الجمهورية الجزائرية الديمقراطية الشعبية

People’s Democratic Republic of Algeria

وزارة التعليم العالي والبحث العلمي

Ministry of Higher Education and Scientific Research

Constantine 1 University Frères Mentouri
Faculty of Natural and Life Sciences

جامعة قسنطينة 1 الإخوة منتوري
كلية علوم الطبيعة والحياة

Department: Applied Biology

قسم: البيولوجيا التطبيقية

Thesis submitted in partial fulfillment of the requirements for the Master’s degree

Domain: Natural and Life Sciences
Field: Biotechnology
Specialty: Bioinformatics

N° d’ordre :
N° de série :

Title:

An Explainable Deep Learning-Based Framework for Diabetic Retinopathy Detection.

Date: 25/06/2025

Submitted by:

AIECH Hadjer
BACHKHAZNADJI Chahinez
BAMIA Hamadoun

Board of Examiners:

Chairperson: Dr. BENHAMDI Asma, (MCA), Constantine 1 Frères Mentouri University
Supervisor: Dr. CHEHILI Hamza, (MCA), Constantine 1 Frères Mentouri University
Examiner: Dr. GHERBOUDJ Amira, (MCA), Constantine 1 Frères Mentouri University

Academic Year
2024 – 2025

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to all those who supported and guided us throughout the development of this thesis.

First and foremost, we are deeply grateful to our supervisor, **Dr CHEHILI Hamza**, whose constant presence, availability, and unwavering support played a pivotal role in shaping this work. His insightful feedback, patience, and constructive criticism consistently pushed us to improve and strive for academic excellence. His dedication and mentorship have been a source of inspiration throughout the research process.

We are especially thankful to **Dr. BENMOUSSA Salah Eddine**, head of the clinic ***"Clinique d'Ophtalmologie Benmoussa"***, for welcoming us for an internship at his clinic and for granting us access to valuable clinical data. His support contributed significantly to the validation and practical relevance of our study.

We also extend our appreciation to the faculty and staff of the **Applied Biology** at **Mentouri Constantine University 1** for providing an enriching academic environment and the necessary resources to conduct our research.

To our families and friends, thank you for your constant encouragement, patience, and emotional support. Your belief in us helped us stay focused and motivated during this journey.

Lastly, we would like to acknowledge the collaboration, commitment, and collective effort of each team member. This thesis reflects not only our shared academic goals but also the strong spirit of teamwork and mutual respect that guided us to its completion.

DEDICATION

This thesis is dedicated with deep gratitude and affection to the people who have shaped and supported me throughout this journey.

To my parents, **Brahim** and **Nadjet**, your unwavering love, sacrifices, and constant encouragement have been the foundation of everything I've achieved. I am forever grateful.

To my siblings, especially my sister **Amal**, thank you for your encouragement, your patience, and for always being there when I needed support.

To my best friends **Selsabil**, **Aya**, and especially to my lifelong friend **Célia**, whose presence and support over 18 years of unbreakable friendship have been a constant source of strength, I am truly fortunate to have you by my side.

To **Chahinez** — Shanèse — my closest friend and my thesis partner, thank you for your dedication, honesty, and unwavering collaboration throughout this journey. Sharing both friendship and hard work with you made this experience all the more meaningful.

To **Wassim**, a dear trusted friend for the past five years, thank you for your kindness, insight, and continued support throughout this process.

I extend my heartfelt thanks to my supervisor, **Dr. Chehili Hamza**, for his valuable guidance, availability, and encouragement during every stage of this work. Your support was essential to the development of this thesis.

I am also deeply grateful to **Dr. Benmoussa SalahEddine**, Head of the clinic *"Clinique d'Ophtalmologie Benmoussa"*, for warmly welcoming us during our internship and for generously providing access to clinical data critical to the validation of our work. Your contribution has been instrumental in bringing this research to life.

To **all of you**, thank you for walking this path with me. This accomplishment is as much yours as it is mine.

Hadjer

DEDICATION

This work is much more than a simple thesis; it is the reflection of a journey filled with effort, sacrifice, and unwavering support. Through these words, I wish to express my deepest gratitude and pay tribute to all those who have played a meaningful role in this journey.

First and foremost, to **myself**, for the perseverance, patience, and hard work I was able to demonstrate despite the challenges encountered. This thesis is the result of countless hours of reflection, moments of doubt, but above all, determination. I am proud of the path I have walked, and I honor myself for staying strong and seeing it through to the end.

To my dear parents, **Ibtissem Nacima** and **Ahmed Yacine**, I would love to express my heartfelt gratitude. Your unconditional love, education, and sacrifices have been the pillars of my growth and development. Your constant support, encouragement, and faith in me gave me the strength and motivation to keep going, no matter what.

I pay a deeply emotional homage to the memory of my grandfather, **Djedou Youcef**, may his soul rest in peace. He was my very first supporter and a foundational figure in my life. His love, encouragement, and pride in me have always been a powerful source of motivation. Though he is no longer physically present, his spirit continues to inspire me every day.

To my beloved sisters, **Yasmine** and **Aziza**, your presence, your support, and our shared moments of laughter and craziness have been a true source of comfort and motivation throughout this journey.

To my maternal grandparents, **Djedou Sallah** and **Mamie Abla**, thank you for your unconditional love, your unwavering support, and your comforting presence in my life.

To my best friends, **Nour El-Yakine** "Kikinou" and **Hadjer** "Geegee", thank you for your moral support, your precious friendship, and all the joyful moments we've shared.

I would also like to extend my sincere gratitude to everyone who contributed to the completion of this thesis. First, I would like to thank **Mr. Chehili Hamza** for his guidance, his invaluable advice, his patience, and his expertise, all of which were essential to the progress of this work. I am also deeply grateful to **Dr. Benmoussa Sallah-Eddine** for his generous support, availability, insightful advice, and trust throughout the development of this thesis.

Chahinez

DEDICATION

With deep gratitude and heartfelt appreciation, I dedicate this work to all those who have walked beside me and lifted me up throughout this journey.

To my beloved parents, **Fatoumata** and **Seydou**, your love, sacrifices, and unwavering faith in me have been my anchor. You taught me the value of hard work, integrity, and perseverance. This achievement is a reflection of your strength and guidance, and I dedicate it first and foremost to you.

To my cherished teachers, **Dr. Chehli H**., **Mme Benhizia H**., **Mme Chellat D**., and **Mme Yasmina D**, your dedication, patience, and belief in my potential have been instrumental in shaping my academic path. Your words of encouragement and your commitment to excellence have left a lasting mark on me, and I am forever thankful.

To **my brothers and sisters**, your love, support, and belief in me have been a constant source of strength. You reminded me, even in difficult times, that I was never alone. Your pride in my work pushed me to aim higher every day.

To **my friends**, thank you for your presence, your understanding, and your kind words during moments of doubt. Your encouragement, even in small gestures, made a big difference.

And above all, to **my team**, we started this journey together, faced every challenge side by side, and supported each other through every high and low. This work is not just the result of individual effort, but a shared triumph. I am proud of what we have accomplished as a team, and I will always carry the memory of our collaboration, trust, and resilience with me. This work is as much yours as it is mine.

To all those who dare to dream, persevere, and grow, may this serve as a reminder that success is rarely a solo path. It is built together, with heart, support, and unity.

Hamadoun

ABSTRACT

Diabetic retinopathy (DR) is a progressive complication of diabetes and remains one of the leading causes of vision impairment and blindness worldwide. Early detection is crucial to prevent irreversible visual damage. However, traditional screening methods are often limited by restricted accessibility, dependence on trained specialists, and variability in diagnostic interpretations. In response to these limitations, recent advancements in deep learning, particularly (CNNs), have shown considerable promise in automating medical image analysis.

This thesis presents the development of a DL model based on the ResNet-50 architecture for the detection of DR in retinal fundus images. A class weighting strategy was employed during training to address the issue of class imbalance commonly observed in DR datasets. Furthermore, the model incorporates Gradient-weighted Class Activation Mapping (Grad-CAM) to generate visual explanations, thereby enhancing the interpretability and transparency of the predictions.

The model was trained using the publicly available mBRSET dataset and validated with real-world clinical images collected during an internship at the clinic "***Clinique D'ophtalmologie Benmoussa".*** The results demonstrate that the proposed approach achieves high accuracy while providing interpretable outputs, making it a potentially valuable tool to support ophthalmologists in the screening and diagnosis of DR.

Keywords:

Diabetic Retinopathy, Deep Learning, Medical Image Analysis, Grad-CAM, Class Imbalance, Retinal Fundus Images, Explainable AI, Ophthalmology.

RÉSUMÉ

La rétinopathie diabétique (RD) est une complication progressive du diabète et demeure l’une des principales causes de déficience visuelle et de cécité dans le monde. La détection précoce est essentielle pour prévenir les dommages visuels irréversibles. Cependant, les méthodes de dépistage traditionnelles sont souvent limitées par l’accessibilité réduite, la dépendance à des spécialistes formés, et une variabilité dans l’interprétation des diagnostics. Face à ces limites, les avancées récentes en apprentissage profond, en particulier les réseaux de neurones convolutifs (CNN), ont montré un grand potentiel pour automatiser l’analyse des images médicales.

Ce mémoire présente le développement d’un modèle d’apprentissage profond basé sur l’architecture ResNet-50 pour la détection de la rétinopathie diabétique à partir d’images du fond d’œil. Une stratégie de pondération des classes a été appliquée durant l’entraînement afin de traiter le déséquilibre des classes souvent observé dans les ensembles de données de la RD. De plus, le modèle intègre la méthode Grad-CAM (Gradient-weighted Class Activation Mapping) pour générer des explications visuelles, améliorant ainsi l’interprétabilité et la transparence des prédictions.

Le modèle a été entraîné à l’aide du jeu de données public mBRSET et validé avec des images cliniques réelles collectées lors d’un stage à la clinique "***Clinique D’ophtalmologie Benmoussa***". Les résultats démontrent que l’approche proposée atteint une haute précision tout en fournissant des sorties interprétables, ce qui en fait un outil potentiellement précieux pour assister les ophtalmologistes dans le dépistage et le diagnostic de la rétinopathie diabétique.

Mots-clés:

Rétinopathie diabétique, Apprentissage profond, Analyse d’images médicales, Grad-CAM, Déséquilibre des classes, Images du fond d’œil, Intelligence artificielle explicable, Ophtalmologie.

الملخص

تُعدّ اعتلالات الشبكية السكري (DR) من المضاعفات التدريجية لداء السكري، وهي من الأسباب الرئيسية لفقدان البصر والعمى على مستوى العالم. ويُعدّ الكشف المبكر عن المرض عاملاً حاسمًا في الوقاية من الأضرار البصرية التي لا يمكن عكسها. ومع ذلك، فإن طرق الفحص التقليدية تعاني من محدودية الوصول، والاعتماد الكبير على الأخصائيين المدربين، بالإضافة إلى التفاوت في تفسير التشخيصات. وفي مواجهة هذه التحديات، برز التعلم العميق، وخاصة الشبكات العصبية الالتفافية(CNNs) ، كحل واعد لأتمتة تحليل الصور الطبية.

يعرض هذا البحث تطوير نموذج للتعلم العميق يعتمد على بنية من أجل الكشف التلقائي عن وجود اعتلال الشبكية السكري بالاعتماد على صور لقاع العين. وقد تم تطبيق إستراتيجية وزن للطبقات أثناء التدريب من أجل معالجة مشكلة عدم التوازن في توزيع الصور بين الدرجات المختلفة. كما تم دمج تقنية Grad-CAM لتوليد تفسيرات مرئية تُسهم في تحسين قابلية الفهم والشفافية لنتائج النموذج.

تم تدريب النموذج باستخدام قاعدة البيانات العامةmBRSET ، وتم التحقق من فعاليته باستخدام صور سريرية حقيقية جُمعت خلال تدريب ميداني في عيادة *‘‘Clinique D’ophtalmologie Benmoussa’’* . أظهرت النتائج أن النموذج المقترح يحقق دقة عالية ويوفر نتائج قابلة للتفسير، مما يجعله أداة واعدة لمساعدة أطباء العيون في الكشف والتشخيص المبكر لاعتلال الشبكية السكري.

الكلمات المفتاحية:

اعتلال الشبكية السكري، التعلم العميق، تحليل الصور الطبية، Grad-CAM، عدم توازن الفئات، صور قاع العين، الذكاء الاصطناعي القابل للتفسير، طب العيون.

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AUC	Area Under the Curve / Area Under the ROC Curve
AAO	American Academy of Ophthalmology
AGEs	Advanced Glycation End Products
AI	Artificial Intelligence
API	Application Programming Interface
CT	Computed Tomography
CSV	Comma-Separated Values (file format)
CUDA	Compute Unified Device Architecture
cuDNN	CUDA Deep Neural Network library
DARs	Diabetic Auto-detection Results (contextually inferred)
DL	Deep Learning
DM	Diabetes Mellitus
DME	Diabetic Macular Edema
DNN	Deep Neural Network
DR	Diabetic Retinopathy
ETDRS	Early Treatment Diabetic Retinopathy Study
FA	Fluorescein Angiography
FDA	Food and Drug Administration
FN	False Negative
FP	False Positive
Grad-CAM	Gradient-weighted Class Activation Mapping
GPU	Graphics Processing Unit
HPC	High-Performance Computing

ICDR	International Clinical Diabetic Retinopathy (classification system)
IDx-DR	Intelligent Retinal Imaging System for Diabetic Retinopathy
IRMAs	Intraretinal Microvascular Abnormalities
JPEG	Joint Photographic Experts Group (image format)
LIME	Local Interpretable Model-Agnostic Explanations
ML	Machine Learning
mBRSET	Mobile Brazilian Retinal Screening Eye Test (Dataset)
MRI	Magnetic Resonance Imaging
NFS	Network File System
NLP	Natural Language Processing
NPDR	Non-Proliferative Diabetic Retinopathy
NVA	Neovascularization of the Angle
NVD	Neovascularization of the Disc
NVE	Neovascularization Elsewhere
NVI	Neovascularization of the Iris
OCT	Optical Coherence Tomography
OCTA	Optical Coherence Tomography Angiography
PDR	Proliferative Diabetic Retinopathy
PKC	Protein Kinase C
ReLU	Rectified Linear Unit (activation function)
ResNet-50	Residual Network with 50 layers
ROC	Receiver Operating Characteristic

RSG-Net	Residual Squeeze-and-Excitation Grouping Network
sbatch	SLURM batch job submission command
SHAP	SHapley Additive exPlanations
SLURM	Simple Linux Utility for Resource Management
SMOTE	Synthetic Minority Over-sampling Technique
SSH	Secure Shell
TN	True Negative
TP	True Positive
UB2-HPC	University of Batna 2 - High Performance Computing Cluster
VGG16	Visual Geometry Group 16-layer CNN architecture
VEGF	Vascular Endothelial Growth Factor
VRAM	Video Random Access Memory
XAI	Explainable Artificial Intelligence

TABLE OF CONTENTS

INTRODUCTION

General Introduction

Background

Diabetes mellitus is a metabolic disorder characterized by insufficient insulin production or reduced insulin sensitivity, resulting in elevated blood glucose levels[1]. According to the 2025 IDF Diabetes Atlas, an estimated 589 million adults (1 in 9) worldwide are living with diabetes, a number projected to reach 853 million by 2050[2]. One of the critical complications experienced by patients with DM is diabetic retinopathy (DR), which is characterized by damage to the blood vessels in the retina, the light-sensitive tissue at the back of the eye, often caused by fluid accumulation. Approximately 30% of individuals with diabetes develop DR. If left untreated, the condition can progress to a more advanced and severe stage known as proliferative diabetic retinopathy (PDR), which can result in significant vision loss or even blindness. According to the American Academy of Ophthalmology (AAO), DR is primarily classified into two stages based on retinal signs: non-proliferative diabetic retinopathy (NPDR) and PDR[3].

Since DR can progress significantly before any noticeable impact on vision occurs, early diagnosis and timely treatment can lower the risk of vision loss by around 57%. As a result, routine screening and regular follow-up are crucial for individuals with diabetes, particularly those who are middle-aged or older. Nevertheless, several studies have shown that many diabetic patients do not undergo the recommended annual eye examinations, often due to lengthy screening procedures, absence of symptoms, and limited access to retinal specialists. This has led to the increasing use of artificial intelligence (AI), especially deep learning, to automate the analysis of retinal images[4].

Motivation

The integration of AI and DL into medical imaging has marked the beginning of a new era of transformation in healthcare[5]. In recent years, Convolutional Neural Networks (CNNs) are central to DL in medical imaging, performing effectively in tasks such as image classification, segmentation, and anomaly detection. Their capacity to automatically extract spatial and hierarchical features makes them ideal for diagnostic purposes. CNNs have also been applied to the analysis of retinal images for DR screening and diagnosis. Multiple studies have shown that these models can detect DR with high sensitivity and specificity, comparable to human experts, while also reducing reliance on extensive human resources[6].

Despite their powerful capabilities, DL based AI models often lack transparency, creating a “black-box” gap that makes it difficult for clinicians to understand the diagnostic reasoning behind their outputs. This opacity raises important questions about how to provide clear and convincing explanations for AI decisions, which is essential for building trust and confidence among healthcare professionals. To overcome these challenges, much research has focused on developing Explainable Artificial Intelligence (XAI) techniques that simplify AI models and improve their interpretability[7] [8]. One method that supports this is Gradient-weighted Class Activation Mapping (Grad-CAM), which is a popular XAI method that creates visual explanations by showing which parts of an image have the biggest impact on a model’s predictions[9].

This thesis is motivated by the need for an accurate, efficient, and interpretable DL model for DR detection.

Problem Statement

Early detection of DR is crucial but often hindered by limited access to screening and reliance on time-consuming manual diagnosis. While DL models offer high accuracy, their lack of interpretability reduces clinical trust. There is a need for an accurate and explainable DL-based approach to support effective and trustworthy DR detection.

Research Objectives

This thesis aims to:

- Develop a DL model based on ResNet-50 for detecting DR from retinal fundus images.
- Implement a weighting strategy during training to address class imbalance.
- To validate the model's performance using both a public dataset (mBRSET) and clinical data from the clinic *"Clinique D'ophtalmologhie Benmoussa"*.
- To incorporate explainability in the model's predictions using the Grad-CAM technique.
- Incorporate Grad-CAM to provide visual explanations for the model's predictions and improve clinical interpretability.

Research Methodology

This study follows a structured methodology comprising:

- **Data Collection**: The primary training data was obtained from the Mobile Brazilian Retinal Dataset (mBRSET) dataset, a Brazilian dataset with labeled retinal images. Clinical validation was performed using a real-world dataset collected during an internship at the clinic *"Clinique D'ophtalmologie Benmoussa"*.
- **Preprocessing**: Images were resized and normalized. Class imbalance was addressed using a weighting method during training rather than oversampling or augmentation.
- **Model Development**: A DL architecture based on ResNet-50 was employed, leveraging transfer learning.
- **Explainability**: Grad-CAM was used to generate visual explanations of the model's predictions.

Thesis Organization

This thesis is organized into four chapters as follows:

- **Introduction:**

This section provides the overall background, motivation, problem statement, research objectives, methodology overview, and thesis organization.

- **Chapter 1: Diabetes and Diabetic Retinopathy**

This chapter introduces diabetes mellitus and its ocular complication, DR. It covers key risk factors, the stages of disease progression, and highlights the importance of early detection. Screening methods such as fundus photography, fluorescein angiography, and OCT are discussed, along with the emerging role of AI in diagnosis.

- **Chapter 2: Deep Learning and XAI for Diabetic Retinopathy Detection**

This chapter introduces the fundamental concepts of AI, ML, and DL, with a particular focus on their applications in medical imaging. It provides an in-depth overview of CNNs, highlighting the ResNet-50 architecture used in this study. Additionally, the chapter discusses the importance of XAI techniques, such as Grad-CAM, to interpret and understand the decision-making process of DL models in a clinical context.

- **Chapter 3: Materials and Methods**

Details the datasets used, preprocessing techniques, the model architecture (ResNet-50), training setup, the use of a weighting method for imbalance, and the evaluation methodology.

- **Chapter 4: Results and Discussion**

Presents and analyzes the model's performance on both the mBRSET and "*Clinique D'ophtalmologie Benmoussa*" dataset, discusses the results, limitations of the study, and potential areas for future work.

Chapter 1: Diabetes & Diabetic Retinopathy

Chapter 1: Diabetes & Diabetic Retinopathy

1.1 Introduction

While DL algorithms have been proved for detecting DR, in the real automation scenario there happen to be some practical issues: the problem of data quality, algorithm generalization, and, to some extent, explainability. Resolving such problems would allow for the development of viable AI systems assisting clinicians in dispensing care to patients. The given study strives for well setting up an explainable DL model for DR detection related to research and clinical applications.

The chapter starts by surveying the most serious complications that diabetes produces in the human body, emphasizing DR, one of its gravest complications. The conditions and risk factors contributing to the onset of DR are laid out, followed by a detailed treatment of its clinical stages. The chapter then ends with a brief summary of the current screening methods used to detect and follow up on this vision-threatening disease.

1.2 Diabetes Mellitus

Diabetes denotes various metabolic disorders. While elevated blood glucose levels are sustained, it can be termed a diabetic condition. However, in 5-10% of the cases, type 1 diabetes is characterized by autoimmune destruction of pancreatic β-cells and thus requires insulin administration for life. The much more common 90-95% variety, type 2 diabetes manifests itself by insulin resistance, together with an insufficient compensatory release of insulin [10]. The other types include gestational diabetes, diagnosed during pregnancy, and rare types brought on by monogenic syndromes, drugs, or other specific causes. One can simply say that the underlining causes differ for different types of diabetes. Type 1 diabetes is instigated by an autoimmune reaction destroying those beta cells that produce insulin in the pancreas, leading to an outright absence of insulin. Type 2 diabetes shall be the one occurring because of progressive development of resistance to insulin along with inadequate insulin response. Obesity, lack of physical activity, heredity, aging, certain ethnicities, history of gestational diabetes, are all hastening factors in this regard [11].

Currently available tests for diabetes screening include fasting plasma glucose, random plasma glucose, oral glucose tolerance test, and HbA1c. Apart from screening through one of these tests, early diagnosis has also been agreed on as a factor contributing towards prevention or delay of diabetes complications [12].

1.3 The Anatomy of The Eye

The eye is an inexpressibly complex organ involved with the transformation of light into rays that it can then interpret as images. Primarily, it consists of three kinds of lenses: the type that comprises the outer fibrous coat, i.e., the sclera and cornea; the hemorrhagic layer in the middle, named the uvai by some or the iris plane by others; and finally, the retinal inner coat at the posterior part.

Light rays come through the cornea and enter a small space in front called the anterior chamber, which is filled with aqueous humor. The light is then passed through the pupil, which is controlled by the iris; then through the lens; then through the vitreous humor; to finally be received in the retina sitting at the posterior of the eye [13].

In the case of DR, the retina is primarily an organization of multi-layer neural tissue that performs the phototransduction process, from light into electrical signals. Photoreceptors such as rods and cones are present in the outermost layer, and bipolar and ganglion cells, along with glial supporting cells in the inner layer, are responsible for complicated visual processing that finally conveys information into the optic nerves. Blood supply of the retina is given by two sources: from the inner layer by the central retinal artery and from the outer layer by the choroidal circulation. The retinal capillaries are in such a way that they do not possess any autonomic control mechanism, the very condition that makes them susceptible to damage from chronic hyperglycemia [12].

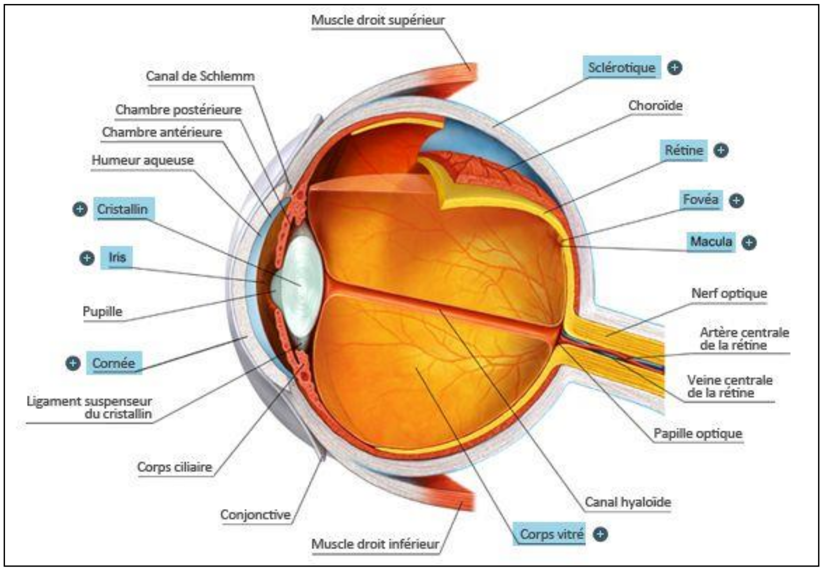


Figure 1 *Anatomy of the eye [14].*

1.4 Diabetic Retinopathy

DR, a microangiopathy, destroys retinal vessels, being a serious long-term complication of type I and II diabetes [15]. Being a slowly progressing disease, it injures small arteries, supplying the retina, and produces varied retinal lesions that may progress with time to visual diminution and even to blindness as a sequel to hyperglycemia [16]. At the molecular level, sustained hyperglycemia induces multiple deleterious effects on the retina via the polyol pathway, formation of advanced glycation end products, and PKCactivation, among others [17]. Histological changes include thickening of the capillary basement membrane, loss of pericytes, increased permeability of vessels, and capillary closure [18].

Considering angiogenesis, DR is broadly separated into two major clinical phases. Thus, NPDR includes early microvascular changes such as microaneurysms, intraretinal hemorrhages, hard exudate, cotton-wool spots, but no new blood vessel formation [19]. The classical definition of PDR is the existence of abnormal blood vessels growing either on the surface of the retina or into the vitreous, induced by an ischemic stimulus [20]. It represents another possible stage of DR, while characterized by fluid build-up in the central macula, leading to one of the primary causes of impairment of vision [21]. DR bears immense clinical significance while one of the important signs aggravating diabetes microvascular complications. Interestingly, the presence and intensity of DR are observed to run parallel with the presence of other diabetes-related diseases, like diabetic nephropathy and diabetic neuropathy, engendering an urgent need for holistic management of diabetes and commenced ophthalmologic screening of every diabetic patient [15].

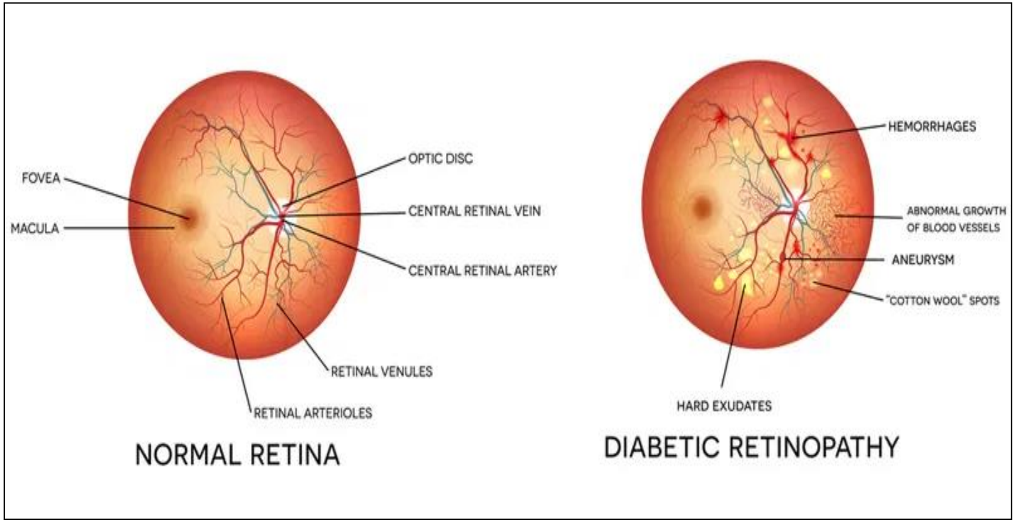


***Figure 2** Distinction between DR and a normal retina*[22].

1.5 Why is Diabetic Retinopathy (DR) an Important Public Health Issue?

The worldwide prevalence of DR makes it the primary reason for vision loss among adults in their working years [23]. The prevalence of DR among diabetic patients reaches one-third while vision-threatening stages affect 10% of diabetic patients [13]. The diabetic population worldwide shows a 34.6% prevalence of DR while 10.2% of these cases threaten visual function [24].

The economic burden of DR creates substantial social and financial challenges. The worldwide healthcare systems face substantial financial pressure from direct medical expenses needed to screen and treat DR and provide vision rehabilitation services. The economic impact becomes more severe due to indirect costs which result from decreased productivity and early retirement and increased need for supportive care services. The annual financial burden of DR in the United States amounts to $4.7 billion according to Zhang et al. (2017) [22].

The silent and typically unnoticeable development of DR creates difficulties because patients usually seek medical attention only after their vision becomes permanently damaged. The situation demonstrates why screening programs need to become the top public health priority. Multiple obstacles affect traditional screening approaches because there is a lack of ophthalmology specialists and geographic barriers to specialized care and inconsistent diagnostic interpretation [23].

1.6 Causes and Risk Factors

It is deemed as a prominent cause of blindness in diabetic patients worldwide. Being a major microvascular complication of diabetes, progressive damage to the vessels in the retina is seen [25]. The pathogenic mechanism behind this ocular pathology is baffling; thus, many biochemical events take place during chronic hyperglycemia, ultimately leading to structural and functional changes in the retinal microvasculature [26]. The stages of DR are known as Mild, Moderate, and Severe NPDR Stage before getting to the stage of PDR. The symptoms of the disease are as under [23] [13]:

- Difficulty seeing objects from a distance
- Blurry vision, Blindness
- Floaters (tiny dark spots or strings that float in your field of vision)
- Obstructed vision at night or in low-light conditions
- Colors appear faded or washed out
- Gradual vision loss

DR is a significant chronical disease affecting the blood vessels in the retina, the light-sensitive tissue at the back of the eye [13]. Blood vessels of the retina, if damaged, may weaken or grossly enlarge. Tiny microaneurysms or bulges may arise if the blood vessels in the eye are compromised. These incite swelling in the neighborhood tissues and leak fluid into the retina. This can cause ocular hemorrhage, and excessive formation of new blood vessels can block blood flow to the retina [26]. Numerous factors have been found to play a role in the incidence and course of DR, including:

- **Hyperglycemia**: It is the effect of chronically high blood sugar levels on the retina's blood vessels.
- **Hypertension**: Increased blood pressure would worsen the vascular damage.
- **Genetic predisposition**: Some individuals, who have a family history of DR, are more likely to develop the disease [24] [22].

1.7 Hyperglycemia as the Primary Driver

First and actually the primary step in inducing injurious metabolic change in retinal vessels is persistent hyperglycemia. One of the pathways leads to the polyol flux, whereby excess intracellular glucose is converted into sorbitol. This causes osmotic imbalance, thickening of the basement membrane of the capillaries, oxidative injury, and degeneration of pericytes [27]. On top of that, extended high blood glucose levels induce the formation of AGEs that modify the extracellular matrix and bind to receptors on target cells, releasing proinflammatory cytokines that threaten vascular integrity [18]. Hyperglycemia further induces the activation of PKC β isoform to yield more thickening of the basement membrane and endothelial cell proliferation, also increasing vascular permeability [28]. Such events synergistically produce excessive reactive oxygen species that initiate oxidative stress, inflammation, and apoptotic signaling to destroy retinal vascular cells [29].

1.8 Hypertension as an Aggravating Co-factor

The presence of hypertension if the retinal artery is already vulnerable brings about an increased amount of mechanical stress, which hastens DR development. Higher pressures accelerate the leakage through the capillaries and also aggravate endothelial dysfunction. Insisting on an adequate blood pressure control has been shown to reduce the progression of DR by 34% and the risk for vision loss by 47% [30].

1.9 Genetic Predisposition

The fact remains that genetic predisposition accounts for 25 to 50% of the variability that exists between individuals in the potential risk of DR [31]. Thus, genome-wide association studies have partly identified candidate genes in the areas of VEGF signaling, inflammatory response, and glucose metabolism in explaining inter-individual variability for retinopathy in patients with an almost similar duration of diabetes and glycemic control [32].

Being aware of the various parameters of DR would put us in the position to frame holistic risk-assessment models and bases for prevention accordingly.

1.10 Stages of Diabetic Retinopathy

The progress of DR is clear. It is usually classified into two main stages: NPDR as the first stage and PDR as the latter stage. This classification along with the assessment of diabetic macular edema (DME) is useful for guiding clinical management decisions.

1.10.1 Non-Proliferative Diabetic Retinopathy (NPDR)

In this phase, early retinopathy comes into being and so to say, the older walls do not concern themselves with aberrant neovascularization; cerebral lacunae of neovascularization feature the abnormal vessels of this phase. NPDR is further divided into [27], according to the size and severity of the retinal lesions:

- **Mild NPDR**

  The appearance of one microaneurysm, an outpouching from a weakened capillary wall that appears as tiny round red dots on a fundus examination, is said to be the first clinically identifiable stage of DR. Most patients, however, retain full visual acuity and suffer from no symptoms [20].

- **Moderate NPDR**

  Some other microaneurysms progress even further and change, and these changes would include venous beading or localized venous dilations, intraretinal hemorrhages (dot and blot type) and soft exudates or cotton wool spots (a sign of a neural layer infarction). Although these changes may only affect well-defined regions of retina, they are still characteristic changes of progressive retinal ischemia [21].

- **Severe NPDR**
  Using the "4-2-1 rule" severe intraretinal hemorrhages and microaneurysms in four quadrants, can either venous beading in two or more quadrants, and/or intraretinal microvascular abnormalities (IRMAs) in at least one quadrant this pre-proliferative stage represents severe retinal ischemia. There is a 52% chance severe NPDR will progress to PDR in one year if left untreated [25].

1.10.2 Proliferative Diabetic Retinopathy (PDR)

During the advanced stages of DR, PDR is characterized by new vessels developing as a consequence of retinal ischemia and primarily under the influence of VEGF. These neovessels could develop at the anterior chamber angle (NVA), on the iris (NVI), anywhere else on the retina (NVE), or on the optic disc (NVD). Few other features of PDR include neovascularization, fibrovascular proliferation, vitreous hemorrhage, tractional retinal detachment, neovascular glaucoma [14].

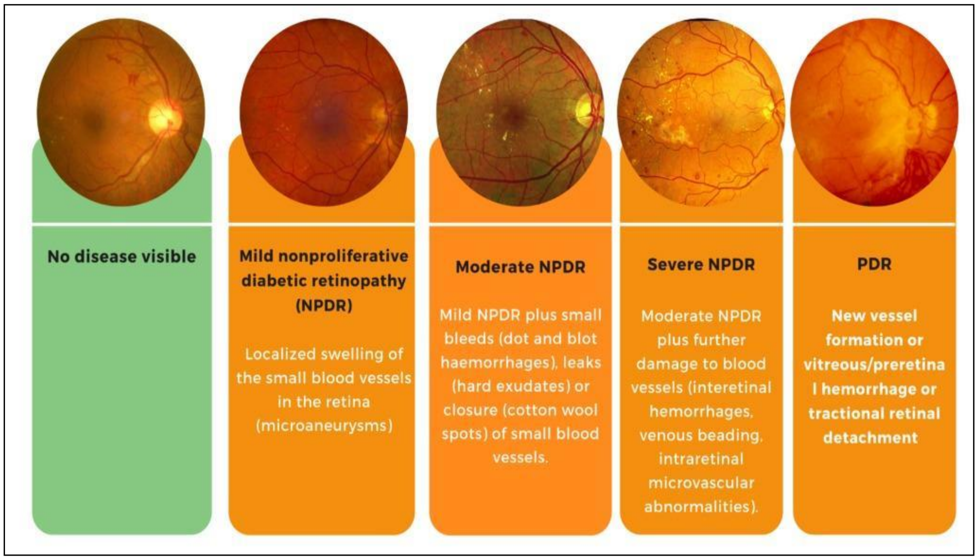

No disease visible
Mild nonproliferative diabetic retinopathy (NPDR)
Localized swelling of the small blood vessels in the retina (microaneurysms)
Moderate NPDR
Mild NPDR plus small bleeds (dot and blot haemorrhages), leaks (hard exudates) or closure (cotton wool spots) of small blood vessels.
Severe NPDR
Moderate NPDR plus further damage to blood vessels (interetinal hemorrhages, venous beading, intraretinal microvascular abnormalities).
PDR
New vessel formation or vitreous/preretinal hemorrhage or tractional retinal detachment

***Figure 3** Different stages DR [32].*

The following are the commonly used clinical stage descriptions of DR:

- **No Disease**: The retina reveals no abnormalities
- **Mild NPDR**: Small microaneurysms present
- **Moderate NPDR**: Increased hemorrhages and leakage
- **Severe NPDR**: Widespread bleeding and capillary dropout
- **PDR**: Abnormal blood vessel growth

Since each stage of diabetic retinopathy (DR) presents distinct morphological characteristics, understanding the progression of the disease is important. This knowledge supports the development of AI models capable of reliably detecting the presence of DR in retinal images.

1.11 Current Screening Methods

DR has to be caught early to minimize one's chances of becoming blind. Current screening techniques are all types of imaging-based methods, one has its advantage in a few lines, and the other has limitations on those lines.

1.11.1 Fundus Photography

Fundus photography is perhaps the most predominantly offered method for DR-screening. It is a non-invasive procedure whereby a beam is projected onto the retina to form two-dimensional images with which the physician can detect DR signs like vascular anomalies, hemorrhages, and exudates. Images are usually taken focusing on an area ranging from 30° to 45° encompassing the macula and the optic disc. The results of the Early Treatment DR Study (ETDRS) established a standardized method (the seven stereoscopic fields) for looking at the peripheral retina extensively since this is where major lesions may develop [15]. New technologies such as ultra-widefield imaging can now capture a single picture of up to 200 degrees, thus improving peripheral lesion diagnoses. Non-mydriatic methods provide more comfort to the patient and much speed, but such means fail to give good-quality pictures in case of very small pupils or lens opacities.

However, there are a number of problems introduced by the images being two-dimensional, with one major problem being that retinal edema cannot be diagnosed by the present systems of photography. Some additional problems are the generation of low-quality images caused by cataracts or motion artifacts when the patient moves. Their interpretation needs to be done by a specialist, for which mass screening becomes almost impossible.

1.11.2 Fluorescein Angiography

Fluorescent angiography (FA) is an encompassing dynamic technique that involves the injection of the dye intravenously, following which timed photographs are shot as the dye passes through the retinal arteries. Basically, it reveals quick functional features of the retinal vessels that would go undetected with normal imaging. The fluorescein angiography shows the microvascular disease in its early stages, as microaneurysms appear as bright fluorescent spots much earlier and more often than in the fundus photographs. It also documents the areas of capillary non-perfusion (retinal ischemia), areas of leakage from breakdown of the blood-retinal barrier, and areas of neovascularization related to PDR.

Although it offers excellent diagnostic accuracy, it has limited choice for routine screening due to its invasive procedure that entails injecting dye into the veins, with the dye possibly causing nausea in around 5–10% of the cases or, more rarely, severe anaphylaxis. Moreover, it also takes time, perhaps the procedure involved tends to be lengthy in execution or so frequently requires special equipment and, also, trained personnel, which makes it unsuitable for the screening on a global level [29].

1.11.3 Optical Coherence Tomography (OCT)

Since the dawn of vitreoretinal surgery, OCT imaging has evolved into a critical tool in adjunct with medical/surgical retinal diagnoses. It offers cross-sectional images at very high resolution of the retinal layers of the fundus. OCT images are formed with the interference of light waves, very much analogous to the way ultrasonography takes images of organs and tissues with lateral displacements in micrometer precision. The treatment of DME is well supported by OCT examination because retinal thickness can be measured accurately in the presence of excess fluid and/or structural abnormalities. It may show early neurodegenerative changes that cannot be appreciated by fundus photography. Currently, in recent years, OCT Angiography (OCTA) has been used to evaluate the retinal and choroidal vasculature without dye injection by detecting motion of blood cells, that is to say, it complements detection of microaneurysms, areas of capillary dropout, and neovascular formations with structural OCT, conventional FA, and overcomes hazards associated with the use of fluorescein. OCTA is a novel imaging modality that allows for non-invasive imaging of retinal and choroidal vessels by recording motion contrast generated from blood flow. OCTA images neovascular networks, microaneurysms, and capillary dropout and thus creates an imaging technique for retinal disorders that steers clear of risks associated with the dye-based technique [29].

The clinical utilization of OCT is complicated by the pricing intricacies of the instruments and other such considerations, which comprise a longer examination period in comparison to fundus photography and the need for the training of various technologists in image capture and interpretation. That is why OCT is generally only used as an additional procedure in the further investigation of cases where fundus imaging is the primary screening.

1.11.4 Manual vs. AI-Assisted Diagnosis

Traditional DR screening is still a manual review of retinal images by trained medical personnel. The ophthalmic exam is recognized as the diagnostic gold standard; however, this method of screening cannot be expanded, as there is an acute shortage of specialists in the epidemiologically significant geographic areas that are home to diabetic patients. Some programs have developed ways around this problem by employing graders (who are not physicians) trained to look at fundus photographs in certain conditions. Although these methods do increase the amount of screening being accomplished in poorly resourced areas somewhat, the negatives are that there is no certainty involved in the diagnosis (the graders feared this uncertainty more than educated physicians!), and any protocol must continue to undergo strict quality-controls and refresher training.

AI has come out in modern society and can be one of the screening alternatives, in particular. Deep neural networks allow the analysis of fundus pictures for the possible detection of DR with a sensitivity and specificity that can rival even those of a human expert. These systems have various advantages [30]:

- The data stands alone without any explanation.
- The results stay impartial since the experiment's operator's emotional state and tiredness never influenced the data.

- The explanation provided here utilizes computer-scale detection for broad applications.
- The costs of the test should hold special pertinence in areas that are seriously ill.

A landmark study showed that these DARs were diagnosed by means of a very broad gamut of human demographics and image acquisition settings with a sensitivity of 97.5% and specificity of 93.4% by DL. This level of performance achieves what most junior graders fail to produce and approximates expert performance [30]. Afterward, regulatory approvals have been granted to various detection algorithms. The FDA-approved first autonomous AI system for clinical use IDx-DR had met safety and efficacy standards based on a pivotal clinical trial, with sensitivity and specificity values of 87.2% and 90.7%, respectively, for the diagnosis of more-than-mild DR [31].

Medical practice faces various hurdles when AI gets introduced since it needs to fit within current systems and manage atypical occurrences while sustaining accurate diagnostic results across all patient categories and establish proper regulatory controls. Specialists could handle cases through a combination of AI pre-screening for standard cases along with their manual assessment of challenging cases which would serve as an effective method to optimize specialist resources.

1.12 Conclusion

DR is a complex and progressive eye disorder occurring due to many unfavorable conditions. One must establish the causality in the occurrence and development of this pathological status from a variety of conditions such as chronic hyperglycemia, hardening of the arteries, or even genetic components. Then, of course, we have the classification of DR according to its temporal course into early and late; NPDR and PDR, for purposes of diagnosis and clinical management.

Though benefits arise with screening methods, nevertheless a number of problems limit their availability, including clarity of interpretation and lack of scaling. Similar concerns are faced when ophthalmic care specialists are not within reasonable accessibility. Since grading has to be done manually, it is highly subjective, coupled with biases or opinions; those norms and standards of opining actually take away from adequate, consistent care. Thus, if we view screen modeling from a different perspective, a model is developed whose efficiency for screening is enhanced, along with an efficient way of increasing the accessibility of eyesight care.

AI remains an essential set of tools in ophthalmology to implement automatic DR detection and grading mechanisms. Therefore, at any place, at any time, the system will perform the DR detection fast, reliably, and in a scalable way, even in the remotest locations. However, issues such as data representativeness, human explainability of the model, and smooth integration into existing workflows need to be considered if it is to become part of everyday medical practice.

Chapter 2: Deep Learning and Explainable AI for Diabetic Retinopathy Detection

Chapter 2: Deep Learning and Explainable AI for Diabetic Retinopathy Detection

2.1 Introduction

AI, especially DL, has greatly improved how medical images are analyzed. These technologies allow computers to automatically understand and recognize patterns in complex images with high accuracy [33]. In the case of DR, DL models can detect signs of the disease in eye images as well as trained doctors can, making them useful for early screening and diagnosis on a large scale [31].

In this chapter, we will introduce the main ideas behind AI, ML, and DL, focusing on how they are used in medical imaging. Special attention is given to Convolutional Neural Networks (CNNs), used in our study, and their applications in DR Detection. The chapter also highlights the importance of understanding how these models make decisions using XAI tools like Grad-CAM.

2.2 Artificial Intelligence (AI) and Its Subfields

2.2.1 Overview of Artificial Intelligence

AI is any technique that aims to enable computers to mimic human behavior, including ML, natural language processing (NLP), language synthesis, computer vision, robotics, sensor analysis, optimization and simulation [34].

2.2.2 Machine Learning: Concepts and Approaches

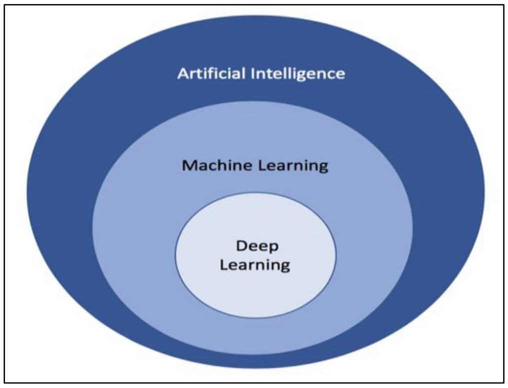


Figure 4 *AI composition [35].*

ML is a branch of AI that enables computers to learn from data and make decisions without being manually programmed. Its main goal is to develop software that can analyze and learn from information on its own[36]. As shown in ***Figure 5***, ML algorithms are divided into three categories: supervised learning, unsupervised learning, and reinforcement learning [37].

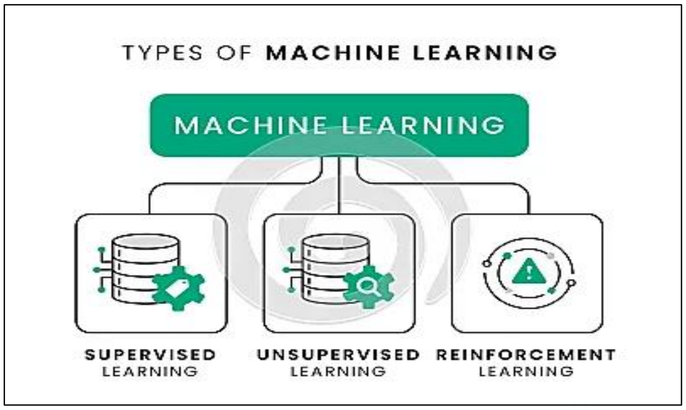


Figure 5 *Types of Machine Learning* [38].

a) Supervised Learning

Supervised learning is a type of ML where the model learns from labeled data. This means the data includes both the input and the correct output. The model uses this information to understand the relationship between them. After training, it can use what it learned to make predictions on new data[36].

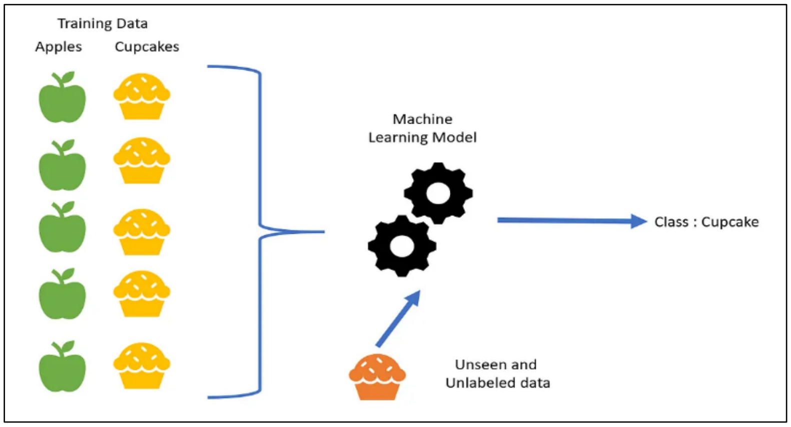


Figure 6 *Supervised learning example* [39].

For example, in ***Figure 6***, we see two categories: apples and cupcakes. We have several training examples for each. The algorithm learns the features of both classes and can then predict the class of a new object [39].

b) Unsupervised Learning

The model learns by analyzing the data and discovering hidden patterns or structures. When it receives a dataset, it automatically organizes similar data into groups. However, it does not assign names to these groups. For example, it can separate apples from mangoes, but it cannot label which group is apples or which is mangoes [37].

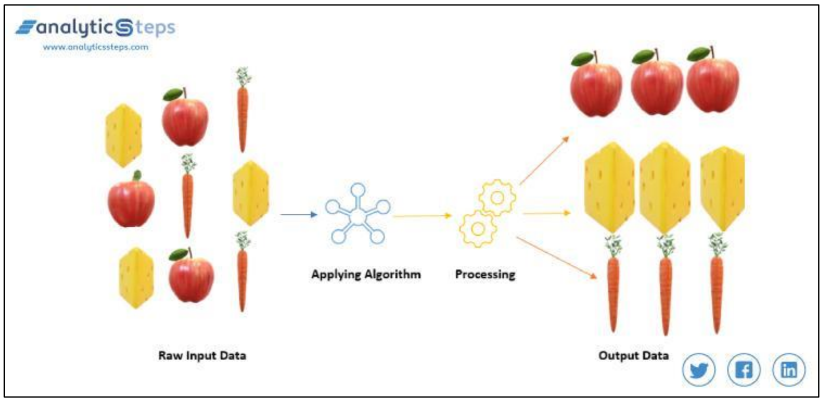


***Figure 7** Unsupervised learning example [40].*

As mentioned earlier, in unsupervised learning, the data is not labeled. So, in ***Figure 7*** if we give raw input like apple, carrot, and cheese, the model will group them based on similarities, but it won't know which group is apple because there are no labels. However, any new data will be placed into one of the existing groups automatically [40].

c) Reinforcement Learning

Reinforcement Learning is a type of ML where software or machines learn the best actions to take in a certain situation to improve their performance. It doesn't use labeled data or known outcomes, instead, it learns through trial and error. When the model makes a good decision, it gets a positive reward. When it makes a bad one, it receives a negative response. Over time, it learns which actions to take and which to avoid[40].

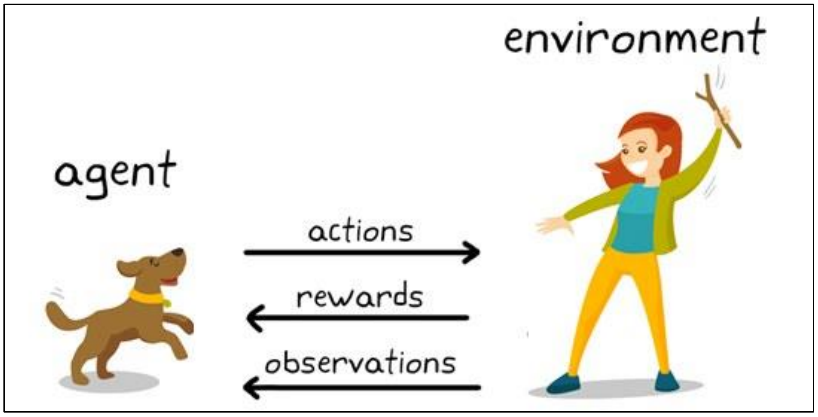


***Figure 8** Reinforcement learning in dog training [41].*

In the example shown in ***Figure 8***, reinforcement learning is illustrated through training a dog (the agent) to perform tasks in its environment, which includes both surroundings and the trainer. The trainer gives a command (observation), the dog responds (action), and receives a reward like a treat or a toy if the response is correct. Early actions may be random, but over time, the dog learns which behaviors lead to rewards. The goal is to refine the dog's behavior (policy) to consistently take the right actions and maximize rewards. Once trained, the dog responds correctly to commands [41].

2.2.3 Deep Learning: A Subset of Machine Learning

DL is a subset of ML that relies on artificial neural networks to learn autonomously, make informed decisions, and evaluate prediction accuracy without human input. As these models continuously process data, they accumulate knowledge over time and form conclusions by analyzing new inputs, referencing learned data, and generating responses [42].

Unlike traditional ML methods, DL methods require significantly less manual intervention. Rather than relying on manually designed features, a process that is often complex and time consuming, DL models automatically learn relevant features from the data itself. Moreover, DL algorithms are generally more effective than traditional ML methods, especially as the size of the dataset grows[43].

This capability has made DL particularly successful in areas such as speech recognition, natural language processing, and, importantly, image analysis in the medical field [33]. Its flexible and multilayered structure enables it to automatically learn complex patterns from large datasets, often outperforming traditional ML models [34] [42].

2.2.4 Difference between AI vs ML vs DL

***Table 01**. Difference between AI vs ML vs DL [44].*

Category	Artificial Intelligence (AI)	Machine Learning (ML)	Deep Learning (DL)
Scope	A broad discipline that includes all types of intelligent behavior in systems.	Concentrates on enabling systems to learn patterns and insights from data.	A more advanced area within ML that uses neural networks.
Goal	Replicate human cognitive abilities.	Enable systems to learn automatically from data without explicit programming.	Extract complex features and automate decision-making processes.
Techniques Used	Incorporates rule-based logic, ML techniques, and DL models.	Uses models such as regression, decision trees, and other learning algorithms.	Relies heavily on neural network architectures like CNNs.
Data Dependency	Can function with minimal data (e.g., predefined rules).	Needs well-organized, structured datasets for effective training.	Demands large datasets to train sophisticated models accurately.
Use Cases	Examples include virtual assistants, smart robots, and self-driving systems.	Common in detecting fraud, filtering content, and recommending products.	Widely applied in tasks like facial recognition, language processing, and autonomous vehicles.

AI, ML, and DL are closely related but differ in scope, methodology, and application.

2.3 Deep Learning in Medical Imaging

2.3.1 Applications of Deep Learning

Medical image analysis focuses on the processing, interpretation, and examination of medical images. In recent years, the rise of DL algorithms has significantly transformed this field, as these methods are increasingly used to improve the diagnosis, treatment, and monitoring of various medical conditions.

DL, a subset of ML, involves training algorithms to learn patterns from large volumes of data. In the context of medical image analysis, DL models can automatically detect and classify abnormalities in different types of medical images, such as X-rays, MRI scans, CT scans, and ultrasound images.

In this context, DL models support the identification and diagnosis of various medical conditions, including tumors, lesions, anatomical irregularities, and pathological changes. They are also useful in assessing disease progression, evaluating treatment outcomes, and predicting prognosis. By automatically extracting important features from medical images, these models enable more efficient and accurate analysis.

Techniques such CNNs are extensively applied to tasks like image segmentation, object detection, disease classification, and image reconstruction, and in identifying conditions such as lung cancer in CT scans, DR in fundus images [45] [6].

This technique will be explored in detail in subsequent sections due to its relevance in our work.

2.3.2 Challenges and Limitations

Although DL has shown great promise in medical imaging, its adoption in clinical settings faces several interconnected challenges:

a) **Data Scarcity and Quality**

Gathering large, diverse, and well-annotated medical imaging datasets is often limited by privacy laws, the high expense of expert annotations, and logistical barriers. This lack of sufficient data limits the ability to develop reliable models and can result in poor generalization across different patient groups or imaging systems, potentially producing unbalanced or inaccurate outcomes [6].

b) **Model Interpretability**

DL models frequently function as “black boxes,” they offer limited transparency into how decisions are made, which can reduce clinician trust and make it challenging to validate or explain AI-based results [6].

c) Computational Demands

Training and implementing DNN demand significant computational power and specialized hardware, which may be impractical in clinical environments with limited resources [6].

d) Algorithmic Bias and Generalization

Models developed using limited or unrepresentative datasets might struggle to perform well across varied demographic groups or imaging conditions, potentially leading to unequal care and diagnostic mistakes [6].

2.4 Convolutional Neural Networks for Diabetic Retinopathy Detection

2.4.1 Convolutional Neural Networks and their Fundamentals

A CNN is a specialized form of artificial neural network specifically designed to handle structured grid-like data, such as images and videos. CNNs have significantly advanced the field of computer vision and are commonly applied in tasks like image classification, object detection, and image segmentation [46].

Here are the main components of a CNN:

- **Convolutional Layers**: Apply learnable filters to the input using the convolution operation to extract features.
- **Activation Functions**: Introduce non-linearity into the network, allowing it to model complex patterns in the data.
- **Pooling Layers**: Divide feature maps into smaller regions to reduce dimensionality and computational cost.
- **Fully Connected Layers**: Link each neuron in one layer to all neurons in the next layer for final decision-making.
- **Padding**: Add extra pixels around the input to preserve spatial dimensions after convolution.
- **Stride**: A larger stride reduces the size of the output feature maps by skipping more input positions.
- **Skip Connections**: Enable certain layers to be bypassed, making it easier to train deeper networks and improving gradient flow [47].

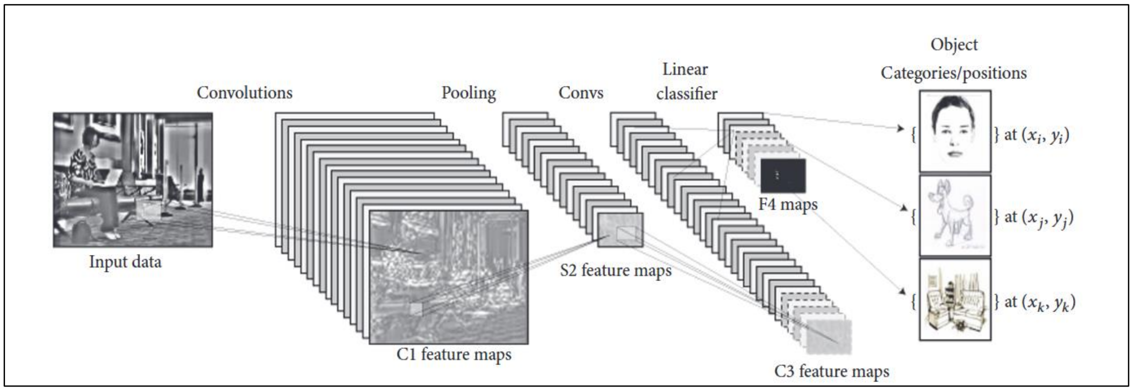


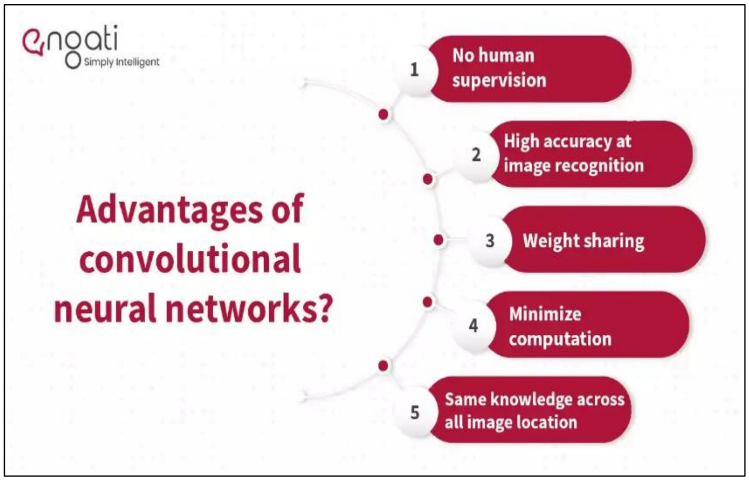
Figure 9 *CNN architecture for visual recognition [47].*

2.4.2 Advantages and Disadvantages of Convolutional Neural Networks

a) Advantages

CNNs offer several key benefits in the field of image processing and recognition:

- They do not require extensive human intervention for feature selection.
- Features are extracted automatically during training.
- CNNs deliver high accuracy in tasks such as image classification and object detection.
- The concept of weight sharing reduces the number of parameters and enhances efficiency.
- They are computationally efficient due to fewer parameters.
- Learned features can be applied consistently across different regions of an image.
- They are well-suited to processing large-scale datasets.
- CNNs learn in a hierarchical fashion, identifying simple to complex features across layers [47].



Advantages of convolutional neural networks?

1 No human supervision

2 High accuracy at image recognition

3 Weight sharing

4 Minimize computation

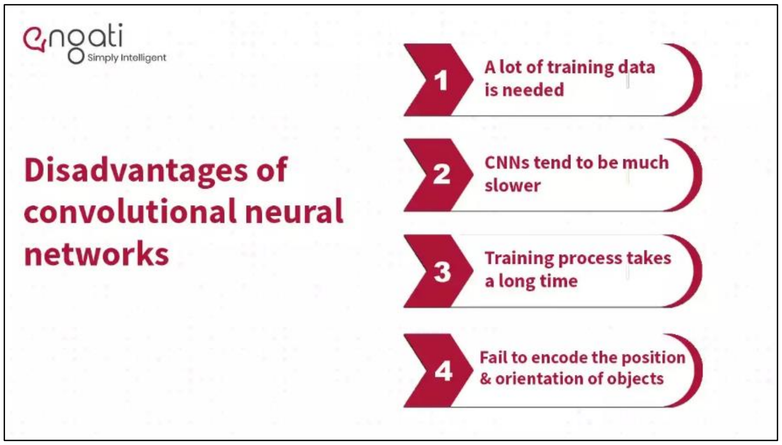
5 Same knowledge across all image location

Figure 10 *Advantages of CNN [47].*

b) Disadvantages

Despite their advantages, CNNs also have several limitations:

- They require substantial computational power for training and inference.
- A large volume of labeled data is often needed for effective learning.
- They can consume significant memory resources.
- The inner workings of CNNs are often hard to interpret, making them less transparent.
- They are not naturally suited for analyzing sequential or time-series data.
- Processing speed can be slow, particularly with deep architectures.
- Training a CNN model can be time-intensive [47].



engati Simply Intelligent

Disadvantages of convolutional neural networks

1 A lot of training data is needed

2 CNNs tend to be much slower

3 Training process takes a long time

4 Fail to encode the position & orientation of objects

Figure 11 *Disadvantages of CNN [47].*

2.4.3 Convolutional Neural Networks Applications in Diabetic Retinopathy

Improvements in computer-aided design and analysis have been driven by the incorporation of advanced ML algorithms and data processing methods, with the goal of increasing detection accuracy. Among the widely used ML techniques in image processing tasks is the CNN [48]. Their main uses involve:

a) Automated Severity Classification

CNNs are widely used to automatically classify the severity of DR from retinal images, enabling faster and more accurate diagnosis. Several architectures have been fine-tuned or developed from scratch for this task [31].

- **RSG-Net**: Stands for Residual Squeeze-and-Excitation Grouping Network, which is a custom CNN, demonstrated strong performance in classifying DR, reaching a testing accuracy of 99.36%, with 99.79% specificity and 99.41% sensitivity across four severity levels. When distinguishing between DR and non-DR cases (binary classification), it achieved 99.37% accuracy, 100% sensitivity, and 98.62% specificity [49].

b) Deep Feature Extraction and Image Classification

CNNs are capable of capturing layered features from retinal images, which helps in accurately detecting patterns and lesions related to DR. Several advanced CNN architectures include:

- **ResNet50**: It reached a training accuracy of 99.37% and a validation accuracy of 71.64%.
- **InceptionV3**: It reached a training accuracy of 99.03% and a validation accuracy of 73.72%.
- **EfficientNetB4**: It has achieved a training accuracy of 99.37% and a validation accuracy of 79.11%.
- **DenseNet201**: It has achieved a training accuracy of 99.58% and a validation accuracy of 76.80%.

These models achieved training accuracies exceeding 99% and validation accuracies around 79%, showing strong ability to generalize on large datasets like Kaggle EyePACS [50].

c) Handling Class Imbalance and Data Scarcity

In DR detection, datasets often suffer from imbalanced class distributions, where some disease stages are underrepresented. This challenge can reduce the performance of CNN-based models. To address it, researchers apply techniques such as flipping, rotation, zooming, and adjustment of color, contrast and brightness to improve model learning and classification accuracy.

- **RSG-Net**: Enhanced its classification performance by applying data augmentation techniques, such as rotating and flipping retinal images, to artificially increase the number of training samples. Additionally, it employed a focal loss function, which emphasizes learning from underrepresented or difficult examples, helping to address class imbalance. As a result of these strategies, the model achieved a high specificity of 99.41%, indicating strong reliability in correctly identifying non-diseased retinal images [49].

d) Explainability in CNN-based DR Detection

CNNs outperform traditional methods in image tasks by learning complex features, but their lack of transparency has driven interest in XAI to clarify their predictions [51]. This section will be discussed in detail in the following sections because of its importance to this study.

- **Gradient-weighted Class Activation Mapping (Grad-CAM)**: Grad-CAM, an explainability algorithm, was applied to highlight DR indicators in fundus images, enhancing the model's interpretability and reaching an accuracy of 98.6% on the EyePACS dataset [52].

Despite the impressive performance of CNN and other DL models, their clinical adoption remains limited due to a lack of interpretability. To address this, the following section discusses the role of XAI in making these models more transparent and trustworthy for medical practitioners.

2.5 Explainable Artificial Intelligence in Medical Imaging

2.5.1 The Need for Explainability in Medical AI

AI is changing healthcare by improving diagnosis, predicting patient risks, streamlining processes, and personalizing treatments. But as AI systems become more complex, a key question arises: Can we trust AI if we don't understand how it makes decisions? That is where XAI plays a crucial role.

Unlike traditional "black-box" models that give predictions without explanations, XAI makes the decision process clear. In healthcare, where decisions affect lives, this transparency is not just helpful, it is essential [53].

This lack of transparency raises several concerns in medicine:

- **Enhancing Patient Safety**: AI is helpful but can make mistakes. Explainability shows why a mistake happened, like bad data or wrong patterns, so we can fix it and keep patients safe.
- **Building Clinician Trust**: Doctors need reasons to trust diagnoses. If AI just says "This patient is at high risk" without explaining why, doctors won't trust it. When AI explains its reasons, doctors are more likely to use it.
- **Regulatory and Ethical Compliance**: Healthcare AI must follow strict rules. Groups like the FDA want AI decisions to be clear and open. XAI helps meet these important rules.
- **Detecting and Reducing Bias**: Medical data can be unfair due to things like race or gender. AI might learn these biases by mistake. XAI helps spot and fix these unfair influences [53].

2.5.2 Common Techniques for Explainable AI (XAI)

With the growing use of AI in healthcare, it is essential to ensure interpretability, understanding not only what the model predicts but also why it makes those predictions. Common approaches to achieve this include post hoc interpretability methods such as Grad-CAM, Local Interpretable Model-Agnostic Explanations (LIME), and Shapley Additive Explanations (SHAP) [54].

- **LIME (Local Interpretable Model-Agnostic Explanations)**: It is a powerful technique that helps explain how ML models make individual predictions. It builds a simple model around one prediction to show which features influenced it most. LIME works by slightly changing the input and seeing how the output changes. It can be used with any model and gives clear, visual explanations that are easy to understand [55].

- **SHAP (SHapley Additive exPlanations)**: It is a method that shows how each feature affects a model's prediction. It's based on Shapley values from game theory, which fairly measure feature importance. SHAP explains both single predictions (local) and overall feature importance (global). It works with any model and can handle complex ones like DNN [55].
- **Grad-CAM (Gradient-weighted Class Activation Mapping)**: It is a method that helps show which parts of an image a neural network focuses on when making a prediction. It creates a heatmap highlighting the important regions by using gradients from the last convolutional layer. Grad-CAM is especially useful for image tasks like classification or object detection. It helps us understand and trust the model by showing why it made a certain decision, which is very helpful in areas like medical imaging [55].

2.5.3 Grad-CAM for Visualizing CNN Decisions in DR Detection

Grad-CAM is a widely used method for visualizing and explaining the decisions of CNNs, particularly in DR detection. It works by generating heatmaps that highlight the regions of a retinal image that most influence the model's prediction, such as lesions, hemorrhages, or microaneurysms.

These visual cues help doctors understand and verify the AI's decisions, addressing the issue of DL models often being perceived as "black boxes." Grad-CAM uses the gradients of the target class flowing into the last convolutional layer of a CNN to compute the importance of each region.

Studies using CNN architectures like VGG16 have shown that Grad-CAM can achieve reasonable accuracy (around 72%) while effectively localizing critical DR features. Compared to models like Vision Transformers, Grad-CAM tends to produce clearer and more interpretable heatmaps with CNNs, although detecting very small lesions remains a challenge.

Additionally, combining Grad-CAM with ensemble models such as Xception, DenseNet121, and InceptionV3 has improved both the accuracy and interpretability of DR detection systems. Overall, Grad-CAM plays a key role in bridging the gap between AI predictions and clinical understanding, supporting the integration of AI tools into medical practice for early and reliable DR diagnosis. [56] [57] [58] [59].

2.6 Conclusion

This chapter presented a comprehensive overview of AI, ML, and DL, with a particular focus on their applications in medical imaging and DR detection. We explored the structure and principles behind CNNs, emphasizing their strengths in analyzing visual medical data. Among these, the ResNet-50 architecture was discussed in detail due to its high performance in DR classification tasks.

In addition, the chapter highlighted the growing need for transparency in AI-driven healthcare tools, introducing XAI methods such as LIME, SHAP, and Grad-CAM. These tools enhance the interpretability of DL models and help build trust among clinicians by providing visual or statistical explanations of model predictions.

Altogether, this chapter lays the conceptual and technical groundwork for the upcoming methodology section, where we describe the dataset, preprocessing techniques, model training process, and evaluation metrics used in this study.

Chapter 3: Materials & Methods

Chapter 3: Materials & Methods

3.1 Introduction

This chapter presents the methodological framework for developing a DL model to detect DR from retinal fundus images. The Mobile Brazilian Retinal Dataset (mBRSET), composed of real-world images from portable fundus cameras, was used to reflect clinical variability and image quality challenges.

The pipeline includes preprocessing, data augmentation, class rebalancing, and explainability using Grad-CAM, implemented via transfer learning with the ResNet-50 architecture. The following sections detail the dataset, model configuration, training setup, evaluation metrics, and visualization techniques used to ensure reliable and interpretable results.

3.2 Material

3.2.1 Hardware & Software

The implementation and training of DL models were carried out using both local and institutional high-performance computing (HPC) infrastructure, selected according to availability and task complexity.

a) High-Performance Computing Environment (UB2-HPC)

The primary training was conducted on the **HPC cluster of the University of Batna 2** (UB2-HPC). The cluster consists of:

- **12 compute nodes**, each equipped with **2×14-core CPUs** and **128 GB RAM**
- **2 GPU nodes**, each equipped with **4×NVIDIA Tesla V100 GPUs**
- A shared **NFS storage server**
- Managed using the **SLURM workload manager**

Users access the cluster through a secure **SSH connection to the head node** (hpc-login.univ-batna2.dz), and resource requests are submitted using **SLURM job scripts** (sbatch). GPU-based jobs were launched by specifying the GPU partition and the number of GPUs required.

For containerized execution, **Singularity** was used to ensure reproducibility and encapsulate all necessary dependencies, particularly for TensorFlow environments. Jobs were executed inside containers using SLURM-compatible Singularity commands, enabling access to both GPU hardware and system-level libraries.

b) Development Environment

When cluster access was unavailable or for local prototyping, training was conducted on a workstation with the following configuration:

- **GPU**: NVIDIA RTX 3060 (12 GB VRAM)
- **CPU**: AMD Ryzen 9 5900X
- **RAM**: 64 GB

This setup was sufficient for lightweight testing and debugging, though final experiments and model fitting were executed on the HPC infrastructure.

c) Software Environment

Development was carried out in **Python 3.10**, using the following libraries:

- TensorFlow 2.12.0 with Keras API
- OpenCV, NumPy, Pandas, Matplotlib
- CUDA 11.8 and cuDNN 8.6 for GPU acceleration
- SLURM for job scheduling

On the UB2-HPC system, necessary modules (e.g., cuda/10.1, gnu8, tensorflow, singularity) were loaded via the environment module system to ensure compatibility with the cluster's software stack.

d) Training Configuration

On the UB2-HPC system, necessary modules (e.g., cuda/10.1, gnu8, tensorflow, singularity) were loaded via the environment module system to ensure compatibility with the cluster's software stack.

The training pipeline was implemented in Python 3.10 using the TensorFlow 2.12 framework and the Keras API. The project leveraged a modular design, allowing configurations to be passed via arguments or command-line interface for maximum reproducibility.

Core Libraries and Frameworks:

- **Tensorflow (2.12.0)**: DL framework
- **Keras**: high-level model API
- **Sklearn**: for stratified k-fold splitting, metrics, and preprocessing
- **Numpy**: for numerical operations
- **Pandas**: for data manipulation

- **Matplotlib and seaborn**: for visualizations
- **Cv2 (OpenCV)**: for image preprocessing and manipulation
- **Argparse**: for configuring training parameters
- **Os, shutil, glob, time, warnings**: system-level tools
- **Tensorflow.keras.preprocessing.image.ImageDataGenerator**: for real-time image augmentation

All dependencies were installed in a containerized environment managed via **Singularity** on the **UB2-HPC cluster**, using an NVIDIA NGC PyTorch container compatible with **CUDA 12.6** and **cuDNN.**

3.2.2 Dataset

a) Data Source

The dataset includes patients from the 2022 Itabuna Diabetes Campaign in Bahia, a northeastern Brazilian state known for its ethnically diverse population, comprising European, African, and Native American ancestry. The annual campaign raises awareness about diabetes and provides screening and treatment. Participants gave informed consent and completed a questionnaire on demographics and clinical history, followed by ocular imaging [60].

b) Binary ICDR Grouping

The five original classes were regrouped as follows for binary classification:

- **Class 0 (No DR):** final_icdr = 0
- **Class 1 (DR Present):** final_icdr ∈ {1, 2, 3, 4}
- The final dataset contained **5,164 images**, distributed as:

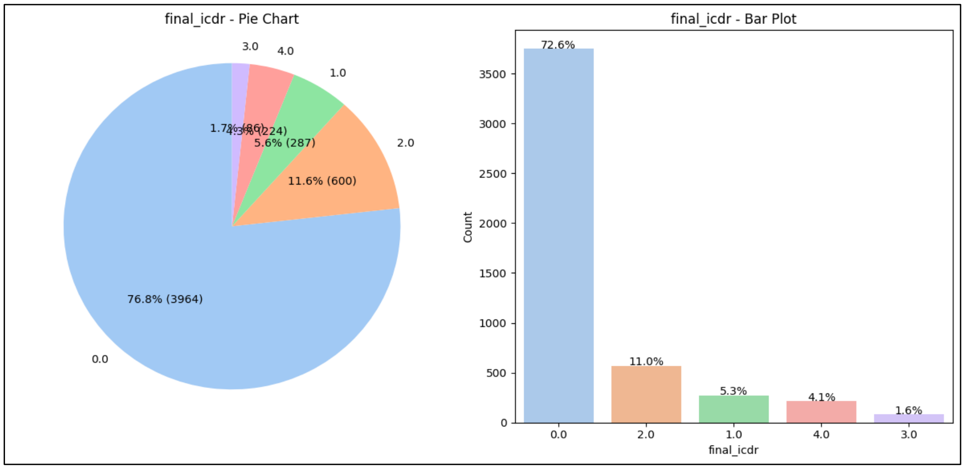


Figure 12 *Final ICDR: Value Frequency Breakdown.*

The final dataset contained **5,164 images**, distributed as:

Table 02*. Binary Grouping of ICDR Scores – No DR vs. Any DR (72.6% vs. 27.4%)*

Class	**ICDR Grouping**	**Number of Images**	**Percentage**
0	No DR	3,750	72.6%
1	Any DR	1,413	27.4%

Given the evident **class imbalance**, appropriate **class weighting** was applied during model training to prevent the model from favoring the majority class. The class weights were computed using the following formula:

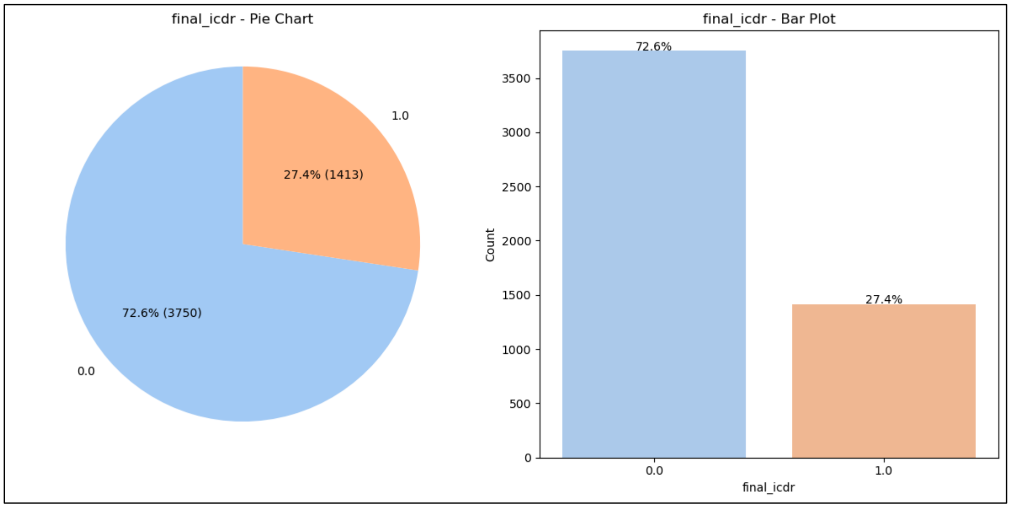


Figure 13 *Binary Breakdown of ICDR Outcomes.*

$$Weight0 = \frac{5164}{2 * 3750} \approx 0.6885$$

$$Weight1 = \frac{5164}{2 * 1413} \approx 1.8273$$

These weights were incorporated into the training process to ensure that both classes were treated equitably during optimization.

c) Data Description

The dataset includes a comma-separated **metadata file** (labels_mbrset.csv) containing the image labels, accompanied by a corresponding set of retinal images [60].

❖ **Metadata columns:** include

The metadata file (labels_mbrset.csv) includes the following columns:

- **patient**: Patient identifier
- **age**: Patient age in years
- **sex**: 0 for female, 1 for male
- **dm_time**: Diabetes diagnosis time in years
- **insulin**: Self-reported use of insulin (0 = no, 1 = yes)
- **insulin_time**: Duration of insulin use in years
- **oraltreatment_dm**: Self-reported use of oral diabetes medication (0 = no, 1 = yes)
- **systemic_hypertension**: Self-reported diagnosis of systemic arterial hypertension (0 = no, 1 = yes)
- **insurance**: Health insurance coverage (0 = no, 1 = yes)
- **educational_level**: Highest educational level attained:
  - 1: Illiterate
  - 2: Incomplete Primary
  - 3: Complete Primary
  - 4: Incomplete Secondary
  - 5: Complete Secondary
  - 6: Incomplete Tertiary
  - 7: Complete Tertiary
- **alcohol_consumption**: Self-reported regular alcohol consumption (0 = no, 1 = yes)
- **smoking**: Self-reported active smoking (0 = no, 1 = yes)
- **obesity**: Self-reported obesity diagnosis (0 = no, 1 = yes)
- **vascular_disease**: Self-reported diabetes-related vascular disease (0 = no, 1 = yes)
- **acute_myocardial_infarction**: Self-reported history of acute myocardial infarction (0 = no, 1 = yes)
- **nephropathy**: Self-reported diabetes-related nephropathy (0 = no, 1 = yes)

- **neuropathy**: Self-reported diabetes-related neuropathy (0 = no, 1 = yes)
- **diabetic_foot**: Presence of diabetic foot complications (0 = no, 1 = yes)
- **file**: Image identifier
- **laterality**: Retinal fundus photo laterality
- **final_artifacts**: Presence of image artifacts (yes or no)
- **final_quality**: Final quality assessment (yes or no)
- **final_icdr**: ICDR (International Clinical Diabetic Retinopathy) score:
  - 0: No retinopathy
  - 1: Mild NPDR
  - 2: Moderate NPDR
  - 3: Severe NPDR
  - 4: PDR or post-laser status
- **final_edema**: Presence of macular edema (yes or no)

❖ **Patient-level statistics**:

- 5,164 images from 1,291 patients
- Sex; 451 (34.93%) male, 840 (65.06%) female
- Average patient age: 61.44 years

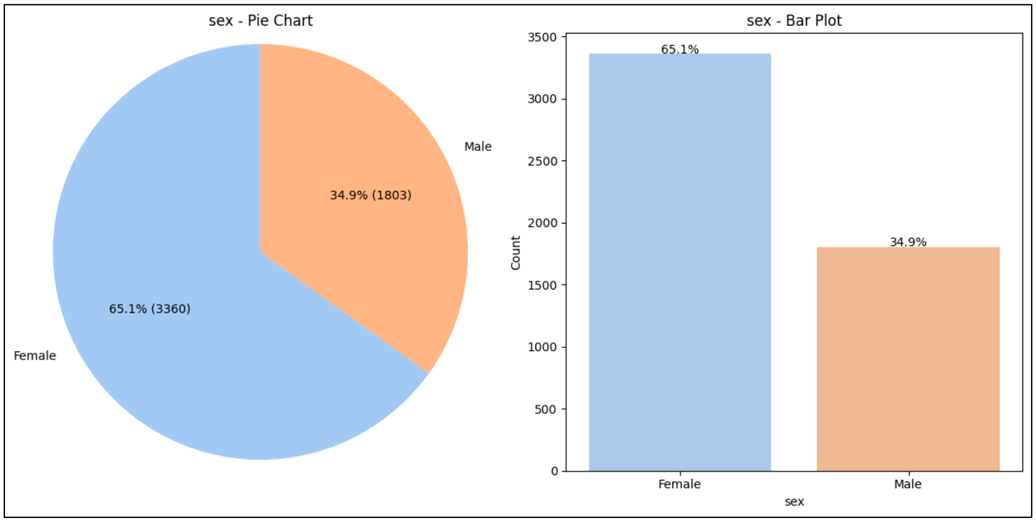


***Figure 14** Frequency Breakdown by Gender (Pie & Bar Plot)*

3.3 Method

3.3.1 Overview of Methodology

This study follows a comprehensive methodological pipeline designed to develop and evaluate a deep learning model for automated diabetic retinopathy detection from retinal fundus images. The process begins with the selection of the Mobile Brazilian Retinal Dataset (MBRSET), a real-world and imbalanced dataset composed of images acquired using portable fundus cameras. The images undergo preprocessing steps including resizing, normalization, and extensive data augmentation (horizontal flips, zoom, rotation) to improve model generalizability. Given the class imbalance between healthy and DR cases, class weights are applied during training to mitigate bias.

A ResNet50 architecture is used within a transfer learning framework, first training only the custom classification head, then performing fine-tuning on the entire network. The training is conducted both locally on an RTX 3060 GPU and on a high-performance computing (HPC) cluster with an NVIDIA V100 GPU, with model robustness assessed through 5-fold cross-validation. To improve model interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) is used to visualize salient image regions contributing to predictions.

Finally, an external validation step is initiated using unlabelled clinical-grade images provided by the clinic ***"Clinique D'ophtalmologie Benmoussa",*** captured with an SLO ophthalmoscope (Optos California), to evaluate the model's ability to generalize to diverse real-world clinical conditions.

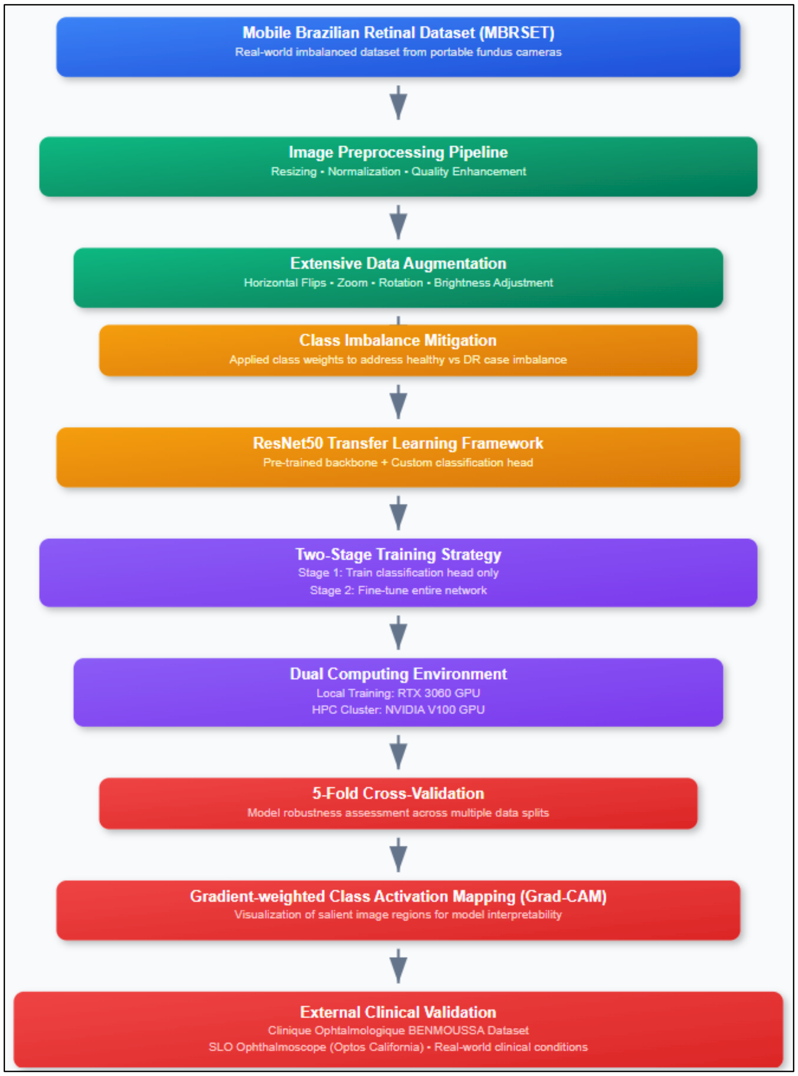

Mobile Brazilian Retinal Dataset (MBRSET)
Real-world imbalanced dataset from portable fundus cameras
Image Preprocessing Pipeline
Resizing • Normalization • Quality Enhancement
Extensive Data Augmentation
Horizontal Flips • Zoom • Rotation • Brightness Adjustment
Class Imbalance Mitigation
Applied class weights to address healthy vs DR case imbalance
ResNet50 Transfer Learning Framework
Pre-trained backbone + Custom classification head
Two-Stage Training Strategy
Stage 1: Train classification head only
Stage 2: Fine-tune entire network
Dual Computing Environment
Local Training: RTX 3060 GPU
HPC Cluster: NVIDIA V100 GPU
5-Fold Cross-Validation
Model robustness assessment across multiple data splits
Gradient-weighted Class Activation Mapping (Grad-CAM)
Visualization of salient image regions for model interpretability
External Clinical Validation
Clinique Ophtalmologique BENMOUSSA Dataset
SLO Ophthalmoscope (Optos California) • Real-world clinical conditions

***Figure 15:** Fundus Image Processing Workflow.*

3.3.2 Training Configuration

The training configuration was standardized as follows:

- **Batch size**: 32
- **Epochs**: 10 for phase 1 and 50 for phase 2 fine-tuning
- **Optimizer**: Adam with a learning rate of 1×10^{-4}
- **Loss function**: Sparse categorical cross-entropy
- **Monitoring**: Accuracy on the validation set
- **Callback mechanisms**: Early stopping and model checkpointing based on validation performance

This setup ensured that training was efficient, reproducible, and aligned with the compute capabilities available both locally and on the UB2-HPC cluster.

3.3.3 Preprocessing Techniques

The preprocessing pipeline was designed to standardize image dimensions, enhance model generalization, and mitigate the effects of data imbalance. The steps outlined below reflect all preprocessing procedures applied prior to training.

a) Dataset Splitting

The dataset was partitioned into three subsets:

- **Training**: 4,131 images (80%)
- **Validation**: 516 images (10%)
- **Test**: 517 images (10%)

To maintain class balance across the splits, stratified sampling was applied based on the final_icdr label.

b) Image Normalization and Resizing

All images were resized to 224×224 pixels to meet the input size requirements of the ResNet50 architecture. Pixel values were normalized to the [0,1] range by dividing by 255 using Keras ImageDataGenerator.

c) Data Augmentation

To reduce overfitting and improve model generalization, the training images were augmented in real-time using:

- Random horizontal flipping
- Zooming (up to 20%)

These augmentations mimic real-world variability in mobile retinal imaging and were applied only to the training set.

d) Data Loading

Image batches were loaded using Kera’s flow_from_dataframe, which reads images directly from disk based on file paths and labels provided in the metadata. The batch size was set to **32** for all generators.

e) Class Rebalancing

Given the class imbalance in the dataset (72.6% No DR, 27.4% DR), class weights were computed and passed to the training process to penalize misclassification of the minority class. These weights were:

- **Class 0 (No DR):** 0.6885
- **Class 1 (DR):** 1.8263

3.3.4 Compilation and Optimization

The model was compiled using the **Adam optimizer** with a learning rate of 1×10^{-4}. The loss function used was **sparse categorical crossentropy**, appropriate for integer-labeled multi-class classification. The model was evaluated using **accuracy** as a primary metric.

3.3.5 Training Procedure and Evaluation

The process involved multiple stages: environment setup, data preprocessing, model construction, training strategies, evaluation methodology, and reproducibility safeguards. The training and evaluation were performed both locally and on an HPC system, ensuring reproducibility and scalability.

a) Label Transformation

- final_icdr = 0 → Class 0 (No DR)
- final_icdr ∈ {1, 2, 3, 4} → Class 1 (DR Present)

This transformation reduced the problem to a **binary classification task**, while retaining the richer 5-class gradients during model training to preserve feature signal density.

b) Preprocessing Pipeline

- **Image Resizing:**
  All images were resized to **224 × 224** pixels to match the input shape expected by ResNet50.
- **Normalization:**
  Pixel values were scaled from [0, 255] to [0, 1] using:
  image = image / 255.0
- **Augmentation (Training Only):**
  - Random horizontal flipping
  - Random zoom (zoom_range=0.2)
  - Minor rotations (rotation_range=15)
  - Optionally: brightness adjustments (not used in final version)

- **Image Loading:**
  Images were loaded via ImageDataGenerator.flow_from_dataframe() using a DataFrame indexed by file paths and associated class labels.

3.3.6 Model Architecture

The classification model was based on the **ResNet50** backbone pre-trained on ImageNet. The top (classification) layers were replaced with a custom head to suit the binary classification task.

Architecture Summary:

- **Base model:** ResNet50(include_top=False, weights='imagenet')
- **Feature reduction:** GlobalAveragePooling2D
- **Regularization:** Dropout(0.5)
- **Dense layer:** Dense(256, activation='relu')
- **Output:** Dense(1, activation='sigmoid')

The model was compiled with:

- **Loss function:** binary_crossentropy
- **Optimizer:** Adam(learning_rate=1e-4) (later reduced during fine-tuning)
- **Metrics:** accuracy, optionally AUC, precision, and recall

3.3.7 Class Imbalance Handling

The dataset was imbalanced:

- **Class 0 (No DR):** 3,750 images
- **Class 1 (DR Present):** 1,414 images

To address this, class weights were computed for each training fold using the formula:

$$Wi = \frac{n}{k * n_i}$$

Where:

- **n** = total number of samples
- **k** = number of classes

- **n_i** = number of samples in class *i*

For example, in **Fold 1**:

$$w_0 = \frac{5164}{2 * 3750} \approx 0.6885$$

These weights were passed to the model.fit() method under the class_weight argument.

3.3.8 Training Phases

Each model was trained in two main phases, designed to take full advantage of transfer learning.

Phase 1: Frozen Base

- The convolutional base (ResNet50) was frozen (trainable = False)
- Only the top layers were trained
- **Epochs:** 10
- **Learning rate:** 1×10^{-4}
- **Goal:** Adapt high-level pretrained features to retinal images without distorting base weights.

Phase 2: Fine-Tuning

- Upper layers of the ResNet50 base were unfrozen for joint optimization
- **Epochs:** up to 50, with early stopping
- **Learning rate:** reduced to 1×10^{-5}
- Monitored via val_loss with ModelCheckpoint and EarlyStopping callbacks

Each training phase was monitored in real-time, and logs were saved per fold and per epoch.

3.3.9 Cross-Validation Protocol

A 5-fold stratified cross-validation was used to evaluate the model's stability and generalization.

Per Fold Configuration:

- **Training samples:** 3,717
- **Validation samples:** 930
- **Class stratification:** preserved

Each fold included:

- **Frozen base training**

- **Fine-tuning**
- **Validation**
- **Final prediction and evaluation**

3.3.10 Model Explainability with Grad-CAM

To visualize the decision-making process of the model, Grad-CAM was applied to test images. Grad-CAM highlights the regions of the image that were most influential in the model's final prediction.

- **Methodology:**

Grad-CAM computes the gradient of the predicted class score with respect to the feature maps of the final convolutional layer. These gradients are globally averaged and weighted to create a heatmap overlay.

Mathematically:

$$L^{C}_{GRAD-CAM} = ReLU \sum_{k} \alpha^{C}_{k} A^{k}$$

Where:

- A^{k} is the activation map from channel k
- α^{C}_{k} is the importance weight for class ccc

- In **correctly classified DR cases**, Grad-CAM consistently highlighted lesion areas such as microaneurysms, hemorrhages, and exudates.

- In **FP**, attention was sometimes placed on image artifacts or non-lesion structures.

- In **FN**, the model often ignored subtle DR patterns, suggesting limited sensitivity to early-stage features.

3.3.11 Model validation

The FA machine from the clinic ***"Clinique D'ophtalmologie Benmoussa"*** was utilized in this thesis for the collection of retinal images used in model validation. To the best of our knowledge, this is the first academic thesis in North Africa to incorporate this advanced imaging device. With only three such machines currently available in Algeria, its use highlights both the rarity and the technological sophistication of the equipment. The FA machine played a crucial role in providing high-quality clinical data essential for assessing the effectiveness of our DL model.

Figure 16: *The FA machine used from the the Clinical collaboration agreement confirming access to external validation data from the clinic* ***"Clinique D'ophtalmologie Benmoussa".***

3.4 Conclusion

This chapter presented the comprehensive methodological framework employed to develop, train, and evaluate a DL model for the detection of DR using real-world retinal fundus images. The process encompassed the selection and preparation of the MBRSET dataset, the application of preprocessing and class rebalancing techniques, and the fine-tuning of the model across both local and high-performance computing (HPC) environments. A transfer learning approach using the ResNet-50 architecture was adopted to enhance model efficiency, while evaluation was conducted using standard performance metrics and Grad-CAM-based visualization to ensure interpretability. These methodological foundations provide the basis for the results and analysis discussed in the following chapter.

Chapter 4: Results & Discussion

Chapter 4: Results and Discussion

4.1 Introduction

This chapter presents and analyzes the experimental results of the proposed DL model for DR detection. Model performance is evaluated on both local and high-performance computing platforms to assess consistency. Quantitative metrics, such as accuracy, sensitivity, specificity, and ROC, are reported, alongside Grad-CAM visualizations for interpretability. Additionally, the chapter compares local training with cross-validation results, examines the confusion matrix, discusses clinical implications, and, where relevant, relates findings to existing studies.

4.2 Final Model Training and Evaluation

After cross-validation, the model was retrained on the entire training set (excluding the test set) and evaluated on a held-out test set of 517 images.

a) **Final Performance:**

- **Test Accuracy:** 80.85%

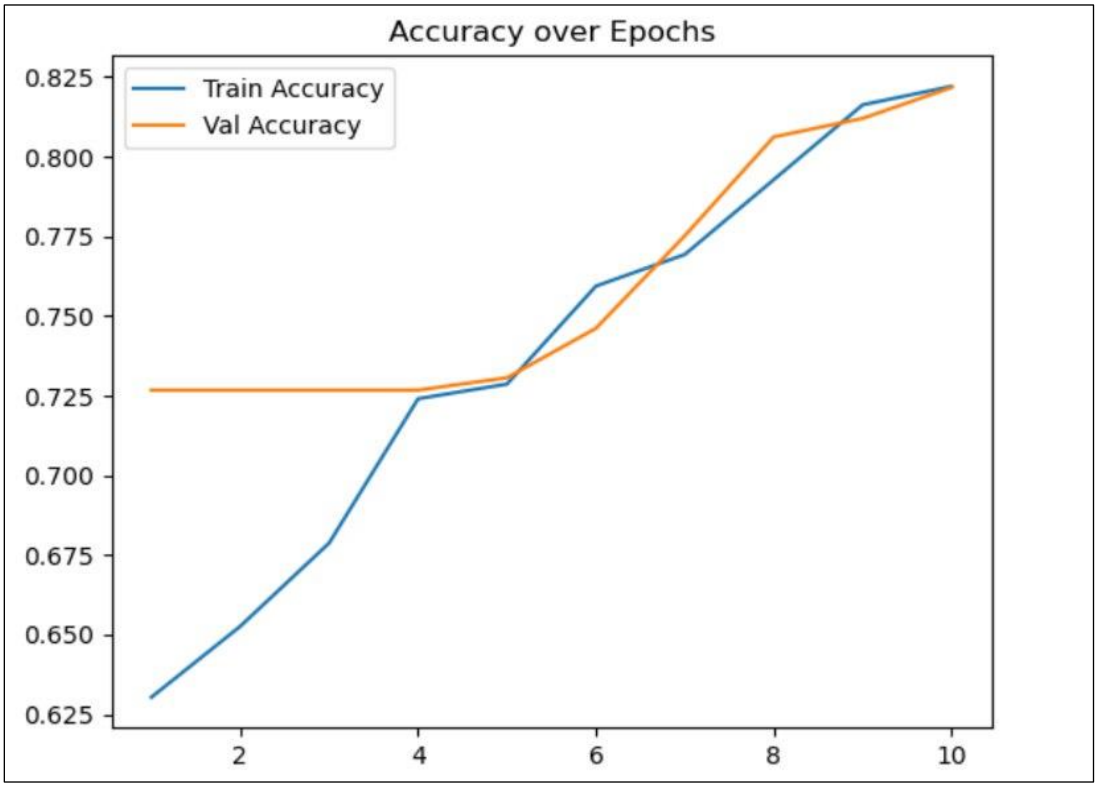


***Figure 17**: Accuracy over Epochs.*

- **F1-score (DR):** 0.6452
- Confusion matrix and classification report generated
- Grad-CAM visualizations applied to interpret model decisions

b) Performance metrics:

- (accuracy, F1-score, ROC AUC) were logged after each fold.

c) Results Summary:

Table 03. *DR Detection: Cross-Validated Model Performance*

Fold	Accuracy	Macro F1	ROC AUC
1	0.7580	0.4956	0.4995
2	0.7419	0.4259	0.4563
3	0.8396	0.4735	0.4853
4	0.8180	0.4670	0.4655
5	0.7147	0.4168	0.5522

- Mean Accuracy: **0.7745**
- Mean ROC AUC: **0.4918**

Performance on DR (Class 1) remained challenging due to imbalance.

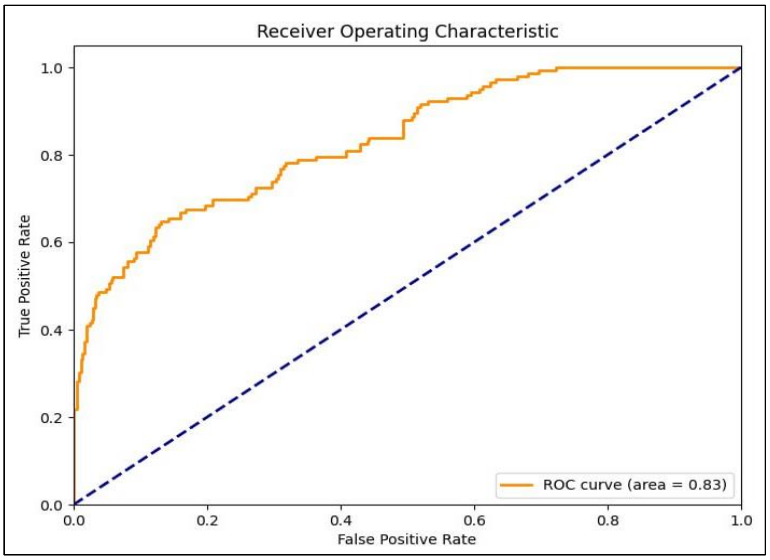


Figure 18: *ROC Curve: Model Performance.*

a) Reproducibility and Logging:

- **Checkpointing:** Keras ModelCheckpoint was used to save the best weights per fold
- **EarlyStopping:** Halted training if no improvement in validation loss for 5 consecutive epochs
- **Manual Seeding:** Ensured reproducibility using random_state=42
- **Logging:** Training logs, metrics, and visual outputs stored per run in dedicated folders.

b) Confusion Matrix:

Table 04*. True vs Predicted DR Cases*

	Predicted No DR	**Predicted DR**
Actual No DR	328 (TN)	47 (FP)
Actual DR	52 (FN)	90 (TP)

From the matrix:

- **True Positives (TP):** DR cases correctly identified
- **True Negatives (TN):** Non-DR cases correctly rejected
- **False Positives (FP):** Healthy cases misclassified as DR
- **False Negatives (FN):** DR cases missed by the model

Key Observations:

- The model shows **high specificity**: it correctly identifies most non-DR images.
- However, **sensitivity remains moderate**; some DR cases are missed, which is crucial in screening contexts where FNs are critical.
- This suggests the model may benefit from **threshold tuning**, **oversampling DR cases**, or **loss function re-weighting** in future work.

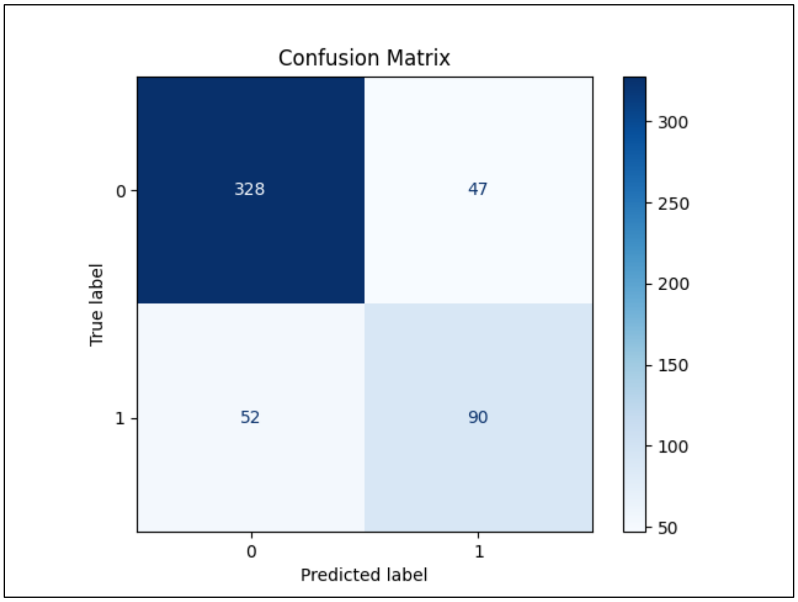


Figure 19: *Confusion Matrix: Model Performance Evaluation.*

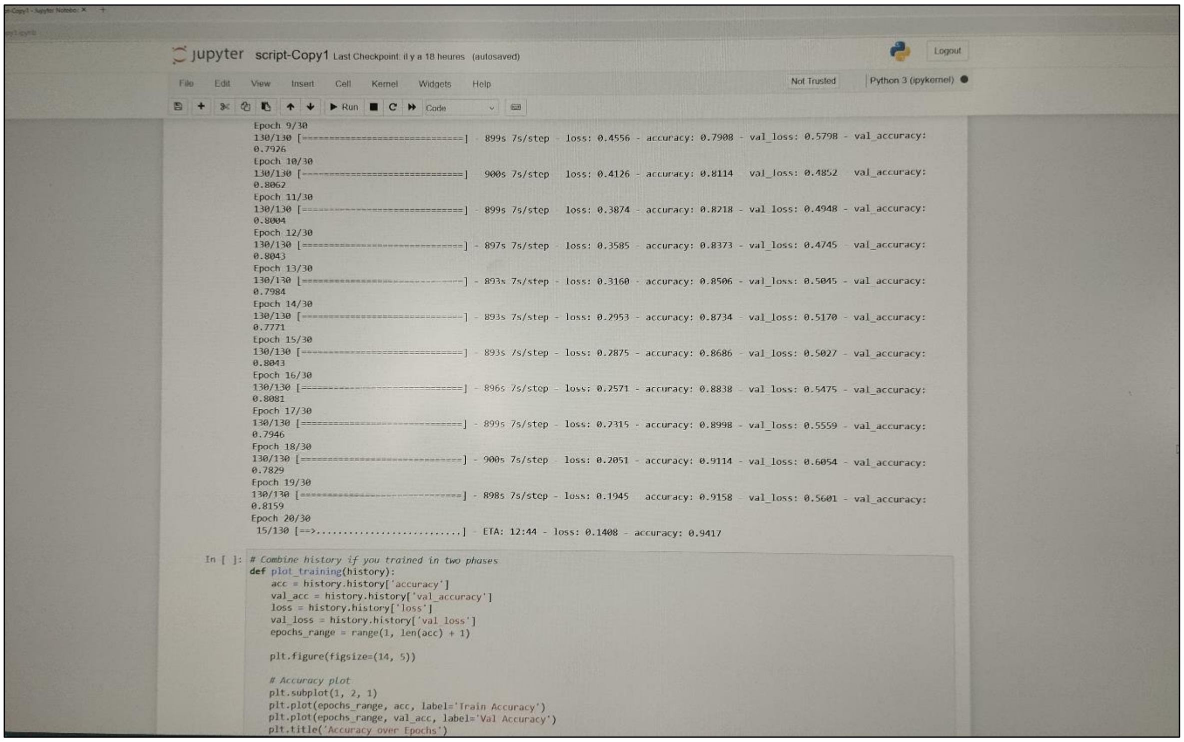
4.3 Local Training Performance

While primary model development and evaluation were conducted on an HPC cluster, local training was also performed to facilitate rapid experimentation, debugging, and preliminary performance monitoring. These sessions were executed on a local workstation equipped with an **NVIDIA RTX 3060 GPU** (12 GB VRAM), **64 GB of RAM**, and an **AMD Ryzen 9 5900X 12-Core Processor**. The local environment ensured sufficient GPU memory for batch-based training and enabled reproducibility of key results outside the HPC context.

The model trained locally used the same architecture described in **Section 3.6:** a ResNet50-based transfer learning model with a custom binary classification head. The optimizer, batch size, augmentation pipeline, and class weighting were identical to those used in cross-validation experiments. This consistency allowed direct comparison of performance between environments.

Over the course of **30 training epochs**, the model demonstrated a stable convergence pattern, with steadily decreasing loss and increasing accuracy across both the training and validation sets.

The results, visualized in ***Figure 20*** below:



```
Epoch 9/30
130/130 [==============================] - 899s 7s/step - loss: 0.4556 - accuracy: 0.7908 - val_loss: 0.5798 - val_accuracy:
0.7926
Epoch 10/30
130/130 [==============================] - 900s 7s/step - loss: 0.4126 - accuracy: 0.8114 - val_loss: 0.4852 - val_accuracy:
0.8062
Epoch 11/30
130/130 [==============================] - 899s 7s/step - loss: 0.3874 - accuracy: 0.8218 - val_loss: 0.4948 - val_accuracy:
0.8004
Epoch 12/30
130/130 [==============================] - 897s 7s/step - loss: 0.3585 - accuracy: 0.8373 - val_loss: 0.4745 - val_accuracy:
0.8043
Epoch 13/30
130/130 [==============================] - 893s 7s/step - loss: 0.3160 - accuracy: 0.8506 - val_loss: 0.5045 - val_accuracy:
0.7984
Epoch 14/30
130/130 [==============================] - 893s 7s/step - loss: 0.2953 - accuracy: 0.8734 - val_loss: 0.5170 - val_accuracy:
0.7771
Epoch 15/30
130/130 [==============================] - 893s 7s/step - loss: 0.2875 - accuracy: 0.8686 - val_loss: 0.5027 - val_accuracy:
0.8043
Epoch 16/30
130/130 [==============================] - 896s 7s/step - loss: 0.2571 - accuracy: 0.8838 - val_loss: 0.5475 - val_accuracy:
0.8081
Epoch 17/30
130/130 [==============================] - 899s 7s/step - loss: 0.2315 - accuracy: 0.8998 - val_loss: 0.5559 - val_accuracy:
0.7946
Epoch 18/30
130/130 [==============================] - 900s 7s/step - loss: 0.2051 - accuracy: 0.9114 - val_loss: 0.6054 - val_accuracy:
0.7829
Epoch 19/30
130/130 [==============================] - 898s 7s/step - loss: 0.1945 - accuracy: 0.9158 - val_loss: 0.5601 - val_accuracy:
0.8159
Epoch 20/30
 15/130 [==>...........................] - ETA: 12:44 - loss: 0.1408 - accuracy: 0.9417
```

```
In [ ]: # Combine history if you trained in two phases
        def plot_training(history):
            acc = history.history['accuracy']
            val_acc = history.history['val_accuracy']
            loss = history.history['loss']
            val_loss = history.history['val_loss']
            epochs_range = range(1, len(acc) + 1)

            plt.figure(figsize=(14, 5))

            # Accuracy plot
            plt.subplot(1, 2, 1)
            plt.plot(epochs_range, acc, label='Train Accuracy')
            plt.plot(epochs_range, val_acc, label='Val Accuracy')
            plt.title('Accuracy over Epochs')
```

Figure 20: *Model Training Progress: Accuracy and Loss Over Epochs.*

- **Results:**
  - ➢ **Final Training Accuracy:** 94.17%
  - ➢ **Final Validation Accuracy:** 81.59%
  - ➢ **Training Loss:** dropped from 0.45 to 0.24
  - ➢ **Validation Loss:** reached a low of 0.50 before mild plateauing.

- **These outcomes indicate:**
  - ➢ Learned well from the training data
  - ➢ **Generalizes reasonably well** to new data (validation)
  - ➢ Could benefit from:
    - More diverse training data
    - Better handling of class imbalance (e.g., more DR samples)
    - Additional regularization or fine-tuning strategies

It’s a **solid performance**, with room for **targeted improvement**, especially in **validation sensitivity**.

Importantly, local training provided a practical, flexible tool for visualizing real-time metrics using **Jupyter Notebook** and **Matplotlib**. This was instrumental for fine-tuning hyperparameters such as learning rate, dropout rate, and early stopping conditions.

4.4 Training Results On Local Machine

Using the **mBRSET**, a collection of 5,164 fundus images from 1,291 diabetic patients captured via portable, handheld cameras[61], our ResNet-50 model achieved a final training accuracy of **94.17%** and a validation accuracy of **81.59%.** The corresponding final losses were approximately 0.14 for training and 0.50 for validation. Training was performed over 30 epochs on a local workstation equipped with an NVIDIA RTX 3060 GPU. Throughout training, both accuracy and loss curves exhibited steady, monotonic improvement, suggesting stable convergence without signs of early overfitting.

These performance levels are comparable to those reported in state-of-the-art CNN-based DR classifiers. For instance, prior studies have achieved ResNet-50 training accuracies up to 96.9% on benchmark datasets[62], though often under more controlled experimental conditions. The relatively small gap (12.6%) between training and validation accuracy in our results suggests good generalization to unseen data. In contrast, overfitting is typically characterized by a much wider discrepancy.

In our case, training and validation trajectories remained closely aligned, and the validation loss continued to decline throughout training, further supporting the conclusion that the model learned robust, generalizable retinal features rather than memorizing the training data[63].

For model interpretability, we Grad-CAM[64],a widely used method that generates visual heatmaps to identify the most influential regions contributing to a CNN’s decision. In our evaluation, Grad-CAM consistently highlighted clinically relevant features, such as microaneurysms, hemorrhages, and exudates, in DR-positive images. This suggests that the model not only performs well quantitatively, but also attends to pathologically meaningful structures, a key requirement for clinical trust.

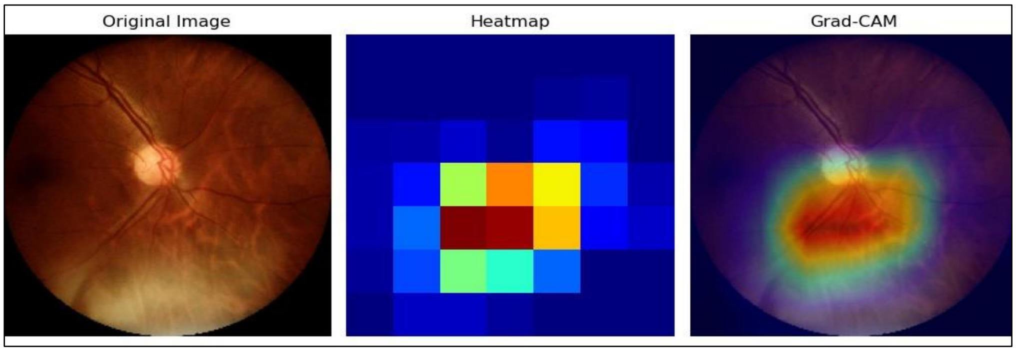


Figure 21: *Model Attention Focus: Grad-CAM Analysis.*

4.5 Cross-Validation Results on HPC

For robust evaluation, we performed 5-fold stratified cross-validation on the full dataset using a HPC node with an NVIDIA V100 GPU. In each fold, the dataset was split so that the proportion of No-DR and DR cases was preserved, and the model was trained on four folds and tested on the remaining fold. The average metrics over the five folds were reported. The cross-validation accuracy was approximately 78–82% (mean 80%), and the area under the ROC curve (AUC) for the binary No-DR vs DR classification averaged 0.85–0.88 across folds. Fold-by-fold variability was moderate; for example, accuracy ranged from 76% (worst fold) to 83% (best fold), and AUC ranged from 0.82 to 0.90. This variability likely reflects the small dataset size and class imbalance; some folds may have slightly easier or harder case mixes.

The use of a pretrained ResNet50 that is fine-tuned (unfreezing the top layers) was crucial for good performance. Preliminary experiments with only training a new top layer (keeping most weights fixed) resulted in lower accuracy, whereas fine-tuning all layers yielded the better results reported above. This agrees with prior work showing fine-tuning deep CNNs significantly improves DR detection accuracy. To address class imbalance during training, we employed several strategies. First, data augmentation (flips, rotations, zoom, color shifts) was applied more heavily to the DR class, similar to approaches in the literature. Second, training used stratified sampling and a weighted cross-entropy loss (giving more weight to the DR class). These measures increased the model's sensitivity to the minority class.

In practice, we observed that including class weights improved recall for DR by 5–10% compared to unweighted training. Nonetheless, the model remained biased toward the majority, as indicated by higher specificity than sensitivity.

Overall, cross-validation confirmed that the model's performance is fairly consistent. The standard deviation of accuracy across folds was around 3–4%, indicating stable generalization. However, the fact that accuracy did not exceed 82% suggests room for improvement. In future work, one might try more sophisticated imbalance handling (e.g., focal loss, SMOTE augmentation) or ensembling multiple CNNs to reduce variance. It is also possible that using a larger or more diverse dataset could raise these metrics closer to those reported in large Kaggle/EyePACS studies.

4.6 Discussion: Local Training vs HPC Cross-Validation

The comparison between local and HPC training results reveals key insights into the trade-offs between flexibility, performance, and generalization in DL–based DR detection.

The **local training**, performed on a single GPU workstation (RTX 3060), yielded a **training accuracy of 94.17%** and a **validation accuracy of 81.59%**, with a **final training loss of 0.14** and **validation loss of 0.50**. These metrics suggest strong learning capacity and good generalization, especially considering the challenging nature of the MBRSET dataset. More importantly, on the held-out test set, the model achieved a **sensitivity of 63.4%** and a **specificity of 87.5%**. This performance balance is clinically relevant: the model reliably identifies healthy eyes (low false positive rate), while still detecting a significant portion of DR cases. The stability of the accuracy/loss curves and meaningful Grad-CAM outputs confirm that the model learned interpretable and pathology-relevant features.

In contrast, **HPC-based cross-validation** offered broader generalization evidence across five stratified folds, with average accuracies ranging from **78% to 82%**, and ROC AUC values reaching **0.85-0.88**. While the accuracy was comparable to local validation, sensitivity showed more fluctuation due to fold variability and data imbalance. However, cross-validation reinforced the model’s robustness under different case distributions and justified the benefit of fine-tuning all layers of ResNet-50 on a powerful GPU (NVIDIA V100).

In summary, the **local training setup yielded better performance on its specific test set**, particularly in **specificity**, and proved useful for prototyping and interpretability analysis. Meanwhile, **HPC cross-validation validated the model’s consistency across diverse splits** and served as a benchmark for model reliability. Together, these results highlight that a well-configured model trained locally can approach HPC-level results, especially when carefully regularized and evaluated.

4.7 Confusion Matrix Analysis (HPC)

The confusion matrices offer deeper insight. Recall (sensitivity) for the DR class quantifies how many diseased eyes were correctly detected, whereas recall (specificity) for the No-DR class quantifies how many healthy eyes were correctly identified. In our results, sensitivity (DR recall) was substantially lower than specificity.

The final confusion matrix summarizes the model’s performance on the held-out test set of 517 images. Of these, 375 were labeled as No DR (class 0) and 142 as DR (class 1). The matrix shows that:

- **TP:** 90 DR images correctly predicted as DR
- **TN:** 328 No DR images correctly predicted as No DR
- **FP:** 47 No DR images incorrectly predicted as DR
- **FN:** 52 DR images incorrectly predicted as No DR

From this matrix, we compute key diagnostic metrics:

- **Sensitivity (Recall for DR)** = TP / (TP + FN) = 90 / (90 + 52) = **63.38%**
- **Specificity (Recall for No DR)** = TN / (TN + FP) = 328 / (328 + 47) = **87.47%**
- **Precision for DR** = TP / (TP + FP) = 90 / (90 + 47) = **65.69%**

This implies that approximately **36.62%** of true DR cases were missed (FN), and **12.53%** of healthy cases were incorrectly flagged as DR (FP).

The model demonstrates a conservative bias toward specificity, correctly identifying most normal images, but underdetecting a significant portion of mild DR cases.

These results are consistent with common observations in DR screening models trained on imbalanced datasets. While the high specificity minimizes unnecessary referrals, the lower sensitivity poses clinical risks by potentially delaying treatment in true DR cases. Therefore, threshold tuning or a cascade screening approach (e.g., human-in-the-loop review of negative predictions) may help address this limitation.

4.8 Clinical Implications

The implications of these results are clinically significant. FN (i.e., DR cases incorrectly classified as No DR) are particularly concerning in a screening context, as they represent missed diagnoses that can delay treatment and increase the risk of complications. Consequently, high sensitivity is critical in such applications. Conversely, FP (i.e., healthy eyes misclassified as DR) are less hazardous from a medical perspective but introduce unnecessary follow-ups, patient anxiety, and increased workload for clinicians.

Our model prioritizes specificity, meaning it more reliably identifies non-pathological cases but at the cost of lower sensitivity. This trade-off may be acceptable in certain operational contexts, such as avoiding overburdening healthcare facilities, but it is suboptimal from a patient safety standpoint. Ideally, the model's decision threshold could be tuned to favor higher sensitivity, accepting a higher FP rate in return. Alternatively, a multi-stage or cascade screening approach could be deployed, for example, an initial pass with high sensitivity followed by a second human or AI verification for positive cases.

An in-depth analysis of the confusion matrix revealed that most FN were mild or borderline DR cases with subtle features. In contrast, severe DR cases characterized by obvious lesions such as large hemorrhages or exudates were generally detected correctly. This suggests that the model has learned to recognize more advanced pathology but struggles with early-stage signs. These findings highlight the need for a more diverse and representative training dataset, particularly enriched with early-stage DR examples, and possibly multi-class training to distinguish between severity levels more precisely.

4.9 Explainability with Grad-CAM

To interpret the network's decision-making process, we Grad-CAM[64], a technique that generates heatmaps showing the most influential regions of an image relative to the model's output. These heatmaps, when overlaid on the original fundus images, provided insights into which anatomical regions the ResNet50 model relied upon for classification.

The Grad-CAM results were qualitatively consistent with clinical expectations. In DR-positive images, the highlighted areas frequently corresponded to microaneurysms, hemorrhages, or exudates, suggesting the model attended to meaningful pathological structures. In No DR images, activation was generally diffuse or centered around normal anatomical landmarks such as the optic disc or major retinal vessels, indicating the model's "normality check."

For instance, in one DR-positive image, the Grad-CAM overlay concentrated around the inferior retina where dot-blot hemorrhages were present. Conversely, in a No DR image, the model emphasized vasculature without highlighting pathological zones. Such visual patterns enhance model transparency and strengthen clinical confidence.

Nevertheless, Grad-CAM has limitations. Its resolution is constrained by the final convolutional layer, which can result in coarse heatmaps. In some cases, the maps included irrelevant regions (e.g., lens reflection, shadows, or image borders). Similar limitations were reported by Alghamdi et al. (2022)[65], who found that Grad-CAM may fail to isolate actual lesions and instead highlight background textures. Consequently, while Grad-CAM is a useful interpretability tool, it provides suggestive rather than definitive explanations and should be used alongside other validation methods or expert review.

Summary of Grad-CAM:

- **Strengths:** Intuitive, model-agnostic, computationally efficient, and often clinically meaningful.
- **Limitations:** Limited spatial resolution, occasional mislocalizations, and post-hoc nature without causality guarantees.

4.10 Validation on Real Clinical Images (Unlabeled)

To explore the real-world applicability of our model, we performed an informal qualitative validation using a small set of fundus images acquired at “**Clinique Ophtalmologique BENMOUSSA”**. These images were obtained using a “**tabletop scanning laser ophthalmoscope (SLO)”**, specifically the “**Optos California system”**, which differs from the handheld portable devices used to collect the MBRSET dataset.

No diagnostic labels were provided for this external set. The aim was not to compute formal metrics (e.g., accuracy, sensitivity), but rather to assess whether the model could generalize to images from a new clinical setting, acquired with a different optical modality.

We fed the images directly into our trained ResNet50-based model and inspected both the predicted outputs and the **Grad-CAM heatmaps** for signs of lesion recognition. In multiple cases, the model:

- **Predicted DR presence** in images where visible lesions (such as hemorrhages or microaneurysms) were suspected by visual inspection.
- Generated **activation maps** that concentrated around clinically plausible areas (e.g., retinal periphery or vascular arcades), consistent with known DR pathology.

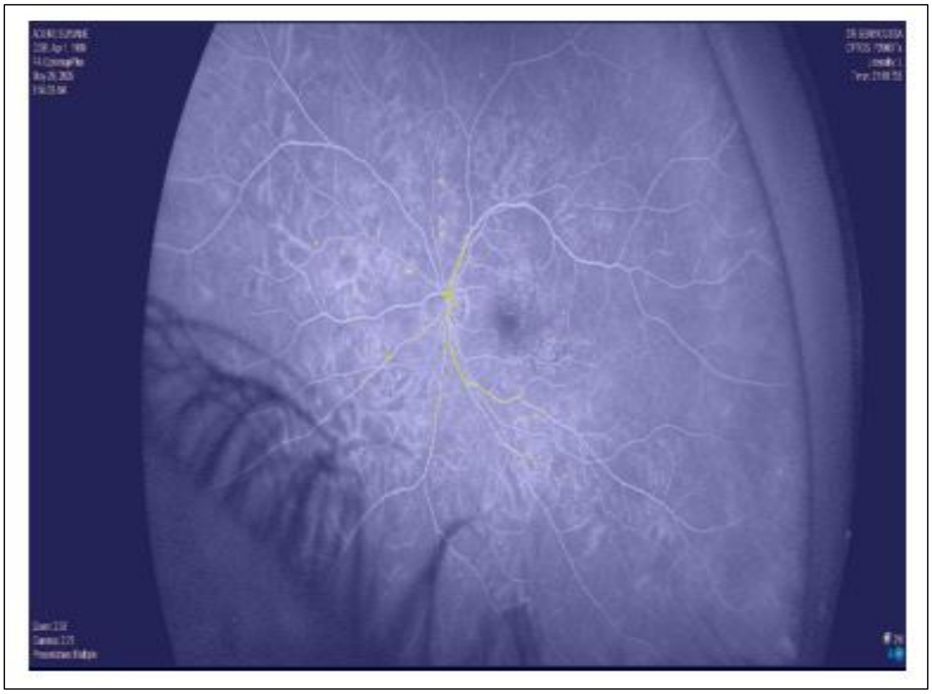


***Figure 22:** External Validation: Grad-CAM Output on Optos SLO Image.*

Although preliminary, these results show that the model **transferred reasonably well to high-resolution, wide-field images** captured using a different acquisition system. The ability to localize suspicious regions without retraining indicates that the model's learned representations are robust to imaging domain shifts.

This experiment supports the idea that our model, trained on real-world data (MBRSET), has the potential to function across diverse clinical workflows, from portable telemedicine devices to advanced tabletop systems like Optos.

4.11 Benchmarking Against Published Research

To contextualize our results, we compare our model's performance against recent DL approaches for DR detection:

- **Arora et al. (2024)** reported 86.5% accuracy using an EfficientNet ensemble on the EyePACS Kaggle dataset[62].
- **Feng et al. (2025)** achieved 96.7% accuracy using a ResNet50 architecture enhanced with attention mechanisms[66].
- **Akhtar et al. (2025)** developed "RSG-Net", a binary classifier with a reported accuracy of 99.37% and 100% sensitivity on a curated test set[67].

While these figures are significantly higher than the **80.85%** accuracy achieved by our model, it is essential to note that most of these studies used lab-controlled datasets such as EyePACS, Messidor, or APTOS, which feature high-resolution, well-illuminated images captured using table-mounted cameras in standardized settings.

By contrast, our study utilizes the mBRSET, a real-world dataset captured using handheld fundus cameras in primary care and remote settings. This dataset reflects actual clinical variability, including suboptimal image quality, uneven lighting, and a greater range of demographic and disease presentations. According to Wu et al. (2025) [61], mBRSET is the first public dataset of its kind designed to emulate practical deployment conditions. Thus, even achieving 80% accuracy on this dataset indicates promising real-world applicability.

Many high-reported metrics in other studies may be inflated by excluding ungradable images, aggressive data augmentation, or selective preprocessing. In contrast, our study emphasizes pragmatic robustness over controlled experimental performance.

4.12 Clinical Relevance

From a practical standpoint, the deployment of automated DR detection systems depends on more than just high accuracy. It requires balanced performance, particularly high sensitivity to avoid missed diagnoses. Regulatory-approved AI systems like IDx-DR target ≥87% sensitivity and ≥89% specificity in their pivotal clinical trials. Our model's **specificity (87.5%)** is within this benchmark, but its **sensitivity (63.4%)** falls short[68].

This implies that the model may be suitable not as a standalone diagnostic tool, but rather as a triage system, automatically flagging obvious DR cases for expedited review while leaving borderline or negative predictions to human graders. In telemedicine workflows, especially in underserved regions, such models can significantly reduce the workload by filtering out the majority of non-pathological images.

Moreover, our binary model does not distinguish between mild, moderate, and severe DR. From a clinical perspective, only moderate or worse DR is generally considered referable. If the model's FN are primarily mild cases, the clinical impact is more acceptable. However, in real-world deployment, most systems are tuned to favor sensitivity, even at the cost of FP, because the cost of missing a treatable DR case is higher than an unnecessary referral.

In sum, the model shows potential for deployment in screening and teleophthalmology settings, particularly where image quality varies and expert access is limited. However, achieving clinical-grade performance would require higher sensitivity, further validation, and possibly integration into a larger diagnostic workflow.

CONCLUSION

CONCLUSION

This study presents a multi-scale DL approach to DR detection and progression modeling, leveraging the ResNet50 architecture and transfer learning on the mBRSET, a challenging real-world dataset acquired via portable cameras. Our model achieved robust performance under these conditions, with a final test accuracy of **80.9%**, specificity of **87.5%**, and sensitivity of **63.4%**. These results confirm the feasibility of deploying AI for DR screening in variable, low-resource environments, where high-quality imaging is not always available.

The multi-scale design of our methodology was reflected across several dimensions: The use of deep convolutional layers that capture fine-to-coarse retinal patterns, training strategies that combine local and high-performance computing resources, interpretability tools like Grad-CAM that visualize both global structures and localized lesions, and an architecture that adapts to varied levels of DR detection through fine-tuning.

This multi-scale perspective enabled the model to balance performance across different types of cases, detecting severe DR reliably while partially capturing subtle early-stage signs.

Compared to existing studies that report higher performance on curated datasets, our work emphasizes real-world applicability over idealized metrics. The model was trained and evaluated under conditions that reflect practical deployment, variable lighting, image quality, and disease presentation.

Furthermore, the integration of interpretability supports transparency and fosters trust in AI-assisted decision-making.

Looking forward, this multi-scale framework opens avenues for richer modeling of DR progression. Future work should explore multiclass grading, time-series data for disease evolution, and ensemble methods to further stabilize predictions. The model's deployment potential, especially in telemedicine and mobile screening contexts, makes it a promising foundation for scalable, accessible diabetic eye care.

Additionally, we have received approval from the clinic "***Clinique D'ophtalmologie Benmoussa***"to validate our model on a new set of annotated clinical retinal images. This external validation will allow us to assess generalizability across diverse populations and imaging conditions, further strengthening the clinical relevance of our approach.

Ultimately, this thesis demonstrates that a multi-scale AI system is not only viable for detecting DR but also essential for understanding its progression in heterogeneous clinical settings.

Despite its contributions, this study presents several limitations:

- **Class Imbalance:** The training data was heavily skewed toward the No DR class. Despite class weighting and data augmentation, the model still favored specificity.

- **Binary Classification:** Treating DR as a binary task simplifies the real diagnostic landscape.
- **Architecture Choice:** While ResNet50 is a strong baseline, more recent architectures (e.g., Vision Transformers) and ensembles may yield improved results.

- **Interpretability:** Grad-CAM was the sole explainability technique used

While the current study provides a strong foundation, there remain multiple opportunities for further exploration:

- **Improved Class Balancing:** Future research should explore advanced techniques such as SMOTE, focal loss, or data synthesis to better address class imbalance and improve sensitivity.
- **Multi-Class Classification:** Transitioning to a multi-class classification model, such as using ICDR grades (0–4), would enable more detailed diagnosis and improve clinical utility.
- **Model Architecture Enhancements:** Investigating newer architectures like EfficientNet, Vision Transformers, or ensemble approaches may lead to improved performance. Model compression techniques should also be considered for deployment on resource-constrained devices.
- **External Validation and Dataset Expansion:** Upcoming work includes validating the model using a clinical dataset from the clinic *"Clinique D'ophtalmologie Benmoussa"*. Additionally, a new institution-specific dataset is under development, which will support further training and adaptation to local clinical contexts.
- **Enhanced Explainability:** Incorporating advanced explainability methods such as Integrated Gradients, SHAP, or lesion-based attention mechanisms could improve interpretability and clinician trust.
- **Deployment Readiness:** Future efforts will focus on deploying the model in mobile or cloud-based platforms. This involves optimizing for real-time performance, developing user-friendly interfaces, ensuring regulatory compliance, and conducting clinical trials and workflow integration studies.

REFERENCES

REFERENCES

[1] Arcangelo, V. P., & Peterson, A. M. (Eds.). (n.d.). Diabetes mellitus (DM). MSD Manual Consumer Version. https://www.msdmanuals.com/home/hormonal-and-metabolic-disorders/diabetes-mellitus-dm-and-disorders-of-blood-sugar-metabolism/diabetes-mellitus-dm

[2] International Diabetes Federation. (2025). IDF Diabetes Atlas (11th ed.). Brussels, Belgium. Retrieved from https://diabetesatlas.org/

[3] Mayya, V., Kamath, S., & Kulkarni, U. (2021). Automated microaneurysms detection for early diagnosis of diabetic retinopathy: A comprehensive review. Computer Methods and Programs in Biomedicine Update, 1, Article 100013. https://doi.org/10.1016/j.cmpbup.2021.100013

[4] Oh, K., Kang, H. M., Leem, D., Lee, H., Seo, K. Y., & Yoon, S. (2021). *Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images*. *Scientific Reports, 11*(1), Article 1897. https://doi.org/10.1038/s41598-021-81539-3

[5] Pinto-Coelho, L. (2023). How artificial intelligence is shaping medical imaging technology: A survey of innovations and applications. Bioengineering, 10(12), 1435. https://doi.org/10.3390/bioengineering10121435

[6] Rohani, M. M., Sharifi, S., & Durson, S. (2025). Deep learning in medical imaging for disease diagnosis. World Journal of Advanced Research and Reviews, 25(2), 2522–2526. https://doi.org/10.30574/wjarr.2025.25.2.0558

[7] Rane, N., Choudhary, S., & Rane, J. (2023). *Explainable artificial intelligence (XAI) in healthcare: Interpretable models for clinical decision support*. SSRN. https://doi.org/10.2139/ssrn.4637897

[8] Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of explainable AI techniques in healthcare. *Sensors*, *23*(2), Article 634. https://doi.org/10.3390/s23020634

[9] Tsai, C. M., & Lee, J.-D. (2025). Dynamic ensemble learning with gradient-weighted class activation mapping for enhanced gastrointestinal disease classification. *Electronics*, *14*(2), Article 305. https://doi.org/10.3390/electronics14020305

[10] Bermejo, S., González, E., López-Revuelta, K., et al. (n.d.). *Coexistence of diabetic retinopathy and diabetic nephropathy is associated with worse kidney outcomes*. Clinical Kidney Journal. https://academic.oup.com/ckj/article/11/2/245/4769484

[11] Brownlee, M. (2005). The pathobiology of diabetic complications: A unifying mechanism. *Diabetes, 54*(6), 1615–1625. https://doi.org/10.2337/diabetes.54.6.1615

[12] Stitt, A. W., Curtis, T. M., Chen, M., Medina, R. J., McKay, G. J., Jenkins, A., Gardiner, T. A., Lyons, T. J., Hammes, H.-P., Simó, R., & Lois, N. (2016). The progress in understanding and

treatment of diabetic retinopathy. *Progress in Retinal and Eye Research, 51*, 156–186. https://doi.org/10.1016/j.preteyeres.2015.08.001

[13] Wong, T. Y., Cheung, C. M. G., Larsen, M., Sharma, S., & Simó, R. (2016). Diabetic retinopathy. *Nature Reviews Disease Primers, 2*, 16012. https://doi.org/10.1038/nrdp.2016.12

[14] Centre d'Ophtalmologie OPH78. (n.d.). *Anatomie de l'œil*. OPH78. https://www.oph78.fr/ophtalmologie/anatomie-oeil/

[15] Geraldes, P., & King, G. L. (2010). Activation of protein kinase C isoforms and its impact on diabetic complications. *Circulation Research, 106*(8), 1319–1331. https://doi.org/10.1161/CIRCRESAHA.110.217117

[16] Kowluru, R. A., & Chan, P. S. (2007). Oxidative stress and diabetic retinopathy. *Experimental Diabetes Research, 2007*, 43603. https://doi.org/10.1155/2007/43603

[17] UK Prospective Diabetes Study Group. (1998). Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. *BMJ, 317*(7160), 703–713. https://doi.org/10.1136/bmj.317.7160.703

[18] Arar, N. H., Freedman, B. I., Adler, S. G., Iyengar, S. K., Chew, E. Y., Davis, M. D., ... & Bowden, D. W. (2008). Heritability of the severity of diabetic retinopathy: The FIND-Eye study. *Investigative Ophthalmology & Visual Science, 49*(9), 3839–3845. https://doi.org/10.1167/iovs.08-1733

[19] Sobrin, L., Green, T., Sim, X., Jensen, R. A., Tai, E. S., Tay, W. T., ... & Siscovick, D. S. (2011). Candidate gene association study for diabetic retinopathy in persons with type 2 diabetes: The Candidate gene Association Resource (CARe). *Investigative Ophthalmology & Visual Science, 52*(10), 7593–7602. https://doi.org/10.1167/iovs.11-7614

[20] Wu, L., Fernandez-Loaiza, P., Sauma, J., Hernandez-Bogantes, E., & Masis, M. (2013). Classification of diabetic retinopathy and diabetic macular edema. *World Journal of Diabetes, 4*(6), 290–294. https://doi.org/10.4239/wjd.v4.i6.290

[21] Solomon, S. D., Chew, E., Duh, E. J., Sobrin, L., Sun, J. K., VanderBeek, B. L., Wykoff, C. C., & Gardner, T. W. (2017). Diabetic retinopathy: A position statement by the American Diabetes Association. *Diabetes Care, 40*(3), 412–418. https://doi.org/10.2337/dc16-2641

[22] Zhang, X., Low, S., Kumari, N., Wang, J., Ang, K., Yeo, D., Yip, C. C., Tavintharan, S., Sum, C. F., & Lim, S. C. (2017). Direct medical cost associated with diabetic retinopathy severity in type 2 diabetes in Singapore. *PLOS ONE, 12*(7), e0180949. https://doi.org/10.1371/journal.pone.0180949

[23] Yau, J. W. Y., Rogers, S. L., Kawasaki, R., Lamoureux, E. L., Kowalski, J. W., Bek, T., ... & Wong, T. Y. (2012). Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care, 35*(3), 556–564. https://doi.org/10.2337/dc11-1909

[24] Lee, R., Wong, T. Y., & Sabanayagam, C. (2015). Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye and Vision*, *2*, 17. https://doi.org/10.1186/s40662-015-0026-2

[25] Early Treatment Diabetic Retinopathy Study Research Group. (1991). Early photocoagulation for diabetic retinopathy: ETDRS report number 9. *Ophthalmology*, *98*(5 Suppl), 766–785. https://doi.org/10.1016/S0161-6420(13)38011-7

[26] Cheung, N., Mitchell, P., & Wong, T. Y. (2010). Diabetic retinopathy. *The Lancet*, *376*(9735), 124–136. https://doi.org/10.1016/S0140-6736(09)62124-3

[27] Eye7 Eye Hospital. (n.d.). *Diabetic retinopathy – Stages, risks & best treatment in Delhi*. Retrieved May 19, 2025, from https://www.eye7.in/retina/diabetic-retinopathy/

[28] Thind Eye Hospital. (n.d.). *Stages of diabetic retinopathy*. https://www.thindeyehospital.org/diabetic-retinopathy/

[29] Stanga, P. E., Boyd, S., & Hamilton, A. M. (2003). Diabetic retinopathy: A review. *Diabetic Medicine*, *20*(4), 247–257. https://doi.org/10.1046/j.1464-5491.2003.00983.x

[30] Schmidt-Erfurth, U., Garcia-Arumi, J., Bandello, F., Berg, K., Chakravarthy, U., Gerendas, B. S., Jonas, J., Larsen, M., Tadayoni, R., & Loewenstein, A. (2017). Guidelines for the management of diabetic macular edema by the European Society of Retina Specialists (EURETINA). *Ophthalmologica*, *237*(4), 185–222. https://doi.org/10.1159/000458539

[31] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, *316*(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

[32] Ipp, E., Liljenquist, D., Bode, B., et al. (2021). Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. *JAMA Network Open*, *4*(11), e2134254. https://doi.org/10.1001/jamanetworkopen.2021.34254

[33] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005

[34] Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., López García, Á., Heredia, I., Malík, P., & Hluchý, L. (2019). Machine learning and deep learning frameworks and libraries for large-scale data mining: A survey. *Artificial Intelligence Review*, *52*(1), 77–124. https://doi.org/10.1007/s10462-018-09679-z

[35] IDM Magazine. (n.d.). What is deep learning and how is it different from machine learning. https://idm.net.au/article/0012488-what-deep-learning-and-how-it-different-machine-learning

[36] Ibrahim, I., & Abdulazeez, A. (2021). The role of machine learning algorithms for diagnosing diseases. *Journal of Applied Science and Technology Trends*, *2*(1), 10–19. https://doi.org/10.38094/jastt20179

[37] Edureka. (n.d.). What is machine learning? | Types of machine learning. Retrieved May 21, 2025, from https://www.edureka.co/blog/what-is-machine-learning/

[38] Abnoohani. (n.d.). Machine learning types: Supervised vs unsupervised vs reinforcement in a glance [Vector illustration]. *CartoonDealer*. https://cartoondealer.com/image/297519663/machine-learning-types-supervised-vs-unsupervised-vs-reinforcement-glance-vector-editable-stroke-colors.html

[39] Tümer, C. (2018, August 26). Machine learning basics with examples — Part 2: Supervised learning. *Medium*. https://medium.com/@canburaktumer/machine-learning-basics-with-examples-part-2-supervised-learning-e2b740ff014c

[40] Balodi, T. (2019, September 6). Introduction to machine learning: Supervised, unsupervised and reinforcement learning. *Analytics Steps*. https://www.analyticssteps.com/blogs/introduction-machine-learning-supervised-and-unsupervised-learning

[41] Tzorakoleftherakis, E. (2019, October). Three things to know about reinforcement learning. *KDnuggets*. https://www.kdnuggets.com/2019/10/mathworks-reinforcement-learning.html

[42] Holdsworth, J., & Scapicchio, M. (2024, June 17). What is deep learning? *IBM*. https://www.ibm.com/think/topics/deep-learning

[43] Tsiknakis, N., Theodoropoulos, D., Manikis, G., Ktistakis, E., Boutsora, O., Berto, A., Scarpa, F., Scarpa, A., Fotiadis, D. I., & Marias, K. (2021). Deep learning for diabetic retinopathy detection and classification based on fundus images: A review. *Computers in Biology and Medicine*, *135*, 104599. https://doi.org/10.1016/j.compbiomed.2021.104599

[44] Gupta, M. (2025, January 27). Difference between AI vs ML vs DL. *Applied AI Course*. https://www.appliedaicourse.com/blog/ai-vs-ml-vs-dl/

[45] Li, M., Jiang, Y., Zhang, Y., & Zhu, H. (2023). Medical image analysis using deep learning algorithms. *Frontiers in Public Health*, *11*, 1273253. https://doi.org/10.3389/fpubh.2023.1273253

[46] DAGsHub. (n.d.). Convolutional neural network. https://dagshub.com/glossary/convolutional-neural-network/

[47] Engati. (n.d.). Convolutional neural network. https://www.engati.com/glossary/convolutional-neural-network

[48] Kumar, R., & Singh, A. (2024). A review on convolutional neural network. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(3), 123–130.

https://www.ijisae.org/index.php/IJISAE/article/view/5732/4475

[49] Akhtar, S., Aftab, S., Ali, O., Ahmad, M., Khan, M. A., Abbas, S., & Ghazal, T. M. (2025). A deep learning based model for diabetic retinopathy grading. *Scientific Reports*, *15*, Article 3763. https://doi.org/10.1038/s41598-025-87171-9

[50] Das, D., Biswas, S. K., & Bandyopadhyay, S. (2023). Detection of diabetic retinopathy using convolutional neural networks for feature extraction and classification (DRFEC). *Multimedia Tools and Applications*, *82*(19), 29943–30001. https://doi.org/10.1007/s11042-022-14165-4

[51] Ibrahim, R., & Shafiq, M. O. (2023). Explainable convolutional neural networks: A taxonomy, review, and future directions. *ACM Computing Surveys*, *55*(10), Article 206. https://doi.org/10.1145/3563691

[52] Chetoui, M., & Akhloufi, M. A. (2020). Explainable end-to-end deep learning for diabetic retinopathy detection across multiple datasets. *Journal of Medical Imaging*, *7*(4), 044503. https://doi.org/10.1117/1.JMI.7.4.044503

[53] Pawar, P. (2025, April 30). The role of explainable AI in healthcare. *Dr. D. Y. Patil School of Science and Technology*. https://dypsst.dpu.edu.in/blogs/explainable-ai-healthcare-role

[54] Ennab, M., & Mcheick, H. (2025). Advancing AI interpretability in medical imaging: A comparative analysis of pixel-level interpretability and Grad-CAM models. *Machine Learning and Knowledge Extraction*, *7*(1), 12. https://doi.org/10.3390/make7010012

[55] Raghuvanshi, S. (2024, February 1). Machine explainability: A guide to LIME, SHAP, and Gradcam. *Medium*. https://suryansh-raghuvanshi.medium.com/machine-explainability-a-guide-to-lime-shap-and-gradcam-60f6265f365f

[56] Said, Z., Ben-Bouazza, F.-E., & Mekkour, M. (2024). Towards interpretable diabetic retinopathy detection: Combining multi-CNN models with Grad-CAM. *International Journal of Advanced Computer Science and Applications*, *15*(10), 1088–1095. https://doi.org/10.14569/IJACSA.2024.01510111

[57] Marapelli, B., Carie, C. A., Ashish, Lal, B., & Bhaskar, K. A. (2025). Exploring explainable artificial intelligence techniques for diabetic retinopathy detection using Grad-CAM and VGG16 framework. *Journal of Neonatal Surgery*, *14*(23S). https://www.jneonatalsurg.com/index.php/jns/article/view/5749

[58] Alqutayfi, A., Almattar, W., Al-Azani, S., Khan, F. A., Al Qahtani, A., Alageel, S., & Alzahrani, M. (2025). Explainable disease classification: Exploring Grad-CAM analysis of CNNs and ViTs. *Journal of Advances in Information Technology*, *16*(2), 264–273. https://doi.org/10.12720/jait.16.2.264-273

[59] Van Craenendonck, T., Elen, B., Gerrits, N., & De Boever, P. (2020). Systematic comparison of heatmapping techniques in deep learning in the context of diabetic retinopathy

lesion detection. *Translational Vision Science & Technology, 9*(2), Article 64. https://doi.org/10.1167/tvst.9.2.64

[60] Nakayama, L. F., Zago Ribeiro, L., Restrepo, D., Santos Barboza, N., Dias Fiterman, R., Vieira Sousa, M. L., Pereira, A. D. A., Regatieri, C. V. S., Malerbi, F. K., & Andrade, R. E. (2024). *mBRSET, a Mobile Brazilian Retinal Dataset (version 1.0)* [Data set]. PhysioNet. https://doi.org/10.13026/qxpd-1y65

[61] Wu, C., Restrepo, D., Nakayama, L. F., Zago Ribeiro, L., Shuai, Z., Santos Barboza, N., Dias Fiterman, R., Vieira Sousa, M. L., Malerbi, F. K., & Andrade, R. E. (2025). A portable retina fundus photos dataset for clinical, demographic, and diabetic retinopathy prediction. *Scientific Data, 12*, Article 46. https://doi.org/10.1038/s41597-025-04627-3

[62] Arora, L., Singh, S. K., Kumar, S., Gupta, H., Alhalabi, W., Arya, V., Bansal, S., Chui, K. T., & Gupta, B. B. (2024). Ensemble deep learning and EfficientNet for accurate diagnosis of diabetic retinopathy. Scientific Reports, 14(1), Article 30554. https://doi.org/10.1038/s41598-024-81132-4

[63] Netguru. (n.d.). *Overfitting: Artificial intelligence explained*. Netguru Glossary. Retrieved June 12, 2025, from https://www.netguru.com/glossary/overfitting-artificial-intelligence-explained

[64] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

[65] Alghamdi, H. S. (2022). Towards explainable deep neural networks for the automatic detection of diabetic retinopathy. Applied Sciences, 12(19), Article 9435. https://doi.org/10.3390/app12199435

[66] Feng, M., Cai, Y., & Yan, S. (2025). Enhanced ResNet50 for diabetic retinopathy classification: External attention and modified residual branch. *Mathematics, 13*(10), Article 1557. https://doi.org/10.3390/math13101557

[67] Akhtar, S., Aftab, S., Ali, O., Ahmad, M., Khan, M. A., Abbas, S., & Ghazal, T. M. (2025). A deep learning based model for diabetic retinopathy grading. Scientific Reports, 15(1), Article 3763. https://doi.org/10.1038/s41598-025-87171-9

[68] U.S. Food and Drug Administration. (2018). *De Novo classification request for IDx-DR: FDA review memorandum* (DEN180001). https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf