Fall 2023

XAI-Driven CNN for Diabetic Retinopathy Detection

Vikas Shenoy Pete

XAI-Driven CNN for Diabetic Retinopathy Detection

A Project

Presented to

The Faculty of the Department of Computer Science

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Vikas Shenoy Pete

December 2023

The Designated Project Committee Approves the Project Titled

XAI-Driven CNN for Diabetic Retinopathy Detection

by

Vikas Shenoy Pete

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

SAN JOSÉ STATE UNIVERSITY

December 2023

Dr. Genya Ishigaki	Department of Computer Science
Dr. Fabio Di Troia	Department of Computer Science
Dr. Navrati Saxena	Department of Computer Science

ABSTRACT

XAI-Driven CNN for Diabetic Retinopathy Detection

by Vikas Shenoy Pete

Diabetes, a chronic metabolic disorder, poses a significant health threat with potentially severe consequences, including diabetic retinopathy, a leading cause of blindness. In this project, we tackle this threat by developing a Convolutional Neural Network (CNN) to support the diagnosis based on eye images. The aim is early detection and intervention to mitigate the effects of diabetes on eye health. To enhance transparency and interpretability, we incorporate explainable AI techniques. This research not only contributes to the early diagnosis of diabetic eye disease but also advances our understanding of how deep learning models arrive at their decisions, fostering trust and clinical applicability in healthcare diagnostics.

Our results show that our CNN model performs exceptionally well in classifying ocular images, attaining a 91% accuracy rate. Furthermore, we implemented explainable AI techniques, such as LIME (Local Interpretable Model-agnostic Explanations), which improves the transparency of our model's decision-making. The areas of interest in the eye images were clarified for us by LIME, which enhanced our understanding of the model's predictions. The high accuracy and interpretability of our approach demonstrate its potential for clinical applications and the broader field of healthcare diagnostics.

***Keywords* - Convolutional Neural Networks, Explainable Artificial Intelligence, Local Interpretable Model-agnostic Explanations, Medical Image Classification.**

ACKNOWLEDGMENTS

I want to express my gratitude to Dr. Genya Ishigaki for his exceptional guidance and unwavering support throughout my master's program. His dedication to my academic growth has had a lasting impact on my path, and his support was invaluable throughout my research project.

I am also grateful to Dr. Fabio Di Troia and Dr. Navrati Saxena for their valuable time and perceptive remarks as committee members. The success of my research was greatly aided by their assistance.

In conclusion, I would like to thank everyone who has supported me during my academic career. Your encouragement and support made it possible for me to achieve what I have.

TABLE OF CONTENTS

CHAPTER

APPENDIX

CHAPTER 1

Introduction

Reduced synthesis or use of insulin is a common long-term metabolic disease known as diabetes, which has grown to be a significant global health concern [1]. Diabetes is a complex disease that significantly affects many organ systems, including the eyes. As a result, its benefits extend well beyond regulating blood sugar levels and overall wellness. One of the most severe consequences is diabetic retinopathy, a disorder that poses a serious risk to vision and has emerged as a major issue in the field of ophthalmology. Irreversible blindness could result from the condition if it is not detected and treated in its early stages. In light of the seriousness of the condition and the potentially disastrous consequences of a diabetic retinopathy diagnosis, this project looks into the application of cutting-edge technology to enhance early detection and, eventually, the standard of care for diabetics.

It is impossible to overestimate the effects of diabetic retinopathy. Millions of people worldwide still suffer from diabetes, and the key to preventing vision loss and enhancing the quality of life for those who are affected is early detection of abnormalities in the eyes [2]. When diabetic retinopathy reaches an advanced stage, it usually advances silently and frequently shows no symptoms at all. This emphasizes how important it is to have a strong diagnostic tool that can spot anomalies while they are still in the early stages[3]. Early intervention not only lessens the risk of vision loss but also considerably lessens the psychological and financial strain that advanced diabetic eye problems have on patients and healthcare systems. The quest for early detection is extremely valuable as it can save people's eyes from suffering from the severe effects of diabetic retinopathy.

CNNs, or convolutional neural networks, have become a very useful technique for classifying images. They are perfectly suited for jobs like medical image analysis because of their capacity to identify intricate patterns and features inside images. CNNs will be used to examine a dataset of eye photographs in the context of diabetic eye diagnosis in order to differentiate between photos that show normal eye conditions and those that indicate anomalies connected to diabetes [4]. In order to guarantee the CNN model's high level of accuracy and dependability in its predictions, this research will examine the model's architecture, training, and optimization. Furthermore, an essential part of this project will be using explainable AI approaches to make the model's decision-making process comprehensible. This will guarantee openness and promote confidence in the model's clinical applications.

The significance of interpretability and transparency is increased when CNNs are used for medical diagnosis. It is essential to comprehend the model's methodology, particularly in a clinical context where the choices the model makes may have an influence on patient care. Therefore, this project will incorporate explainable AI techniques in tandem with CNN development to clarify the reasoning behind the model's predictions. This improves the reliability of the model and provides information about the characteristics and patterns in the images that influence the diagnosis. Healthcare professionals can validate and comprehend the diagnostic recommendations when the decision-making process of CNNs is made interpretable, which ultimately results in more informed and efficient clinical decisions. Advanced machine learning and explainable AI together have the potential to transform the diagnosis of diabetic eye disease, providing a glimmer of hope for those who suffer from the condition and furthering the field of medical image analysis as a whole.

Taking into account the previously mentioned factors, this project tries to achieve two major goals to build a trustworthy machine learning-based diagnosis tool. Creating and refining a CNN model that can correctly categorize eye images into normal and diabetes-related conditions is the main objective. In addition, by incorporating explainable AI techniques, the project seeks to improve the model's decision-making process's transparency and interpretability. By doing this, it aims to close the gap that exists between clinical practice and cutting-edge technology, guaranteeing the accuracy and reliability of AI-based diagnostics. This effort is in line with the pressing need for trustworthy early detection techniques for diabetic eye disease and represents a breakthrough in the worldwide battle against vision-threatening complications resulting from diabetes. The project's outcomes hold the promise of positively impacting healthcare and the lives of those afflicted by this prevalent chronic condition.

The structure of this report is as follows. In Chapter 2, we lay a thorough foundation for our research by reviewing relevant literature and introducing the principles of Explainable AI (XAI). Then, in chapter 3, we discuss the overall methodology by first introducing concepts of CNNs, and the theory behind the explainable AI techniques and then finally covering the details of the experiments conducted. Then, in Chapter 4 we go over the various experimental results obtained from CNNs and XAI methods. Finally, chapter 5 brings the work to a close and offers some suggestions for future improvements, this is our final chapter

CHAPTER 2

Background

2.1 Related Works

The field of medical image analysis has experienced a substantial transformation due to the swift progress made in artificial intelligence and deep learning. The incorporation of Convolutional Neural Networks (CNNs), which have proven to have exceptional abilities in the analysis of complex medical images, is central to this transformation. Explainable Artificial Intelligence (XAI) has become an important field of study due to the inherent difficulties in deciphering the decisions made by these deep learning models. In order to improve AI models' dependability and credibility in crucial applications like healthcare, XAI aims to make their decision-making processes transparent.

In order to overcome the black-box nature of these models and attempt to make their judgments comprehensible and justified, researchers have investigated a variety of approaches for integrating XAI into CNNs, particularly in the context of medical diagnostics. This methodology not only conforms to the ethical implications of utilizing AI in delicate domains but also aids healthcare practitioners in comprehending and establishing confidence in AI-supported diagnosis and interventions.

This chapter explores the latest research and developments on the integration of XAI in CNN-based medical image analysis. The strategies it looks at range from building CNN architectures from scratch with XAI integrated to fine-tuning pre-trained models with explainability in mind. We hope to offer a thorough grasp of how XAI is being applied to demystify CNN decisions in medical imaging by examining these methods.

This summary will provide an understanding of the state-of-the-art approaches currently in use, highlighting their advantages, disadvantages, and possible directions for further research. Studies that have used XAI to analyze different kinds of medical images are included in our investigation, and we discuss the subtleties of each method and how they apply to actual medical situations. An important step toward the use of AI in healthcare that is more accountable, transparent, and efficient is the incorporation of XAI into CNN models. We aim to provide a comprehensive understanding of this quickly developing field by combining these recent developments, helping practitioners and researchers better grasp the opportunities and difficulties associated with applying XAI to medical image analysis using CNNs.

Convolutional Neural Networks (CNNs) and Explainable Artificial Intelligence (XAI) are transforming medical image analysis by helping us comprehend and interpret complex medical data in a way that has never been seen before. The use of CNNs in medical imaging is primarily defined by three main strategies, as described in the research done by Dutta et al. [5]: training CNNs from scratch, utilizing pre-trained CNN features, and combining unattended pre-training with supervised fine-tuning. This method emphasizes how important transfer learning is to improving the effectiveness and precision of medical image classification. The study goes into more detail about each tactic, looking at its benefits and drawbacks in different medical imaging scenarios. As a result, it provides researchers and practitioners in the field with a useful manual.

Furthermore, Volkov et al. [6] offer a thorough summary of the use of XAI techniques in medical image analysis. It explores the current status of XAI technologies with an emphasis on improving the interpretability and transparency of AI models, especially CNNs, in medical settings. This paper explores promising future directions in the field of XAI while highlighting its technical aspects and potential benefits, such as increased trust among medical professionals and improved diagnostic accuracy.

Parallel to this, Moradi et al. [7] provides a comprehensive six-category XAI architecture for classifying deep learning-based medical image analysis and interpretability methods. This paper categorically outlines the different interpretability methods and XAI approaches, underlining the importance of explanation and technical methods in medical imaging. It elaborates on how these methods can be effectively applied in real-world scenarios, offering detailed insights into their practical implications and potential to improve patient outcomes.

Additionally, Dharshini et al. [8] explore the application of deep-learning convolutional neural networks (DCNNs), which are frequently thought of as "black-box" predictors, in medical imaging. This study emphasizes the emerging field of XAI and its function in elucidating AI models' decision-making processes, thereby promoting trustworthiness and dependability in medical diagnosis and treatment planning. The authors provide case studies that demonstrate how XAI has improved the interpretability and acceptability of AI tools in clinical settings by clearly illuminating the operation of DCNNs.

Additionally, Papanastasopoulos et al. [9] surveys more than 200 papers and categorizes them using an XAI framework. This paper offers a comprehensive overview of the state-of-the-art in medical imaging and offers insights into the trends and future prospects for XAI. The survey underscores the swift expansion of XAI applications within the field of medical imaging and pinpoints crucial domains requiring additional investigation and enhancement, thereby steering forthcoming progress in this field.

Furthermore, Gilhuijs et al. [10] discusses how AI is widely used in a variety of fields, including biomedical imaging. It emphasizes the significance of explainability in AI applications in biomedical settings and makes the case that progress in the field depends on our ability to comprehend AI decision-making processes. This paper advocates for a more responsible approach to AI development and deployment by discussing the ethical implications of AI in healthcare as well as the need for transparent algorithms in sensitive medical applications.

Yang et al. [11] emphasize the difficulties in creating CNNs from the ground up and offer to fine-tune pre-trained CNNs as a workable substitute. This method takes into account the variations between natural and medical images and suggests ways to successfully modify CNNs for use in medical settings. The paper then examines several case studies that show the applicability and efficacy of this strategy by showing how it has significantly improved the analysis of intricate medical datasets.

Moreover, Tajbakhsh et al. [12] offers a thorough examination of CNN's uses in medical imaging, addressing topics like large-scale image processing and brain MRI analysis. The various ways that CNNs can enhance medical image analysis procedures are highlighted in this review. In order to handle the increasing complexity and volume of medical data, it also addresses potential future directions for CNN research in medical imaging, highlighting the necessity for more reliable, flexible, and effective CNN architectures.

A novel diagnosis platform utilizing a DCNN was created in the study done by Kshatri et al. [13] to help radiologists differentiate COVID-19 pneumonia from other kinds. This method's average accuracy was very high, demonstrating how XAI can improve medical diagnostics and COVID-19 screening on a large scale. This study illustrates the wider applications of XAI in public health, especially in situations requiring quick responses, like pandemic outbreaks, in addition to proving the models' technical viability.

2.2 Introduction to Explainable AI (XAI)

One of the most significant advances in the field of artificial intelligence is explainable AI, or XAI for short. The need to close the gap between the decisions made by AI models and human understanding has become increasingly apparent as these models especially complex machine learning models like deep neural networks acquire more and more capabilities at a faster rate [14]. Interpretability is the crux of the matter with AI systems. Although these sophisticated models have shown incredible accuracy and predictive power in a variety of applications, they frequently operate as 'black boxes,' making it difficult to understand how they arrive at their conclusions. There are significant ramifications for this lack of transparency. Knowing and believing in the logic underlying AI-driven decisions is critical in domains where AI is making critical decisions, such as autonomous systems, financial risk assessment, and healthcare diagnosis [15]. Hence, explainable AI emerges in response to these demands, providing a resolution to the enduring problem of rendering AI systems more understandable, transparent, and morally upright.

Among the many benefits of explainable AI is the improvement of trust and accountability. As artificial intelligence (AI) applications become more and more integrated into our daily lives, the decision-making processes they use must adhere to ethical and legal requirements. XAI gives the tools to examine AI results and make sure they are impartial, fair, and just by offering insights into the reasoning behind particular decisions [16]. Transparency in AI has significant practical implications in addition to these ethical ones. Validating and understanding AI-generated outputs whether it's a recommendation for investments or a life-saving medical diagnosis benefits users and stakeholders. Furthermore, XAI makes it easier to identify and correct model biases and discrepancies, which helps create AI solutions that are fairer and more equitable [17]. The path to achieving XAI is characterized by continuous innovation and research, but it holds the potential to transform the field of AI applications and foster trust and accountability in the era of sophisticated machine learning.

2.2.1 Importance of Interpretability

In the field of artificial intelligence, interpretability is extremely important. This is especially true in industries as vital as healthcare and finance, where AI applications are essential to the processes that lead to decisions that have an immediate effect on people and organizations. Understanding the process and reasoning behind an AI system's recommendation or prediction is not only valuable but also crucial in these high-stakes domains [18].

The consequences of AI choices in healthcare are frequently life-or-death decisions. Transparency in AI systems can be crucial for a variety of purposes, including disease diagnosis, treatment planning, and risk assessment. Healthcare providers, physicians, and patients need to understand the rationale behind the recommendations made by these systems in addition to having faith in them. Medical professionals can trust AI to be a helpful tool when making important decisions because of interpretability. Additionally, it enhances patients' overall experience by empowering them to be knowledgeable about and confident in their healthcare journeys [19].

Moreover, interpretability is imperative in guaranteeing that AI models conform to legal and ethical guidelines. It makes sure AI systems don't discriminate based on race, gender, or socioeconomic status by enabling the identification of biases. Interpretability allows for the development of more equitable and just AI solutions by highlighting any inconsistencies or unfairness in the model's behaviour. Interpretability is crucial, as demonstrated by the use of AI in areas like hiring and criminal justice, where justice and transparency are primary concerns [20].

Furthermore, it is impossible to overestimate the contribution interpretability makes to innovation and teamwork in these fields. Transparency in AI systems facilitates more effective collaboration between researchers and practitioners from different disciplines, bringing diverse perspectives that improve the robustness and equity of AI solutions [21]. Clear understanding enables this multidisciplinary approach, which has the potential to advance AI applications and ultimately benefit society as a whole.

Interpretability also complies with the growing public demand for accountability and transparency in technology. There is a growing expectation for technology to be not only efficient but also responsible and understandable as society grows more aware of the potential risks and benefits of artificial intelligence. The significance of interpretability is further highlighted by this shift in society, guaranteeing that AI systems are created and implemented in a way that is not only technically sound but also socially responsible and consistent with public values.

In conclusion, interpretability is a practical, ethical, and legal requirement for AI applications. It is not just a convenience. A key component of AI deployment in delicate industries like healthcare and finance, where decisions have a direct impact on people's lives and significant financial investments, is its role in empowering stakeholders, guaranteeing ethical compliance, and promoting fair and unbiased AI solutions.

CHAPTER 3

Methodology

3.1 Convolutional Neural Networks

Convolutional neural networks, or CNNs, are a major advancement in machine learning, especially in the area of image analysis. Deep neural networks, of which CNNs are a subclass, were developed specifically to process and extract meaningful information from visual data. They are therefore excellent for a variety of tasks, including object detection and image classification. What sets CNNs apart from traditional neural networks is their ability to automatically extract relevant features from an image's raw pixel values. Layers of convolution and pooling operations enable the network to identify complex patterns, edges, textures, and higher-level visual structures. Because of their hierarchical structure, CNNs can progressively collect and process data at various levels of abstraction, which improves their capacity to recognize minute details in [22]. Consequently, a plethora of cutting-edge applications, including facial recognition software and medical image analysis, rely on CNNs as their foundation, fundamentally altering the way people interact with and process visual information.

At the heart of CNNs lies the concept of convolution, a fundamental operation that mimics the human visual system's ability to perceive features within images. CNNs perform convolutions by swiping tiny filters, referred to as kernels, across the input image in a methodical manner in order to identify unique patterns. Edges, corners, and textures are a few examples of these patterns; these elements are fundamental for identifying more intricate objects and structures. Convolution produces a feature map that indicates where these patterns are found in the input data. These feature maps' spatial dimensions are further reduced by pooling layers, which usually come after convolution layers [23]. This helps to preserve crucial information while lightening the computational burden. CNNs are able to automatically learn and extract features from images in situations where manual feature engineering would be impractical or not feasible. This is made possible by the combination of convolution and pooling operations.

Typically, convolutional neural networks (CNNs) have multiple layers, each of which is responsible for processing and comprehending visual input in a particular way. The fundamental layers of a CNN architecture consist of,

1. Input Layer: A CNN's input layer is where data enters the system. It takes the input image's raw pixel values and provides the first set of data for network processing. The input layer serves as the basis for later feature extraction and classification, and its dimensions match those of the input images.

2. Convolutional Layers: Important components of CNNs, convolutional layers identify patterns and features in the input data. These tiers utilize adaptive filters that traverse the input and execute convolutions to recognize pertinent attributes like borders, patterns, and intricate configurations. Multiple filters are used by convolutional layers to capture distinct features at different spatial scales.

3. Activation Layer: Activation Layers are vital because they add non-linearity to the network. The ability of the network to model complex relationships within the data depends on this non-linearity. Some examples of activation functions generally, used are:

   (a) The Rectified Linear Unit (ReLU), which substitutes zeros for negative values to enable the network to learn complex patterns effectively, is one of the most widely used activation functions in this layer. Beyond ReLU, various activation functions are employed based on specific needs.

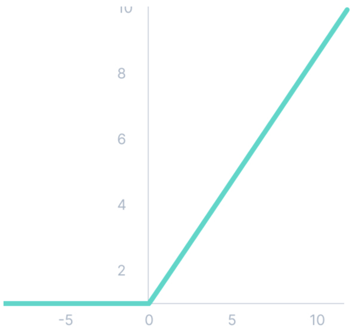

Figure 1: ReLU function from [24]

(b) The Sigmoid function reduces output values to a range of 0 to 1, making it appropriate for binary classification tasks.

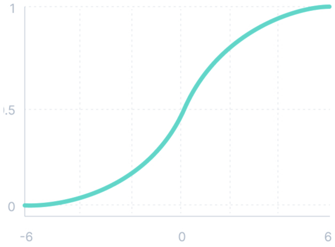

Figure 2: Sigmoid function from [24]

(c) For zero-centered data, the Hyperbolic Tangent (Tanh) function is suitable since it maps values to a range between -1 and 1.

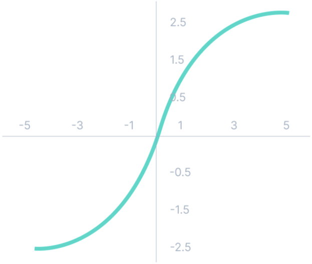

Figure 3: Tanh function from [24]

(d) Leaky ReLU allows small, non-zero gradients for negative inputs, thereby resolving the vanishing gradient issue with traditional ReLU

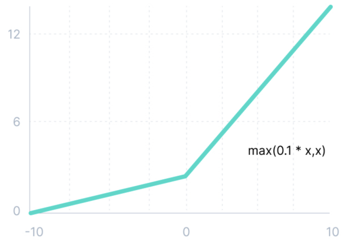

Figure 4: Leaky ReLU function from [24]

These varied activation functions give the network flexibility in defining its non-linear activation properties, enabling it to adjust to various tasks and data distributions.

4. Pooling (Subsampling) Layers: Pooling layers follow convolutional layers and serve to downsample the feature maps produced earlier. Pooling layers assist in controlling computational complexity while preserving crucial information by lowering spatial dimensions [25]. To downsample the data, max pooling and average pooling are popular methods.

5. Fully Connected Layers: Fully connected layers connect every neuron in one layer to every neuron in the subsequent layer. Based on the features retrieved by preceding layers, these layers allow the network to generate high-level predictions. Fully connected layers combine extracted features in classification tasks to identify the most likely class for the input data.

6. Output Layer: The output layer produces the final classification or prediction. The number of classes in a classification task is reflected in the number of neurons in this layer. In order to calculate class probabilities and base the

final prediction on the probability with the highest value, softmax activation is frequently utilized.

3.2 Theoretical Foundations of Explainable AI Techniques

3.2.1 Local Interpretable Model-Agnostic Explanations (LIME)

LIME, or local interpretable model-agnostic explanations, is a potent method in the area of explainable AI (XAI). The goal of LIME is to provide clear and understandable justifications for the predictions generated by intricate machine learning models. It works especially well when a model produces complicated or "black-box" output that is challenging to interpret. LIME's model independence is one of its most noteworthy qualities. It applies to any machine learning model, regardless of the architecture or underlying algorithm. LIME is a flexible tool for interpretability across various models and domains because of its adaptability.

The various steps involved in LIME are as follows:

1. Local Interpretability: The foundational idea of LIME's methodology is local interpretability. When an explanation is needed for a particular prediction, LIME concentrates on the immediate area around the data point of interest. Rather than trying to explain the entire behaviour of the model, this local perspective makes sure that the explanation is specific to the particular features of each instance.

2. Data Perturbation and Sampling: LIME introduces controlled changes to the data point's features in order to perturb it and produce explanations. A newly created dataset reflecting the local variations surrounding the instance is the outcome of this perturbation process. To evaluate the effect of various feature combinations on the model's predictions, randomness in data perturbation must be introduced. The perturbed data space is effectively explored through the

application of sampling techniques.

3. Surrogate Model Creation: Using the perturbed data space as a guide, LIME builds a surrogate model that roughly mimics the behaviour of the complex AI model. This stand-in model is typically a linear model that is chosen for its interpretability. The surrogate model's coefficients are meticulously calibrated to accurately represent the impact of every characteristic on the AI model's forecasts for the particular case [26]. Because of its model-agnostic nature, LIME can be used with a wide range of machine-learning models.

4. Locally Faithful and Interpretable explanations: The LIME process culminates in explanations that are both locally faithful and interpretable. These justifications provide a clear understanding of the reasons behind the AI model's decision-making for the particular data point in question by highlighting the contributions of individual features to the particular prediction.

After establishing the theoretical foundation of LIME, we now explore the mathematical foundation. For every X, there exists an interpretable binary vector of dimension d', denoted by X'. The presence or absence of a word (also known as the Bag of Words) is the interpretable vector X' for text data. The presence or lack of an image patch or superpixel in image data is represented by the interpretable vector X'. A contiguous patch of comparable pixels is all that a superpixel is. In tabular data, we perform feature binning if the feature is real-valued, and X' is the one-hot encoding of that feature if it is categorical. Any comprehensible model, such as decision trees or linear models, is the surrogate model $g \in G$. $\Omega(g)$ is a measure of surrogate model complexity. For instance, in the linear model, complexity rises with the number of non-zero weights. In a similar vein, the model's complexity rises with the decision tree's depth [27].

Next, $\Pi_x(Z)$ would be proximity calculation linking x and z, where Π_x is the measure of the locality of X. Local fidelity states that the surrogate function g() should resemble f() as much as possible in the vicinity of X. The degree to which the function g() approximates f() is indicated by the Loss function $\Lambda(f, g, \Pi_x)$.

$$\epsilon(x) = \underset{g \in G}{\operatorname{argmin}} \ \Lambda(f, g, \Pi_x) + \Omega(g), \tag{1}$$

$$\Lambda(f, g, \Pi_x) = \underset{z, z' \in Z}{\operatorname{argmin}} \ \Pi_x(z)(f(z) - g(z'))^2, \tag{2}$$

where $\Pi_x(z) = \exp\left(-\frac{D(x,z)^2}{\sigma^2}\right)$ is an exponential kernel, and g(z') $=$ (w_g)z', z' is an interpretable feature corresponding to z.

3.3 XAI in Healthcare

One of the most promising industries to use Explainable AI (XAI) is the healthcare sector. With the development of sophisticated machine learning models, XAI is essential for tackling important opportunities and problems in the healthcare industry. Enhancing the Diagnostic Accuracy is one such chance.

XAI has the power to fundamentally alter medical diagnosis, making it a disruptive force in the healthcare industry. One of the most intriguing applications of XAI in medicine is the analysis of medical imaging data from modalities like MRIs, CT scans, and X-rays. These diagnostic tools are vital for detecting a wide range of diseases, including diabetes, cancer, fractures, heart issues, and neurological conditions. Many stakeholders rely on the decisions made by AI models to be understandable, even though these models have demonstrated an impressive ability to process and interpret these images.

For radiologists and clinicians, interpretability is synonymous with trust and validation. It is insufficient for an AI model to produce a diagnosis on its own when it offers a quick evaluation of a medical image. Healthcare workers must comprehend how the model arrived at its conclusion in order to seamlessly integrate AI into clinical workflows. They look for process transparency, and interpretability offers the required context. By carefully examining the areas of interest in an image and understanding the characteristics or patterns that influenced the AI's judgment, radiologists and clinicians can diagnose patients with greater confidence and knowledge. In the end, this collaborative approach to medical diagnosis produces more accurate and trustworthy diagnoses by fusing the knowledge of medical specialists with the analytical capabilities of AI.

Patients, too, stand to benefit significantly from the interpretability offered by XAI in medical imaging. The results of diagnostic procedures have the potential to change people's lives. A medical image could show that there is a serious ailment that needs to be treated right away, or it could show that there is a less serious problem that just needs to be watched over. In either scenario, patients need to know why they were given a specific diagnosis. By providing patients with an understanding of the logic underlying the AI's judgments, XAI empowers them. Patients are more confident in the suggested treatment plan as a result of this increased transparency. When patients can understand the reasoning behind medical advice, they are more likely to trust it. Maintaining patient compliance, lowering anxiety, and fostering a positive patient experience all depend on this trust.

In essence, XAI fosters a partnership between medical professionals, patients, and AI systems, with the common goal of achieving the most accurate and beneficial medical diagnoses. XAI's contribution to improving diagnostic precision extends beyond the field of medical imaging. It covers a wide range of healthcare scenarios, such as the interpretation of laboratory test results, patient outcome prediction using predictive analytics, and even the detection of possible drug interactions. XAI positions itself as a driving force behind more accurate, responsible, and transparent healthcare practices by offering a transparent and easily understandable explanation for its decisions.

3.4 Experimental Setup

3.4.1 Dataset

The Ocular Disease Intelligent Recognition (ODIR) database, a structured ophthalmic repository with 5,000 patient records, was obtained from Kaggle for this research [28]. The patient's age, gender, colour fundus photos of both the left and right eyes, and the diagnostic keywords supplied by the doctors are among the some of the details included in these records.

The purpose of this dataset is to depict a "real-life" set of patient data that Shanggong Medical Technology Co., Ltd. gathered from various Chinese hospitals and medical facilities. These institutions use a variety of cameras on the market, including Canon, Zeiss, and Kowa, to take fundus images, which produce images with different resolutions.

The key fields within the dataset are, the left-eye retinal fundus images are identified by the "Left-Fundus" field in the ODIR dataset. The "Right-Fundus" field has retinal fundus images, just like the "Left-Fundus," but they depict the view from the right eye in this instance. Then, descriptive terms or phrases related to the diagnostic conclusions found in the left-eye retinal images are entered into the "Left-Diagnostic Keywords" field. The terms or phrases in the "Right-Diagnostic Keywords" field, like in the "Left-Diagnostic Keywords," are descriptive and related to the diagnostic results found in the retinal images of the right eye.

ODIR addresses a range of ocular conditions, such as hypertension, age-related macular degeneration (AMD), glaucoma, and diabetic retinopathy. The dataset has two subsets: one for the left eye and one for the right. The distribution of ocular diseases within the dataset is shown visually in the graph below. It provides insight into the makeup of the medical photos used in the study by illuminating the prevalence of different eye conditions.

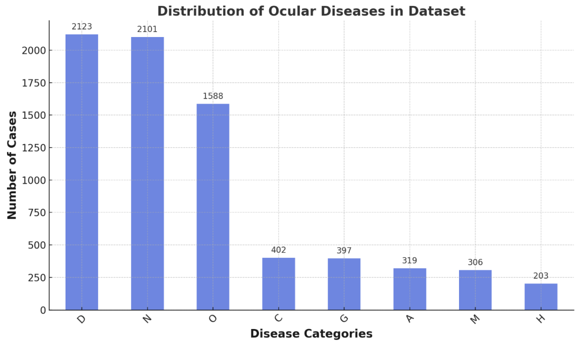

Figure 5: Dataset distribution

3.4.2 Data Preprocessing

In the process of preparing the ocular dataset for analysis and machine learning, it became evident that two key categories, ’Diabetic’ and ’Normal,’ held the majority of images. Several data preprocessing procedures were applied to these images in recognition of the significance of these categories. The initial goal was to correctly categorize the pictures using the diagnostic keywords connected to every single eye image. Images of the left and right eyes that were categorized as "Normal Fundus" were located and extracted using diagnostic keywords. One result of the process was a list of image names classified as "Normal."Images of the left and right eyes were separated and their labels bearing the diagnostic term "retinopathy" were extracted. Consistent with the "Normal" images, this process additionally generated a list of image names labelled as "Diabetic."This categorization procedure attempted to treat left and right images as distinct entities, rather than grouping individual images into the "Diabetic" and "Normal" categories. This made it possible to analyze the dataset in a more specialized and focused way.

Next, in order to prepare our ocular dataset for machine learning analysis, data normalization is an essential preprocessing step. All of the images’ pixel values must be rescaled to a common range, usually between 0 and 1. Because it guarantees that biases in our machine learning model are not introduced by pixel intensity variations across different images, this process is essential. Images with higher pixel values are kept from controlling the learning process by levelling the playing field for the model through data normalization. This is a crucial step because consistent feature interpretation is crucial in the analysis of medical images. Stabilizing the training process with the help of normalized data also contributes to more dependable outcomes.

To standardize the dimensions of every image in our ocular dataset, we also performed data resizing in addition to data normalization. Every image was resized to 224 x 224 pixels, which was the standard size. This resizing has useful advantages in addition to guaranteeing that the images can be fed into a convolutional neural network (CNN). Consistently sized images reduce the processing burden during training, resulting in a more efficient and controllable process. The 224x224 pixel size is a compromise between limiting computational resources and maintaining enough image detail for precise classification. This preprocessing stage helps our machine learning model to be more robust and efficient overall, which makes it capable of handling a variety of ocular images.

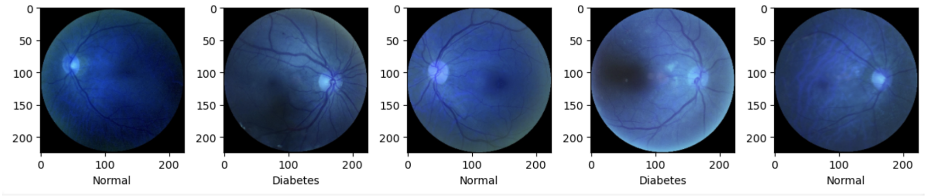

Figure 6: Dataset visulization

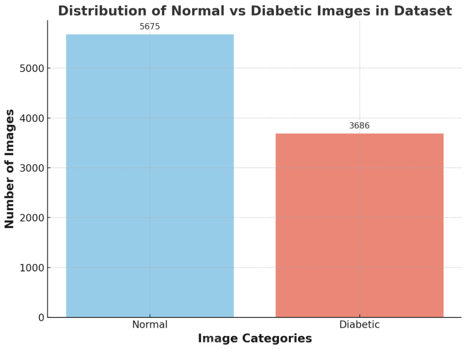

Figure 7: Dataset Imbalance

Due to the imbalanced distribution of diabetic and normal images, we sought to address this issue through data augmentation. The diabetic image category contained fewer samples compared to the normal category, which could potentially lead to a bias in the model's predictions. To mitigate this, we employed data augmentation techniques to artificially increase the size of the diabetic image dataset. This process involved generating new, slightly modified versions of existing diabetic images, effectively expanding the dataset and balancing the number of samples between the two categories. The augmented images retained the key characteristics necessary for accurate classification while introducing variability that aids in training a robust model.

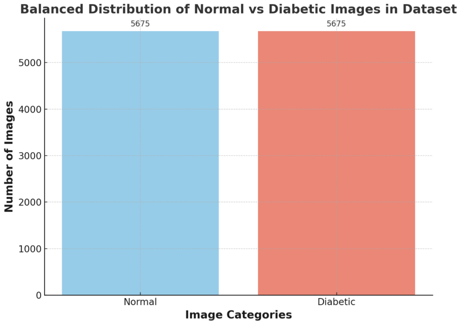

Figure 8: Dataset after oversampling

3.4.3 CNN Training

To train and evaluate our Convolutional Neural Network (CNN), we carefully split our dataset into three subsets: the training set, the testing set, and the validation set. The training set, which gets 70% of our dataset, serves as the foundation for the learning process of our model. This is a typical machine learning procedure. Using the training set, CNN can recognize underlying patterns and features in the images. To evaluate the model's generalization to new, unseen data, we also reserve 20% of the dataset as the testing set, which serves as an unobserved benchmark. Finally, a 10% portion of the dataset was dedicated to the validation set, which is crucial for hyperparameter tuning and model selection. This separation strategy ensures that our CNN is trained, validated, and tested on distinct data subsets, contributing to robust and reliable performance evaluation.

To achieve the best outcome for every set of features, CNN has multiple hyperparameters that can be optimized. We utilize the Grid Search module found in the sklearnlibrary to find these. To achieve this, we first give several different values for each parameter we wish to tune. After that, the module runs a comprehensive search using every possible combination of the parameters. The most accurate combination gets chosen. The fact that it finds the best combination by doing cross-validation over five folds for each combination is also very helpful. In other words, the training set fed into GridSearch will verify the various combinations across five distinct intermediate test sets derived from the original training set. This module's cross-validation feature helps make the results more reliable and broadly applicable, as our experiments typically involve fewer samples per label. The parameters that we tested for the various classifiers are displayed in the Table below.

Table 1: Hyperparameters Tested

Hyperparameter	**Tested Values**
Learning Rate	0.1, 0.01, 0.001
Epochs	40, 50, 60
Activation Functions	Tanh, ReLU
Optimizer	Adam, SGD

3.4.4 LIME

An important factor in improving the interpretability and transparency of our CNN model is LIME. In the context of our ocular disease classification experiments, we explore the useful implementation of LIME in this section.The LIME model provides us with a robust framework for producing interpretable, localized explanations for each individual prediction. We utilize LIME to get insights into why our model makes specific predictions for particular instances by picking particular examples from our dataset, like fundus images. The process entails creating perturbed samples and perturbing the chosen instance. We use our CNN model, which is in charge of classifying ocular diseases, to predict both the perturbed samples and the original instance. Next, LIME builds a surrogate model that approximates the intricate CNN behaviour in a nearby neighbourhood surrounding the selected instance. We can decipher the surrogate model's coefficients and feature importances, providing insight into the variables that shaped the model's choice for the chosen example.

Several useful advantages of LIME are demonstrated in our experiments,

1. Interpretability: It improves our comprehension of our CNN model's decision-making process by giving us concise, understandable explanations for each prediction.

2. Transparency: The use of LIME improves our deep learning model's transparency, making it more dependable and approachable, especially in the field of medicine.

CHAPTER 4

Results

4.1 CNN Evaluations

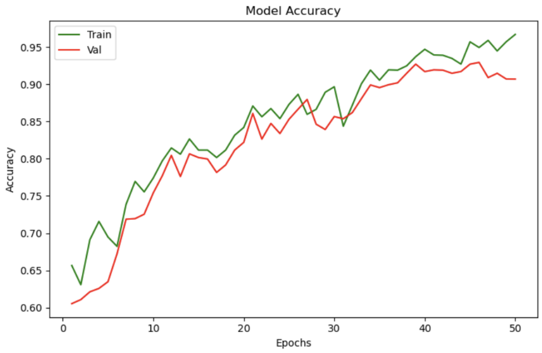

Figure 9: Model Accuracy

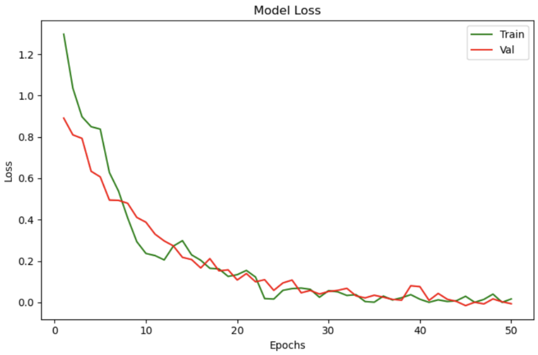

Figure 10: Model Loss

Four important performance metrics were tracked during the training phase of our Convolutional Neural Network (CNN) model: training accuracy and loss, as well as validation accuracy and loss on an independent dataset. These measures are crucial for assessing the model's effectiveness and spotting any overfitting or underfitting problems.

Our CNN successfully learned to identify patterns and features in the ocular images as evidenced by the training accuracy and validation accuracy both increased steadily throughout the training process. This improvement is a testament to the model's capacity to capture the complexities of the dataset. Simultaneously, we observed the training loss and validation loss, which quantify the difference between the expected and true labels in corresponding data sets. The model's capacity to reduce errors and enhance its predictive accuracy was demonstrated by the steady decline in both kinds of losses during training.

An extensive assessment of the model's performance in differentiating between diabetic and normal eye images can be found in the classification report table below. This report is a crucial tool for assessing the model's ability to correctly identify each class. Here, recall indicates the model's ability to correctly classify every image in the correct target class, whereas precision gauges the model's accuracy of its prediction for the target class. The F1-score, a reasonable measure of overall accuracy, is obtained by taking the harmonic mean of recall and precision. With a precision of 0.91 for both classes, the model correctly classifies images as "Normal" or "Diabetes" 91% of the time. In contrast, the recall for the "Diabetes" class is 0.93 and for the "Normal" class is 0.90. This means that the model correctly identifies 90% of real "Normal" images and 93% of real "Diabetes" images. Furthermore, the F1 scores are 0.92 and 0.91 for the "Diabetes" class and the "Normal" class, respectively. Last but not least, the model's overall accuracy of 0.91 shows that it can correctly categorize 91% of the images into the relevant classes. When combined, these metrics show how well the model can classify images, suggesting that it could be a valuable tool for medical image analysis pertaining to diabetic retinopathy.

Table 2: Classification Report

	Precision	**Recall**	**F1-score**	**Support**
Normal	0.91	0.93	0.92	1071
Diabetes	0.92	0.91	0.91	1199
Accuracy			0.91	2270

The confusion matrix provides a comprehensive examination of the model's predictions, showing the percentage of cases correctly classified as "Normal" or "Diabetes," in addition to the cases that were misclassified. It assists in determining misclassifications and evaluating the accuracy of the model. It shows how many false negatives were incorrectly classified as the opposite class, how many false positives were incorrectly classified as "Diabetes" when they were actually "Normal," how many false negatives were incorrectly classified as "Normal" when they were actually "Diabetes," and how many true positives were correctly classified as either "Normal" or "Diabetes."

Out of the 1199 total "Normal" images, 1097 were correctly classified as "Normal", on the other hand, 102 cases of "Normal" as "Diabetes," representing a 9% error rate. A 92% success rate was achieved for the 'Diabetes' class, with 983 cases correctly identified. An 8% misclassification rate for this category was achieved by mistakenly labelling 88 cases as "Normal". This shows that the model predicts the data for both classes with a high degree of accuracy, but it also shows that there are false positives and false negatives, pointing to potential areas for improvement, especially in lowering the number of cases that are misdiagnosed.

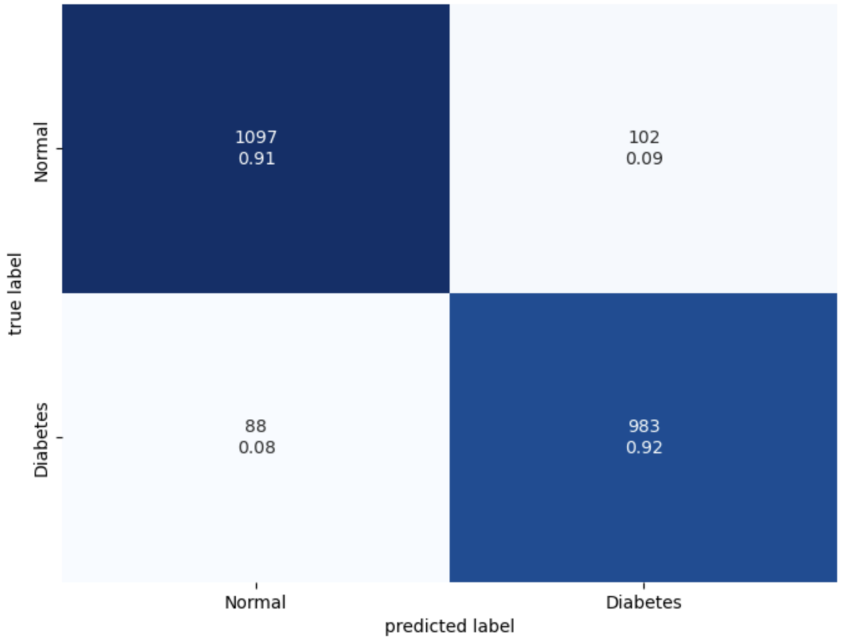

Figure 11: Confusion Matrix

4.1.1 CNN Sensitivity Evaluation

Sensitivity is a crucial indicator of a model's clinical utility in medical data analysis, especially when it comes to identifying conditions like diabetic retinopathy. Sensitivity in this context refers to the model's capacity to accurately distinguish "true positives"—cases of diabetic eye conditions—from the total number of cases that are actually diagnosed with the disease. Sensitivity is computed as follows:

$$Sensitivity = TruePositives/(TruePositives + FalseNegatives) \tag{3}$$

In order to accurately diagnose patients and provide them with the care they need, high sensitivity is crucial in medical diagnostics. As can be seen by looking at the high recall rate of 0.93 for the 'Diabetes' class, our CNN model showed excellent sensitivity. This shows that the model correctly detected 93% of the real cases of diabetes, which is an important reduction in the possibility of missing patients who require medical attention. Even so, there are some false negatives, which emphasizes the continuous difficulty in reaching perfect diagnostic accuracy. The sensitivity is not perfect. Effective medical image analysis is centred on striking a balance between maximizing true positive rates and minimizing false negatives.

It is also necessary to consider the model's sensitivity in light of its overall performance metrics, such as F1-score and precision. With a precision of 0.91, the model is 91% accurate 91% of the time when it classifies an image as "Diabetic." On the other hand, the marginally reduced recall of the 'Normal' class at 0.90, which also indicates the sensitivity of the model in recognizing normal cases, implies that there is room for improvement in accurately recognizing every normal case. The model's balanced performance in terms of accuracy and reliability is confirmed by the F1 scores for both classes, which are near the precision and recall values. However, as the confusion matrix illustrates, the existence of false positives and negatives indicates a crucial area requiring additional refinement. Continuous efforts to improve these metrics, including sensitivity, are essential in medical imaging, where the cost of misdiagnosis can be high. Although our CNN model's sensitivity level offers a promising starting point, it also emphasizes the necessity of ongoing improvements in model validation and training in order to improve the model's diagnostic accuracy.

4.2 LIME Evaluations

This section explores the evaluation results obtained from LIME for correct and incorrect classifications of our CNN model.

4.2.1 Correct Predictions

This section examines CNN's prediction of accurately classifying diabetic ocular conditions. The third image shows how the highlighted areas in the mask match the actual features in the eye image by combining the LIME mask with the original.

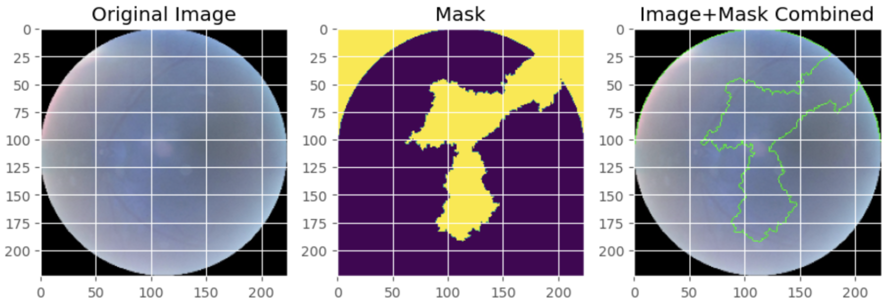

Figure 12: Positive pixels of Correct classification

Moving on in the same diabetic image, the second image in Figure 13 displays the LIME mask, which highlights areas in the original image's top left and bottom where the CNN identified unfavourable features. Despite this, our model identified the image as a diabetic eye with success. Ultimately, the third image provides a comprehensive view of the regions that affected the model's classification decision by combining the LIME mask with the original image. This illustrates how CNN can still produce precise classification even in cases where specific image regions point to a different classification.

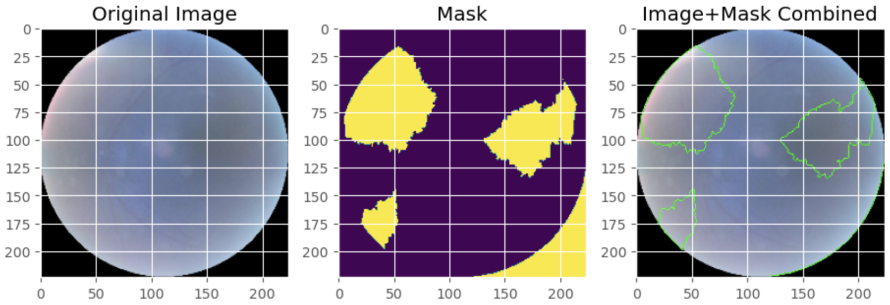

Figure 13: Negative pixels of Correct classification

4.2.2 Incorrect Predictions

In this section, we explore the CNN's analysis of a diabetic eye that was incorrectly classified as normal. The first image in Figure 14 is the original image of another diabetic eye, again acting as the unmodified baseline for our analysis. The regions that had a positive influence on the Convolutional Neural Network (CNN) model's classification of the image as a normal eye are indicated by the top pixels in the second image of the figure below. The CNN identified these areas as having strong characteristics linked to a normal eye, which resulted in the model classifying them incorrectly. The phrase "positive impact" denotes that these areas showed characteristics that CNN considered typical of a normal eye when it conducted its analysis.

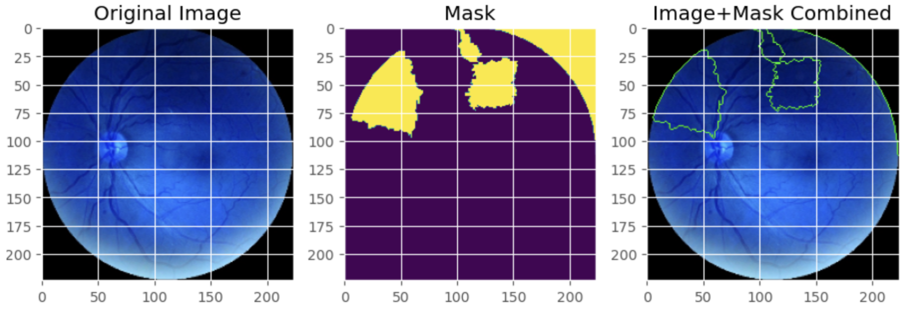

Figure 14: Positive pixels of Incorrect classification

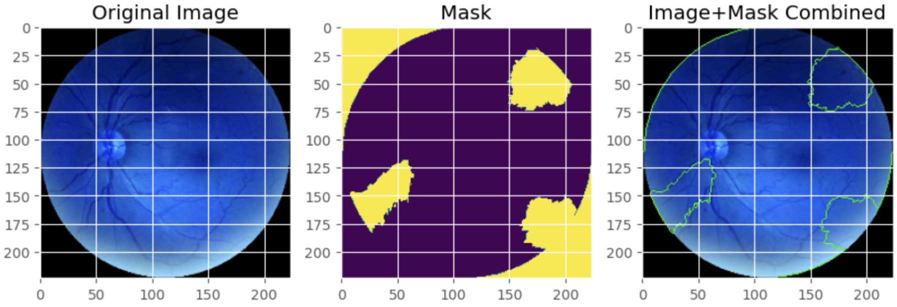

Figure 15: Negative pixels of Incorrect classification

Then, in Figure 15, the mask image highlights the 3 corner regions of the eye which negatively affected the CNN in its incorrect classification. These regions were identified by the CNN to have contained features associated with a diabetic eye. However, despite the presence of these features, CNN decided to classify the eye as normal, potentially due to the limited influence of these corner regions on the overall classification decision. This intriguing finding emphasizes CNN's intricate decision-making process and emphasizes the significance of comprehending both the image's positive and negative aspects that influence its ultimate classification.

In cases where doctors diagnose images as normal but CNN classifies them as diabetic, medical professionals should investigate the areas that CNN flags more closely. The reason for this reevaluation is that the AI may have identified early warning indicators of diabetes that may have been missed by the doctors. This kind of examination may reveal disease indicators in their early stages, which if missed could result in postponing treatment. On the other hand, when a doctor diagnoses a patient with diabetes but the CNN classifies the image as normal, analyzing these differences can offer important insights for fine-tuning the AI model. Medical professionals can provide additional information to the AI to help it learn more by pointing out the precise characteristics or patterns that it missed. Through this iterative process of incorporating expert feedback, the accuracy of the AI model is continuously improved as it evolves to more closely resemble the complex understanding that doctors bring to diagnoses.

Both methods use the AI's analytical power to enhance and supplement medical knowledge while reinforcing the comprehensiveness of clinical evaluations. By identifying and treating diabetic conditions in their early stages, this collaborative approach combining AI and human judgment could result in more precise diagnoses and prompt interventions, improving patient treatment.

CHAPTER 5

Conclusion and Future Works

In this work, we conducted a thorough investigation into Convolutional Neural Networks (CNNs) for the purpose of classifying images related to eye diseases. Using the power of CNNs, we developed a robust model that can distinguish between images of normal and diabetic retinopathy with high reliability. The outcomes of our experiments demonstrated potential and validated the utility of CNNs in the healthcare sector, with a 91% classification accuracy. Furthermore, we investigated Explainable Artificial Intelligence (XAI) by demonstrating our model's decision-making process using LIME (Local Interpretable Model-agnostic Explanations). LIME's astute examination of the image's regions of interest allowed for a deeper understanding of the model's predictions. By doing this, we were able to pinpoint regions where positive and negative pixel contributions were present, deciphering the complex patterns the model was using to classify data. We also found examples of both accurate and inaccurate labelling by the model in LIME. Through pixel contribution visualization, we were able to discern the model's strong points and weak points and learn why specific predictions were made. This sophisticated comprehension lays the path for upcoming enhancements and performance adjustments of the model.

There are numerous directions that future research and development could go as we advance. First off, our model has proven to be capable of binary classifying images of normal and diabetic retinopathy. The crucial next step is to increase its capacity to manage multi-class classifications, which include a range of ocular conditions. Furthermore, there is potential for our model to be implemented more widely in healthcare facilities. Integration with clinical workflows would be necessary for real-world deployment in order to guarantee smooth communication between medical personnel and the AI system. This entails attending to security and privacy issues as well as following legal mandates. Our work also provides an avenue to investigate further XAI methods in the context of interpretable AI. We can improve our model's transparency and clinicians' depth of insight by exploring and applying cutting-edge interpretability techniques.

In summary, our study not only makes a significant contribution to the field of ocular disease image classification but also emphasizes how crucial interpretability is for AI-driven medical interventions. We have made progress toward clearer and more accurate diagnoses by fusing the power of CNNs with XAI, which will ultimately help patients and medical professionals combat visual impairments.

LIST OF REFERENCES

[1] W. Wang and A. Lo, ‘‘Diabetic retinopathy: Pathophysiology and treatments,’’ *International Journal of Molecular Sciences*, vol. 19, no. 6, p. 1816, 2018.

[2] A. Rastogi, T. Z. Rizvi, and D. Deeba Kanan, ‘‘Diabetic retinopathy - an ensemble approach,’’ in *2023 International Conference on Computer, Electrical Communication Engineering (ICCECE)*, 2023, pp. 1--10.

[3] F. A. Suwandi, P. Ricky Kurnianda, and A. A. Santoso Gunawan, ‘‘A systematic literature review: Diabetic retinopathy detection using deep learning,’’ in *2023 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2023, pp. 12--17.

[4] M. Jena, S. P. Mishra, and D. Mishra, ‘‘Detection of diabetic retinopathy images using a fully convolutional neural network,’’ in *2018 2nd International Conference on Data Science and Business Analytics (ICDSBA)*, 2018, pp. 523--527.

[5] P. Dutta, P. Upadhyay, M. De, and R. Khalkar, ‘‘Medical image analysis using deep convolutional neural networks: Cnn architectures and transfer learning,’’ in *2020 International Conference on Inventive Computation Technologies (ICICT)*, 2020, pp. 175--180.

[6] E. N. Volkov and A. N. Averkin, ‘‘Explainable artificial intelligence in medical image analysis: State of the art and prospects,’’ in *2023 XXVI International Conference on Soft Computing and Measurements (SCM)*, 2023, pp. 134--137.

[7] L. Moradi, B. Kalantar, E. H. Zaryabi, A. A. Halin, and N. Ueda, ‘‘On the use of xai for cnn model interpretation: A remote sensing case study,’’ in *2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2022, pp. 1--5.

[8] P. D. S, R. Kumar K, V. S, N. K, and A. K, ‘‘An overview of interpretability techniques for explainable artificial intelligence (xai) in deep learning-based medical image analysis,’’ in *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, 2023, pp. 175--182.

[9] Z. Papanastasopoulos, R. K. Samala, H.-P. Chan, L. Hadjiiski, C. Paramagul, M. A. Helvie, and C. H. Neal, ‘‘Explainable AI for medical imaging: deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI,’’ in *Medical Imaging 2020: Computer-Aided Diagnosis*, H. K. Hahn and M. A. Mazurowski, Eds., vol. 11314, International Society for Optics and Photonics. SPIE, 2020, p. 113140Z. [Online]. Available: https://doi.org/10.1117/12.2549298

[10] 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*, vol. 79, p. 102470, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841522001177

[11] G. Yang, A. Rao, C. Fernandez-Maloigne, V. Calhoun, and G. Menegaz, ‘‘Explainable ai (xai) in biomedical signal and image processing: Promises and challenges,’’ in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 1531--1535.

[12] 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, may 2016. [Online]. Available: https://doi.org/10.1109%2Ftmi.2016.2535302

[13] S. S. Kshatri and D. Singh, ‘‘Convolutional neural network in medical image analysis: A review,’’ *Archives of Computational Methods in Engineering*, vol. 30, no. 4, pp. 2793--2810, 2023.

[14] U. Pawar, D. O’Shea, S. Rea, and R. O’Reilly, ‘‘Explainable ai in healthcare,’’ in *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, 2020, pp. 1--2.

[15] Y. Liu, ‘‘Artificial intelligence and machine learning based financial risk network assessment model,’’ in *2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT)*, 2023, pp. 158--163.

[16] G. Vilone and L. Longo, ‘‘Explainable artificial intelligence: a systematic review,’’ 05 2020.

[17] G. P. Reddy and Y. V. P. Kumar, ‘‘Explainable ai (xai): Explained,’’ in *2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, 2023, pp. 1--6.

[18] M. A. Ahmad, A. Teredesai, and C. Eckert, ‘‘Interpretable machine learning in healthcare,’’ in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, 2018, pp. 447--447.

[19] S. Bharati, M. R. H. Mondal, and P. Podder, ‘‘A review on explainable artificial intelligence for healthcare: Why, how, and when?’’ *IEEE Transactions on Artificial Intelligence*, pp. 1--15, 2023. [Online]. Available: https://doi.org/10.1109%2Ftai.2023.3266418

[20] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[21] S. Gaikwad, ‘‘Study on artificial intelligence in healthcare,’’ in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, 2021, pp. 1165--1169.

[22] A. Ajit, K. Acharya, and A. Samanta, ‘‘A review of convolutional neural networks,’’ in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 2020, pp. 1--5.

[23] S. Albawi, T. A. Mohammed, and S. Al-Zawi, ‘‘Understanding of a convolutional neural network,’’ in *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1--6.

[24] P. Baheti. ‘‘Activation functions in neural networks.’’ 2021. [Online]. Available: https://www.v7labs.com/blog/neural-networks-activation-functions

[25] S. Albawi, T. A. Mohammed, and S. Al-Zawi, ‘‘Understanding of a convolutional neural network,’’ in *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1--6.

[26] V. Mathivanan. ‘‘Everything you need to know about lime.’’ 2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2022/07/everything-you-need-to-know-about-lime/

[27] M. T. Ribeiro, S. Singh, and C. Guestrin, ‘‘"why should i trust you?": Explaining the predictions of any classifier,’’ 2016.

[28] Kaggle. ‘‘Ocular disease recognition.’’ 2020. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k

APPENDIX A

Undersampling Experiment

A.1 Data Preprocessing

The model's capacity to distinguish between "normal" and "diabetic" eye images was improved earlier in the project when we used data augmentation. One of the oversampling techniques on our unbalanced dataset produced a balanced dataset, which ultimately produced positive results. In order to attain a balanced representation of both classes in our dataset, we are currently investigating the effects of undersampling as a substitute approach.

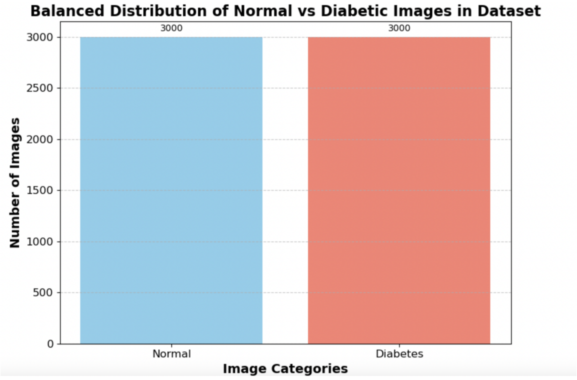

Figure A.16: Dataset after undersampling

We lowered the quantity of "normal" eye images from 5675 to 3000 for each category, as shown in Figure A.16, to put them on par with the "diabetic" images. The 'normal' class images were chosen at random to achieve this, guaranteeing that each image had an equal chance of being included and preserving the sample's diversity. The balanced dataset that is produced, as shown, should give a more accurate representation of the model's performance because it is not skewed by the previously asymmetric class distribution.

A.2 CNN training

We used grid search for thorough hyperparameter optimization in the CNN model's training phase, which is similar to our approach when investigating oversampling techniques. The table below displays the range of hyperparameters that we experimented with, which includes changes to the activation functions, optimizers, epochs, and learning rate. Five-fold cross-validation was used to make sure that the performance of our model was robustly validated. By dividing the dataset into five parts and iteratively using one part for training and the other for validation, this technique made it easier to evaluate the consistency and efficacy of the model across a variety of data samples. The combination of hyperparameters that yielded consistently promising accuracy and f-1 score across all folds was chosen.

Table A.3: Hyperparameters Tested

Hyperparameter	**Tested Values**
Learning Rate	0.1, 0.01, 0.001
Epochs	30, 40, 50
Activation Functions	Tanh, ReLU
Optimizer	Adam, SGD

A.3 Results

A.3.1 CNN Evaluations

Our A.17 accuracy graph shows an initial positive trend in the CNN results, with training and validation accuracies increasing over time but reaching a plateau in the 80%–86% range. This pattern indicates that although the model performs well in terms of initial generalization, it reaches a performance ceiling of approximately 86% accuracy. Our figure A.17, loss graph shows a consistent decline in both the training and validation losses. Together, this suggests that the model is capable of learning until a particular moment.

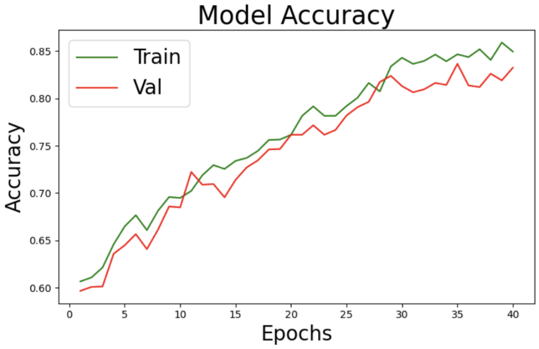

Figure A.17: Model Accuracy

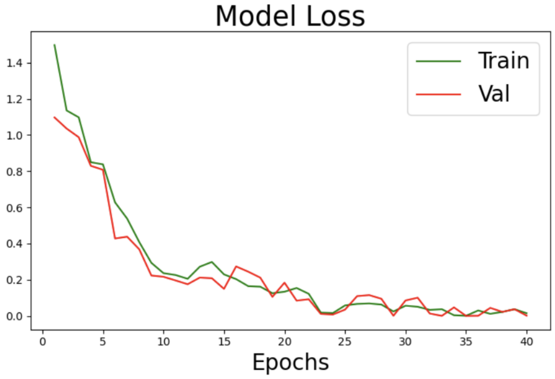

Figure A.18: Model Loss

As per the classification report below, our CNN model exhibits a moderate level of effectiveness. Its precision, recall, and F1-score all fall within the low 80s for both the 'Normal' and 'Diabetic' categories. Specifically, the 'Diabetic' class exhibits an F1-score of 0.83 over 624 images, whereas the 'Normal' class yields an F1-score of 0.82 over 576 images with a precision of 0.82 and recall of 0.81. Furthermore, the model can predict 83 out of every 100 images with an accuracy of 83% overall. Though these numbers suggest a well-balanced model, the scores in the 80s also indicate our model requires improvement, particularly in medical image classification where higher scores are crucial.

Table A.4: Classification Report

	Precision	**Recall**	**F1-score**	**Support**
Normal	0.82	0.81	0.82	576
Diabetes	0.83	0.84	0.83	624
Accuracy			0.83	1200

A visual and numerical depiction of the CNN model's performance in classifying images as "Normal" or "Diabetic" can be found in the confusion matrix shown in Figure A.19. 109 photos were mistakenly labelled as "Diabetic" (false positives), with a normalized value of 0.18, whereas 515 correctly identified (true positives) out of the total images classified as "Normal," according to the matrix. On the other hand, the 'Diabetic' class contained 475 images that were correctly identified as 'Diabetic' (true negatives), with a normalized value of 0.83, and 101 incorrectly classified as 'Normal' (false negatives). This breakdown shows a higher true positive rate for both classes, but it also shows a significant percentage of false positives and negatives, indicating that although the model is reasonably accurate, its precision and recall could be enhanced.

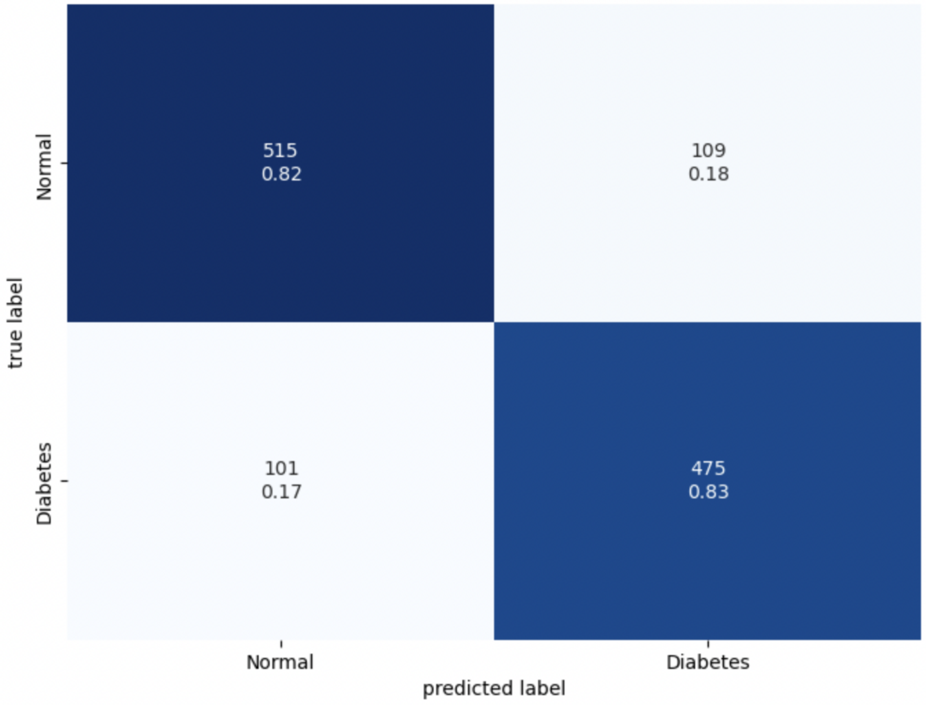

Figure A.19: Confusion Matrix

A.3.2 LIME Evaluations

A.3.2.1 Correct Predictions

In this section, we explore the CNN's prediction of correctly classified diabetic eye. The first image of Figure A.20 displays the original diabetic eye image from the dataset and will be used as the unaltered baseline for our analysis. Our LIME mask highlights the regions in the second image, mainly on the right side, where the CNN found positive features associated with diabetic retinopathy in the original image.

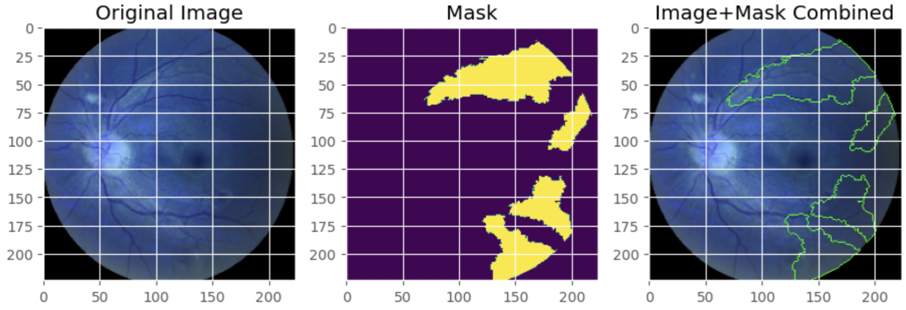

Figure A.20: Positive pixels of Correct classification

Proceeding with the diabetic image, the second image in Figure A.21 shows the LIME mask, emphasizing regions in the top right and bottom of the original image where the CNN detected unfavorable features. In spite of this, our model successfully recognized the image as a diabetic eye. Finally, by fusing the LIME mask with the original image, the third image offers a thorough perspective of the areas that influenced the model's classification choice.

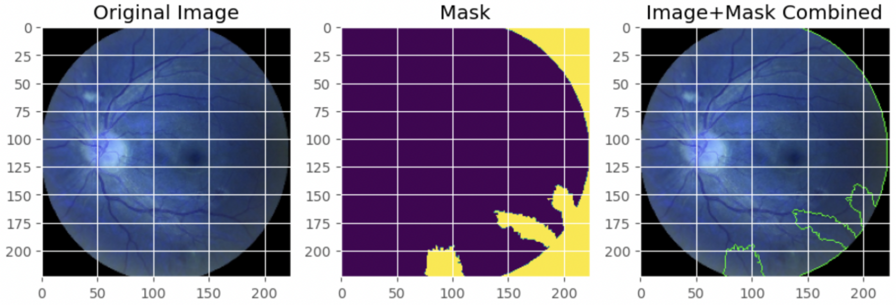

Figure A.21: Negative pixels of Correct classification

A.3.2.2 Incorrect Predictions

We examine CNN's analysis of a diabetic eye that was mistakenly identified as normal in this section. In the second image of the Figure A.22, the top right, top left, and bottom right pixels represent the regions that positively impacted the model's classification of the image as a normal eye. The model misclassified these areas because the CNN determined that they had strong characteristics associated with a normal eye. The term "positive impact" indicates that during CNN's analysis, these areas displayed traits that were thought to be typical of a normal eye.

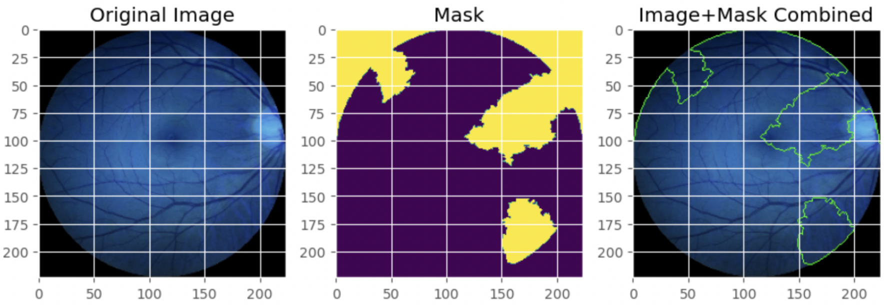

Figure A.22: Positive pixels of Incorrect classification

Next, the mask image in Figure A.23's second image draws attention to the upper and lower left eye regions, which had a negative impact on CNN's inaccurate classification. CNN determined that these areas included characteristics typical of a diabetic eye. CNN chose to classify the eye as normal despite the presence of these features, possibly as a result of these regions' limited impact on the final classification choice.

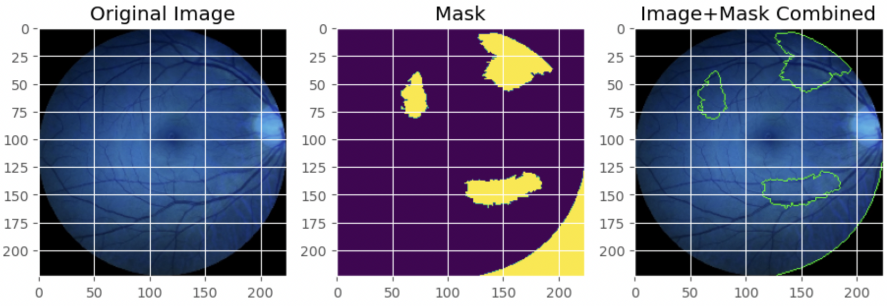

Figure A.23: Negative pixels of Incorrect classification

Conclusively, our investigation into CNN models utilizing LIME analysis, supported by an undersampling approach, has yielded significant knowledge regarding the characteristics that are most significant in classifying eye images into two categories:

"Normal" and "Diabetic." Our comprehension of the internal workings of the model has improved with the identification of the critical pixels that both favorably and unfavorably influence the decisions made by the model. On the other hand, the 80–85% range is where the overall accuracy and associated performance metrics—like recall, precision, and F1-score—have stabilized. These are respectable numbers overall, but they don't meet the high standards for accuracy that are common in the medical field, where incorrect classification can have particularly costly consequences. Notably, our prior work with oversampling produced better results, suggesting that oversampling is more effective than undersampling in this particular situation. In the context of eye image classification, where error margins are small and accuracy is highly valued, this comparison unequivocally suggests that oversampling is the better method.