor in image classification, as elucidated by [49]. The loss curve of the pre-trained DenseNet in Figure 4.4 indicated that the model learned efficiently well using the Gaussian weigted contrast images. The loss initially decreased, however, remain stagnant. Approximately near to epoch 20, the learning rate decreased with the help of ReduceLROnPlateau method and the loss started to decread again. Approximately at epoch 80, second activation of ReduceLROnPlateau method was able to reduce the learning rate once more. In contrast to accuracy curve (Figure 4.5), it was indicated that the accuracy increased as well as epoch increased.

The fine-tuned DenseNet model was also evaluated using confusion matrix and the ROC AUC curves. The confusion matrix in Figure 4.6 showed that the fine-tuned DenseNet model was able to efficiently predict non-referable class (c0) while decently well on referable class (c1). The ROC AUC curve in Figure 4.7 also showed the efficiency of the fine-tuned DenseNet model using the Gaussian weighted contrast images.

4.3.2 Quantitative Saliency-oriented Evaluation Metric Performance

After the evaluation of the fine-tuned DenseNet model, the model was then used for model inferencing. Model inferencing is a process of using the fine-tuned model to infer a result from another new, unseen data. In this case,

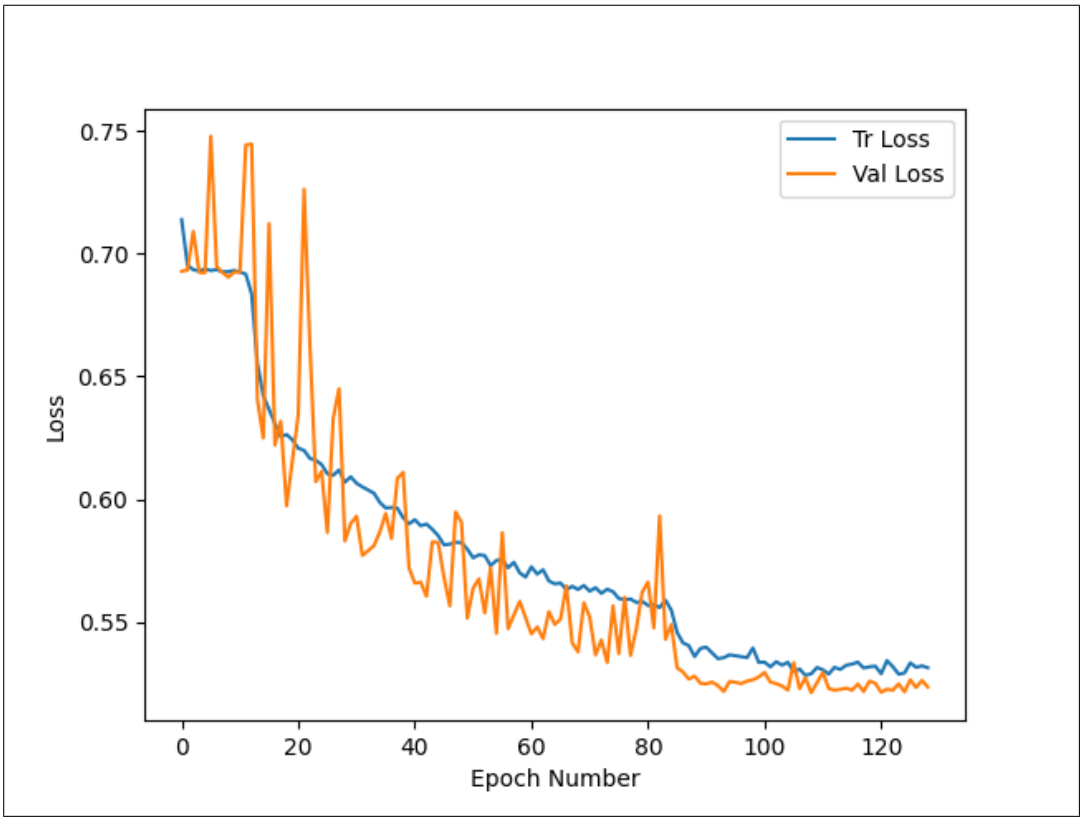


Figure 4.4: *Training and testing loss of the fine-tuned DenseNet model using the Gaussian weight contrast images.*

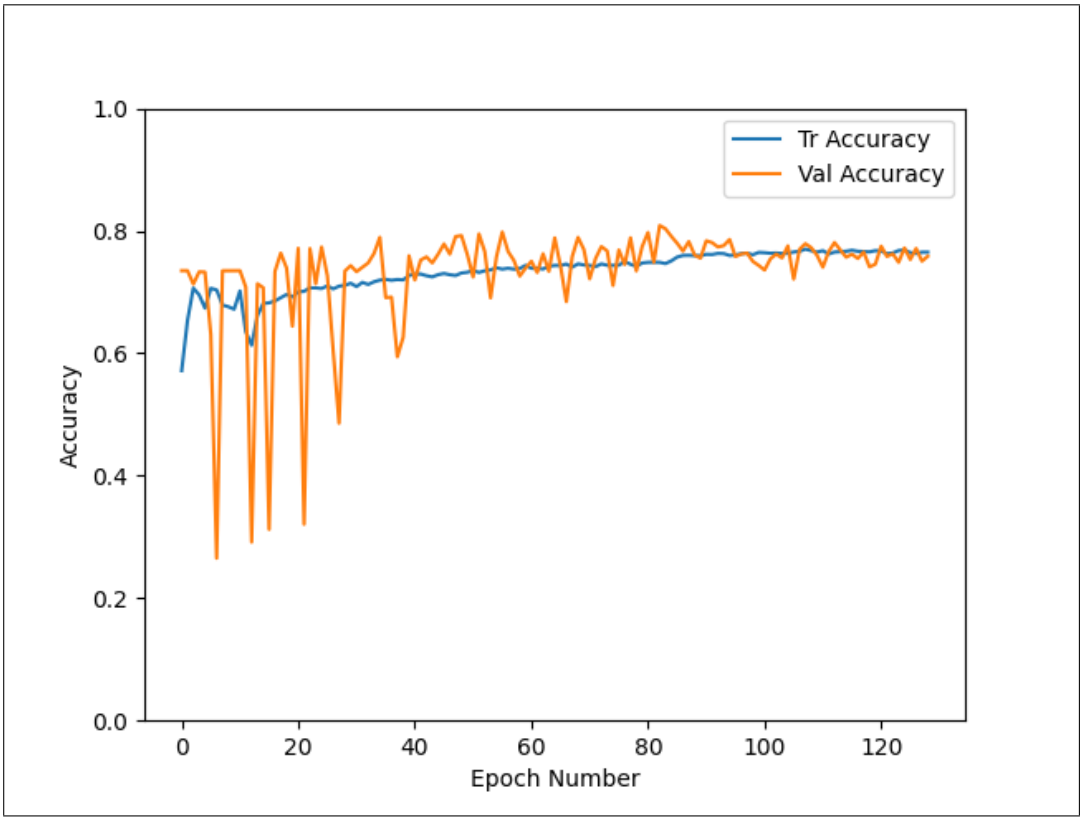


Figure 4.5: *Training and testing accuracy of the fine-tuned DenseNet model using the Gaussian weight contrast images.*

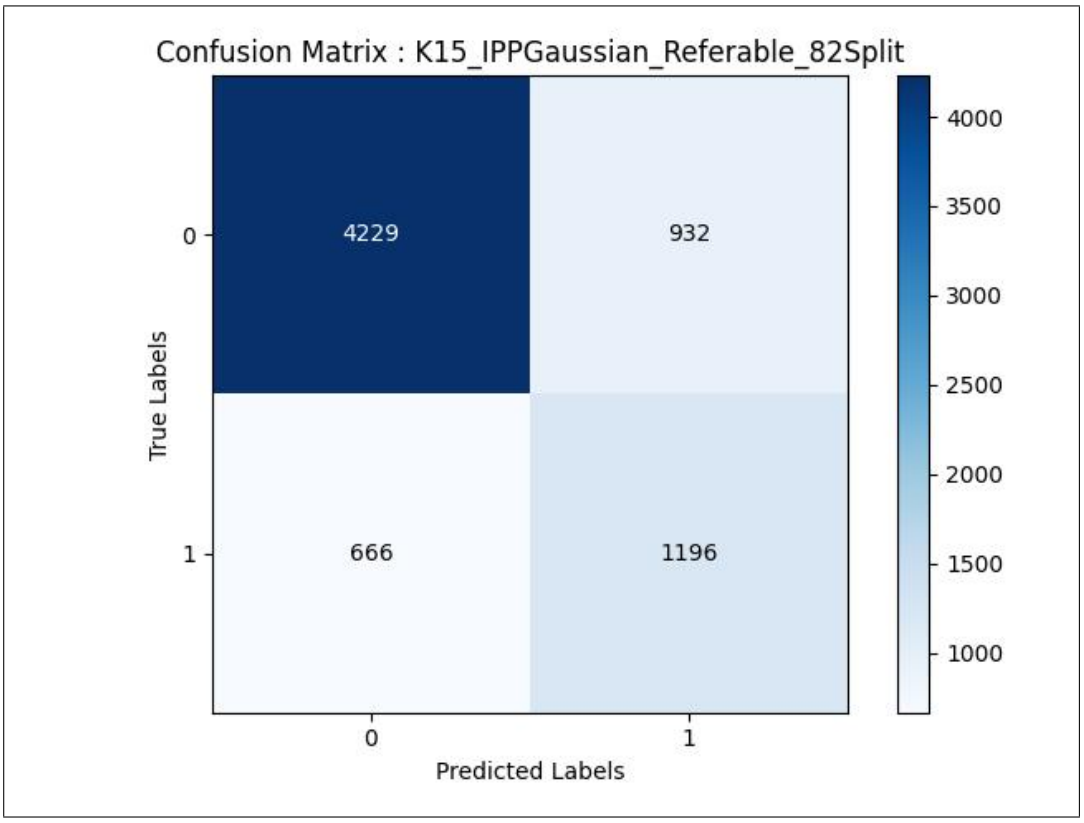
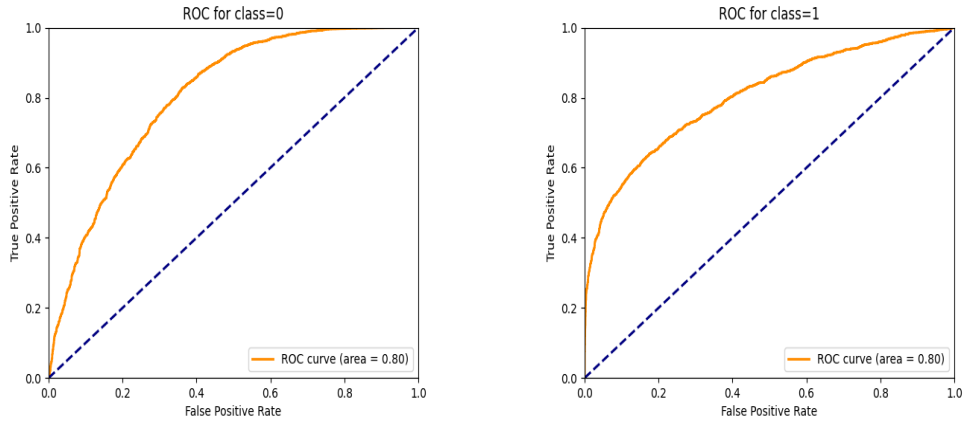


Figure 4.6: *The confusion matrix of the fine-tuned DenseNet model using the Gaussian weight contrast images.*



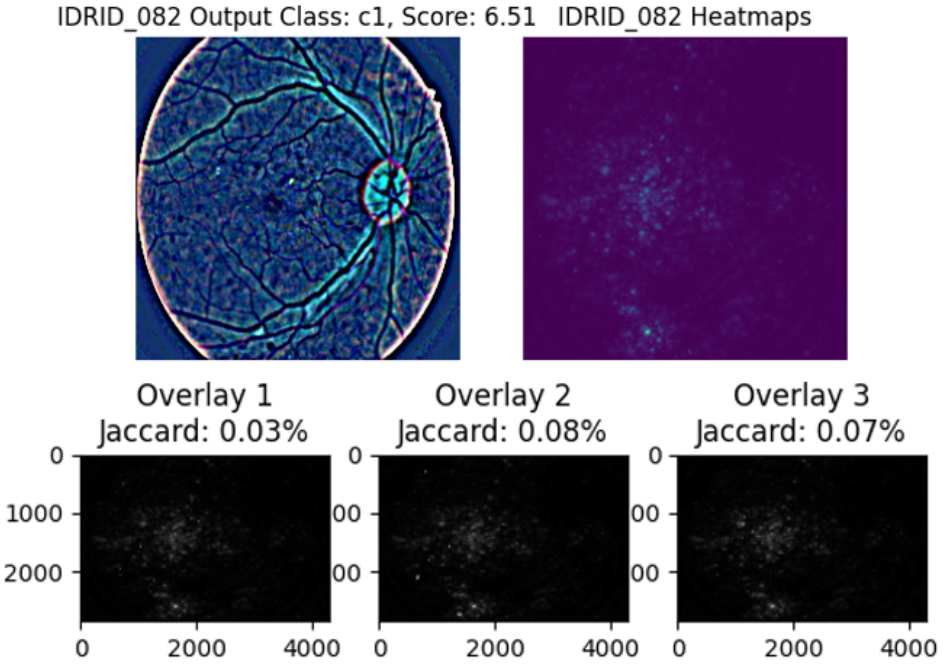
(a) *AUC ROC curve for non-referable class (Class 0)* **(b)** *AUC ROC curve for referable class (Class 1)*

Figure 4.7: *AUC ROC curve for non-referable and referable classes using the Gaussian weighted contrast images.*

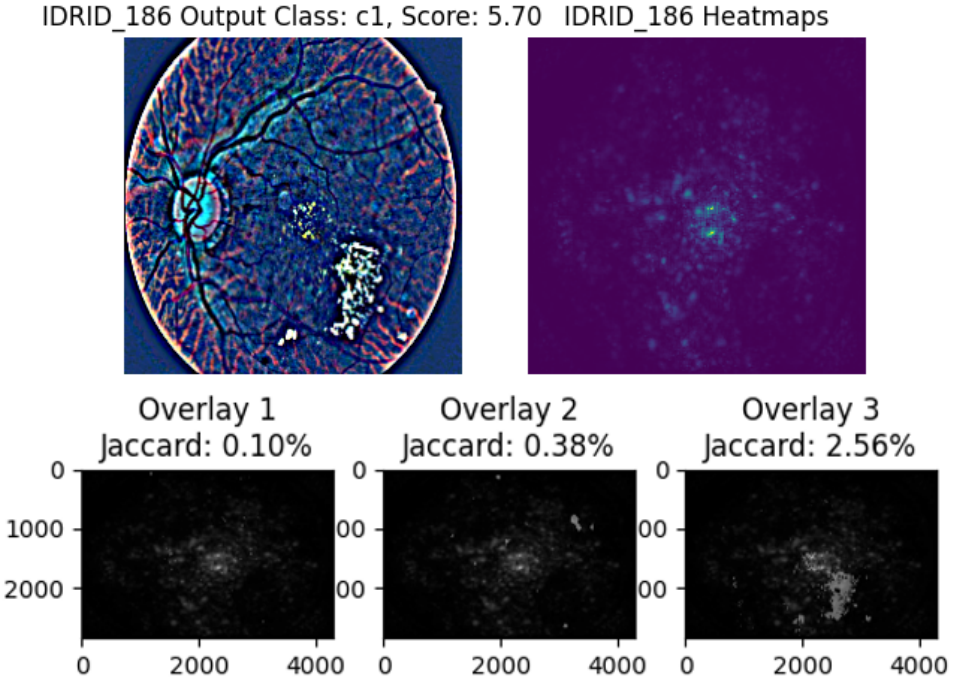
a number of images from the Indian Diabetic Retinoapthy Image Dataset (IDRID) was used. The fine-tuned model infers and produces the prediction class and a score of a given image. The higher the score, the accurate the prediction is. Having a prediction score may not be trustworthy, thus the gradient-based visual explanation method Sensitivity Analysis (SA) was used to generate heatmaps to offer insights into the predictions subsequent to the fine-tuned DenseNet model. Figure 4.8 showed a few example results that the fine-tuned model inferred the IDRID images.

The visual elucidation yielded by the gradient-based visual explanation method Sensitivity Analysis (SA) demonstrated that the fine-tuned DenseNet model was capable of acquiring visual features pertinent to the classification task that was undertaken during the training phase. In Figure 4.8, it can be observed that the heatmaps were able to accentuate pixels that played a significant role in the inferencing process. Pixels exhibiting brighter coloration are indicative of their contribution to the final inferencing outcome.

Next, the generated heatmaps and the pixel-level annotated ground truth images of diabetic lesions provided by the IDRID dataset were overlayed together. With the use of the proposed quantitative saliency-oriented evaluation metric, the method measured the congruence between the heatmaps and the IDRiD pixel-level annotated ground truth masks. Scores were given to indicate the validity of the generated heatmaps. For example, the white regions in the overlay images indicated that those were the features that the fine-tuned DenseNet model look at while infer the image's respective classes. Nonetheless, it was also notable that the grayer regions inidicated the possibility of some noisy pixels emerged within the visualization. This is true to IDRID image 186 and IDRID image 246.

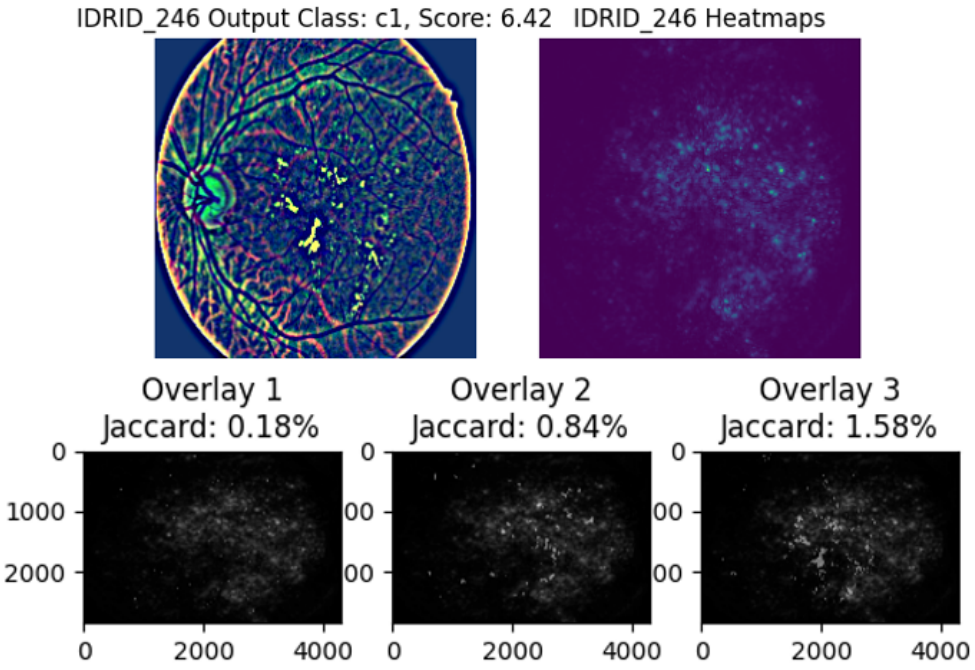


(a) *IDRID Image: IDRID 082*



(b) *IDRID Image: IDRID 186*

Figure 4.8: *Example results of prediction score with heatmaps and quantitative saliency-oriented evaluation score.*



(c) *IDRID Image: IDRID 246*

Figure 4.8: *Example results of prediction score with heatmaps and quantitative saliency-oriented evaluation score.*

Chapter 5

SLIC-G — An Optimised Image Preprocessing Method

5.1 Introduction

High-quality medical imagery used as input data to fine-tune Convolutional Neural Networks (CNNs) have shown to contribute positively to image classification performance [50–52]. To procure such quality in medical images, pixel-based image preprocessing methods such as Contrast Limited Adaptive Histogram Equalization (CLAHE) [142] has been the preferred method to preprocess Diabetic Retinopathy (DR) fundus photography images and those images preprocessed by it improved classification outcomes [142]. Despite the effectivenss of pixel-based image preprocessing methods, two crucial challenges inherent in them: (1) the discretization of image information and (2) the computational inefficiency of numerous algorithms when confronted with expansive image dimensions containing a high pixel count [55, 56].

Prompted by the limitations associated with pixels, this chapter endeavored to present an innovative image preprocessing method called SLIC-G, that employed Simple Linear Iterative Clustering (SLIC) superpixel segmentation and a Gaussian smoothing operation. The SLIC superpixel segmentation partitions the images into regions that provide meaningful descriptions with sig-

nificantly less data, as opposed to utilizing all pixels in an image [55, 56]. This method also reduces local redundancy within an image while preserving local structures and circumventing noisy, excessively segmented fragments [55, 56]. The Gaussian smoothing operation enhances image sharpness by eliminating image noise. To the best of our knowledge, there exists a lack of research implementing superpixel segmentation as an image preprocessing tool specifically on Diabetic Retinopathy (DR) fundus photographs.

5.2 Methods & Materials

5.2.1 Datasets

The Asia Pacific Tele-Ophthalmology Society 2019 (APTOS2019) dataset was used for fine-tuning the pre-trained DenseNet model while the Indian Diabetic Retinopathy Image Dataset (IDRID) was used for model inferencing. The rationale of using a new and recent dataset was because the data distribution of APTOS 2019 was significantly smaller as compared to the Kaggle Diabetic Retinopathy Dataset 2015 (KDRD2015) (refer to Table 3.3). With a smaller dataset, the opprtunity was given to adopt a data-centric approach on how to use this small dataset efficiently following and answering the questionnaires in the DC-Check checklist and produce on-par classification results.

The APTOS2019 dataset comprised of a total of 3,662 fundus photgraphy images. Technicians from the Aravind Eye Hospital in India travelled to rural areas to capture fundus photography images and then rely on highly trained doctors to review the images and provide diagnosis. Similar to the KDRD2015, the APTOS2019 fundus photography images were reviewed and categorized in to five imbalanced classes. The APTOS2019 dataset was split into 80% training set and 20% testing set. The data distribution of the APTOS2019 dataset was illustrated in Table 5.1.

Table 5.1: Data distribution of the APTOS 2019 (APTOS2019).

Class	Category	Number of Images
0	No DR	1,805
1	Mild DR	370
2	Moderate DR	999
3	Severe DR	193
4	Proliferative DR	295
	Total	**3,662**

5.2.2 Data Preprocessing

To solve the aforementioned problem of pixel-based image preprocessing methods, this section explained the proposed image preprocessing method called SLIC-G. SLIC-G is a method that involved the fusion of Simple Linear Iterative Clustering (SLIC) superpixel segmentation and Gaussian smoothing kernel.

The SLIC superpixel segmentation method is the most well-liked and effective superpixel segmentation technique [146]. This method is straightforward and effective for creating useful image regions. In order to produce uniformly sized and compact superpixels that accurately depict the organization of objects in an image, clustering was used. Each pixel in the SLIC method was represented by a five-dimensional vector. The color was represented by the vector's three components, and its position by its two components. The CIELAB color space, which provides a good similarity measure for color perception, is used to represent the color of the pixels. The key contribution of SLIC superpixel segmentation is that the vector components can effectively differentiate objects in an image [8]. Similarity between pixels is defined as:

$$D = d_c + m\frac{d_s}{s} \tag{5.1}$$

, where d_c and d_s are the color and spatial differences. Given two points with co-ordinates (x_i, y_j) and (x_i, y_j) with the color components (l_i, a_i, b_i) and

(l_j, a_j, b_j), we have that:

$$d_s(p_i, p_j) = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \tag{5.2}$$

$$d_c(p_i, p_j) = \sqrt{(l_j - l_i)^2 + (a_j - a_i)^2 + (b_j - b_i)^2} \tag{5.3}$$

The distance d_s is normalized using the constant s. In SLIC, the initial centre of a region is defined in a regular grid with sides s. As a result, the normalization d_s/s defines the distance in terms of number of interval lengths. The distance adds more to the similarity increasing the compactness as the grid size is reduced. Independent of grid size, the constant m determines how compact the system is. The greater the value of m the more special proximity is emphasized.

The SLIC superpixel segmentation requires simple three steps to create the regions:

1. Creates initial regions according to a parameter that defines the desired number of superpixels.

2. Performs region clustering to aggregate pixels to the regions according to the similarity criteria.

3. Reinforces connectivity.

Figure 5.1 depicted the operational sequence by which the proposed image preprocessing method was applied to both datasets. The images were presented in the RGB color space. In a practical context, ophthalmologists diagnose the severity of the disease by identifying crucial lesions within the images through visual perception. Consequently, all images underwent a transformation of color space. The images were converted from the RGB color space to the CIELAB color space, which more closely aligns with human perception [170]. Thus, the CIELAB color space was adopted to replicate the real-world use

case. The L* channel symbolizes lightness, ranging from black to white on a scale of zero to 100, while the a* and b* channels denote dchromaticity without any specific numeric limitations.

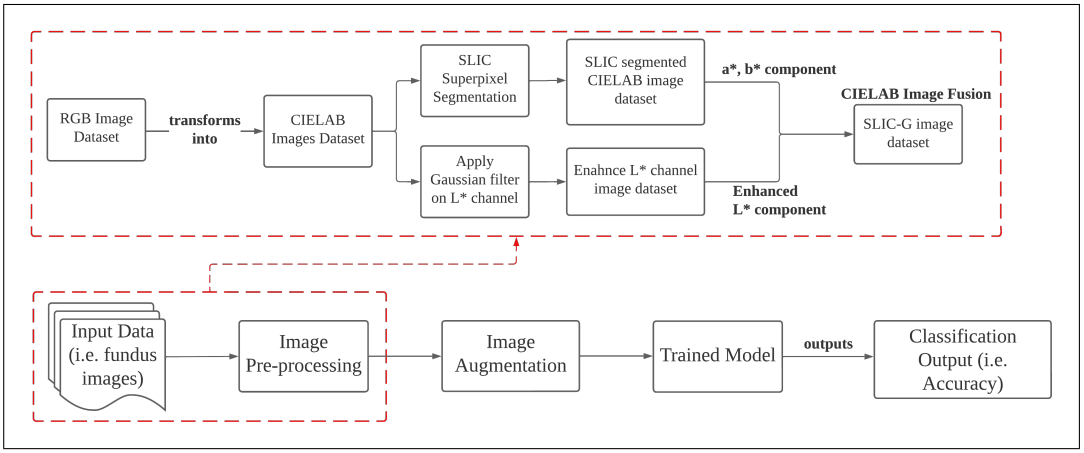


Figure 5.1: *A workflow of the proposed image pre-processing module.*

Following the conversion of all images to the CIELAB color space, two separate operations were executed. Firstly, the SLIC superpixel segmentation will extract local clusters utilizing both spectral and spatial similarity measures. Subsequently, a Gaussian kernel will traverse the L* channel of all images to accentuate the features within these images. Finally, the a* and b* channels of the SLIC segmented images and the Gaussian-filtered L* channel will be merged to form the SLIC-G images. Figure 5.2 visually represents a retinal image that has undergone the proposed hybrid image processing method and displays the resulting outcome.

5.2.3 Data Augmentation

The Asia Pacific Tele-Ophthalmology Society 2019 (APTOS2019) displayed significant data imbalance. To counteract the risk of model overfitting, strategies involving data augmentation were implemented. In this study, the data augmentation methodologies employed consisted of image resizing, random image cropping, and random flipping horizontally and vertically with a probability of 0.5, as well as random rotations ranging from 0 to 360 degrees. RandAugment method was used as well.

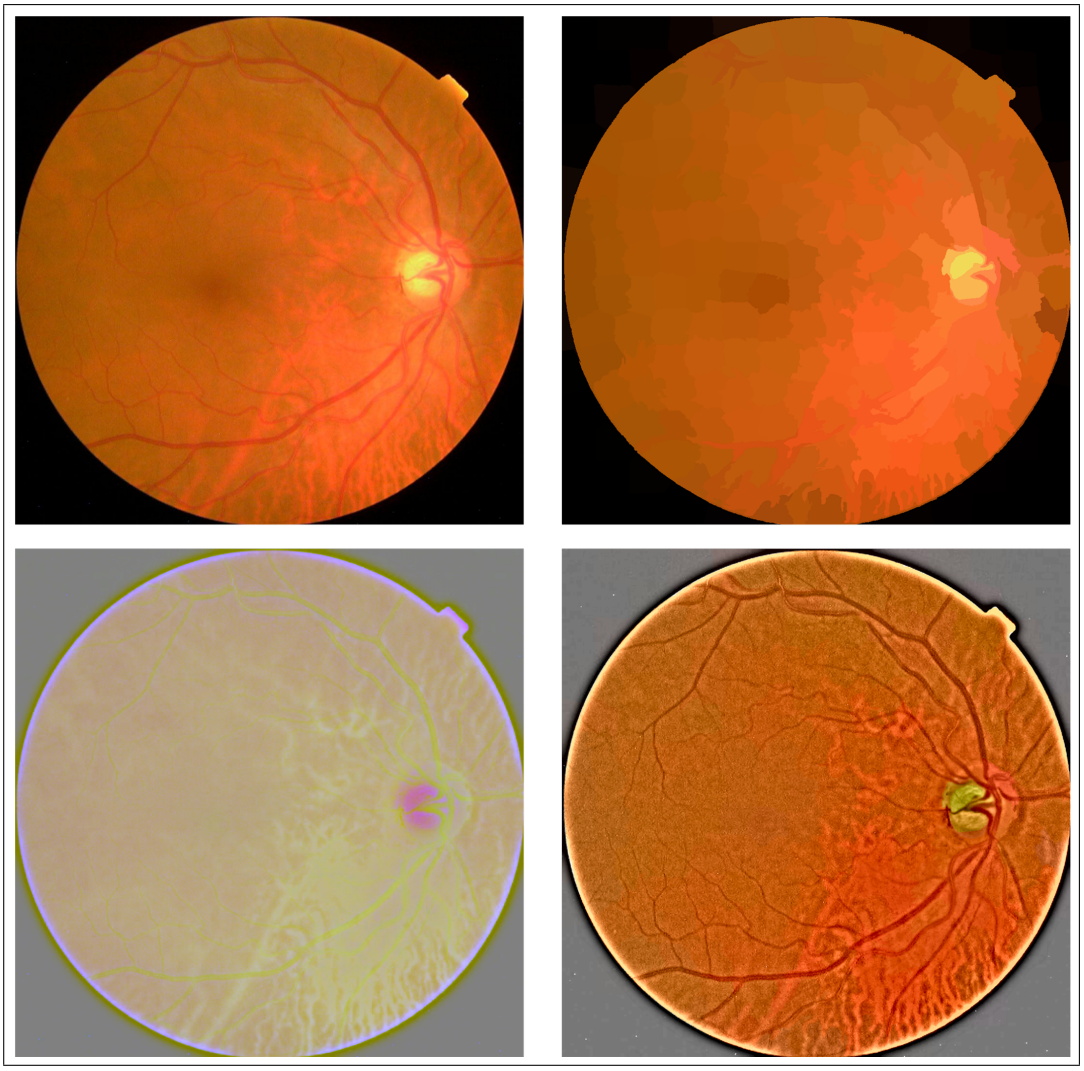


Figure 5.2: *A sample retinal image that applied with the proposed hybrid image processing image method. (From top left to bottom right: Original retinal image, SLIC segmented retinal image, Gaussian-filtered CIELAB retinal image, SLIC-G retinal image).*

5.2.4 Implementation Details

Model Building and Training

The machine learning procedure started off by carefully analyzing the procured APTOS2019 dataset. Here, the DC-Check checklist was adopted and a series of questions were carefully answered. The DC-Check checklist was helpful in identifying the necessary considerations on improving the quality of the dataset. The checklist can be viewed in the Appendix.

Once the dataset was thoroughly analyzed and the checklist was filled, then started the implementation. The dataset went through image preprocessing phase and was preprocessed using the proposed SLIC-G image preprocessing method as explained in Section 5.2.2. Furthermore, the APTOS2019 dataset was imbalanced in nature, thus data augmentation was applied to the dataset as described in Section 5.2.3.

The pre-trained DenseNet model was used and transfer learning method was adopted to fine-tuned the model. After several iterations through the experimental try-outs, the following hyperparameters illustrated in Table 5.2 were used for model building and training.

In this study, the use case was treated as a multiclass classification problem. A weighted categorical cross-entropy loss was utilized as the preferred loss function. Aside from data augmentation to tackle data imbalance, class weights were incorporated into the CNN training as well. The weight of a class, W_c, was defined as follows:

$$Class\,Weights, W_c = \frac{Number\,of\,images\,of\,the\,largest\,class}{Number\,of\,images\,of\,class\,c} \tag{5.4}$$

The class weights were taken into consideration during the update of the models' weights through backpropagation. The minority classes were given a higher weight compared to the majority classes. Thus, the class weights reduced the risk of model overfitting. Moreover, the Adaptive Momentum Optimization

Table 5.2: The model hyperparameters that best fit the model classification.

Hyperparameters	**Values**
Batch Size	16
Loss Function	Cross Entropy Loss
Optimizer	- Adam - ReduceLROnPlateau - EarlyStopping
Initial Learning Rate	0.003

(Adam) was employed for model training. The learning rate was initially set to 0.003. The model was trained over a span of 200 epochs. Throughout the process, two optimization methodologies, namely ReduceLROnPlateau [173] and EarlyStopping [174] were employed. The ReduceLROnPlateau optimization methodology diminishes the learning rate when a monitored metric ceases to improve, whereas the EarlyStopping methodology halts the model training when a monitored metric no longer shows signs of improvement. Both these optimization methodologies monitor the validation loss metric. If the validation loss does not exhibit improvement over 10 epochs, the ReduceLROnPlateau method decreases the learning rate by a factor of 0.1. In case the validation loss fails to show progress with the new learning rate after 20 epochs, the model training is halted, implying the application of early stopping.

After the evaluation of the fine-tuned DenseNet model, the model was used as model inferencing using unseen images such as the IDRID fundus photography images to make predictions. The fine-tuned model tried to accurately classify the IDRID images. Moreover, the proposed quantitative saliency-oriented evaluation metric was used to explain the fine-tuned model's decision making.

5.3 Result

5.3.1 Model Performance

Figure 5.3 illustrated the loss and the accuracy curve of fine-tuned DenseNet model using the APTOS2019 SLIC-G dataset. Based on observation, the

training and the testing loss were constantly decreasing. This implied that the fine-tuned DenseNet model was learning. Approximately at epoch 30, the training and testing loss decreased again. This phenomenon suggested that the learning rate had been reduced by the ReduceLROnPlateau optimization method due to no improvement on the testing loss. At around epoch 50, the model experienced overfitting, the EarlyStopping optimization method was triggered, and the training stopped at epoch 70. As for the accuracy curve, it was observed that the training and testing accuracy gradually increased. Similar to the phenomenon of second improvement of loss, the accuracy curve also had second improvement at approximately at epoch 30.

Figure 5.4 illustrated the confusion matrix of the SLIC-G APTOS2019 dataset. Based on observation, the fine-tune DenseNet model efficiently identified the images in Class 0 (No DR) and Class 3 (Moderate DR). However, early stages of Diabetic Retinopathy (DR) such as Class 1 (Mild DR) was difficult to detect irrespective of whether that class falls under minority category or not. In addition, ROC-AUC performance metrics were used to analyze the model. Figure 5.5 illustrated the ROC-AUC curve for different classes. The ROC-AUC curves that drawn above the threshold suggested that the model performed well. Despite the confusion matrix result, it is confident that the model is efficient in classifying each classes.

Table 5.3 shows the comparative analysis between the experiment result and the result from prior published work that uses APTOS 2019 dataset. The hyper-parameters employed in the experiments conducted by the researchers cited in [65, 152, 175] were not clearly outlined. Consequently, the hyper-parameters for this proposed method were meticulously adjusted to optimize the outcomes. The values of the hyper-parameters were initially set at a baseline and then incrementally raised until no further improvements in the results were observed. The results of the proposed method were then evaluated against those of the deep networks referenced in [65, 152, 175] using performance metrics such as accuracy, Cohen's kappa, and the geometric mean.

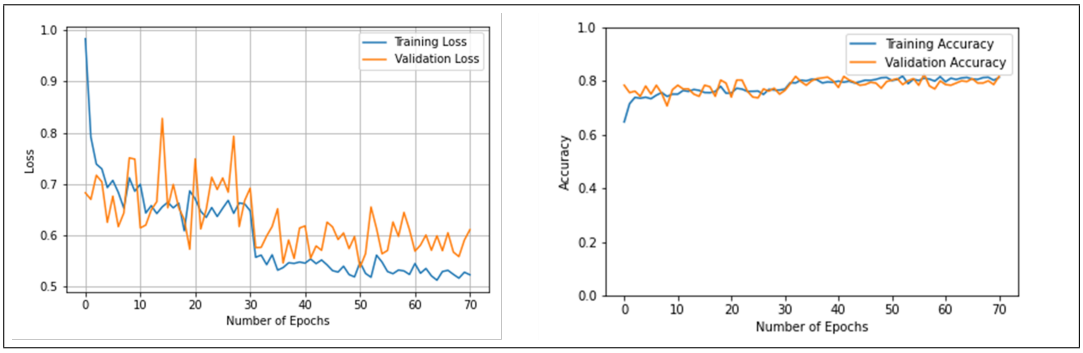


Figure 5.3: *The training and testing loss and accuracy curve on SLIC-G APTOS 2019 dataset.*

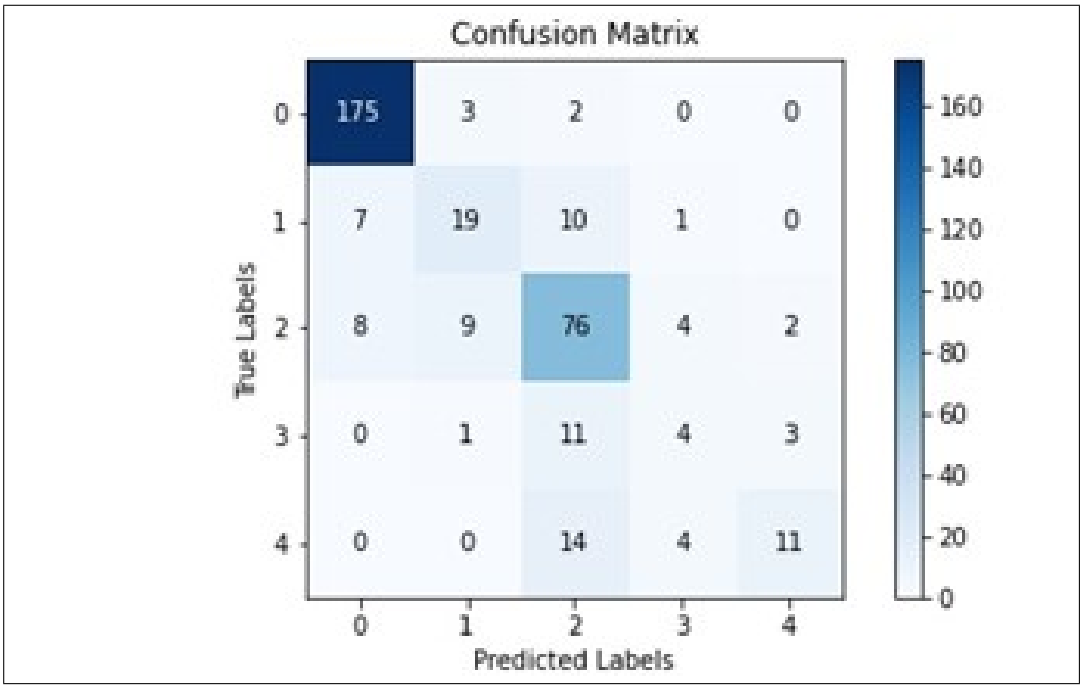


Figure 5.4: *Training and testing loss curve and testing accuracy curve on SLIC-G APTOS 2019 dataset.*

Table 5.3: The comparative analysis between the proposed method and other state-of-the-art pre-trained CNNs.

Methods	Accuracy	Cohen's Kappa	Geometric Mean
MobileNet [175]	0.790	0.770	N/A
VGG16 [65]	0.800	0.787	N/A
ResNet50 [152]	0.746	0.786	N/A
Proposed Method	**0.816**	**0.795**	**0.790**

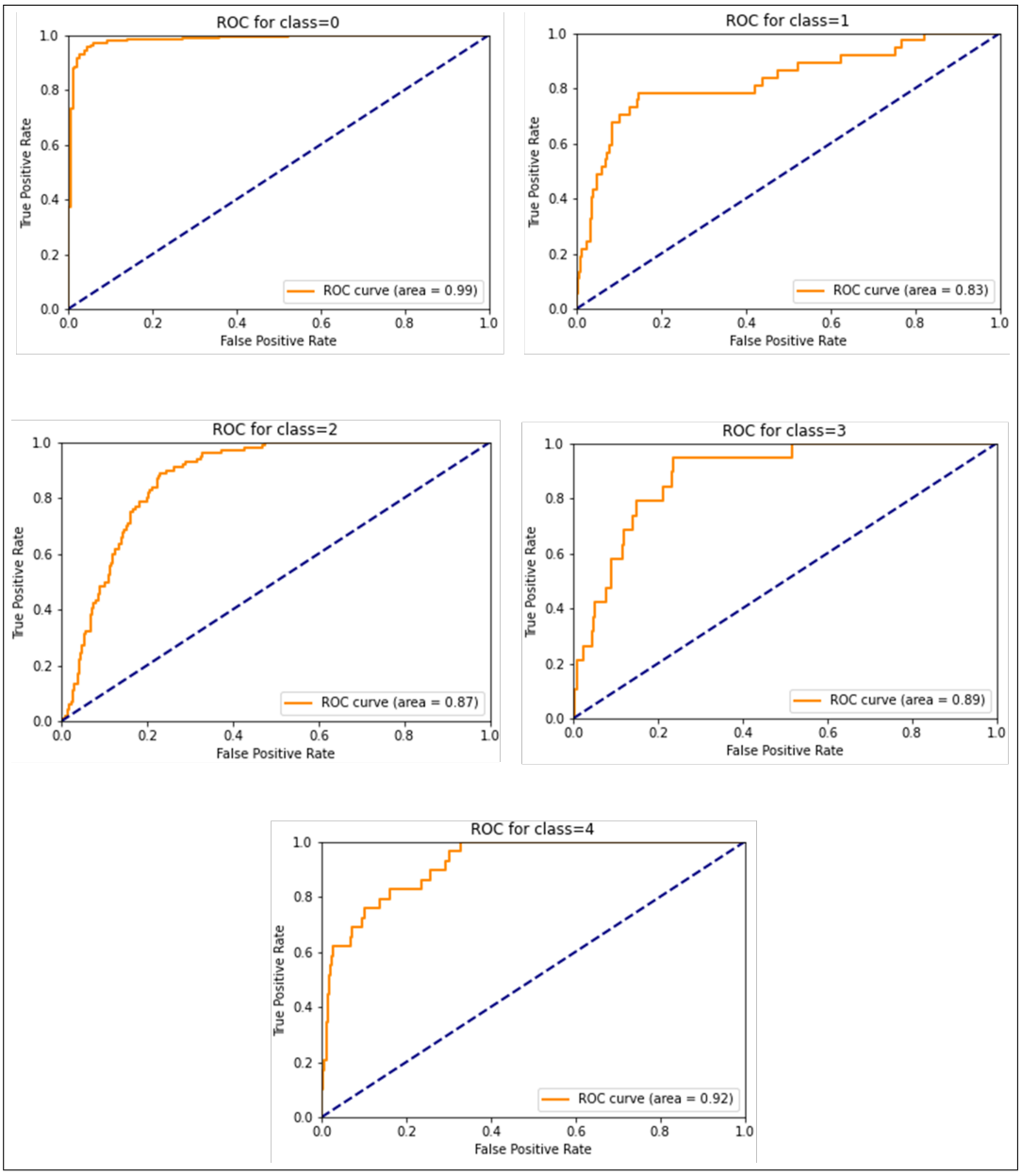
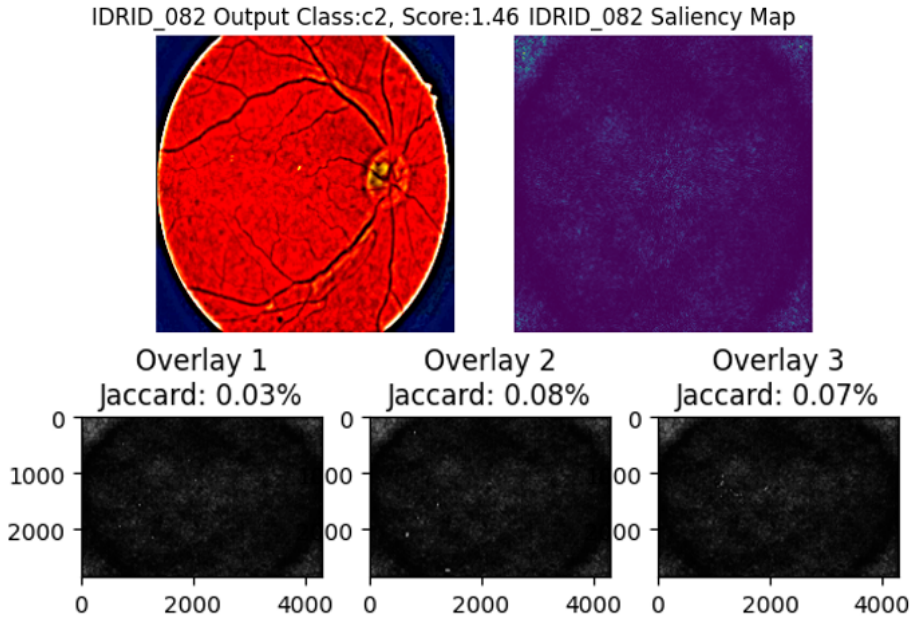


Figure 5.5: *Training and testing loss curve and testing accuracy curve on SLIC-G APTOS 2019 dataset.*

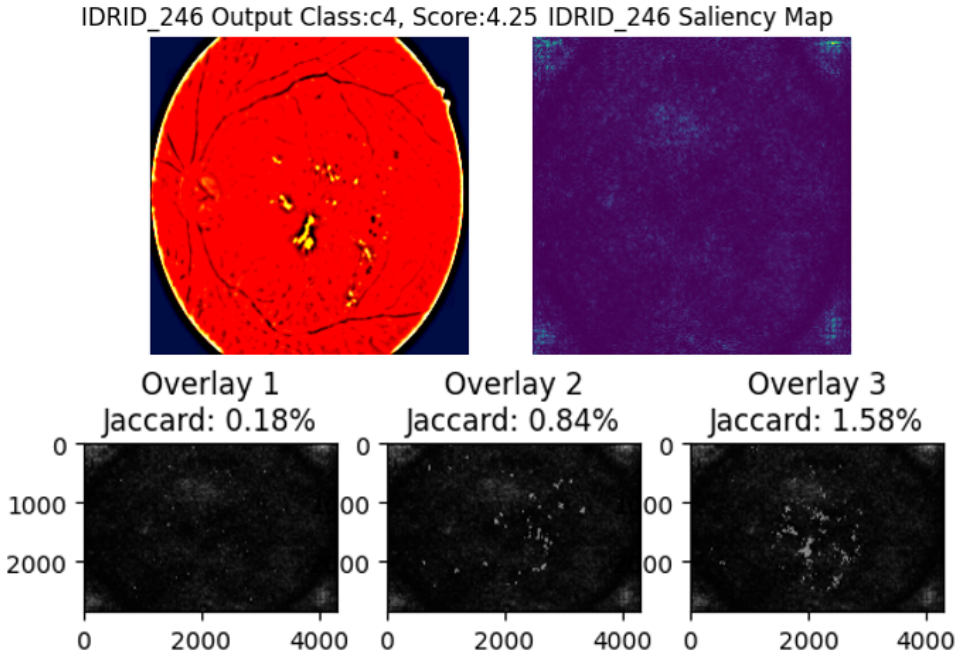
5.3.2 Quantitative Saliency-oriented Evaluation Metric Performance

The visual elucidation yielded by the gradient-based visual explanation method Sensitivity Analysis (SA) demonstrated that the fine-tuned DenseNet model was capable of acquiring visual features pertinent to the classification task that was undertaken during the training phase. In Figure 5.6, it can be observed that the heatmaps were able to accentuate pixels that played a significant role in the inferencing process. Pixels exhibiting brighter coloration are indicative of their contribution to the final inferencing outcome.

Next, the generated heatmaps and the pixel-level annotated ground truth images of diabetic lesions provided by the IDRID dataset were overlayed together. With the use of the proposed quantitative saliency-oriented evaluation metric, the method measured the congruence between the heatmaps and the IDRiD pixel-level annotated ground truth masks. Scores were given to indicate the validity of the generated heatmaps. For example, the white regions in the overlay images indicated that those were the features that the fine-tuned DenseNet model look at while infer the image's respective classes. However, it can be observed that the amount of generated heatmaps was higher than the result in Figure 4.8. Significant heatmaps were accentuated around the corner of the images. This is true to IDRID image 186 and IDRID image 246.

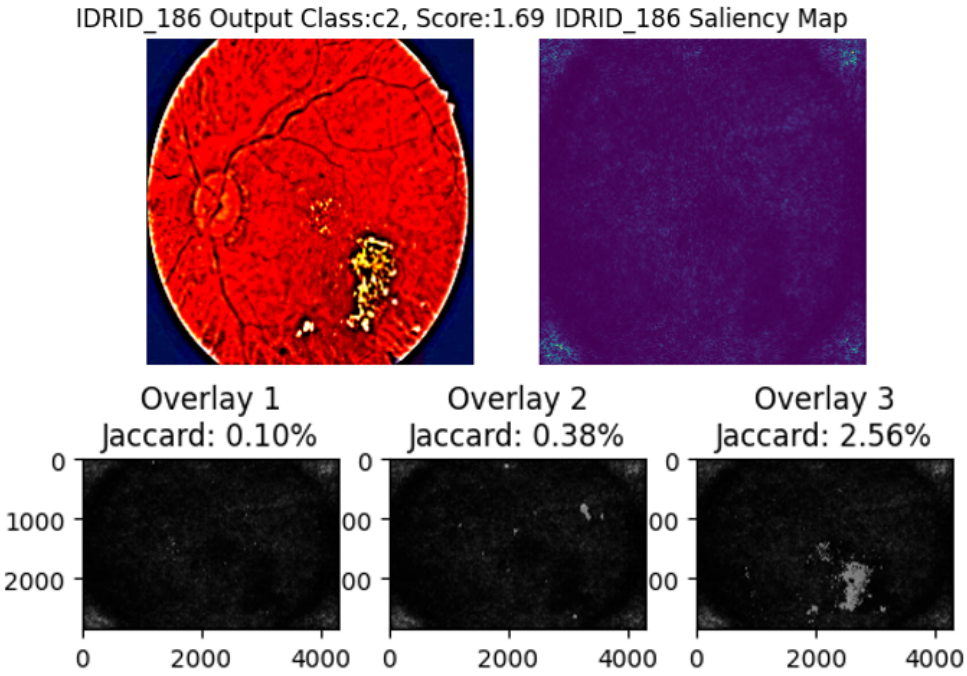


(a) *IDRID Image: IDRID 082*



(b) *IDRID Image: IDRID 186*

Figure 5.6: *Example results of prediction score with heatmaps and quantitative saliency-oriented evaluation score.*



(c) *IDRID Image: IDRID 246*

Figure 5.6: *Example results of prediction score with heatmaps and quantitative saliency-oriented evaluation score.*

Chapter 6

General Discussion

6.1 Research Summary

The objective of this research was to tackle several obstacles impeding the effective implementation of Diabetic Retinopathy (DR) image classification via supervised learning methods. Specifically, deep learning architectures like Convolutional Neural Networks (CNNs) presented a challenge in terms of explainability, offering little to no transparency in how they arrived at decisions [11–18]. To improve transparency, Explainable Artificial Intelligence (XAI) approaches, including gradient-based visual explanation methods, were utilized to produce attribution maps that illuminate crucial areas influencing model decisions [28–30]. Nonetheless, these gradient-based methods exhibited a limitation: the generated heatmaps maps lacked validity due to the absence of a standardized quantitative evaluation metric [12, 13, 18, 31, 32].

To address this issue, Chapter 4 presented a novel, cost-efficient, and effective quantitative evaluation metric that was saliency-oriented. This metric served to offer a quantitative score that reinforced the credibility of these heatmaps. The score empirically validated the accuracy of the heatmaps in emphasizing important regions. Therefore, this evaluation framework confirmed that the analysis of heatmaps was not merely based on subjective interpretation but was also substantiated by empirical data.

Additionally, existing scholarly work has identified a critical challenge in the realm of Explainable Artificial Intelligence (XAI), which was the system's reliance on the caliber of the training data [37–39]. Factors such as data bias, data incompleteness, and inaccuracies can jeopardize the quality and thus the trustworthiness of the explanation. The consequences of using subpar data are evident not only in the system's performance but also in its capacity to furnish dependable explanations [40]. To address this, a novel image preprocessing technique named SLIC-G was proposed, incorporating Simple Linear Iterative Clustering (SLIC) superpixel segmentation with a Gaussian smoothing operation to enhance image quality. Experimental outcomes revealed that the utilization of this preprocessing approach led to improvements in the performance of the fine-tuned DenseNet model when compared to results presented in existing literature.

This chapter presents a discussion of the methods proposed, examining both their innovative aspects and limitations. It also outlined potential avenues for future research to mitigate some of these identified shortcomings.

6.2 Validity of Visual Explanation Methods

In Chapter 4, the architecture of DenseNet was employed as the underlying pre-trained model for feature extraction from fundus photography images. Fine-tuning of this DenseNet model was conducted using the Kaggle Diabetic Retinopathy Dataset 2015 (KDRD2015). Four distinct image preprocessing techniques were applied to the dataset before utilizing it for model fine-tuning. The empirical results, as delineated in Table 4.4, using the fundus photography images preprocessed by the Gaussian Weighted Contrast image preprocessing method achieved the highest model accuracy at 77%. This finding underscores that the critical role of image preprocessing in the training of deep learning models, corroborating the observations made in [49]. Furthermore, using Gaussian Weight Contrast method to preprocess fundus photography images

allows the diabetic lesions to amplify, allowing the pre-trained model to better extract features for model training, rather than using the conventional Green channel extraction method [168, 169].

In model inferencing phase, the fine-tuned model was able to infer the Asia Pacific Tele-Ophthalmology Society 2019 (APTOS2019) images. The prediction score, along wih generated heatmaps by the gradient-based visual explanation Sensitivity Analysis (SA) were able to accentuate the areas of importance that contributed to the prediction. To tackle the lack of validation strategy, as highlighted in the Problem Statement (Section 1.2), the proposed cost-effective and efficient quantitative saliency-oriented evaluation metric was able to empirically validate the heatmaps.

In the study cited as [102], the adoption of Sensitivity Analysis (SA) was beneficial in improving the interpretability of the model's predictive process by accentuating critical regions within images that significantly impact the model's classification decisions for diabetic retinopathy. This enhancement in transparency is notably valuable in scenarios where grasping the credibility behind the model's choices is as critical as the predictions themselves, particularly when the model is tasked with accurately categorizing the severity of diabetic retinopathy by relying on key features (i.e. diabetic lesions) present in the images. Comparatively, the proposed saliency-oriented evaluation method extended these benefits by not only highlighting areas of importance but also providing a framework for heatmap validation. This approach involved coupling Jaccard Similarity method with XAI methods like SA enabled us to explicitly validate the pixels that significantly influence classification. The quantitative element offered evidence that the highlighted areas are truly significant and not artefacts of the model or the data processing.

Nonetheless, it was noted that although the suggested quantitative saliency-based evaluation method succeeded in generating empirical scores that support the credibility of the heatmaps, the resulting Jaccard Similarity scores were modest. This could be attributed to the Jaccard Similarity metric comparing

two images pixel-by-pixel — the heatmap generated by the SA image and the ground truth image at the pixel level of diabetic lesions. Given that the regions affected by diabetic lesions are typically small, this could influence the Jaccard Similarity score. This phenomenon presents an opportunity for future work to investigate the correlation between the area of the diabetic lesions and Jaccard Similarity scores. The correllation between the generated heatmaps and the quantitative scores opens up a new research problem. Medical image analysis often involves results based on correlation rather than causation. Causality refers to the study of how things influence one another, how causes lead to effects, and how this relationship can be mathematically described [176]. With the inclusion of heatmaps and quantitative validity scores, researchers or machine learning engineers are able to use them to improve model performance and minimize biases, as prototypes might still be sensitive to differences in training populations, potentially hindering generalizability [177].

6.3 Superpixel-based Image Preprocessing

In Chapter 5, the issue of data quality, which has implications for the effectiveness of Convolutional Neural Networks (CNNs), was tackled in real-world scenarios, medical images display a significant degree of variability, noise, and temporal fluctuations [49]. These variations can stem from various factors, including disparate imaging protocols, patient cohorts, and environmental variables, among others. Prior research has indicated that image preprocessing plays a vital role in mitigating noisy artifacts and enhancing image classification performance [50–52]. Techniques for image preprocessing at the pixel-level—such as Contrast Limited Adaptive Histogram Equalization (CLAHE) [142], Gaussian filtering, green channel extraction, and grayscale conversion—were utilized to enhance image quality [53, 54]. However, these pixel-based approaches were constrained by inherent challenges in digital image preprocessing, notably, information discretization and computational limi-

tations when dealing with high-resolution images with an abundant pixel count [55, 56].

In addition, another issue was that the machine learning community has predominantly honed algorithms for state-of-the-art CNNs, primarily employing benchmark datasets. This focus has often come at the expense of overlooking critical elements pertinent to real-world applications, such as data curation, quality assessment, and subgroup characterization [3]. There is a lack of standardized guidelines detailing the design intricacies of data-centric machine learning pipelines.

The proposed solution in Chapter 5 has twofolds. First, an innovative image preprocessing technique called SLIC-G was developed. This method integrated Simple Linear Iterative Clustering (SLIC) for superpixel segmentation with Gaussian smoothing operation to mitigate the challenges associated with pixel-based image preprocessing. Secondly, a structured methodology, named DC-Check, was implemented. This checklist aimed to underscore the importance of data-centric considerations throughout the entire machine learning process, from the data acquisition stage through to the deployment of the final model.

The Asia Pacific Tele-Ophthalmology Society 2019 (APTOS2019) and Indian Diabetic Retinopathy Image Dataset (IDRID) were used in the study. Before model building and training procedure starts, the APTOS2019 first underwent analysis using the proposed structured methodology DC-Check. The DC-Check provided a series of questions that need to be answered fully to ensure all considerations were met in order to produce high-quality dataset. Next, the datasets underwent image preprocessing using the proposed SLIC-G image preprocessing method. The fundus photography images preprocessed by the proposed SLIC-G was then used to fine-tune the proposed backbone pre-trained DenseNet.

The experimental results in Section 5.3 showed that the training and validation loss followed a downward trend while the training and validation accu-

racy exhibited an upward trend (Figure 5.3. This indicated that the fine-tuned DenseNet model is genuinely learning. The comparative analysis in Table 5.3 unequivocally demonstrated that the proposed method surpassed other state-of-the-art CNNs, such as MobileNet, VGG, and ResNet. The combination of SLIC superpixel and Gaussian filtering operation, as opposed to standard image processing, enhanced accuracy score, ROC AUC score, and geometric mean. The utilization of Gaussian filtering operation aids in reducing image noise, and by transforming the image into the CIE L* a* b* color space, it yields better results after combining with the SLIC superpixel.

The proposed method also performed well due to its utilization of the CNN DenseNet [4]. The DenseNet model proved to be effective owing to its architectural design. DenseNet excels in circumventing the vanishing gradient problem and enabling feature reuse, achieving state-of-the-art results on the ImageNet, CIFAR, and SVHN data sets. The architecture of DenseNet consists of dense blocks where each layer receives concatenated data from all preceding layers, thus simplifying the residual network pattern. Furthermore, it necessitates fewer parameters than other convolutional networks. In contrast to ResNets, which also employ skip connections, DenseNets concatenate the output and input feature maps. Owing to input concatenation, the ease of access to learned maps for corresponding layers is enhanced. Fine-tuning on the pre-trained DenseNet aids in preserving salient image features while concurrently learning the diabetic retinopathy classification features.

Regrettably, as demonstrated in the confusion matrix presented in Figure 5.4, the performance of Class 1 (Mild DR), Class 3 (Severe DR), and Class 4 (Proliferative DR) was merely adequate compared to Class 0 (No DR) and Class 2 (Moderate DR). It can be posited that this is due to the fact that images from Class 1 (Mild DR) contain a minimal quantity of lesions, which leads to the CNN's confusion in differentiating between images from Class 1 (Mild DR) and Class 0 (No DR) [102, 166]. This explanation is also applicable to Class 3 (Severe DR) and Class 4 (Proliferative DR), as these two classes

exhibit similar lesion characteristics [102, 166].

In model inferencing phase, the fine-tuned model was able to infer the Asia Pacific Tele-Ophthalmology Society 2019 (APTOS2019) images. The prediction score, along wih generated heatmaps by the gradient-based visual explanation Sensitivity Analysis (SA) were able to accentuate the areas of importance that contributed to the prediction. To tackle the lack of formal validation strategy, as highlighted in the Problem Statement (Section 1.2), the proposed cost-effective and efficient quantitative saliency-oriented evaluation metric was able to empirically validate the heatmaps. While the proposed quantitative saliency-oriented evaluation metric was able to validate the generated heatmaps, there were a lot more heatmaps generated as compared to Figure 4.8. Other pixels of the heatmaps implies that those could be noise during the image preprocessing phase. This is true as the authors in [178–180] stated that some image preprocessing methods are able to reduce the image quality.

6.4 Limitations & Future Work

Although the proposed methodologies demonstrated promising results and generally outperformed other state-of-the-art and commonly employed methods, it is important to acknowledge a few limitations that warrant attention and could be addressed in future work.

Gradient-based visual explanation methods have become a popular tool for elucidating the opaque nature of Convolutional Neural Networks (CNNs) in medical image analysis. However, researchers are increasingly exploring alternative directions in XAI, with a focus on connecting XAI to causality [181]. Medical image analysis often involves results based on correlation rather than causation. Causality refers to the study of how things influence one another, how causes lead to effects, and how this relationship can be mathematically described [176]. Current XAI methods aimed to minimize biases, as prototypes

might still be sensitive to differences in training populations, potentially hindering generalizability. The authors in [182] provided an example of detecting dataset bias using XAI. In their study, they employed visual explanation to distinguish X-rays of patients who were COVID-19-positive from those who were COVID-19-negative. They found that the high performance of deep learning models was primarily attributed to the composition of the datasets rather than the actual detection of COVID-19 in the X-rays. Another noteworthy example of using causality is the work by [183], who eliminated bias through causality, resulting in unbiased predictions of prognosis for patients with lung cancer.

In medical image analysis, one of the most significant limiting factors is the lack of large, high-quality medical image datasets for model training. It is widely acknowledged that data is the primary driver for improving AI-powered systems. The quality of data not only affects algorithm performance but also influences prediction accuracy [184]. The machine learning community has predominantly focused on optimizing algorithms for state-of-the-art CNNs on benchmark datasets, often neglecting crucial factors relevant in real-world scenarios, such as data curation, data quality assessment and cleaning, and subgroup characterization. Consequently, adopting a data-centric approach is essential. The machine learning community should shift its focus from model-centric AI to data-centric AI.

To achieve data-centric AI, [42] suggested that employing foundational models, which are large models trained on massive amounts of unlabeled data and adaptable to various tasks. These large models include large language models [123, 124] and Stable Diffusion [185]. As models become sufficiently powerful, they can feasibly perform numerous data-centric AI tasks such as data labeling [124] and data augmentation [125]. The recent trend of foundational models has the potential to fundamentally alter our understanding of data. Instead of storing raw data values in datasets, the model itself can serve as container of data since it conveys information. Foundational models blur the boundary between data and model, but their training still heavily relies on

large and high-quality datasets. Secondly, data-centric AI can be achieved by exploring reinforcement learning. Reinforcement learning is a machine learning method that trains intelligent agents to optimize rewards without any initial data [186–188]. This unique learning paradigm alternates between generating data with the model and training the model with self-generated data. Similar to foundational models, the advancement of reinforcement learning could blur the boundary between data and model. Furthermore, reinforcement learning has already been widely adopted in several data-centric AI sub-goals, such as data labeling [189, 190], data preparation [191], data reduction [192], and data augmentation [150].

Furthermore, this study can also be conceptualized as a regression task because the output categories possess an inherent order, such as varying degrees of severity. This ordering facilitates the model's ability to differentiate between problematic classes within the confusion matrix. For example, in a binary classification scenario where the classes "A"and "B"are assigned numerical labels "0"and "1"respectively, employing regression techniques can enhance the prediction of these numerical labels. Finally, the datasets utilized in this research were not real-time datasets but rather open-sourced datasets employed by other research works. Almost all prior works that focused on DR image classification used these open-sourced datasets as benchmarks datasets. Although using benchmark datasets facilitates result comparability, such comparisons can be questionable due to the neglect of other factors. As [193] highlighted, comparing deep learning methods is challenging, as these methods may exhibit different performance when subjected to various preprocessing methods, different hardware, varying ratios between training and test set sizes, or alternative network configurations. It is recommended to conduct more comprehensive experiments using both benchmark datasets and real-time datasets, along with additional state-of-the-art methods.

Chapter 7

Conclusion

The opaque nature of Convolutional Neural Networks (CNNs) and the scarcity of quality data present significant obstacles that can impede the widespread adoption of Artificial Intelligence (AI) in various domains. To address these challenges, this thesis builds upon prior research from different perspectives. The contribution of this work can be divided into two main parts, each aimed at addressing one or more of the challenges of adopting CNNs in medical image analysis.

Upon examining the existing literature on Explainable Artificial Intelligence (XAI) and conducting a review of previous studies, it was concluded that gradient-based visual explanation methods are commonly utilized in medical image analysis. These methods are capable of generating visual explanation maps (i.e. heatmaps) that demonstrate the significance of image pixels to the model classification. The primary advantage of these methods is their ease of implementation. Moreover, another advantage is their reduced computational cost in comparison to perturbation-based methods. Despite their simplicity and computational efficiency, gradient-based visual explanation frequently produce ambiguous explanations, complicating their qualitative evaluation. Additionally, the absence of a ground-truth explanation presents a challenge for quantitative evaluation. Therefore, one contribution of this study was to propose a quantitative saliency-oriented evaluation metric for assessing the visual

explanation heat maps generated by gradient-based visual explanation. The experimental results indicated that the proposed evaluation metric can quantify the localized significant pixels that contributed to the image classification prediction.

This study further explored the obstacles associated with the implementation of CNNs by examining pixel-based image processing techniques. Although methods such as Contrast Limited Adaptive Histogram Equalization (CLAHE) have been extensively employed to enhance medical imagery, they have not yet addressed two critical issues inherent in pixel processing: the discretization of image information and the computational impracticality of numerous algorithms when managing vast image dimensions with a high pixel count. Consequently, this research introduces a superpixel-based image processing approach known as SLIC-G. The proposed SLIC-G method facilitates image segmentation into regions by taking into account similarity measures defined by perceptual features. The underlying objective is to acquire regions that convey meaningful descriptions while utilizing significantly less data than is required when employing all the pixels in an image. The rationale is that by diminishing the number of primitives, redundancy is minimized, thereby reducing the intricacy of recognition tasks. Moreover, superpixels supplant the inflexible pixel structure by demarcating regions that preserve meaning in the image, consequently enabling the regions to provide insights into the structure of the scene and simplifying other processing tasks compared to the use of individual image pixels. Experimental outcomes demonstrated that the proposed image processing technique allows CNNs to yield superior image classification results in comparison to previous literature. As a result, it can be surmised that segmenting an image into regions that represent meaningful descriptions diminishes the complexity of the image processing procedure and facilitates enhanced image classification outcomes.

References

[1] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proceedings of 2010 IEEE international symposium on circuits and systems*. IEEE, 2010, pp. 253–256.

[2] IBM. (2024) convolutionlayer. [Online]. Available: https://www.ibm.com/topics/convolutional-neural-networks

[3] N. Seedat, F. Imrie, and M. van der Schaar, "Dc-check: A data-centric ai checklist to guide the development of reliable machine learning systems," *arXiv preprint arXiv:2211.05764*, 2022.

[4] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[6] P. P. Shinde and S. Shah, "A review of machine learning and deep learning applications," in *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*. IEEE, 2018, pp. 1–6.

[7] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of big data*, vol. 2, no. 1, pp. 1–21, 2015.

[8] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of digital imaging*, vol. 32, pp. 582–596, 2019.

[9] D. Sarvamangala and R. V. Kulkarni, "Convolutional neural networks in medical image understanding: a survey," *Evolutionary intelligence*, vol. 15, no. 1, pp. 1–22, 2022.

[10] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC medical imaging*, vol. 22, no. 1, p. 69, 2022.

[11] D. Jin, E. Sergeeva, W.-H. Weng, G. Chauhan, and P. Szolovits, "Explainable deep learning in healthcare: A methodological survey from an attribution view," *WIREs Mechanisms of Disease*, vol. 14, no. 3, p. e1548, 2022.

[12] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," *Computers in biology and medicine*, vol. 140, p. 105111, 2022.

[13] B. H. Van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (xai) in deep learning-based medical image analysis," *Medical Image Analysis*, p. 102470, 2022.

[14] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of Imaging*, vol. 6, no. 6, p. 52, 2020.

[15] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[16] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[17] A. Rai, "Explainable ai: From black box to glass box," *Journal of the Academy of Marketing Science*, vol. 48, pp. 137–141, 2020.

[18] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[19] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[20] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[21] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, "A historical perspective of explainable artificial intelligence," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 1, p. e1391, 2021.

[22] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[23] L. Arras, F. Horn, G. Montavon, K.-R. Müller, and W. Samek, "" what is relevant in a text document?": An interpretable machine learning approach," *PloS one*, vol. 12, no. 8, p. e0181142, 2017.

[24] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[25] Q. Zhang, W. Wang, and S.-C. Zhu, "Examining cnn representations with respect to dataset bias," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[26] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 427–436.

[27] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, 2021.

[28] J. D. Fuhrman, N. Gorre, Q. Hu, H. Li, I. El Naqa, and M. L. Giger, "A review of explainable and interpretable ai with applications in covid-19 imaging," *Medical Physics*, vol. 49, no. 1, pp. 1–14, 2022.

[29] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," *Information Fusion*, vol. 77, pp. 29–52, 2022.

[30] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable ai: A brief survey on history, research areas, approaches and challenges," in *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*. Springer, 2019, pp. 563–574.

[31] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning–a brief history, state-of-the-art and challenges," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2020, pp. 417–431.

[32] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Gradient-based attribution methods," *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 169–191, 2019.

[33] G. Ras, N. Xie, M. Van Gerven, and D. Doran, "Explainable deep learning: A field guide for the uninitiated," *Journal of Artificial Intelligence Research*, vol. 73, pp. 329–396, 2022.

[34] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Communications of the ACM*, vol. 63, no. 1, pp. 68–77, 2019.

[35] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021.

[36] N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, M. Gidwani *et al.*, "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging," *Radiology: Artificial Intelligence*, vol. 3, no. 6, p. e200267, 2021.

[37] L. Bertossi and F. Geerts, "Data quality and explainable ai," *Journal of Data and Information Quality (JDIQ)*, vol. 12, no. 2, pp. 1–9, 2020.

[38] W. Saeed and C. Omlin, "Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities," *Knowledge-Based Systems*, vol. 263, p. 110273, 2023.

[39] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies," *Journal of biomedical informatics*, vol. 113, p. 103655, 2021.

[40] E. Reiter, "Natural language generation," *The handbook of computational linguistics and natural language processing*, pp. 574–598, 2010.

[41] T. R. Society. (2019) Explainable ai: the basics. [Online]. Available: https://royalsociety.org/topics-policy/projects/explainable-ai/

[42] D. Zha, Z. P. Bhat, K.-H. Lai, F. Yang, Z. Jiang, S. Zhong, and X. Hu, "Data-centric artificial intelligence: A survey," *arXiv preprint arXiv:2303.10158*, 2023.

[43] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, ""everyone wants to do the model work, not the data work": Data cascades in high-stakes ai," in *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–15.

[44] W. Liang, G. A. Tadesse, D. Ho, L. Fei-Fei, M. Zaharia, C. Zhang, and J. Zou, "Advances, challenges and opportunities in creating data for trustworthy ai," *Nature Machine Intelligence*, vol. 4, no. 8, pp. 669–677, 2022.

[45] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[46] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[47] Y. Li, D. Gu, Z. Wen, F. Jiang, and S. Liu, "Classify and explain: An interpretable convolutional neural network for lung cancer diagnosis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1065–1069.

[48] K. H. Lee, C. Park, J. Oh, and N. Kwak, "Lfi-cam: Learning feature importance for better visual explanation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1355–1363.

[49] B. Goyal, S. Agrawal, and B. Sohi, "Noise issues prevailing in various types of medical images," *Biomedical & Pharmacology Journal*, vol. 11, no. 3, p. 1227, 2018.

[50] B. Basavaprasad and M. Ravi, "A study on the importance of image processing and its applications," *IJRET: International Journal of Research in Engineering and Technology*, vol. 3, no. 1, 2014.

[51] M. Mateen, J. Wen, M. Hassan, N. Nasrullah, S. Sun, and S. Hayat, "Automatic detection of diabetic retinopathy: a review on datasets, methods and evaluation metrics," *IEEE Access*, vol. 8, pp. 48 784–48 811, 2020.

[52] T. Huang, J. Burnett, and A. Deczky, "The importance of phase in image processing filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 6, pp. 529–542, 1975.

[53] R. Sarki, K. Ahmed, H. Wang, and Y. Zhang, "Automatic detection of diabetic eye disease through deep learning using fundus images: a survey," *IEEE access*, vol. 8, pp. 151 133–151 149, 2020.

[54] A. Ahmad, A. B. Mansoor, R. Mumtaz, M. Khan, and S. Mirza, "Image processing and classification in diabetic retinopathy: A review," in *2014 5th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2014, pp. 1–6.

[55] D. Stutz, A. Hermans, and B. Leibe, "Superpixels: An evaluation of the state-of-the-art," *Computer Vision and Image Understanding*, vol. 166, pp. 1–27, 2018.

[56] M. Nixon and A. Aguado, *Feature extraction and image processing for computer vision*. Academic press, 2019.

[57] Kaggle. (2023) Find open datasets and machine learning projects. [Online]. Available: https://www.kaggle.com/datasets

[58] Google. (2023) Google datasets search. [Online]. Available: https://datasetsearch.research.google.com/

[59] I. Kandel and M. Castelli, "Transfer learning with convolutional neural networks for diabetic retinopathy image classification. a review," *Applied Sciences*, vol. 10, no. 6, p. 2021, 2020.

[60] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," *Advances in neural information processing systems*, vol. 2, 1989.

[61] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[62] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Neural networks for perception*. Elsevier, 1992, pp. 65–93.

[63] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 111–118.

[64] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*. IEEE, 2012, pp. 3304–3308.

[65] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[66] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[67] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[68] IEEE. (2023) Ieee dataport. [Online]. Available: https://ieee-dataport.org/

[69] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[70] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[71] Z. Huang, X. Du, L. Chen, Y. Li, M. Liu, Y. Chou, and L. Jin, "Convolutional neural network based on complex networks for brain tumor image classification with a modified activation function," *IEEE Access*, vol. 8, pp. 89 281–89 290, 2020.

[72] D. J. Hemanth, J. Anitha, A. Naaji, O. Geman, D. E. Popescu *et al.*, "A modified deep convolutional neural network for abnormal brain image classification," *IEEE Access*, vol. 7, pp. 4275–4283, 2018.

[73] W. H. Organization. (2023) Coronavirus disease (covid-19). [Online]. Available: https://www.who.int/health-topics/coronavirus#tab=tab_1

[74] T. J. H. H. The Johns Hopkins University and J. H. H. System. (2023) What is coronavirus? [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus

[75] X. Yu, S.-H. Wang, and Y.-D. Zhang, "Cgnet: A graph-knowledge embedded convolutional neural network for detection of pneumonia," *Information Processing & Management*, vol. 58, no. 1, p. 102411, 2021.

[76] T. Yan, P. K. Wong, H. Ren, H. Wang, J. Wang, and Y. Li, "Automatic distinction between covid-19 and common pneumonia using multi-scale convolutional neural network on chest ct scans," *Chaos, Solitons & Fractals*, vol. 140, p. 110153, 2020.

[77] T. Goel, R. Murugan, S. Mirjalili, and D. K. Chakrabartty, "Optconet: an optimized convolutional neural network for an automatic diagnosis of covid-19," *Applied Intelligence*, vol. 51, pp. 1351–1366, 2021.

[78] Y. Zhang, F. Li, F. Yuan, K. Zhang, L. Huo, Z. Dong, Y. Lang, Y. Zhang, M. Wang, Z. Gao *et al.*, "Diagnosing chronic atrophic gastritis by gastroscopy using artificial intelligence," *Digestive and Liver Disease*, vol. 52, no. 5, pp. 566–572, 2020.

[79] L. Zhang, Y. Zhang, L. Wang, J. Wang, and Y. Liu, "Diagnosis of gastric lesions through a deep convolutional neural network," *Digestive Endoscopy*, vol. 33, no. 5, pp. 788–796, 2021.

[80] H. Wang, S. Ding, D. Wu, Y. Zhang, and S. Yang, "Smart connected electronic gastroscope system for gastric cancer screening using multi-column convolutional neural networks," *International Journal of Production Research*, vol. 57, no. 21, pp. 6795–6806, 2019.

[81] S.-A. Lee, H. C. Cho, and H.-C. Cho, "A novel approach for increased convolutional neural network performance in gastric-cancer classification using endoscopic images," *IEEE Access*, vol. 9, pp. 51 847–51 854, 2021.

[82] Z. Yu, D. Ni, S. Chen, J. Qin, S. Li, T. Wang, and B. Lei, "Hybrid dermoscopy image classification framework based on deep convolutional neural network and fisher vector," in *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*. IEEE, 2017, pp. 301–304.

[83] U.-O. Dorj, K.-K. Lee, J.-Y. Choi, and M. Lee, "The skin cancer classification using deep convolutional neural network," *Multimedia Tools and Applications*, vol. 77, pp. 9909–9924, 2018.

[84] R. C. Maron, M. Weichenthal, J. S. Utikal, A. Hekler, C. Berking, A. Hauschild, A. H. Enk, S. Haferkamp, J. Klode, D. Schadendorf *et al.*, "Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks," *European Journal of Cancer*, vol. 119, pp. 57–65, 2019.

[85] S. Khan, N. Islam, Z. Jan, I. U. Din, and J. J. C. Rodrigues, "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning," *Pattern Recognition Letters*, vol. 125, pp. 1–6, 2019.

[86] K. Nazeri, A. Aminpour, and M. Ebrahimi, "Two-stage convolutional neural network for breast cancer histology image classification," in *Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15*. Springer, 2018, pp. 717–726.

[87] D. Lévy and A. Jain, "Breast mass classification from mammograms using deep convolutional neural networks," *arXiv preprint arXiv:1612.00542*, 2016.

[88] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International conference on machine learning*. PMLR, 2017, pp. 3145–3153.

[89] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[90] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[91] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[92] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[93] W. M. Gondal, J. M. Köhler, R. Grzeszick, G. A. Fink, and M. Hirsch, "Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 2069–2073.

[94] Z. Wang and J. Yang, "Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation," *arXiv preprint arXiv:1703.10757*, 2017.

[95] M. Ahmad, N. Kasukurthi, and H. Pande, "Deep learning for weak supervision of diabetic retinopathy abnormalities," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 573–577.

[96] H. Jiang, J. Xu, R. Shi, K. Yang, D. Zhang, M. Gao, H. Ma, and W. Qian, "A multi-label deep learning model with interpretable grad-cam for diabetic retinopathy classification," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 1560–1563.

[97] D. Kumar, G. W. Taylor, and A. Wong, "Discovery radiomics with clear-dr: interpretable computer aided diagnosis of diabetic retinopathy," *IEEE Access*, vol. 7, pp. 25 891–25 896, 2019.

[98] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6666–6673.

[99] W. Liao, B. Zou, R. Zhao, Y. Chen, Z. He, and M. Zhou, "Clinical interpretable deep learning model for glaucoma diagnosis," *IEEE journal of biomedical and health informatics*, vol. 24, no. 5, pp. 1405–1412, 2019.

[100] X. Wang, M. Xu, L. Li, Z. Wang, and Z. Guan, "Pathology-aware deep network visualization and its application in glaucoma image synthesis," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*. Springer, 2019, pp. 423–431.

[101] O. Perdomo, H. Rios, F. J. Rodríguez, S. Otálora, F. Meriaudeau, H. Müller, and F. A. González, "Classification of diabetes-related retinal diseases using a deep learning approach in optical coherence tomography," *Computer methods and programs in biomedicine*, vol. 178, pp. 181–189, 2019.

[102] H. Pratt, F. Coenen, S. P. Harding, D. M. Broadbent, and Y. Zheng, "Feature visualisation of classification of diabetic retinopathy using a convolutional neural network," in *CEUR Workshop Proceedings*, vol. 2429, 2019, pp. 23–29.

[103] B. N. Narayanan, R. C. Hardie, M. S. De Silva, and N. K. Kueterman, "Hybrid machine learning architecture for automated detection and grading of retinal images for diabetic retinopathy," *Journal of Medical Imaging*, vol. 7, no. 3, pp. 034 501–034 501, 2020.

[104] Y. Shen, B. Sheng, R. Fang, H. Li, L. Dai, S. Stolte, J. Qin, W. Jia, and D. Shen, "Domain-invariant interpretable fundus image quality assessment," *Medical image analysis*, vol. 61, p. 101654, 2020.

[105] Z. Tu, S. Gao, K. Zhou, X. Chen, H. Fu, Z. Gu, J. Cheng, Z. Yu, and J. Liu, "Sunet: A lesion regularized model for simultaneous diabetic retinopathy and diabetic macular edema grading," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 1378–1382.

[106] X. Wang, H. Chen, A.-R. Ran, L. Luo, P. P. Chan, C. C. Tham, R. T. Chang, S. S. Mannil, C. Y. Cheung, and P.-A. Heng, "Towards multicenter glaucoma oct image screening with semi-supervised joint structure and function multi-task learning," *Medical Image Analysis*, vol. 63, p. 101695, 2020.

[107] K. A. Thakoor, X. Li, E. Tsamis, P. Sajda, and D. C. Hood, "Enhancing the accuracy of glaucoma detection from oct probability maps using convolutional neural networks," in *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2019, pp. 2036–2040.

[108] R. Zhang, S. Tan, R. Wang, S. Manivannan, J. Chen, H. Lin, and W.-S. Zheng, "Biomarker localization by combining cnn classifier and generative adversarial network," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Confer-*

ence, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22. Springer, 2019, pp. 209–217.

[109] J. Martins, J. S. Cardoso, and F. Soares, "Offline computer-aided diagnosis for glaucoma detection using fundus images targeted at mobile devices," *Computer Methods and Programs in Biomedicine*, vol. 192, p. 105341, 2020.

[110] Q. Meng, Y. Hashimoto, and S. Satoh, "How to extract more information with less burden: Fundus image classification and retinal disease localization with ophthalmologist intervention," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3351–3361, 2020.

[111] R. Wang, D. Fan, B. Lv, M. Wang, Q. Zhou, C. Lv, G. Xie, and L. Wang, "Oct image quality evaluation based on deep and shallow features fusion network," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 1561–1564.

[112] Y. Jang, J. Son, K. H. Park, S. J. Park, and K.-H. Jung, "Laterality classification of fundus images using interpretable deep neural network," *Journal of digital imaging*, vol. 31, pp. 923–928, 2018.

[113] G. Quellec, K. Charriere, Y. Boudi, B. Cochener, and M. Lamard, "Deep image mining for diabetic retinopathy screening," *Medical image analysis*, vol. 39, pp. 178–193, 2017.

[114] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu *et al.*, "Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy," *Ophthalmology*, vol. 126, no. 4, pp. 552–564, 2019.

[115] M. Kim, J. C. Han, S. H. Hyun, O. Janssens, S. Van Hoecke, C. Kee, and W. De Neve, "Medinoid: computer-aided diagnosis and localization of glaucoma using deep learning," *Applied Sciences*, vol. 9, no. 15, p. 3064, 2019.

[116] TensorFlow. (2023) Tensorflow. [Online]. Available: https://www.tensorflow.org/

[117] Pytorch. (2023) Pytorch. [Online]. Available: https://pytorch.org/

[118] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[119] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, 2018.

[120] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis, "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–12.

[121] K. Costello and M. Rimol. (2019) Gartner identifies the top strategic technology trends for 2021. [Online]. Available: https://www.gartner.com/en/information-technology/insights/top-technology-trends

[122] L. Aroyo, M. Lease, P. Paritosh, and M. Schaekermann, "Data excellence for ai: why should you care?" *Interactions*, vol. 29, no. 2, pp. 66–69, 2022.

[123] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[124] OpenAI. (2023) Gpt-4 technical report. [Online]. Available: https://cdn.openai.com/papers/gpt-4.pdf

[125] K. M. Yoo, D. Park, J. Kang, S.-W. Lee, and W. Park, "Gpt3mix: Leveraging large-scale language models for text augmentation," *arXiv preprint arXiv:2104.08826*, 2021.

[126] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[127] D. S. Fong, L. Aiello, T. W. Gardner, G. L. King, G. Blankenship, J. D. Cavallerano, F. L. Ferris III, R. Klein, and A. D. Association, "Retinopathy in diabetes," *Diabetes care*, vol. 27, no. suppl_1, pp. s84–s87, 2004.

[128] "Diabetic retinopathy," Dec 2021. [Online]. Available: https://www.nhs.uk/conditions/diabetic-retinopathy/

[129] I. D. Federation. (2023) International diabetes federation. [Online]. Available: https://idf.org/

[130] C. for Disease Control and Prevention. (2023) Centers for disease control and prevention. [Online]. Available: https://www.cdc.gov/

[131] W. H. Organization. (2023) World health organization. [Online]. Available: https://www.who.int/

[132] "8th edition." [Online]. Available: https://diabetesatlas.org/atlas/eighth-edition/

[133] D. S. Fong, L. Aiello, T. W. Gardner, G. L. King, G. Blankenship, J. D. Cavallerano, F. L. Ferris III, R. Klein, and A. D. Association, "Retinopathy in diabetes," *Diabetes care*, vol. 27, no. suppl_1, pp. s84–s87, 2004.

[134] R. Taylor and D. Batey, *Handbook of retinal screening in diabetes: diagnosis and management*. John Wiley & Sons, 2012.

[135] E. T. D. R. S. R. Group *et al.*, "Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified airlie house classification: Etdrs report number 10," *Ophthalmology*, vol. 98, no. 5, pp. 786–806, 1991.

[136] P. H. Scanlon, A. Sallam, and P. Van Wijngaarden, *A practical manual of diabetic retinopathy management*. John Wiley & Sons, 2017.

[137] C. P. Wilkinson, F. L. Ferris III, R. E. Klein, P. P. Lee, C. D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, J. T. Verdaguer *et al.*, "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales," *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, 2003.

[138] Kaggle. (2019) Aptos 2019 blindness detection. [Online]. Available: https://www.kaggle.com/c/aptos2019-blindness-detection

[139] IEEEDataport. (2019) Indian diabetic retinopathy image dataset (idrid). [Online]. Available: https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid

[140] Kaggle. (2015) Diabetic retinopathy detection. [Online]. Available: https://www.kaggle.com/competitions/diabetic-retinopathy-detection/data

[141] W. H. Dean, S. Grant, J. McHugh, O. Bowes, and F. Spencer, "Ophthalmology specialist trainee survey in the united kingdom," *Eye*, vol. 33, no. 6, pp. 917–924, 2019.

[142] A. Mishra, "Contrast limited adaptive histogram equalization (clahe) approach for enhancement of the microstructures of friction stir welded joints," *arXiv preprint arXiv:2109.00886*, 2021.

[143] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355–368, 1987.

[144] W. L. Alyoubi, W. M. Shalash, and M. F. Abulkhair, "Diabetic retinopathy detection through deep learning techniques: A review," *Informatics in Medicine Unlocked*, vol. 20, p. 100377, 2020.

[145] B. Graham. (2023) Competition report (min-pooling) and thank you. [Online]. Available: https://www.kaggle.com/competitions/diabetic-retinopathy-detection/discussion/15801

[146] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels," Tech. Rep., 2010.

[147] A. A. Wei Xiang Lim, Zhiyuan Chen, "A quantitative interpretable convolutional neural network to diagnose diabetic retinopathy," *The 25th International Conference on Image Processing, Computer Vision, & Pattern Recognition*, 2021.

[148] A. Hernández-García and P. König, "Data augmentation instead of explicit regularization," *arXiv preprint arXiv:1806.03852*, 2018.

[149] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[150] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.

[151] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," *Advances in neural information processing systems*, vol. 32, 2019.

[152] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[153] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[154] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[155] H. Deng, X. Qi, and X. Guo, "Diagnostic accuracy of apri, aar, fib-4, fi, king, lok, forns, and fibroindex scores in predicting the presence of esophageal varices in liver cirrhosis: a systematic review and meta-analysis," *Medicine*, vol. 94, no. 42, 2015.

[156] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, 2011.

[157] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030–3043, 2017.

[158] S. Beguería, "Validation and evaluation of predictive models in hazard assessment and risk management," *Natural Hazards*, vol. 37, pp. 315–329, 2006.

[159] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[160] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Rusboost: A hybrid approach to alleviating class imbalance," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 1, pp. 185–197, 2009.

[161] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[162] T. Eelbode, J. Bertels, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index," *IEEE transactions on medical imaging*, vol. 39, no. 11, pp. 3679–3690, 2020.

[163] EyePACS. (2023) Committed to preserving vision for all. [Online]. Available: https://www.eyepacs.com/

[164] G. Mushtaq and F. Siddiqui, "Detection of diabetic retinopathy using deep learning methodology," in *IOP conference series: materials science and engineering*, vol. 1070, no. 1. IOP Publishing, 2021, p. 012049.

[165] R. Pires, S. Avila, J. Wainer, E. Valle, M. D. Abramoff, and A. Rocha, "A data-driven approach to referable diabetic retinopathy detection," *Artificial intelligence in medicine*, vol. 96, pp. 93–106, 2019.

[166] M. Saini and S. Susan, "Diabetic retinopathy screening using deep learning for multi-class imbalanced datasets," *Computers in Biology and Medicine*, vol. 149, p. 105989, 2022.

[167] S. G. Karst, J. Lammer, S. H. Radwan, H. Kwak, P. S. Silva, S. A. Burns, L. P. Aiello, and J. K. Sun, "Characterization of in vivo retinal lesions of diabetic retinopathy using adaptive optics scanning laser ophthalmoscopy," *International journal of endocrinology*, vol. 2018, 2018.

[168] G. Indumathi and V. Sathananthavathi, "Microaneurysms detection for early diagnosis of diabetic retinopathy using shape and steerable gaussian features," in *Telemedicine Technologies*. Elsevier, 2019, pp. 57–69.

[169] A. Avinash, P. Biju, P. Premanath, A. Thomas, and V. P. Gopi, "An improved method for automated detection of microaneurysm in retinal fundus images," in *Edge-of-Things in Personalized Healthcare Support Systems*. Elsevier, 2022, pp. 173–186.

[170] J. W. Woods, *Multidimensional signal, image, and video processing and coding*. Academic press, 2011.

[171] B. Graham, "Kaggle diabetic retinopathy detection competition report," *University of Warwick*, pp. 24–26, 2015.

[172] S. Natarajan, A. Jain, R. Krishnan, A. Rogye, and S. Sivaprasad, "Diagnostic accuracy of community-based diabetic retinopathy screening with an offline artificial intelligence system on a smartphone," *JAMA ophthalmology*, vol. 137, no. 10, pp. 1182–1188, 2019.

[173] PyTorch. (2023) Reducelronplateau. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html

[174] ——. (2023) Earlystopping. [Online]. Available: https://pytorch.org/ignite/generated/ignite.handlers.early_stopping.EarlyStopping.html

[175] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[176] J. Pearl, "Causality: models, reasoning, and inference," 1980.

[177] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv preprint arXiv:2006.11371*, 2020.

[178] A. Kumcu, K. Bombeke, L. Platiša, L. Jovanov, J. Van Looy, and W. Philips, "Performance of four subjective video quality assessment protocols and impact of different rating preprocessing and analysis methods," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 48–63, 2016.

[179] H. Moradmand, S. M. R. Aghamiri, and R. Ghaderi, "Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma," *Journal of applied clinical medical physics*, vol. 21, no. 1, pp. 179–190, 2020.

[180] C. Li, C. Xu, C. Gui, and M. D. Fox, "Distance regularized level set evolution and its application to image segmentation," *IEEE transactions on image processing*, vol. 19, no. 12, pp. 3243–3254, 2010.

[181] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Communications*, vol. 11, no. 1, p. 3673, 2020.

[182] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "Ai for radiographic covid-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, vol. 3, no. 7, pp. 610–619, 2021.

[183] W. van Amsterdam, J. Verhoeff, P. de Jong, T. Leiner, and M. Eijkemans, "Eliminating biasing signals in lung cancer images for prognosis predictions with deep learning," *NPJ digital medicine*, vol. 2, no. 1, p. 122, 2019.

[184] J. Ding, X. Li, X. Kang, and V. N. Gudivada, "A case study of the augmentation and evaluation of training data for deep learning," *Journal of Data and Information Quality (JDIQ)*, vol. 11, no. 4, pp. 1–22, 2019.

[185] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[186] D. Zha, K.-H. Lai, K. Zhou, and X. Hu, "Simplifying deep reinforcement learning via self-supervision," *arXiv preprint arXiv:2106.05526*, 2021.

[187] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[188] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.

[189] D. Zha, K.-H. Lai, M. Wan, and X. Hu, "Meta-aad: Active anomaly detection with deep reinforcement learning," in *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2020, pp. 771–780.

[190] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[191] U. Khurana, H. Samulowitz, and D. Turaga, "Feature engineering for predictive modeling using reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[192] Z. Liu, P. Wei, J. Jiang, W. Cao, J. Bian, and Y. Chang, "Mesa: boost ensemble imbalanced learning with meta-sampler," *Advances in neural information processing systems*, vol. 33, pp. 14 463–14 474, 2020.

[193] A. Aldweesh, A. Derhab, and A. Z. Emam, "Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues," *Knowledge-Based Systems*, vol. 189, p. 105124, 2020.

Appendix A

List of Abbreviations

CNNs	Convolutional Neural Networks
AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
SA	Sensitivity Analysis
GI	Gradient x Input
CAM	Class Activation Map
Grad-CAM	Gradient-based Class Activation Map
Guided Grad-CAM	Guided Gradient-based Class Activation Map
IG	Integrated Gradients
EU	European Union
GDPR	General Data Protection Regulation
DR	Diabetic Retinopathy
SLIC	Simple Linear Iterative Clustering
KDRD2015	Kaggle Diabetic Retinopathy Dataset 2015
APTOS2019	Asia Pacific Tele-Ophthalmology Society 2019
IDRID	Indian Diabetic Retinopathy Image Dataset
CorLoc	Correct Localization
IoU	Intersection over Union
mAP	mean Average Precision
mAR	mean Average Recall
MAE	Mean Average Error
PR	Precision-Recall
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
KC	Kappa Coefficient
OE	Overall Error

Appendix B

DC-Check Checklist A

DC-CHECK

Project Description: Referable Diabetic Retinopathy Image Classification with Quantitative Saliency-oriented Visual Explanation.

Link: To-be-determined

Citation:

DATA
Q1: How did you select, collect or curate your dataset? - *Have you conducted forensics on the dataset (i.e. provenance)?* - *Did you assess the pertinence of the dataset for the task?* - *Is your dataset curation a once-off?*
Data Acquisition: 1. **Kaggle Diabetic Retinopathy Dataset 2015 (KDRD2015)** - Publicly available dataset. Can be procured from Kaggle website or Google Dataset Search. - All fundus photography images were taken and verified by EyePACS. - All images were coloured fundus photography images. - Data distribution of each class is highly imbalanced. 2. **Indian Diabetic Retinopathy Image Dataset (IDRID)** - Publicly available dataset. Can be procured from Google Dataset Search or IEEE Dataport. - All fundus photography images were taken and verified by a specialist at an eye clinic located at Nanded, Maharashtra, India. - All images were coloured fundus photography images. - Data distribution of each class is highly imbalanced. - The dataset contained pixel-level annotated ground truth images of diabetic lesions such as haemorrhage, soft exudates, hard exudates, and microaneurysm.
Q2: What data cleaning and/or pre-processing, if any, has been performed? - *Do you need to "clean" your data?* - *Do you have any missing data?*
- The KDRD2015 dataset will be preprocessed using state-of-the-art Contrast Limited Adaptive Histogram Equalization (CLAHE) on different colour space. 1. Apply CLAHE on Black/White fundus images. 2. Apply CLAHE on Green channel fundus images. 3. Apply CLAHE on Luminance channel fundus images. 4. Apply Gaussian Weighted Contrast on fundus images.

Figure B.1: *Page 1 of DC-Check checklist for experiments in Chapter 4.*

Q3: Has data quality been assessed?

- Completeness, diversity, label quality
- Easy, Hard examples etc

- During preprocessing stage, the black border around the retina will be removed to reduce any unimportant information. While the black border is removed, a function will be used to remove any fundus images that was almost/totally black.

Q4: Have you considered synthetic data?

1. Data augmentation will be used to mitigate model overfitting.

TRAINING

Q5: Have you conducted a model architecture and hyperparameter search?

- *Have you specifically adapted the model for the task?*
- *Is there a potential to incorporate inductive biases?*

Model Architecture:

- Based on prior literature, the primary adoption of pre-trained model is DenseNet. (Justifications were explained in the thesis).

Hyperparameters:

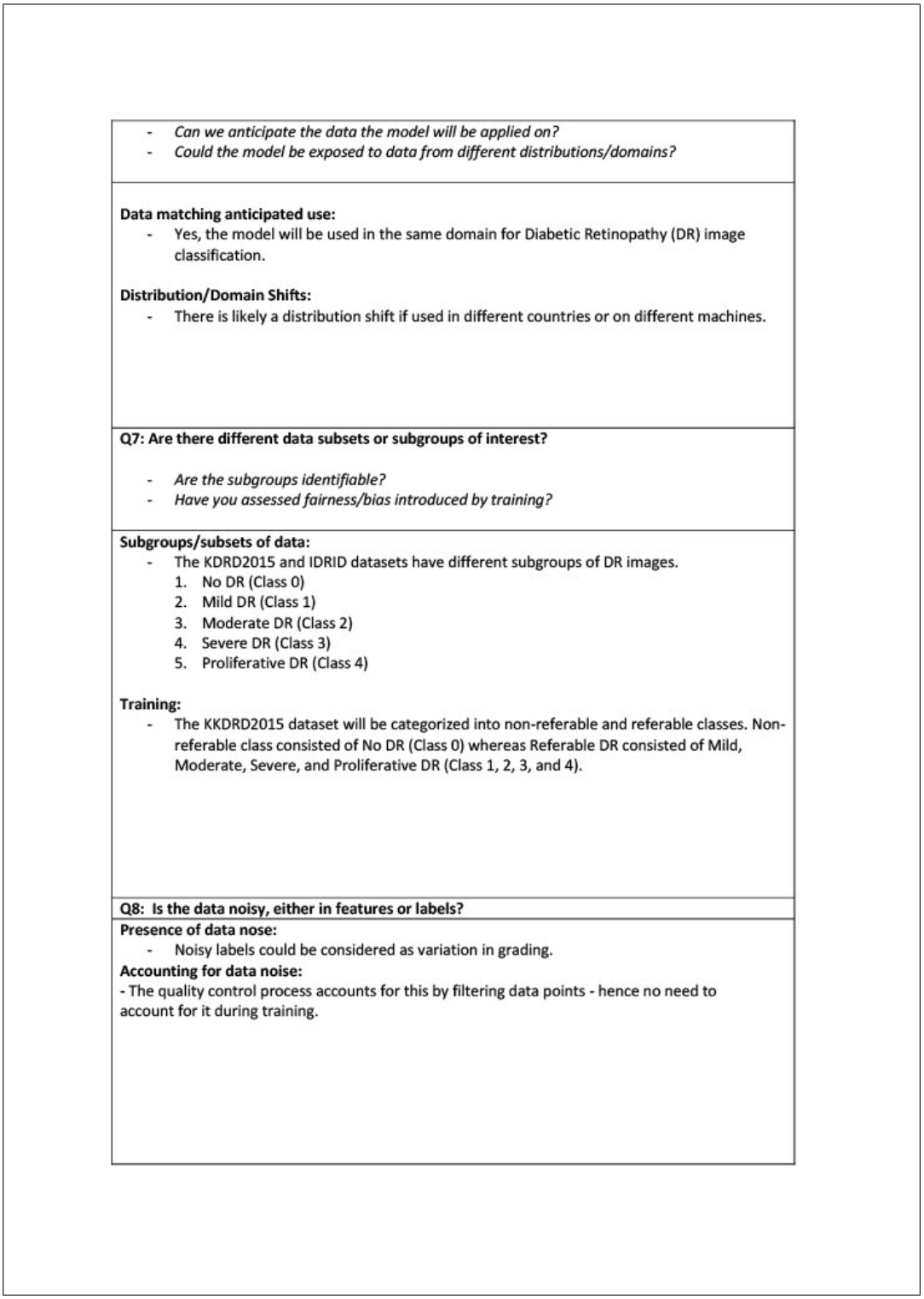
- Categorical cross entropy loss will be used as the loss function.
- Adam optimizer with an initial learning rate of 0.003 will be used.
- ReduceLROnPlateau optimization method will be used to diminish the learning rate if the testing loss plateaus.
- EarlyStopping optimization method will be used to stop the training is the testing accuracy did not improve.
- Number of epochs = 200
- Experiments will be conducted on NVIDIA RTX 3070 GPU machine.

Dataset:

- The KKDRD2015 dataset will split into 80% training set and 20% testing set.
- The IDRID dataset will be used as model inferencing.

Q6: Does the training data match the anticipated use?

Figure B.2: *Page 2 of DC-Check checklist for experiments in Chapter 4.*



- *Can we anticipate the data the model will be applied on?*
- *Could the model be exposed to data from different distributions/domains?*

Data matching anticipated use:

- Yes, the model will be used in the same domain for Diabetic Retinopathy (DR) image classification.

Distribution/Domain Shifts:

- There is likely a distribution shift if used in different countries or on different machines.

Q7: Are there different data subsets or subgroups of interest?

- *Are the subgroups identifiable?*
- *Have you assessed fairness/bias introduced by training?*

Subgroups/subsets of data:

- The KDRD2015 and IDRID datasets have different subgroups of DR images.
  1. No DR (Class 0)
  2. Mild DR (Class 1)
  3. Moderate DR (Class 2)
  4. Severe DR (Class 3)
  5. Proliferative DR (Class 4)

Training:

- The KKDRD2015 dataset will be categorized into non-referable and referable classes. Non-referable class consisted of No DR (Class 0) whereas Referable DR consisted of Mild, Moderate, Severe, and Proliferative DR (Class 1, 2, 3, and 4).

Q8: Is the data noisy, either in features or labels?

Presence of data nose:

- Noisy labels could be considered as variation in grading.

Accounting for data noise:

- The quality control process accounts for this by filtering data points - hence no need to account for it during training.

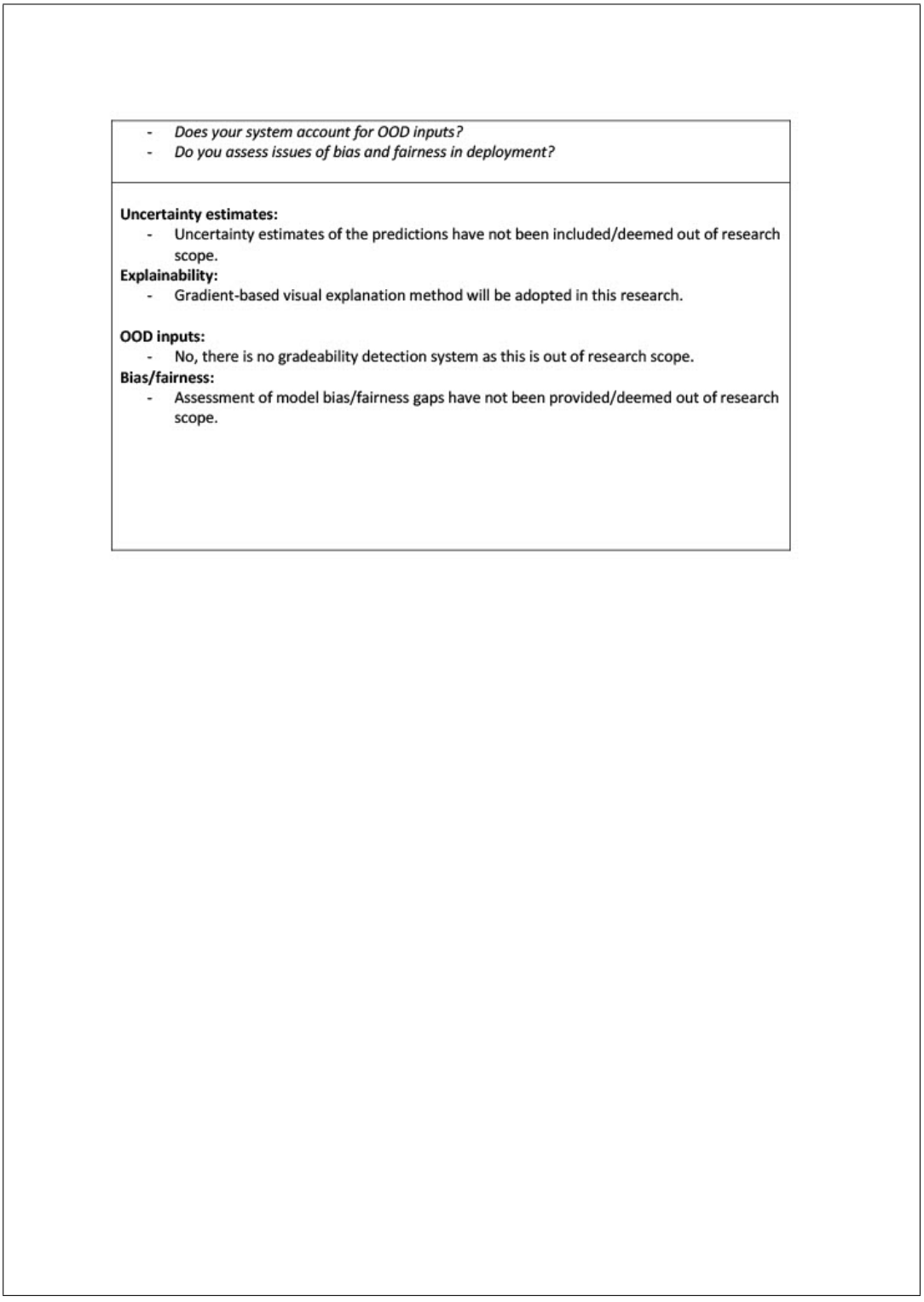
Figure B.3: *Page 3 of DC-Check checklist for experiments in Chapter 4.*

TESTING
Q9: How has the dataset been split for model training and validation? - *Are you using a benchmark dataset?* - *Is the splitting random?*
Data split: - Random split of 80/20 training and testing set.
Q10: How has the model been evaluated (e.g. metrics & stress tests)? - *What metrics are used to assess the model?* - *Has the model been evaluated beyond average?* - *Is there a potential to test specific aspects/sub-groups of the model (i.e. stress tests)?*
Model Performance Evaluation Metrics: - F1-score, Precision, and Recall - Accuracy - Confusion matrix - AUC ROC curve

Figure B.4: *Page 4 of DC-Check checklist for experiments in Chapter 4.*

DEPLOYMENT
Q11: Are you monitoring your model? - *Have you considered dimensionality in the monitoring?* - *Is there lag in ground truth feedback?*
- Currently, there is no model monitoring as this research is still ongoing/not ready for real-world deployment.
Q12: Do you have mechanisms in place to address data shifts? - *Have model updates been considered & when to retrain?* - *Does your system provide actionable feedback upon failure?* - *Are you able to characterize the type of shift?* - *Do system failures inform dataset updates?*
Model update: - Retraining the model is necessary. **System failure feedback:** - System failure feedback will not be available at this moment as this research is still ongoing/not ready for real-world deployment. **Characterizing the type of shift:** - There will be no consideration characterizing the type of shift. **Failure informed dataset updates:** - System failure feedback will not be available at this moment as this research is still ongoing/not ready for real-world deployment.
Q13: Have you incorporated tools to engender model trust? - *Do you require predictive uncertainty estimates?* - *Does your ML system have explainability?*

Figure B.5: *Page 5 of DC-Check checklist for experiments in Chapter 4.*



- *Does your system account for OOD inputs?*
- *Do you assess issues of bias and fairness in deployment?*

Uncertainty estimates:
- Uncertainty estimates of the predictions have not been included/deemed out of research scope.

Explainability:
- Gradient-based visual explanation method will be adopted in this research.

OOD inputs:
- No, there is no gradeability detection system as this is out of research scope.

Bias/fairness:
- Assessment of model bias/fairness gaps have not been provided/deemed out of research scope.

Figure B.6: *Page 6 of DC-Check checklist for experiments in Chapter 4.*

Appendix C

DC-Check Checklist B

DC-CHECK

Project Description: Enhancing Deep Learning Pre-trained Networks on Diabetic Retinopathy Fundus Photography with SLIC-G

Link: To-be-determined

Citation:

DATA
Q1: How did you select, collect or curate your dataset? - *Have you conducted forensics on the dataset (i.e. provenance)?* - *Did you assess the pertinence of the dataset for the task?* - *Is your dataset curation a once-off?*
Data Acquisition: 1. **Asia Pacific Tele-Ophthalmology Society 2019 (APTOS2019)** - Publicly available dataset. Can be procured from Kaggle website or Google Dataset Search. - All fundus photography images were taken and verified by technicians in Aravind Hospital in India. - All images were coloured fundus photography images. - Data distribution of each class is highly imbalanced. 2. **Indian Diabetic Retinopathy Image Dataset (IDRID)** - Publicly available dataset. Can be procured from Google Dataset Search or IEEE Dataport. - All fundus photography images were taken and verified by a specialist at an eye clinic located at Nanded, Maharashtra, India. - All images were coloured fundus photography images. - Data distribution of each class is highly imbalanced. - The dataset contained pixel-level annotated ground truth images of diabetic lesions such as haemorrhage, soft exudates, hard exudates, and microaneurysm.
Q2: What data cleaning and/or pre-processing, if any, has been performed? - *Do you need to "clean" your data?* - *Do you have any missing data?*
- The APTOS2019 will be preprocessed using the proposed image preproceesing method called SLIC-G that combines Simple Linear Iterative Clustering (SLIC) superpixel segmentation and Gaussian smoothing operation.

Figure C.1: *Page 1 of DC-Check checklist for experiments in Chapter 5.*

Q3: Has data quality been assessed?

- Completeness, diversity, label quality
- Easy, Hard examples etc

- During preprocessing stage, the black border around the retina will be removed to reduce any unimportant information. While the black border is removed, a function will be used to remove any fundus images that was almost/totally black.

Q4: Have you considered synthetic data?

1. Data augmentation will be used to mitigate model overfitting.

TRAINING

Q5: Have you conducted a model architecture and hyperparameter search?

- *Have you specifically adapted the model for the task?*
- *Is there a potential to incorporate inductive biases?*

Model Architecture:

- Based on prior literature, the primary adoption of pre-trained model is DenseNet. (Justifications were explained in the thesis).

Hyperparameters:

- Categorical cross entropy loss will be used as the loss function.
- Adam optimizer with an initial learning rate of 0.003 will be used.
- ReduceLROnPlateau optimization method will be used to diminish the learning rate if the testing loss plateaus.
- EarlyStopping optimization method will be used to stop the training is the testing accuracy did not improve.
- Number of epochs = 200
- Experiments will be conducted on NVIDIA RTX 3070 GPU machine.

Dataset:

- The APTOS2019 dataset will split into 80% training set and 20% testing set.
- The IDRID dataset will be used as model inferencing.

Q6: Does the training data match the anticipated use?

- *Can we anticipate the data the model will be applied on?*
- *Could the model be exposed to data from different distributions/domains?*

Figure C.2: *Page 2 of DC-Check checklist for experiments in Chapter 5.*

Data matching anticipated use:

- Yes, the model will be used in the same domain for Diabetic Retinopathy (DR) image classification.

Distribution/Domain Shifts:

- There is likely a distribution shift if used in different countries or on different machines.

Q7: Are there different data subsets or subgroups of interest?

- *Are the subgroups identifiable?*
- *Have you assessed fairness/bias introduced by training?*

Subgroups/subsets of data:

- The KDRD2015 and IDRID datasets have different subgroups of DR images.
  1. No DR (Class 0)
  2. Mild DR (Class 1)
  3. Moderate DR (Class 2)
  4. Severe DR (Class 3)
  5. Proliferative DR (Class 4)

Training:

- The APTOS2019 dataset will be categorized into multi-classes.

Q8: Is the data noisy, either in features or labels?

Presence of data nose:

- Noisy labels could be considered as variation in grading.

Accounting for data noise:

- The quality control process accounts for this by filtering data points - hence no need to account for it during training.

Figure C.3: *Page 3 of DC-Check checklist for experiments in Chapter 5.*

TESTING
Q9: How has the dataset been split for model training and validation? - *Are you using a benchmark dataset?* - *Is the splitting random?*
Data split: - Random split of 80/20 training and testing set.
Q10: How has the model been evaluated (e.g. metrics & stress tests)? - *What metrics are used to assess the model?* - *Has the model been evaluated beyond average?* - *Is there a potential to test specific aspects/sub-groups of the model (i.e. stress tests)?*
Model Performance Evaluation Metrics: - F1-score, Precision, and Recall - Accuracy - Confusion matrix - AUC ROC curve - Geometric Mean

Figure C.4: *Page 4 of DC-Check checklist for experiments in Chapter 5.*

DEPLOYMENT
Q11: Are you monitoring your model? - *Have you considered dimensionality in the monitoring?* - *Is there lag in ground truth feedback?*
- Currently, there is no model monitoring as this research is still ongoing/not ready for real-world deployment.
Q12: Do you have mechanisms in place to address data shifts? - *Have model updates been considered & when to retrain?* - *Does your system provide actionable feedback upon failure?* - *Are you able to characterize the type of shift?* - *Do system failures inform dataset updates?*
Model update: - Retraining the model is necessary. **System failure feedback:** - System failure feedback will not be available at this moment as this research is still ongoing/not ready for real-world deployment. **Characterizing the type of shift:** - There will be no consideration characterizing the type of shift. **Failure informed dataset updates:** - System failure feedback will not be available at this moment as this research is still ongoing/not ready for real-world deployment.
Q13: Have you incorporated tools to engender model trust? - *Do you require predictive uncertainty estimates?* - *Does your ML system have explainability?* - *Does your system account for OOD inputs?* - *Do you assess issues of bias and fairness in deployment?*
Uncertainty estimates:

Figure C.5: *Page 5 of DC-Check checklist for experiments in Chapter 5.*

- Uncertainty estimates of the predictions have not been included/deemed out of research scope.

Explainability:

- Gradient-based visual explanation method will be adopted in this research.

OOD inputs:

- No, there is no gradeability detection system as this is out of research scope.

Bias/fairness:

- Assessment of model bias/fairness gaps have not been provided/deemed out of research scope.

Figure C.6: *Page 6 of DC-Check checklist for experiments in Chapter 5.*