Thesis submitted to the

UNIVERSITY OF MOHAMED BOUDIAF – MSILA

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
DEPARTMENT OF COMPUTER SCIENCE

In partial fulfillment of the requirements for the degree of

Master in **Computer science**

By

DIB , DHIAEDDINE

Title of the thesis

AN EXPLAINABLE DEEP RULE BASED FOR DIABETIC RETINOPATHY CLASSIFICATION

Under the supervision of

MOHAMED CHATRA

Composition of the jury

Dr. Tahar Mehenni	University of Msila	President
Dr. Mohamed Chatra	University of Msila	Reporter
Dr. Meftah Lakhal	University of Msila	Examiner

June, 2026

Abstract:

Diabetes is a chronic disease that affects millions of people worldwide. Among its most serious complications is Diabetic Retinopathy (DR), an eye condition that can lead to vision loss and even blindness in the absence of early diagnosis. Traditional screening methods rely on the manual examination of fundus images, a procedure that is often time-consuming and prone to human error. These complications have encouraged researchers to develop automated diagnostic systems based on Artificial Intelligence and Deep Learning. In this thesis, we propose an automated Diabetic Retinopathy detection system based on the EfficientNet-B3 architecture and transfer learning. The model was trained on a combined EyePACS-APTOS-Messidor dataset containing 53,411 balanced fundus images. Advanced preprocessing techniques, such as image resizing to 288×288 pixels, normalization using ImageNet statistics, and data augmentation, were applied. The obtained results show excellent performance with a Quadratic Weighted Kappa score of 0.9134 and an overall accuracy of 85.01%, demonstrating the effectiveness of the proposed approach for early clinical diagnosis.

Keywords: Diabetic Retinopathy, Deep Learning, EfficientNet-B3, Transfer Learning, Medical Image Classification.

الملخص:

يُعد داء السكري مرضًا مزمنًا يصيب ملايين الأشخاص حول العالم .ومن بين أخطر مضاعفاته اعتلال الشبكية السكري، وهو مرض يصيب العين وقد يؤدي إلى فقدان البصر أو حتى العمى في حال غياب التشخيص المبكر .وتعتمد طرق الكشف التقليدية على الفحص اليدوي لصور قاع العين، وهي عملية غالبًا ما تكون طويلة ومعرضة للأخطاء البشرية . وقد شجّعت هذه المضاعفات الباحثين على تطوير أنظمة تشخيص آلية تعتمد على الذكاء الاصطناعي والتعلم العميق .في هذه الأطروحة، نقترح نظامًا آليًا للكشف عن اعتلال الشبكية السكري يعتمد على بنية EfficientNet-B3وتقنية التعلم بالنقل .تم تدريب النموذج على قاعدة بيانات مدمجة تضم 53,411صورة متوازنة لقاع العين من مجموعات EyePACSو APTOSو .Messidorكما تم تطبيق تقنيات متقدمة للمعالجة المسبقة، مثل إعادة تحجيم الصور إلى 288×288بكسل، والتطبيع باستخدام إحصائيات ImageNet، وزيادة البيانات .أظهرت النتائج المحققة أداءً ممتازًا، حيث بلغ معامل Quadratic Weighted Kappaقيمة 0.9134وبلغت الدقة الإجمالية 85.01%، مما يبرهن على فعالية النهج المقترح في التشخيص السريري المبكر.

الكلمات المفتاحية :اعتلال الشبكية السكري، التعلم العميق، EfficientNet-B3، التعلم بالنقل، تصنيف الصور الطبية.

Résumé:

Le diabète est une maladie chronique qui touche des millions de personnes dans le monde. Parmi ses complications les plus graves figure la rétinopathie diabétique (RD), une affection oculaire pouvant entraîner une perte de vision, voire la cécité, en l'absence d'un diagnostic précoce. Les méthodes traditionnelles de dépistage reposent sur l'examen manuel des images du fond d'œil, une procédure souvent longue et sujette aux erreurs humaines. Ces complications ont encouragé les chercheurs à développer des systèmes automatisés de diagnostic fondés sur l'intelligence artificielle et l'apprentissage profond. Dans cette thèse, nous proposons un système automatisé de détection de la rétinopathie diabétique basé sur l'architecture EfficientNet-B3 et l'apprentissage par transfert. Le modèle a été entraîné sur une base de données combinée EyePACS-APTOS-Messidor contenant 53 411 images de fond d'œil équilibrées. Des techniques avancées de prétraitement, telles que le redimensionnement à 288×288 pixels, la normalisation selon les statistiques d'ImageNet et l'augmentation de données, ont été appliquées. Les résultats obtenus montrent une excellente performance avec un score Quadratic Weighted Kappa de 0,9134 et une précision globale de 85,01 %, démontrant l'efficacité de l'approche proposée pour le diagnostic clinique précoce. Mots-clés : Rétinopathie diabétique, Apprentissage profond, EfficientNet-B3, Apprentissage par transfert, Classification d'images médicales.

Dedications

″ **To my mother**, the eternal source of tenderness and grace, whose continuous prayers have paved my path to success, and whose unconditional love and sacrifices have been my ultimate sanctuary throughout the hardest times. **To my father**, the symbol of strength, honor, and boundless dedication, who spent his life guiding my steps, supporting my ambitions, and teaching me the true meaning of perseverance, patience, and dignity. To you both, my beloved parents, I dedicate this academic work as a small token of my eternal gratitude, love, and respect; you are my greatest pride. To my dear sister, **Sadjida**, and my beloved brothers, **Ayoub and Bara**, the irreplaceable pillars of my life and my constant source of joy, thank you for your unwavering support, your continuous encouragement, and for always standing firmly by my side through every phase of this journey. Last but certainly not least, to my loyal and closest friend, **Mouhoun Abderaaouf**, who has always been much more than a friend—a true brother in every sense of the word—thank you for your sincere brotherhood, your absolute belief in my potential, and for walking through the challenges of life and study right beside me. This achievement is a reflection of your profound presence in my life.″

DIB

Acknowledgments

We are deeply indebted to our thesis supervisor **Chatra Mohamed** whose unlimited steadfast support and inspirations have made this project a great success. In a very special way, we thank him for every support he has rendered unto us to see that we succeed in this challenging study.

Special thanks go to our friends and families who have contained the hectic moments and stress we have been through during the course of the research project.

We thank the school for giving us the grand opportunity to work as a team which has indeed promoted our team work spirit and communication skills.

I would like to thank myself for believing in me, for doing all the hard work, for staying dedicated throughout this journey, for never giving up despite the challenges, and for always striving to achieve my goals. This accomplishment reflects the perseverance, determination, and continuous effort that made this achievement possible.

Contents

List of figures

List of tables

General Introduction

In recent years, artificial intelligence has become an important technology in the healthcare sector, especially in medical image analysis. Deep learning techniques have shown strong ability to analyze complex medical images such as retinal fundus photographs, X-rays, computed tomography scans, and magnetic resonance images. These techniques can support clinicians by detecting visual patterns that may be difficult or time-consuming to identify manually.

Diabetic retinopathy is one of the most serious complications of diabetes mellitus and one of the leading causes of preventable blindness worldwide. The disease affects the retinal blood vessels and progresses gradually, often without noticeable symptoms in its early stages. As the number of diabetic patients continues to increase, regular screening becomes essential for early detection and prevention of irreversible vision loss.

Traditional diabetic retinopathy screening relies mainly on manual examination of retinal fundus images by ophthalmologists or trained graders. Although this approach remains clinically important, it presents several limitations, including time consumption, subjectivity, fatigue, inter-observer variability, and limited availability of specialists. These limitations make large-scale screening difficult, especially in regions with limited medical resources.

In this context, deep learning provides a promising solution for automatic diabetic retinopathy classification. Convolutional neural networks can automatically extract meaningful features from retinal images and classify disease severity levels. This thesis proposes an intelligent system based on EfficientNet-B3 and transfer learning to classify diabetic retinopathy into five severity levels: No DR, Mild DR, Moderate DR, Severe DR, and Proliferative DR.

The proposed system is trained and evaluated using a combined EyePACS-APTOS-Messidor dataset, which provides a diverse collection of retinal fundus images from different sources. The methodology includes image preprocessing, data augmentation, class imbalance handling, model training, and performance evaluation using accuracy, precision, recall, F1-score, confusion matrix, and Quadratic Weighted Kappa. In addition, the project includes a comparative study with other CNN architectures and considers model interpretability using Grad-CAM.

Beyond model development, this work also aims to make the system more practical through a web application prototype. The application allows users to upload retinal images and obtain prediction results through a simple interface. It can also be accessed locally from another device, such as a smartphone, using the host machine IP address on the same network, which makes it suitable for demonstration and local testing.

This thesis is organized as follows. Chapter 1 presents the background of diabetic retinopathy, traditional diagnosis, artificial intelligence, deep learning, related works, and the research gap. Chapter 2 describes the proposed methodology, including dataset preparation, preprocessing, augmentation, model architecture, and training strategy. Chapter 3 presents the experimental results, performance analysis, and comparative evaluation. Chapter 4 describes the web application development and local deployment scenario. Finally, the thesis ends with a general conclusion and possible future improvements.

Chapter 1 Artificial Intelligence & Diabetic Retinopathy

Introduction

In the past decade, the integration of artificial intelligence (AI) into the medical domain has significantly transformed diagnostic processes, particularly in medical image analysis. Medical imaging techniques, including fundus photography, magnetic resonance imaging (MRI), computed tomography (CT), and X-rays, generate large amounts of visual data that require accurate expert interpretation. However, the increasing number of patients, combined with the shortage of specialized physicians, has created an urgent need for automated, accurate, and scalable diagnostic support systems.

Diabetic retinopathy (DR) is one of the most common and severe complications of diabetes mellitus and remains a leading cause of preventable blindness among working-age adults worldwide. It affects the small blood vessels of the retina and progresses gradually, often without noticeable symptoms in the early stages. As a result, early detection and regular screening are essential to prevent irreversible vision loss.

Traditional DR screening relies mainly on manual examination of retinal fundus images by ophthalmologists or trained graders. Although manual diagnosis remains clinically important, it suffers from several limitations, including time consumption, subjectivity, fatigue, inter-observer variability, and limited availability of specialists, especially in regions with restricted access to ophthalmological care. These limitations motivate the development of automated and intelligent diagnostic systems.

Deep learning, especially convolutional neural networks (CNNs), has shown strong potential for automatic diabetic retinopathy classification. CNN-based models can automatically extract hierarchical visual features from retinal images and classify the disease according to its severity levels. Transfer learning further improves performance by allowing models pre-trained on large image datasets to be adapted to medical image classification tasks.

This chapter provides the theoretical and clinical background necessary for understanding the proposed diabetic retinopathy classification system. It presents the definition, causes, and severity stages of DR; discusses traditional diagnostic methods and their limitations; introduces key concepts of artificial intelligence, deep learning, convolutional neural networks, transfer learning, and explainable AI; reviews modern

CNN architectures and related works; identifies the main research gaps; and defines the objectives of this thesis.

Overview of Diabetic Retinopathy (DR)

1.1.1 Definition and Etiology

Diabetic retinopathy is a microvascular complication of diabetes mellitus resulting from chronic hyperglycemia (elevated blood glucose levels). Prolonged exposure to high blood sugar damages the small blood vessels (capillaries) that supply the retina — the light-sensitive tissue at the back of the eye responsible for vision. The damaged vessels may leak fluid, blood, or lipids (exudates), leading to retinal tissue ischemia (lack of oxygen) and subsequent pathological changes **[1]**.

Table 1: The primary causes of DR include:

Cause	Explanation
Chronic hyperglycemia	Persistently high blood glucose levels weaken the vascular endothelium.
Duration of diabetes	The longer a person has diabetes, the higher the risk. After 20 years, nearly 90% of type 1 and 60% of type 2 patients show DR **[2]**.
Dyslipidemia	Abnormal lipid levels accelerate vascular damage.
Hypertension	High blood pressure increases mechanical stress on retinal vessels.
Pregnancy	Pregnant women with diabetes have a higher risk of DR progression.

1.1.2 Stages of Disease Progression

The international clinical classification system for diabetic retinopathy defines five stages based on fundus photography findings [3]. Each stage corresponds to specific observable lesions and indicates the severity of the disease.

- **Stage 0: No Apparent Retinopathy (Healthy)**

  No visible retinal abnormalities. Blood vessels appear normal in caliber, color, and trajectory. The optic disc and macula are healthy. The patient has diabetes but no DR signs.

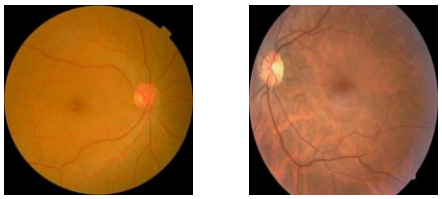

Figure 1: Healthy retina (Stage 0 - No DR)

- **Stage 1: Mild Non-Proliferative Diabetic Retinopathy (NPDR)**

Presence of microaneurysms, which are small, dot-like outpouchings of retinal capillaries. Microaneurysms appear as small red dots measuring 25 to 100 microns scattered in the retina. These are the earliest clinically detectable sign of DR. At this stage, vision is usually unaffected, making automated detection crucial for early intervention.

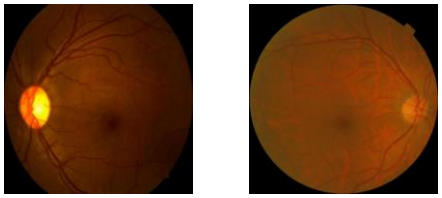

Figure 2: Mild Non-Proliferative DR (Stage 1)

- **Stage 2: Moderate Non-Proliferative Diabetic Retinopathy (NPDR)**

More extensive microaneurysms than in mild NPDR. Additional signs include hemorrhages (small dot-blot or flame-shaped bleeding due to ruptured capillaries), hard exudates (yellow-white lipid deposits with sharp margins indicating lipid leakage), and cotton-wool spots (soft, fluffy white patches representing nerve fiber layer infarction). The patient may still be asymptomatic, but retinal damage is progressing.

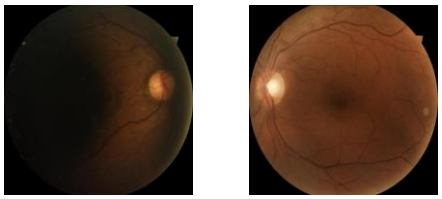

Figure 3: Moderate Non-Proliferative DR (Stage 2)

- **Stage 3: Severe Non-Proliferative Diabetic Retinopathy (NPDR)**

The "4-2-1 rule" is applied. Extensive hemorrhages and microaneurysms in all four quadrants, OR venous beading (irregular vein caliber) in two or more quadrants, OR intraretinal microvascular abnormalities (IRMA) in one or more quadrants. No neovascularization yet. High risk of progression to proliferative DR within 12 months, approximately 50 to 75 percent.

Figure 4: Severe Non-Proliferative DR (Stage 3)

- **Stage 4: Proliferative Diabetic Retinopathy (PDR)**

The most advanced and sight-threatening stage. Neovascularization, which is the growth of abnormal, fragile new blood vessels on the optic disc (NVD) or elsewhere in the retina (NVE). These new vessels are weak and can rupture easily, leading to vitreous hemorrhage (blood leaking into the vitreous cavity), tractional retinal detachment (scar tissue pulling the retina away), and neovascular glaucoma (severe complication causing high intraocular pressure). Additionally, diabetic macular edema (DME) may occur at any stage, characterized by fluid accumulation in the macula, the central vision area **[4]**.

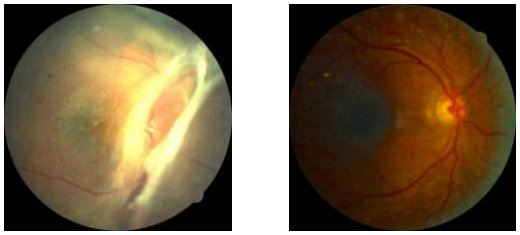

Figure 5: Proliferative DR (Stage 4)

1.1.3 Importance of Early Automated Diagnosis

Early diagnosis of DR is crucial for several reasons.

First, prevention of blindness: Treatment such as laser photocoagulation and anti-VEGF injections is most effective when applied at early stages (NPDR). Once PDR or DME causes irreversible damage, vision loss is permanent [5] .

Second, asymptomatic nature: Patients often have no visual symptoms until the disease is advanced. Regular screening is necessary, but manual screening by ophthalmologists is not feasible for large populations.

Third, global burden: According to the International Diabetes Federation (IDF), approximately 537 million adults had diabetes in 2021, and this number is expected to rise to 783 million by 2045 [6]. Approximately one-third of diabetic patients have DR [7].

Fourth, limitations of manual screening : (discussed in detail in section 1.3) create a strong motivation for automated, AI-based diagnostic systems that can screen large populations quickly, accurately, and cost-effectively [8].

Traditional Diagnostic Methods (Manual Detection)

1.1.4 Manual Examination and Fundus Ophthalmoscopy

The traditional gold standard for DR diagnosis is a comprehensive dilated eye examination performed by an ophthalmologist or trained retina specialist. The main manual methods include:

- **Direct Ophthalmoscopy**: A handheld device allowing the examiner to view the retina through the pupil. Limited field of view and requires significant skill.
- **Indirect Ophthalmoscopy**: Provides a wider field of view using a head-mounted light source and a handheld condensing lens. This is more informative but requires mydriatic drops to dilate the pupil.
- **Fundus Photography with Manual Grading**: High-resolution color fundus images are captured, and a trained human grader (ophthalmologist or certified reader) examines each image to identify lesions (microaneurysms, hemorrhages, exudates, neovascularization) and assign a DR severity level according to standardized scales **[3] .**

1.1.5 Limitations and Problems of Human Diagnosis

Despite being the current reference standard, manual DR diagnosis suffers from several critical disadvantages [8]:

Table 2: Main limitations of traditional manual diagnosis

Limitation	Description
Subjectivity and inter-observer variability	Different ophthalmologists may assign different severity grades to the same fundus image. Studies show moderate to fair agreement (kappa = 0.4–0.6) even among experts.
Intra-observer variability	The same clinician may give different grades at different times due to fatigue or inconsistency.
Human fatigue	Reading hundreds of fundus images per day leads to decreased diagnostic accuracy, especially towards the end of a session.
Time consumption	Manual grading of a single fundus image can take 5–15 minutes, which is unsustainable for large-scale screening programs.
Shortage of specialists	In many low- and middle-income countries (e.g., parts of Africa, Asia, Latin America), the ratio of ophthalmologists to diabetic patients is extremely low (sometimes 1 per 1 million people).

Cost	Manual examination requires expensive equipment and highly trained personnel, making frequent screening unaffordable.
Delayed diagnosis	Patients often wait weeks or months for an appointment, allowing disease progression.

1.1.6 The Need for a Shift Toward Automated Diagnosis

Given the limitations above, there is an urgent global need for **automated, AI-driven diagnostic systems** that can:

- Process thousands of fundus images in minutes rather than hours.
- Provide consistent, reproducible, and objective grading (no variability).
- Operate in primary care settings without requiring an ophthalmologist on-site.
- Enable **telemedicine** and large-scale population screening.
- Reduce healthcare costs while improving early detection rates.

This need has driven the rapid adoption of **artificial intelligence**, particularly **deep learning**, in medical image analysis [9].

Artificial Intelligence Techniques in the Medical Field

1.1.7 Definition of Artificial Intelligence

Artificial Intelligence (AI) is a branch of computer science concerned with building systems capable of performing tasks that typically require human intelligence. These tasks include learning from data, recognizing patterns, understanding language, solving problems, and making decisions.

AI can be divided into narrow AI and general AI. Narrow AI is designed to perform specific tasks, such as image classification, speech recognition, or medical image analysis. General AI, on the other hand, refers to a hypothetical system capable of performing any intellectual task that a human can do. Most current AI applications, including those used in medical imaging, belong to narrow AI.

In practice, AI mainly relies on machine learning (ML) and deep learning (DL), where algorithms learn patterns directly from data instead of following explicitly programmed rules. In healthcare, AI enables the automatic analysis of medical images, such as fundus photographs, X-rays, and CT scans, in order to detect diseases, classify severity, and support clinical decision-making.

1.1.8 Difference Between Machine Learning and Deep Learning

Machine Learning (ML) is a subfield of AI that enables computers to learn patterns from data without being explicitly programmed. Traditional ML requires **handcrafted features** (e.g., texture, shape, color histograms) extracted by human experts, which are then fed into classifiers such as Support Vector Machines (SVM), Random Forests, or k-Nearest Neighbors (k-NN) [10].

Deep Learning (DL) is a subfield of ML inspired by the structure and function of the human brain (neural networks) [11]. Unlike traditional ML, deep learning models automatically learn hierarchical feature representations directly from raw data (e.g., pixel values of images). This eliminates the need for manual feature engineering.

Table 3: Comparison between traditional Machine Learning and Deep Learning

Aspect	Traditional ML	Deep Learning
Feature extraction	Manual, handcrafted	Automatic, learned from data
Performance with large data	Plateaus after a certain point	Improves continuously with more data
Interpretability	Relatively high	Low ("black box")
Computational requirements	Moderate	High (GPUs required)
Medical image analysis	Limited success due to variability in lesions	State-of-the-art results, often surpassing human experts

For DR detection, deep learning has proven vastly superior because retinal lesions are highly variable in size, shape, color, and location — making manual feature engineering impractical.

1.1.9 Fundamentals of Convolutional Neural Networks (CNN)

Convolutional neural networks (CNNs) are the most successful deep learning architecture for image analysis tasks, including classification, segmentation, and detection [12]. A standard CNN consists of several types of layers:

a) **Convolutional Layer**:
   - Applies a set of learnable filters (kernels) to the input image.
   - Each filter detects specific features such as edges, corners, textures, or, in deeper layers, complex structures like microaneurysms or exudates.
   - Produces feature maps.

b) **Activation Function (usually ReLU - Rectified Linear Unit)** :
   - Introduces non-linearity ($f(x) = \max(0, x)$).
   - Allows the network to learn complex patterns.

c) **Pooling Layer (Max or Average Pooling)** :
   - Reduces spatial dimensions (downsampling), decreasing computational load and providing translation invariance.
   - Example: a 2×2 max pooling reduces a 4×4 region to a single maximum value.

d) **Fully Connected (Dense) Layer**:
   - Flattens the high-level features from the last convolutional block and performs the final classification (e.g., assigning the DR grade: 0, 1, 2, 3, or 4).

e) **Output Layer (Softmax Activation)** :
   - Converts the raw outputs (logits) into a probability distribution over the five DR classes.

A typical CNN architecture for DR classification (e.g., ResNet50) may have dozens of convolutional layers, millions of trainable parameters, and is trained using backpropagation and gradient descent [13] .

1.1.10Transfer Learning

Transfer learning is a powerful technique in deep learning where a model pre-trained on a large, general-purpose dataset (usually **ImageNet** — 1.2 million images of 1,000 everyday object categories) is reused (fine-tuned) for a specific target task (e.g., DR classification) with a smaller medical dataset [14] .

Why transfer learning is essential for medical imaging:

- **Limited medical data**: Large labeled medical datasets are rare due to privacy concerns, annotation costs, and the need for expert labeling.
- **Reduced training time**: Training a CNN from scratch on fundus images would take weeks even on powerful GPUs. Transfer learning reduces this to hours.
- **Better performance**: Pre-trained models have already learned general visual features (edges, textures, shapes), which are transferable to medical images.

Common transfer learning approaches:

1. **Feature extraction**: Keep the pre-trained convolutional base frozen, replace the fully connected head, and train only the new classifier.
2. **Fine-tuning**: Unfreeze some of the top layers of the pre-trained model and retrain them on the target dataset with a low learning rate.

Popular CNN architectures used as base models for transfer learning in DR detection include **VGG16, ResNet50, InceptionV3, DenseNet121**, and **EfficientNet** [15].

1.1.11Explainable AI (XAI) in Medical Diagnosis

Despite the strong performance achieved by deep learning models in medical image classification, one of their main limitations is the lack of interpretability. Many deep learning systems are considered black-box models because they can produce accurate predictions without clearly explaining how these decisions were made. In the medical field, this issue is particularly important because clinicians need to understand and trust the reasoning behind an AI-based diagnosis before using it as a decision-support tool.

Explainable Artificial Intelligence (XAI) refers to techniques that aim to make AI models more transparent and understandable. In medical diagnosis, XAI helps answer important questions such as which regions of an image influenced the model's prediction and why a specific class was selected. This is essential for increasing trust, detecting possible model errors, and supporting clinical interpretation.

One of the most commonly used XAI techniques in medical image analysis is Grad-CAM, which stands for Gradient-weighted Class Activation Mapping. Grad-CAM generates a heatmap that highlights the image regions that contributed most to the model's decision. In diabetic retinopathy classification, this can help visualize whether the model focused on clinically relevant retinal lesions such as microaneurysms, hemorrhages, exudates, or neovascularization.

Other explainability techniques, such as LIME and SHAP, can also be used to interpret machine learning models. However, Grad-CAM is particularly suitable for convolutional neural networks because it provides visual explanations directly on the medical image. In this thesis, Grad-CAM is used as an explainability tool to support the interpretation of the EfficientNet-B3 model's predictions and to improve the clinical trustworthiness of the proposed system.

1.1.12Challenges of AI in Medical Imaging

Although artificial intelligence has achieved remarkable progress in medical image analysis, several challenges still limit its full adoption in clinical practice. These challenges must be considered carefully because medical diagnosis requires high reliability, safety, and interpretability.

The first challenge is data availability and quality. Deep learning models require large amounts of annotated data to learn effectively. In the medical domain, collecting

labeled images is difficult because annotation requires expert knowledge, patient data must be protected, and some disease classes are naturally rare.

The second challenge is class imbalance. In diabetic retinopathy datasets, some classes appear more frequently than others. For example, No DR and Moderate DR images are usually more common than Severe or Proliferative DR images. If this imbalance is not handled properly, the model may become biased toward the majority classes and perform poorly on minority classes.

The third challenge is generalization. A deep learning model trained on images from one dataset may not perform equally well on images collected from another hospital, camera, or patient population. In medical imaging, acquisition conditions can vary significantly, including image resolution, illumination, field of view, and noise level. Therefore, a model must be trained and evaluated on diverse data to ensure that it can generalize to unseen clinical cases.

The fourth challenge is interpretability. Deep learning models often produce predictions without clearly explaining the reasons behind them. In clinical practice, this lack of transparency can reduce trust and limit adoption by medical professionals. For this reason, explainability techniques such as Grad-CAM are important because they can highlight the image regions that contributed most to the model's decision.

The fifth challenge is clinical validation and ethical concerns. Before AI-based diagnostic systems can be used in real clinical settings, they must be validated carefully on independent datasets and under real medical conditions. Ethical issues such as patient privacy, algorithmic bias, responsibility for incorrect predictions, and the need for human supervision must also be considered. AI systems should support clinicians rather than replace them, especially in high-risk medical decisions.

In this thesis, several strategies are adopted to address these challenges. Transfer learning is used to improve performance with limited medical data, data augmentation is applied to increase image variability, weighted sampling is used to reduce the effect of class imbalance, and Grad-CAM is considered to improve model interpretability. In addition, the use of a combined EyePACS-APTOS-Messidor dataset helps improve model generalization by exposing the model to retinal images collected from different sources.

Modern CNN Architectures for DR Classification

1.1.13Review of Major Models Used (ResNet, VGG, Inception)

Numerous deep learning models have been applied to diabetic retinopathy classification. Below are the most influential architectures:

Table 4: Comparison of popular CNN architectures for diabetic retinopathy classification

Model	Key Feature	Strengths for DR
VGG16 / VGG19	Simple, uniform architecture (3×3 convolutions stacked deeply).	Easy to implement; good performance but computationally heavy.
ResNet50 / ResNet101	Introduced skip connections (residual blocks) to solve the vanishing gradient problem, allowing extremely deep networks (up to 152 layers) [15]	State-of-the-art for DR detection; widely used in competitions.
InceptionV3 / InceptionResNetV2	Uses inception modules with multiple filter sizes (1×1, 3×3, 5×5) in parallel to capture multi-scale features.	Excellent for detecting lesions of varying sizes (microaneurysms vs. large exudates).
DenseNet121	Each layer connects to all subsequent layers (dense connectivity), improving gradient flow.	Very parameter-efficient; good for small datasets.
EfficientNet	Uses compound scaling of depth, width, and resolution.	Superior accuracy with fewer parameters; modern state-of-the-art.

Most published studies report that ResNet50 and InceptionV3 achieve the highest accuracy (typically 85–95%) for 5-class DR classification, depending on dataset size and preprocessing.

1.1.14Why EfficientNet-B3 is Suitable for Diabetic Retinopathy Classification

Among the CNN architectures used in medical image classification, EfficientNet-B3 was selected as the main backbone of the proposed diabetic retinopathy classification system. This choice is based on its strong balance between classification performance, computational efficiency, and suitability for transfer learning.

EfficientNet is based on a compound scaling strategy. Instead of increasing only the depth, width, or input resolution of the network, EfficientNet scales these three

dimensions in a balanced way. This allows the model to achieve high accuracy while using fewer parameters compared with many traditional CNN architectures.

EfficientNet-B3 is also more parameter-efficient than heavier models such as VGG16. This is important in medical image classification because large models can overfit when the dataset is limited or imbalanced. A model with fewer parameters can be easier to train, faster to fine-tune, and more practical for prototype deployment.

Another important advantage of EfficientNet-B3 is its ability to extract meaningful features from retinal fundus images. Diabetic retinopathy lesions vary in size and appearance. Some lesions, such as microaneurysms, are very small, while others, such as hemorrhages and exudates, are larger and more visible. EfficientNet-B3 can learn hierarchical visual features that help detect both fine details and broader retinal patterns.

EfficientNet-B3 is also compatible with transfer learning because pre-trained weights are available through libraries such as TIMM. This allows the model to benefit from features learned on large-scale image datasets, then adapt them to the diabetic retinopathy classification task. For these reasons, EfficientNet-B3 was chosen as the proposed model in this thesis.

Previous Studies and Research Gap

1.1.15 Recent Studies Based on Combined Diabetic Retinopathy Datasets

The Combined EyePACS-APTOS-Messidor Diabetic Retinopathy Dataset is a large-scale public benchmark created by unifying four major diabetic retinography datasets: the EyePACS Diabetic Retinopathy Detection dataset, the APTOS 2019 Blindness Detection dataset, an APTOS Gaussian filtered version, and the Messidor database [16]. This combined dataset contains a total of 92,501 original JPG images. After manual data augmentation applied in January 2024, the dataset size increased by approximately 55 percent, reaching 143,669 images. All images have been resized to 600 by 600 pixels to reduce disk size from 18.5 gigabytes to 3.8 gigabytes and to avoid repeated augmentation and resizing operations during model training.

The dataset is randomly split into three subsets: 80 percent for training, 10 percent for validation, and 10 percent for testing. The images represent real-world clinical conditions with varying image quality, lighting conditions, and camera types, as they originate from multiple sources including rural India (APTOS), the United States (EyePACS), and France (Messidor). The dataset maintains the standard five severity

classes: 0 for No DR, 1 for Mild DR, 2 for Moderate DR, 3 for Severe DR, and 4 for Proliferative DR.

Several notable studies have used this combined dataset or its original components. Karthik and colleagues in 2019 proposed an ensemble of ResNet50 and InceptionV3 with attention mechanisms on the APTOS 2019 dataset, achieving a quadratic weighted kappa of 0.94. Bodapati and colleagues in 2020 combined DenseNet121 with class activation mapping for explainable diabetic retinopathy grading, reaching 92 percent accuracy on APTOS 2019. Saxena and colleagues in 2021 used EfficientNetB3 with heavy data augmentation including rotation, flipping, and color jitter, and achieved a kappa score of 0.96 on APTOS 2019. More recent work has focused on training on the combined dataset to leverage its larger size and diversity.

Common challenges reported across studies using combined datasets include class imbalance, although the augmented version partially addresses this through balancing techniques, and image quality variability due to different fundus cameras from multiple clinical sites. Despite these challenges, deep learning models trained on larger combined datasets consistently outperform those trained on single-source datasets.

1.1.16Research Gap and Limitations of Prior Work

Although previous studies have shown that deep learning can achieve strong results in diabetic retinopathy classification, several limitations remain. These limitations represent the research gap that motivates the proposed work.

First, many previous studies were trained on single-source datasets such as APTOS 2019 or EyePACS. While these datasets are useful, they may not represent the full variability of real clinical images. Fundus images can differ depending on the camera, acquisition conditions, patient population, and image quality. Therefore, models trained on limited datasets may not generalize well to new data.

Second, class imbalance remains a major problem in diabetic retinopathy datasets. Some severity levels are underrepresented, especially advanced stages such as Severe DR and Proliferative DR. If this issue is not addressed, the model may learn to favor majority classes and fail to correctly detect rare but clinically important cases.

Third, many studies focus mainly on accuracy without giving enough attention to ordinal evaluation metrics. Diabetic retinopathy severity levels are ordered from No DR to Proliferative DR. Therefore, confusing No DR with Proliferative DR is more

serious than confusing Mild DR with Moderate DR. For this reason, Quadratic Weighted Kappa is an important metric for this task.

Fourth, interpretability is often limited in previous works. In clinical applications, it is not enough for a model to provide a prediction; it should also provide visual evidence that helps clinicians understand the decision. Explainability methods such as Grad-CAM are therefore important for improving trust.

Finally, many proposed models remain limited to experimental notebooks and are not integrated into practical applications. This limits their usability for doctors or non-technical users. To address these limitations, this thesis proposes a diabetic retinopathy classification system based on a combined EyePACS-APTOS-Messidor dataset, EfficientNet-B3 transfer learning, class imbalance handling, QWK-based evaluation, Grad-CAM interpretability, and a web application prototype for local testing.

Project Objectives: Developing an Intelligent System for DR Severity Classification

Based on the literature review and the identified research gaps, this project aims to develop an intelligent deep learning system for the automatic classification of diabetic retinopathy severity from retinal fundus images. The proposed system is designed not only to classify the disease into five severity levels, but also to support model interpretability and practical use through a web application prototype.

The main objective of this thesis is to design, implement, and evaluate a deep learning-based diabetic retinopathy classification system using a combined EyePACS-APTOS-Messidor dataset and an EfficientNet-B3 architecture.

The specific objectives of this work are:

First, to study diabetic retinopathy, its clinical severity stages, and the limitations of traditional manual diagnosis.

Second, to analyze the role of artificial intelligence, deep learning, convolutional neural networks, and transfer learning in medical image classification.

Third, to prepare a diverse retinal fundus image dataset by applying appropriate preprocessing, augmentation, and class imbalance handling techniques.

Fourth, to develop a five-class diabetic retinopathy classification model based on EfficientNet-B3 and transfer learning.

Fifth, to evaluate the proposed model using suitable classification metrics, including accuracy, precision, recall, F1-score, confusion matrix, and Quadratic Weighted Kappa.

Sixth, to compare the proposed EfficientNet-B3 approach with other common CNN architectures used in diabetic retinopathy classification.

Seventh, to improve model transparency by using visual explanation techniques such as Grad-CAM.

Eighth, to integrate the trained model into a web application prototype that allows users to upload fundus images and obtain prediction results through a simple interface.

Finally, to demonstrate the possibility of local deployment, where the application can be accessed from another device, such as a smartphone, using the host machine IP address on the same network.

Conclusion

In conclusion, this chapter has presented the fundamental background required to understand the motivation and scope of this research. Diabetic retinopathy was introduced as a progressive and vision-threatening complication of diabetes that requires early and accurate detection in order to prevent irreversible visual impairment. The chapter also discussed traditional diagnostic approaches based on manual examination and fundus image grading. Although these methods remain clinically important, they present several limitations, including subjectivity, inter-observer variability, fatigue, time consumption, and limited availability of ophthalmologists. These limitations justify the need for automated and intelligent diagnostic support systems.

Furthermore, the chapter introduced the role of artificial intelligence in medical image analysis, with particular focus on **machine learning, deep learning, convolutional neural networks, transfer learning, and explainable AI**. Modern CNN architectures such as VGG, ResNet, Inception, DenseNet, and EfficientNet were reviewed, and EfficientNet-B3 was identified as a suitable architecture for this project because of its balance between classification performance and computational efficiency.

The review of previous studies showed that deep learning approaches can achieve strong performance in diabetic retinopathy classification, especially when trained on large and diverse datasets. However, several research gaps remain, including dataset variability, class imbalance, limited interpretability, and lack of practical deployment. These gaps motivated the proposed system, which combines EfficientNet-B3 transfer

learning, preprocessing, augmentation, class imbalance handling, Grad-CAM interpretability, and a web application prototype.

Finally, the objectives of this thesis were defined, focusing on the development of a five-class diabetic retinopathy classification system supported by robust evaluation metrics such as **accuracy, precision, recall, F1-score, confusion matrix, and Quadratic Weighted Kappa.** The following chapter presents the proposed methodology, including dataset preparation, preprocessing, augmentation, model architecture, training strategy, and evaluation protocol.

Chapter 2 Methodology and Implementation of the Intelligent DR Classification System

2.1 Introduction

After establishing the theoretical foundations of diabetic retinopathy, artificial intelligence, and deep learning in Chapter 1, this chapter moves from theory to practice by presenting the complete design and methodology of the intelligent classification system developed in this thesis.

The main problem addressed is that most existing diabetic retinopathy classification models are trained on single-source datasets with limited diversity, which leads to poor generalization when applied to images from different cameras, populations, or clinical settings. Furthermore, the ordinal nature of DR severity, where misclassifying severe DR as mild is far more harmful than misclassifying mild as moderate, is often not properly handled.

To address these limitations, this chapter describes a robust deep learning system based on the EfficientNet-B3 architecture with transfer learning from ImageNet pre-trained weights. The system is trained and validated on a large-scale combined dataset containing 53,411 fundus images after balancing.

The main contributions of this methodology are using a multi-source dataset to improve generalization across different clinical settings, handling class imbalance through a weighted random sampler, adopting quadratic weighted kappa as the primary evaluation metric to account for the ordinal nature of DR grading, and applying extensive data augmentation and regularization techniques to reduce overfitting.

This chapter is organized into ten main sections. Section 2.2 describes the dataset. Section 2.3 covers preprocessing and augmentation. Section 2.4 presents the model architecture. Section 2.5 details the training strategy. Section 2.6 explains class imbalance handling. Section 2.7 defines the evaluation metrics. Section 2.8 describes the experimental environment and implementation details. Section 2.9 presents the experimental results. Section 2.10 summarizes the chapter.

2.2 Global System Overview

The proposed system follows a structured deep learning pipeline for diabetic retinopathy severity classification. The process begins with a retinal fundus image as input. The image is loaded, converted into RGB format, resized, normalized, and transformed into a tensor. During training, data augmentation is applied to increase

image variability and reduce overfitting. The processed image is then passed to the EfficientNet-B3 model, which predicts one of the five diabetic retinopathy severity classes.

The general workflow of the proposed system can be summarized as follows:

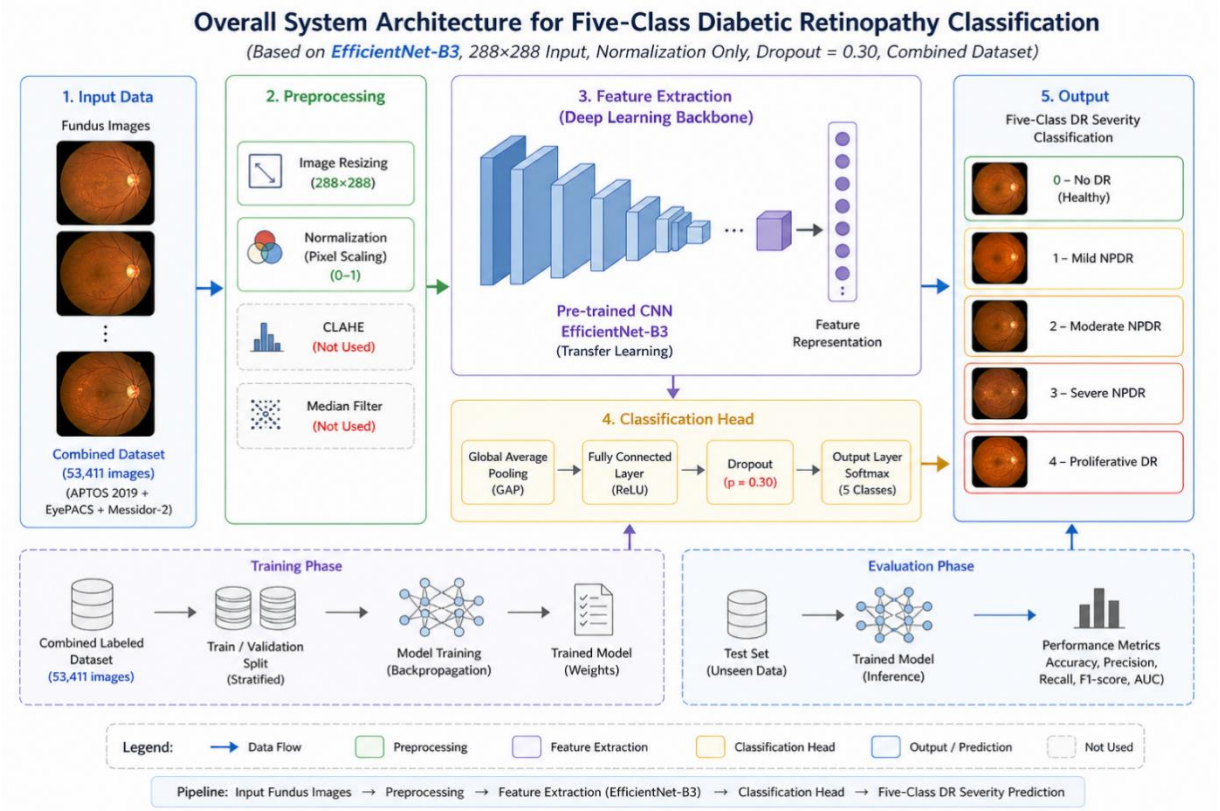

Figure 6: Overall system architecture for five-class diabetic retinopathy classification

2.3 Dataset description

2.3.1 Overview of the Combined Dataset

The dataset used in this thesis is the Combined EyePACS-APTOS-Messidor Diabetic Retinopathy Dataset. It combines retinal fundus images from several public sources, including EyePACS, APTOS 2019, an APTOS Gaussian filtered version, and the Messidor database. This combination provides a diverse collection of retinal images captured under different clinical conditions, camera types, lighting environments, and patient populations.

Using a combined dataset is important because models trained on a single-source dataset may not generalize well to images from other hospitals, cameras, or acquisition conditions. The diversity of the combined dataset helps improve the robustness of the proposed classification system.

2.3.2 Original Dataset Size

The original combined dataset contains 92,501 JPG images. In January 2024, manual data augmentation was applied, increasing the dataset by approximately 55 percent to reach 143,669 images. All images were resized to 600 by 600 pixels, reducing the disk size from 18.5 gigabytes to 3.8 gigabytes.

2.3.3 Dataset Balancing Strategy

For this thesis, a balanced subset was created to address class imbalance. The five DR severity classes were balanced using a maximum of 12,000 images per class. This resulted in the following distribution:

Table 5: Distribution of images per class after balancing

Class	DR Severity	Number of Images
0	No DR	12000
1	Mild DR	12000
2	Moderate DR	12000
3	Severe DR	7936
4	Proliferative DR	9475
Total		53411

This distribution ensures that the three largest classes are equal while preserving the natural clinical rarity of Severe and Proliferative DR.

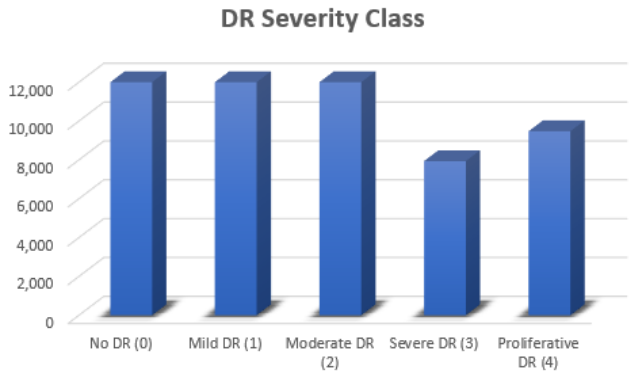

Figure 7: Class distribution after balancing

2.3.4 Train-Validation Split

The dataset was split using **80-20 stratified sampling** to maintain class distribution across subsets. Stratified sampling ensures that each subset has a similar class distribution, preventing bias during training and evaluation.

```
train_df, valid_df = train_test_split(
    df,
    test_size=0.2,
    stratify=df["label"],
    random_state=CFG.seed
)
```

Table 6: Training and validation split (80/20)

Subset	Percentage	Number of Images
Training	80%	42,728
Validation	20%	10,683

No separate testing set was used in this phase; the validation set served for both validation and final evaluation. The test set from the original combined dataset was reserved for future work.

2.3.5 Dataset Challenges and Variability

The combined dataset presents several challenges that make it realistic and valuable for developing robust models. These challenges include:

First, image quality variability due to different fundus cameras and field conditions across the three source datasets (EyePACS from the US, APTOS from rural India, and Messidor from France). Some images contain artifacts such as eyelashes, lens reflections, and uneven illumination.

Second, class imbalance is present in the original datasets. While balancing was applied to equalize the three largest classes (No DR, Mild, Moderate), the Severe and Proliferative DR classes naturally have fewer samples (7,936 and 9,475 respectively).

Third, the ordinal nature of DR severity means that misclassifications between non-adjacent classes (e.g., predicting No DR for Proliferative DR) are more severe than between adjacent classes (e.g., Mild vs Moderate).

Fourth, the dataset contains images with varying resolutions and quality levels, requiring robust preprocessing and augmentation to ensure model generalization.

2.4 Data preprocessing and augmentation

2.4.1 Image Loading and RGB Conversion

All images were loaded using the OpenCV library. Since OpenCV reads images in BGR format by default, each image was converted into RGB format before being passed to the model. This step is necessary because pre-trained CNN models, including EfficientNet-B3, expect RGB input channels.

```
img = cv2.imread(self.paths[idx])
img = cv2.cvtColor(img, cv2.COLOR_BGR2RGB)
```

2.4.2 Image Resizing

All images were resized to 288 × 288 pixels before training. This size was selected as a compromise between preserving important retinal details and reducing

computational cost. Larger images may contain more fine-grained lesion information, but they require more GPU memory and increase training time.

```
A.Resize(CFG.img_size, CFG.img_size)
```

2.4.3 Image Normalization

Image normalization was applied using the standard ImageNet mean and standard deviation values because the EfficientNet-B3 model was initialized with ImageNet pre-trained weights.

- Mean: [0.485, 0.456, 0.406]
- Standard deviation: [0.229, 0.224, 0.225]

```
A.Normalize(mean=(0.485, 0.456, 0.406), std=(0.229, 0.224, 0.225))
```

Normalization helps stabilize training by ensuring that pixel values follow a distribution similar to the data used during pre-training.

2.4.4 Tensor Conversion

After resizing and normalization, each image was converted into a tensor so that it could be processed by PyTorch. Tensor conversion is required because deep learning models operate on numerical tensors rather than raw image files.

2.5 Data Augmentation

2.5.1 Purpose of Data Augmentation

Data augmentation was applied **only to the training set** using the Albumentations library [17]. Its purpose is to increase image variability, reduce overfitting, and improve model generalization. In medical image classification, augmentation is especially useful because collecting large labeled datasets is difficult and expensive.

2.5.2 Applied Transformations

The augmentation operations included horizontal flipping, vertical flipping, affine transformations, rotation, scaling, translation, and random brightness/contrast adjustment. These transformations simulate realistic variations in retinal image acquisition while preserving the clinical meaning of the image.

The following transformations were applied:

```
train_tfms = A.Compose([
    A.Resize(CFG.img_size, CFG.img_size),
    A.HorizontalFlip(p=0.5),
    A.VerticalFlip(p=0.3),
    A.Affine(scale=(0.92, 1.08), translate_percent=(-0.03, 0.03), rotate=(-25, 25), p=0.45),
    A.RandomBrightnessContrast(p=0.3),
    A.Normalize(mean=(0.485, 0.456, 0.406), std=(0.229, 0.224, 0.225)),
    ToTensorV2()
])
```

2.5.3 Validation Transformations

For the validation set, only resizing, normalization, and tensor conversion were applied. No augmentation was used during validation because validation must represent a stable and fair evaluation of the model performance.

2.5.4 Importance of Preprocessing and Augmentation

Preprocessing ensures that all input images have the same format, size, and pixel distribution. Augmentation increases the diversity of the training data and helps the model learn more robust features. Together, these steps reduce the risk of overfitting and improve the model's ability to generalize to unseen retinal images.

2.6 PROPOSED MODEL ARCHITECTURE

2.6.1 Selection of EfficientNet-B3

The **EfficientNet-B3** architecture was selected as the backbone for this system. EfficientNet was introduced by Google researchers in 2019 [18]. It achieves state-of-the-art accuracy on image classification benchmarks while using significantly fewer parameters (12 million) than previous architectures like ResNet50 (25 million) or VGG16 (138 million).

2.6.2 Transfer Learning Strategy

Transfer learning was used by initializing EfficientNet-B3 with ImageNet pre-trained weights. The purpose of transfer learning is to reuse visual features learned from a large general image dataset and adapt them to retinal image classification.

Training a deep CNN from scratch requires a large amount of labeled data and high computational resources. By using a pre-trained model, the system benefits from previously learned features such as edges, textures, shapes, and general visual patterns.

The model was loaded with pre-trained weights from ImageNet:

```python
class DRModel(nn.Module):
    def __init__(self):
        super().__init__()
        self.model = timm.create_model(
            CFG.model_name,
            pretrained=CFG.pretrained,
            num_classes=CFG.num_classes,
            drop_rate=0.30
        )

    def forward(self, x):
        return self.model(x)
```

2.6.3 Model Architecture Details

The EfficientNet-B3 backbone consists of multiple mobile inverted bottleneck convolution (MBConv) blocks with squeeze-and-excitation optimization. The original classification head was replaced with a custom head that uses a dropout rate of 0.30 before the final linear layer to prevent overfitting.

The final output layer uses the softmax activation function, which converts the raw logits into probabilities that sum to 1 for the five DR classes.

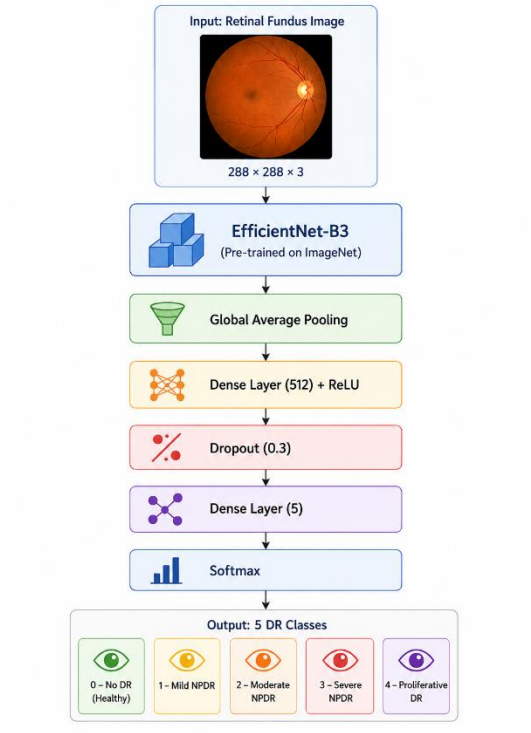

Figure 8: EfficientNet-B3 architecture with custom classification head for 5-class DR classification

2.7 Training strategy

2.7.1 Hyperparameter Configuration

The hyperparameters were organized in a CFG class:

```python
class CFG:
    base = "/kaggle/input/datasets/.../train"
    out_dir = "/kaggle/working/dr_p100_safe"
    model_name = "efficientnet_b3"
    img_size = 288
    num_classes = 5
    epochs = 6
    batch_size = 24
    lr = 2e-4
    weight_decay = 1e-4
    label_smoothing = 0.05
    num_workers = 2
    seed = 42
    patience = 2
    max_per_class = 12000
    pretrained = True
    class_names = ["No DR", "Mild", "Moderate", "Severe", "Proliferative DR"]
```

Table 7: Training configuration and hyperparameters

Parameter	Value
Model Architecture	EfficientNet-B3
Input Image Size	288 × 288 pixels
Batch Size	24
Number of Epochs	6 (with early stopping)
Initial Learning Rate	2×10^{-4}
Optimizer	AdamW
Weight Decay	1×10^{-4}
Loss Function	CrossEntropyLoss with label_smoothing = 0.05
Learning Rate Scheduler	CosineAnnealingLR ($\eta_{min} = 1 \times 10^{-6}$)
Dropout Rate	0.30
Gradient Clipping	max_norm = 1.0

2.7.2 Loss Function with Label Smoothing

The loss function used in this work is CrossEntropyLoss with label smoothing. Label smoothing reduces model overconfidence by assigning a small probability mass to non-target classes. This can improve generalization, especially in medical image classification where boundaries between adjacent severity classes may be subtle.

```python
criterion = nn.CrossEntropyLoss(label_smoothing=CFG.label_smoothing)
```

2.7.3 Optimizer: AdamW

The AdamW optimizer was used because it combines adaptive learning rates with decoupled weight decay. This helps stabilize training and reduce overfitting. The initial learning rate was set to 2×10^{-4} and weight decay was set to 1×10^{-4}.

```
optimizer = torch.optim.AdamW(
    model.parameters(),
    lr=CFG.lr,
    weight_decay=CFG.weight_decay
)
```

2.7.4 Learning Rate Scheduler: Cosine Annealing

A cosine annealing learning rate scheduler was used to gradually reduce the learning rate during training. The learning rate decreased from the initial value to a minimum value of 1×10^{-6}. This helps the model make larger updates at the beginning of training and smaller refinements near the end.

```
scheduler = torch.optim.lr_scheduler.CosineAnnealingLR(
    optimizer,
    T_max=CFG.epochs,
    eta_min=1e-6
)
```

2.7.5 Gradient Clipping and Early Stopping

Gradient clipping was applied with a maximum norm of 1.0. This technique prevents very large gradient values that may destabilize training, especially when fine-tuning deep neural networks.

```
torch.nn.utils.clip_grad_norm_(model.parameters(), 1.0)
```

Early stopping was implemented with patience of 2 epochs, monitoring the validation QWK.

2.8 Handling class imbalance

2.8.1 Class Imbalance Problem

Class imbalance is a common issue in diabetic retinopathy datasets. If the model sees more examples from certain classes, it may become biased toward those classes and fail to detect less frequent but clinically important cases.

2.8.2 Weighted Random Sampler

To reduce the effect of class imbalance, a weighted random sampler was used during training. This sampler assigns higher sampling probabilities to underrepresented classes, allowing the model to see minority classes more frequently during mini-batch creation.

This strategy helps improve learning for Severe DR and Proliferative DR, which are clinically important because they represent sight-threatening disease stages.

```python
def make_sampler(train_df):
    counts = train_df["label"].value_counts().sort_index().values
    class_weights = 1.0 / torch.tensor(counts, dtype=torch.float)
    sample_weights = train_df["label"].map(lambda x: class_weights[x].item()).values
    return WeightedRandomSampler(
        weights=torch.DoubleTensor(sample_weights),
        num_samples=len(sample_weights),
        replacement=True
    )
```

The sampler was used when creating the DataLoader for training:

```python
train_loader = DataLoader(
    DRDataset(train_df, train_tfms),
    batch_size=CFG.batch_size,
    sampler=make_sampler(train_df),
    num_workers=CFG.num_workers,
    pin_memory=True,
    drop_last=True
)
```

2.9 Evaluation metrics

2.9.1 Accuracy

Accuracy measures the percentage of correctly classified images among all validation images. Although useful, accuracy alone is not sufficient for imbalanced and ordinal medical classification problems.

2.9.2 Precision

Precision measures how many images predicted as a given class actually belong to that class. It is useful for evaluating how reliable the model's positive predictions are.

2.9.3 Recall

Recall measures how many true images of a class were correctly detected by the model. In medical diagnosis, recall is especially important because missing a diseased case can have serious clinical consequences.

2.9.4 F1-Score

F1-score combines precision and recall into a single metric. It is useful when there is a need to balance false positives and false negatives.

2.9.5 Confusion Matrix

The confusion matrix shows the relationship between predicted labels and true labels. It helps identify which classes are correctly classified and which classes are commonly confused. In DR classification, confusion between adjacent classes is expected, while confusion between distant classes is more clinically serious.

2.9.6 Quadratic Weighted Kappa (QWK)

Quadratic Weighted Kappa is the primary evaluation metric used in this thesis because diabetic retinopathy severity levels are ordinal. QWK penalizes large classification errors more heavily than small errors. For example, predicting Proliferative DR as No DR is penalized more than predicting Moderate DR as Mild DR.

```
from sklearn.metrics import cohen_kappa_score
kappa = cohen_kappa_score(trues, preds, weights="quadratic")
```

The QWK ranges from -1 to 1, where 1.00 is perfect agreement, 0.90–0.99 is excellent agreement, and 0.80–0.89 is good agreement.

2.10 Experimental environment and implementation details

2.10.1Hardware Environment

The model was trained using a Kaggle GPU environment.

Table 8: Kaggle environment hardware configuration

Component	Specification
GPU	Tesla T4 (16 GB memory)
CPU	Intel Xeon
RAM	32 GB

2.10.2Software Environment and Libraries

The implementation was developed using Python and PyTorch. Several libraries were used for image processing, augmentation, model loading, training, evaluation, and visualization.

Table 9: Software libraries and deep learning frameworks used for implementation

Library	Purpose
PyTorch	Main deep learning framework
TIMM	Pre-trained EfficientNet-B3 loading
Albumentations	Fast image augmentation
OpenCV (cv2)	Image loading and color conversion
NumPy	Numerical operations
Pandas	Data manipulation
Scikit-learn	Metrics computation
Matplotlib	Visualization

2.10.3Training and Validation Functions

Separate functions were implemented for training and validation. The training function performs forward propagation, loss computation, backpropagation, gradient

clipping, optimizer updates, and loss accumulation. The validation function disables gradient computation and evaluates the model on the validation set using loss and QWK. The training function for one epoch:

```
def train_one_epoch():
    model.train()
    total_loss = 0.0
    for imgs, labels in train_loader:
        imgs = imgs.to(device, non_blocking=True)
        labels = labels.to(device, non_blocking=True)
        optimizer.zero_grad(set_to_none=True)
        logits = model(imgs)
        loss = criterion(logits, labels)
        loss.backward()
        torch.nn.utils.clip_grad_norm_(model.parameters(), 1.0)
        optimizer.step()
        total_loss += loss.item()
    return total_loss / len(train_loader)
```

The validation function:

```
@torch.no_grad()
def valid_one_epoch():
    model.eval()
    total_loss = 0.0
    preds, trues = [], []
    for imgs, labels in valid_loader:
        imgs = imgs.to(device, non_blocking=True)
        labels = labels.to(device, non_blocking=True)
        logits = model(imgs)
        loss = criterion(logits, labels)
        pred = logits.argmax(1)
        total_loss += loss.item()
        preds.extend(pred.cpu().numpy())
        trues.extend(labels.cpu().numpy())
    kappa = cohen_kappa_score(trues, preds, weights="quadratic")
    return total_loss / len(valid_loader), kappa, np.array(preds), np.array(trues)
```

2.10.4Training Loop with Checkpointing

The training loop controls the complete optimization process across epochs. At each epoch, the model is trained on the training set and evaluated on the validation set. If the validation QWK improves, the model checkpoint is saved. Early stopping is applied when the validation score does not improve for a predefined number of epochs.

The main training loop with early stopping and model checkpointing:

```
best_kappa = -1
no_improve = 0

for epoch in range(CFG.epochs):
    train_loss = train_one_epoch()
    valid_loss, kappa, preds, trues = valid_one_epoch()
    scheduler.step()

    print(f"Epoch {epoch+1:02d}/{CFG.epochs} | Train Loss: {train_loss:.4f} | Valid Loss: {valid_los
s:.4f} | Kappa: {kappa:.4f}")

    torch.save({"model": model.state_dict(), "epoch": epoch+1, "kappa": kappa}, last_path)

    if kappa > best_kappa:
        best_kappa = kappa
        no_improve = 0
        torch.save({"model": model.state_dict(), "epoch": epoch+1, "kappa": best_kappa}, best_path)
    else:
        no_improve += 1

    if no_improve >= CFG.patience:
        print(f"Early stopping at epoch {epoch+1}")
        break
```

2.10.5 Strengths and Clinical Impact

The model demonstrates high sensitivity for sight-threatening DR (Severe + Proliferative), with recall exceeding 97% for both classes. This makes the system suitable for large-scale screening programs.

2.11 Conclusion

This chapter presented the proposed methodology used to develop the diabetic retinopathy classification model. It described the overall system workflow, starting from retinal fundus image input, followed by preprocessing, data augmentation, model training, and evaluation setup.

The chapter also introduced the combined **EyePACS-APTOS-Messidor dataset** and explained the main preparation steps, including dataset balancing, train-validation splitting, image resizing, RGB conversion, normalization, and tensor conversion. These steps were necessary to prepare the images for deep learning and to improve model generalization.

The proposed model was based on EfficientNet-B3 with transfer learning from ImageNet pre-trained weights. A custom classification head was added to adapt the model to the five diabetic retinopathy severity classes. The training strategy included AdamW optimizer, cross-entropy loss with label smoothing, cosine annealing scheduler, gradient clipping, early stopping, and weighted random sampling to handle class imbalance.

In conclusion, this chapter established the complete methodology and implementation details of the proposed system. The experimental results, confusion matrix analysis, per-class performance, and comparative evaluation are presented in the next chapter.

Chapter 3: Experimental Results and Web Application

Introduction

After presenting the proposed methodology, dataset preparation, preprocessing steps, model architecture, and training strategy in Chapter 2, this chapter presents the experimental results obtained from the proposed diabetic retinopathy classification system. The main objective of this chapter is to evaluate the performance of the trained EfficientNet-B3 model and analyze its ability to classify retinal fundus images into five diabetic retinopathy severity levels: No DR, Mild DR, Moderate DR, Severe DR, and Proliferative DR.

This chapter also presents the web application developed to demonstrate the practical use of the proposed model. The application allows the user to upload a retinal fundus image and automatically obtain the predicted diabetic retinopathy class. Therefore, this chapter links the experimental evaluation of the model with its practical implementation in an interactive system.

The evaluation of the model was carried out using several metrics, including accuracy, precision, recall, F1-score, confusion matrix, and quadratic weighted kappa. These metrics were selected because they provide a complete evaluation of the system. Accuracy gives a general idea about the number of correct predictions, while precision, recall, and F1-score provide detailed information about each class. The confusion matrix helps analyze the distribution of correct and incorrect predictions. In addition, quadratic weighted kappa is particularly important because diabetic retinopathy grading is an ordinal classification problem, where the distance between classes has clinical meaning.

This chapter is organized into two main parts. The first part presents and discusses the experimental results of the EfficientNet-B3 model. The second part describes the developed web application, including its workflow, interface, main features, practical importance, and limitations.

Experimental Setup

The experimental evaluation was conducted using the validation subset of the balanced combined EyePACS-APTOS-Messidor dataset described in Chapter 2. The dataset was divided using an 80/20 stratified split, resulting in 42,728 images for training and 10,683 images for validation.

The proposed EfficientNet-B3 model was trained using the preprocessing, augmentation, and training strategy explained in the previous chapter. In this chapter, the focus is on analyzing the obtained results, including training behavior, final accuracy, quadratic weighted kappa, classification report, and confusion matrix.

The validation set was used to evaluate the ability of the model to classify retinal fundus images into the five diabetic retinopathy severity levels: No DR, Mild DR, Moderate DR, Severe DR, and Proliferative DR.

Table 10 : Summary of the experimental setup

Parameter	**Value**
Dataset	Combined EyePACS-APTOS-Messidor
Number of classes	5
Training images	42,728
Validation images	10,683
Image size	288 × 288 pixels
Model architecture	EfficientNet-B3
Transfer learning	ImageNet pre-trained weights
Optimizer	AdamW
Loss function	Cross-entropy with label smoothing
Scheduler	Cosine annealing
Main metric	Quadratic Weighted Kappa

Training and Validation Results

During training, the model performance was monitored using training loss, validation loss, and quadratic weighted kappa. These indicators were used to evaluate whether the model was learning correctly and whether it was able to generalize to unseen validation images.

Table 11: Training and validation performance across epochs

Epoch	**Train Loss**	**Validation Loss**	**QWK**
1	1.1421	0.8839	0.8061
2	0.8098	0.8121	0.8475
3	0.6658	0.6619	0.8863
4	0.5673	0.6139	0.9009
5	0.5010	0.5843	0.9065
6	0.4648	0.5702	0.9134

The training loss decreased progressively from 1.1421 in the first epoch to 0.4648 in the sixth epoch. This indicates that the model successfully learned discriminative features from the retinal fundus images. The validation loss also decreased from 0.8839 to 0.5702, which shows that the model improved its performance on unseen validation data during training.

The quadratic weighted kappa increased from 0.8061 in the first epoch to 0.9134 in the final epoch. This steady improvement shows that the model became more capable of predicting the correct diabetic retinopathy severity levels while reducing serious classification errors between distant classes.

The continuous decrease in both training and validation loss indicates that the optimization process was stable during the six epochs. The model was able to reduce prediction errors while maintaining good performance on the validation set. This behavior suggests that the use of transfer learning, data augmentation, regularization, and learning rate scheduling helped the model learn useful and generalizable features from the retinal images.

The improvement in QWK is also important because it reflects better agreement between the predicted labels and the true labels. Since diabetic retinopathy classes are ordered by severity, improving QWK means that the model not only increased the number of correct predictions, but also reduced severe errors between distant classes.

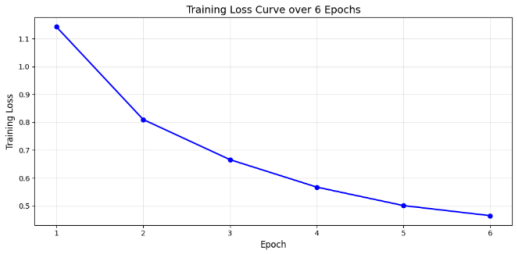

Figure 9: Training loss curve over 6 epochs

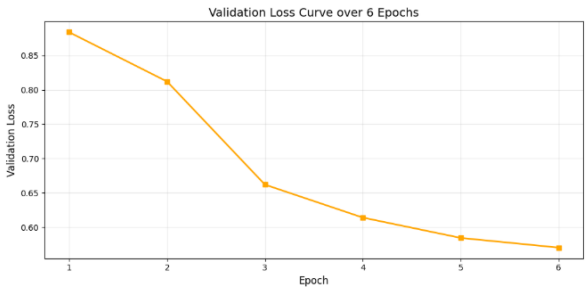

Figure 10: Validation loss curve over 6 epochs

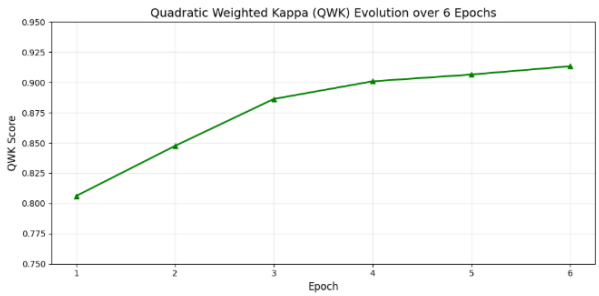

Figure 11: QWK score evolution over 6 epochs

Final Model Performance

At the end of training, the proposed EfficientNet-B3 model achieved a final validation accuracy of **85.01%** and a final quadratic weighted kappa score of **0.9134**.

Metric	Value
Accuracy	85.01%
Quadratic Weighted Kappa	0.9134
Validation images	10,683
Number of classes	5

The achieved accuracy shows that the model correctly classified a large proportion of validation images. However, accuracy alone is not sufficient for diabetic retinopathy grading because the classes are ordered according to disease severity. For this reason, QWK was used as a main evaluation metric.

The QWK score of 0.9134 indicates excellent agreement between the predicted labels and the true labels. This result is important because QWK penalizes large classification errors more heavily than small errors. For example, predicting Moderate DR instead of Mild DR is less serious than predicting No DR instead of Proliferative DR. Therefore, the high QWK value confirms that the model is not only accurate, but also reliable in terms of severity-level prediction.

These final results demonstrate that EfficientNet-B3 was able to learn relevant visual patterns from retinal fundus images. The model achieved strong performance despite the difficulty of the task, especially because diabetic retinopathy classes may have overlapping visual characteristics.

Classification Report

The classification report provides a more detailed evaluation of the model for each diabetic retinopathy class. It includes precision, recall, F1-score, and support. Precision measures how many images predicted as a class actually belong to that class, while recall measures how many images of a real class were correctly detected. The F1-score combines precision and recall into a single value.

DR Class	Precision	Recall	F1-Score	Support
No DR	0.776	0.914	0.839	2400
Mild DR	0.826	0.778	0.801	2400
Moderate DR	0.841	0.673	0.748	2400
Severe DR	0.943	0.985	0.963	1588
Proliferative DR	0.913	0.973	0.942	1895
Weighted Average	0.851	0.850	0.847	10683

The model achieved strong performance across most classes. The highest F1-score was obtained for Severe DR, with a value of 0.963, followed by Proliferative DR with

an F1-score of 0.942. These results are clinically important because Severe and Proliferative DR are advanced stages that may lead to vision loss if not detected and treated in time.

The No DR class achieved a recall of 0.914, which means that the model correctly identified most healthy retinal images. Mild DR achieved balanced performance, with an F1-score of 0.801. However, Moderate DR obtained the lowest recall value, 0.673, indicating that some Moderate DR images were confused with neighboring severity levels.

This difficulty can be explained by the visual similarity between adjacent diabetic retinopathy stages. Moderate DR may share characteristics with both Mild and Severe DR, making it more challenging to classify accurately. In addition, retinal image quality, illumination, and lesion visibility can affect the ability of the model to distinguish between intermediate disease stages.

Overall, the classification report shows that the proposed model is particularly strong in detecting advanced diabetic retinopathy cases. This is a positive result because these classes are the most important from a clinical screening perspective.

Confusion Matrix Analysis

The confusion matrix was used to analyze the distribution of correct and incorrect predictions across the five diabetic retinopathy classes. Unlike global metrics, the confusion matrix provides a detailed view of how the model behaves for each class and which classes are most frequently confused.

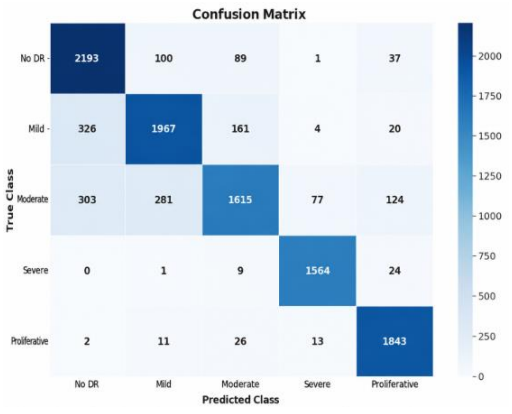

Figure 12: Confusion matrix heatmap

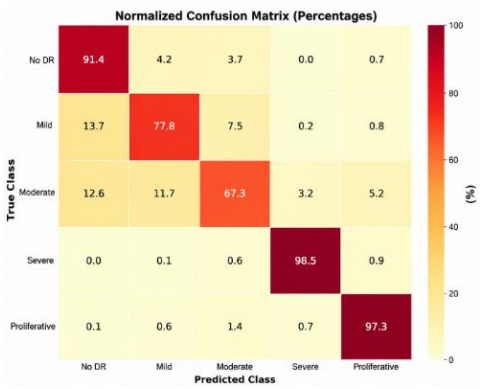

Figure 13: Normalized confusion matrix

The confusion matrix shows that most predictions are located on the diagonal, which indicates that the model correctly classified many validation images. This confirms the strong overall performance observed in the accuracy and QWK results.

Most classification errors occurred between neighboring classes, especially between Mild DR and Moderate DR, or between Moderate DR and Severe DR. This type of error is expected in diabetic retinopathy classification because adjacent severity levels often share similar visual signs, such as microaneurysms, hemorrhages, exudates, and vascular abnormalities.

The normalized confusion matrix provides a clearer view of the classification behavior for each class. It helps identify which classes were well recognized and which classes were more frequently confused. This analysis is useful because overall accuracy does not show the exact distribution of errors across the five diabetic retinopathy stages.

The errors observed between neighboring classes can be explained by the gradual progression of diabetic retinopathy. The difference between Mild and Moderate DR, or between Moderate and Severe DR, may depend on small visual details that are difficult to detect automatically. Therefore, these errors are more understandable than errors betw een very distant classes such as No DR and Proliferative DR.

From a clinical point of view, the most critical errors are those where advanced diabetic retinopathy cases are classified as normal or mild. The results show that the model achieved high recall for Severe DR and Proliferative DR, which reduces the risk of missing sight-threatening cases. This is an important strength of the proposed system.

Discussion of Experimental Results

The experimental results demonstrate that the proposed EfficientNet-B3 model is effective for five-class diabetic retinopathy classification. The final accuracy of 85.01%

and QWK score of 0.9134 show that the model achieved strong classification performance on the validation set.

One of the main strengths of the model is its ability to detect advanced diabetic retinopathy stages. Severe DR and Proliferative DR achieved very high recall values, which is important for medical screening applications. In real-world diabetic retinopathy screening, detecting patients with advanced disease is a priority because these patients may require urgent referral to an ophthalmologist.

The results also show that early and intermediate stages remain more difficult to classify. Mild DR and Moderate DR are challenging because the lesions may be small, subtle, or affected by image quality. In addition, the boundaries between adjacent DR stages are not always visually clear, even for medical experts.

The results obtained in this work confirm the importance of using a diverse dataset and a strong transfer learning architecture. The combined dataset allowed the model to learn from images captured under different acquisition conditions, while EfficientNet-B3 provided effective feature extraction with a reasonable computational cost.

Another important point is that the model performed better on advanced DR stages than on some early or intermediate stages. This can be considered a positive aspect for screening purposes, because advanced stages are more urgent and require faster medical attention. However, improving the detection of early DR remains important for prevention and early treatment.

Overall, the proposed model provides promising results and shows that transfer learning with EfficientNet-B3 can be an effective approach for diabetic retinopathy classification. The results also indicate that the system has potential to support preliminary screening, especially when used as an assistive tool rather than a replacement for medical diagnosis.

Web Application Overview

In addition to training and evaluating the deep learning model, a web application was developed to demonstrate the practical use of the proposed system. The purpose of the application is to provide a simple and interactive interface where users can upload a retinal fundus image and obtain an automatic diabetic retinopathy prediction.

The web application represents the deployment part of the project. Instead of keeping the model only as a training script, the trained model was integrated into an

accessible interface that can be used for demonstration and testing. This makes the system more practical and closer to real-world medical decision-support tools.

The application allows the user to upload a fundus image, processes the image using the same preprocessing steps used during model training, sends it to the trained EfficientNet-B3 model, and displays the predicted diabetic retinopathy class. The result may also include a confidence score and a risk interpretation to make the output easier to understand.

The development of the web application makes the proposed system more complete because it transforms the trained model into an interactive tool. Instead of limiting the work to model training and evaluation, the application demonstrates how the model can be used in a practical environment. This gives the project both a research dimension and an implementation dimension.

The application was designed with simplicity in mind. Since the system is related to medical image classification, the interface focuses on clarity, readability, and direct access to the prediction result. This helps users interact with the system without unnecessary complexity.

Web Application Workflow

The main workflow of the application is as follows:

The uploaded image is converted to RGB format, resized to the required input size, normalized, and transformed into a tensor. After preprocessing, the image is passed to the trained EfficientNet-B3 model. The model produces prediction probabilities for the five diabetic retinopathy classes. The class with the highest probability is selected as the final predicted class.

The final result is then displayed to the user through the web interface. The result page may include the uploaded image, predicted class, confidence score, and risk level. This workflow allows the user to move from image upload to classification result in a simple and direct way.

By following the same preprocessing steps used during training, the application ensures consistency between the training environment and the inference environment. This consistency is important because differences in preprocessing can negatively affect model predictions.

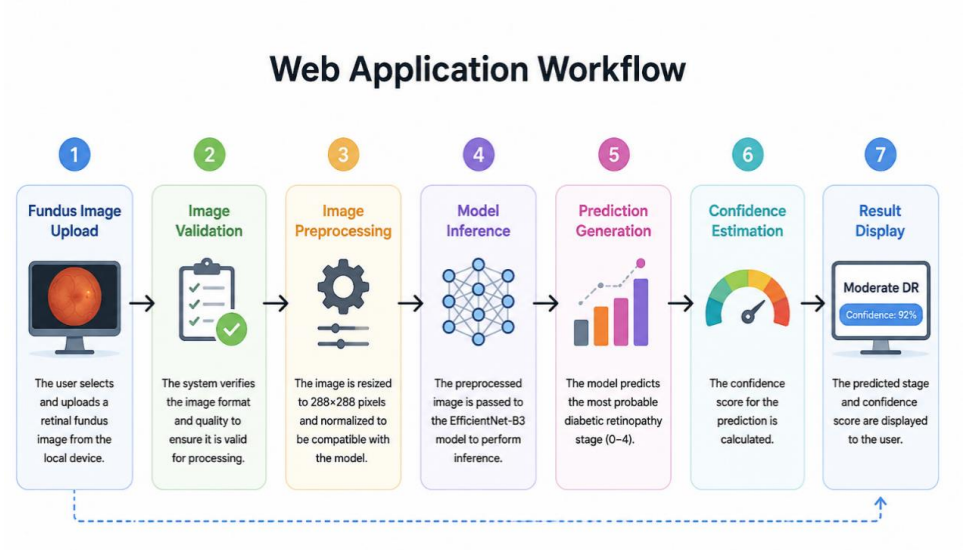

Figure 14: General workflow of the web application

Web Application Interface

The application interface was designed to be simple, clear, and easy to use. Since the system is related to a medical classification task, the interface focuses on readability and clarity rather than unnecessary visual complexity.

The main page allows the user to upload a retinal fundus image. After selecting the image, the user can start the classification process. The result page then presents the output generated by the model.

3.1.1 Home Page

The home page introduces the system and provides the image upload option. The user can select a fundus image and submit it for analysis. This page represents the entry point of the application and is designed to make the process simple for the user.

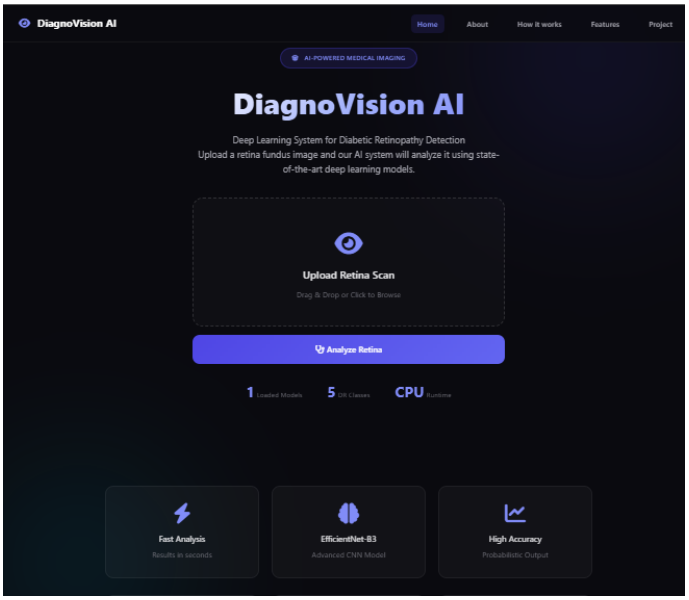

Figure 15: Home page of the web application

3.1.2 Image Upload Interface

The upload interface allows the user to choose an image from the device. This step represents the input stage of the system. The selected image is then sent to the server for preprocessing and classification.

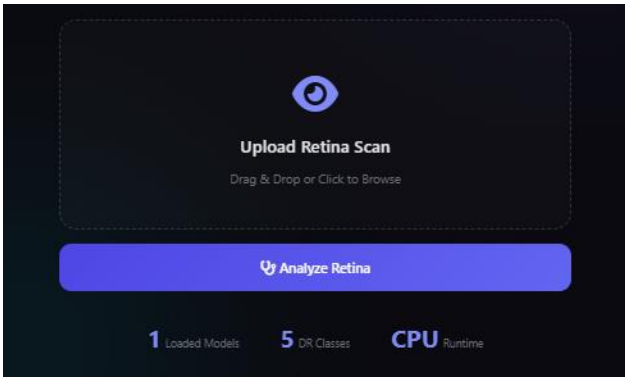

Figure 16: Image upload interface

3.1.3 Prediction Result Page

The result page displays the predicted diabetic retinopathy class. It may also show the confidence score and the clinical risk level associated with the prediction. Presenting the result clearly is important because it helps the user understand the model output quickly.

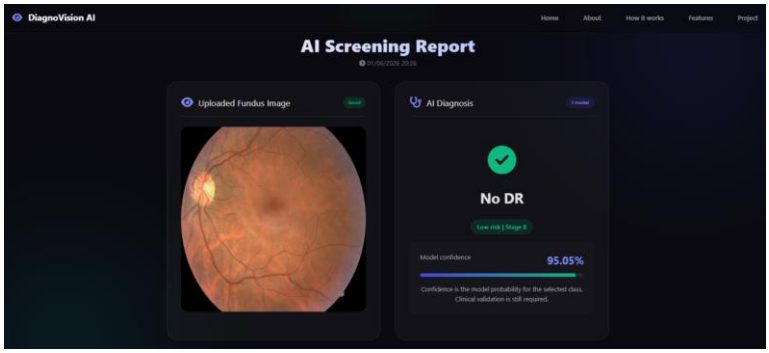

Figure 17: Prediction result page

3.1.4 Attention Map Visualization

The application also provides a visual attention map that highlights the retinal regions that contributed most to the model prediction. This visualization helps make the classification result more interpretable by showing the areas that may contain signs related to diabetic retinopathy, such as lesions, hemorrhages, or abnormal retinal patterns. By displaying both the predicted class and the highlighted regions, the application becomes more transparent and easier to understand.

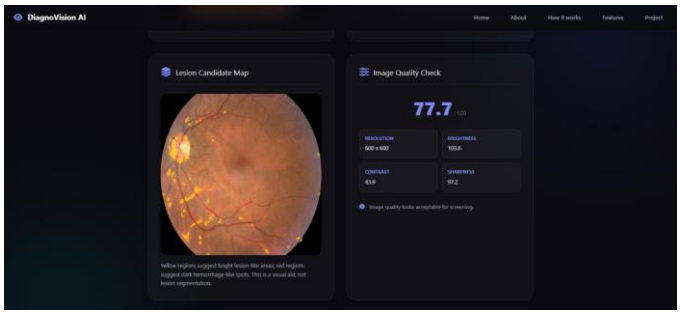

Figure 18: Attention map highlighting the retinal regions involved in the prediction

3.1.5 Result Interpretation for Each DR Stage

After the image is analyzed by the model, the application displays the predicted diabetic retinopathy stage with a confidence score and a risk interpretation. Each possible output is presented in a clear way so that the user can understand the meaning of the prediction.

Predicted Class	Result Displayed in the Application	Risk Level	Interpretation
No DR	No DR	Low risk	No visible sign of diabetic retinopathy was detected by the model.
Mild	Mild	Mild risk	The model detected early signs compatible with mild diabetic retinopathy.
Moderate	Moderate	Moderate risk	The model detected lesions compatible with moderate diabetic retinopathy.

Severe	Severe	High risk	The model detected signs compatible with severe diabetic retinopathy.
Proliferative DR	Proliferative DR	Urgent risk	The model detected signs compatible with advanced proliferative diabetic retinopathy.

The result page presents the model output in a clear and understandable format. For each uploaded retinal image, the application displays the predicted class, confidence score, probability distribution, risk level, and clinical interpretation. This helps the user understand whether the image corresponds to a normal retina, an early stage, or an advanced stage that may require medical attention.

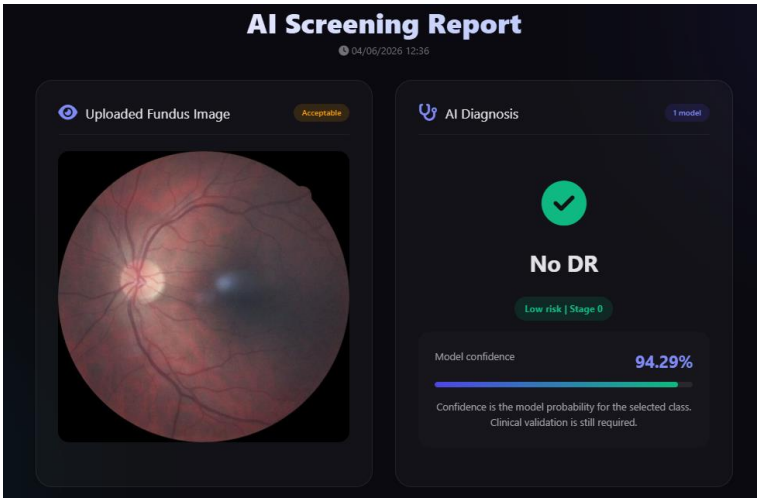

Figure 19: Application result for No DR case

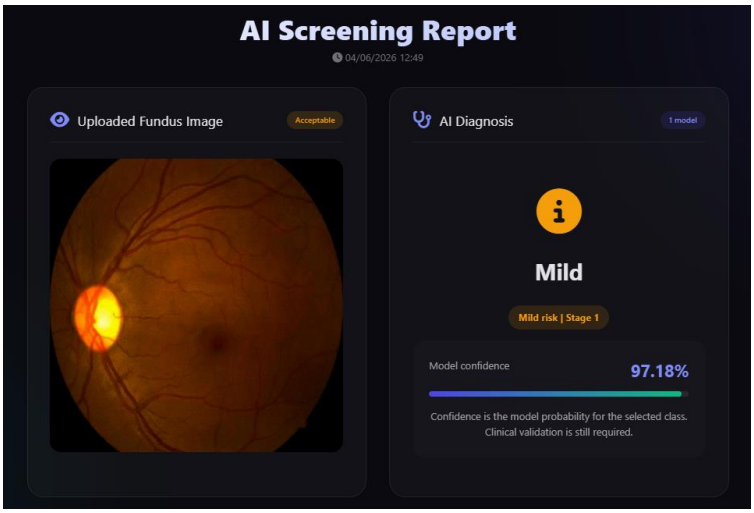

Figure 20: Application result for Mild DR case

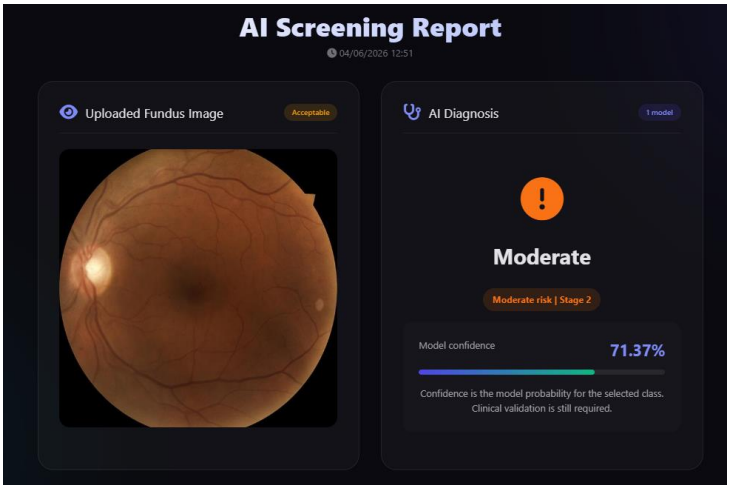

Figure 21: Application result for Moderate DR case

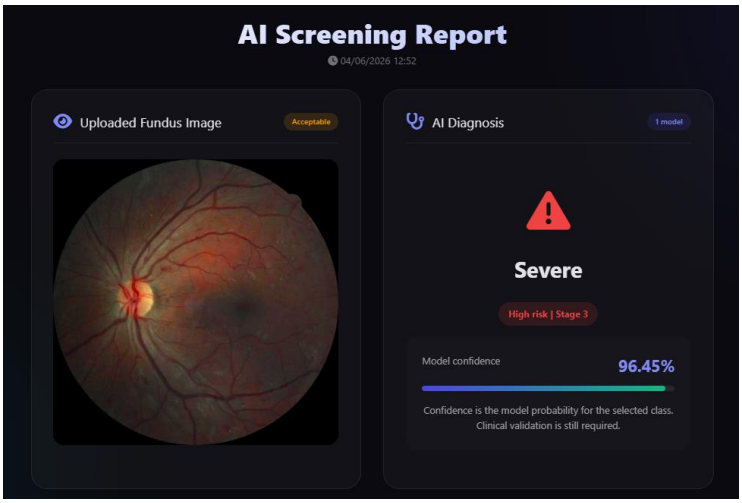

Figure 22: Application result for Severe DR case

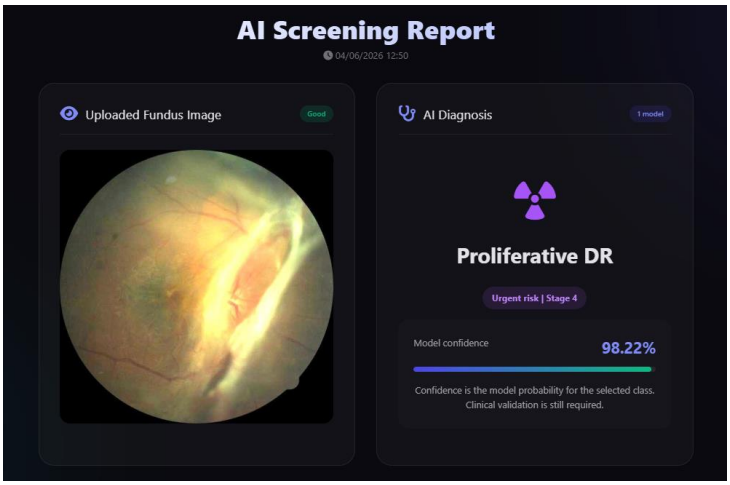

Figure 23: Application result for Proliferative DR case

Main Features of the Web Application

The developed web application includes several important features that support the practical use of the proposed diabetic retinopathy classification system.

Retinal Image Upload

The application allows users to upload fundus images in common formats such as JPG, PNG, JPEG, BMP, TIFF, and JFIF.

AI-Based Diabetic Retinopathy Classification

The system uses a deep learning model to classify diabetic retinopathy into five stages: No DR, Mild, Moderate, Severe, and Proliferative DR.

Prediction Confidence Score

After analysis, the application displays the predicted class with a confidence percentage, showing how certain the model is about its decision.

Class Probability Distribution

The web app shows the probability of each diabetic retinopathy stage, allowing users to see how the model compared all possible classes.

Image Quality Assessment

The system evaluates the uploaded image quality using resolution, brightness, contrast, and sharpness measurements.

Lesion Candidate Map

The application generates a visual map that highlights possible lesion-like regions in the retina, such as bright exudate-like areas and dark hemorrhage-like spots.

Clinical Interpretation

Each prediction is accompanied by a short explanation describing the meaning of the detected stage.

Risk Level Indication

The app assigns a risk level based on the predicted stage, such as Low risk, Moderate risk, High risk, or Urgent risk.

Recommended Next Steps

The system provides guidance and recommendations, such as consulting an ophthalmologist or scheduling follow-up screening.

Printable and Downloadable Report

Users can print the analysis result or download it as a PDF report for documentation or medical review.

User-Friendly Interface

The web application provides a modern and simple interface that makes the screening process easy to use and understand.

Medical Disclaimer

The application clearly states that it is a screening and decision-support tool, not a replacement for diagnosis by an ophthalmologist.

Practical Importance of the Application

The developed application is important because it transforms the trained deep learning model into an interactive tool. In medical image analysis projects, deployment is a key step because it allows the model to be tested and demonstrated in a realistic way.

The application can help show how artificial intelligence may support diabetic retinopathy screening. By automatically analyzing retinal fundus images, such a system could assist healthcare professionals in identifying patients who may require further examination.

Such an application can be useful in awareness, education, and preliminary screening scenarios. It can help demonstrate how artificial intelligence can support medical image analysis and assist in identifying cases that may need further examination. Nevertheless, the output of the system should always be interpreted carefully and should not replace expert medical diagnosis.

The practical value of the application also lies in its accessibility. A web-based interface allows users to interact with the model more easily than a programming environment. This makes the system more understandable for non-technical users and more suitable for demonstration during project presentation.

However, the application should be considered as a decision-support tool and not as a replacement for a medical specialist. The final diagnosis must always be confirmed by an ophthalmologist or qualified healthcare professional. The main objective of the application is to demonstrate the potential of deep learning in assisting diabetic retinopathy detection.

Conclusion

This chapter presented the experimental results and web application implementation of the proposed diabetic retinopathy classification system. The EfficientNet-B3 model

was evaluated on a validation set of 10,683 retinal fundus images using accuracy, precision, recall, F1-score, confusion matrix, and quadratic weighted kappa.

The proposed model achieved a **final validation accuracy of 85.01%** and a **QWK score of 0.9134**. These results indicate strong classification performance and excellent agreement between predicted and true labels. The model performed particularly well in detecting Severe and Proliferative DR, which are the most clinically critical stages.

The analysis showed that most classification errors occurred between neighboring diabetic retinopathy severity levels. This is expected because adjacent stages may share similar visual characteristics and may be difficult to distinguish, especially when image quality is low or lesions are subtle.

The chapter also presented the developed web application, which allows users to upload retinal fundus images and obtain automatic diabetic retinopathy predictions. The application demonstrates the practical integration of the trained model into an interactive system and highlights the potential of deep learning for supporting diabetic retinopathy screening.

In conclusion, the results confirm the effectiveness of the proposed EfficientNet-B3 model and show that the developed application can serve as a useful demonstration of an AI-based diabetic retinopathy classification system. The next section presents the general conclusion of the thesis, summarizes the main contributions, and suggests future improvements.

General Conclusion

Diabetic retinopathy is one of the most serious complications of diabetes and remains a major cause of vision loss when it is not detected and treated at an early stage. Regular screening is therefore essential to identify retinal abnormalities before the disease progresses to advanced stages. However, manual analysis of retinal fundus images can be time-consuming and depends heavily on the availability of experienced specialists. For this reason, artificial intelligence and deep learning techniques have become promising tools for supporting diabetic retinopathy detection and classification.

This thesis proposed an intelligent diabetic retinopathy classification system based on deep learning. The main objective was to develop a model capable of classifying retinal fundus images into five severity levels: No DR, Mild DR, Moderate DR, Severe DR, and Proliferative DR. To achieve this objective, a combined EyePACS-APTOS-Messidor dataset was used in order to increase data diversity and improve the generalization ability of the model.

The proposed methodology included several important steps, such as dataset balancing, train-validation splitting, image preprocessing, data augmentation, and transfer learning. The EfficientNet-B3 architecture was selected as the main classification model because it provides a good balance between classification performance and computational efficiency. The model was trained using AdamW optimizer, cross-entropy loss with label smoothing, cosine annealing learning rate scheduling, gradient clipping, early stopping, and weighted random sampling to handle class imbalance.

The experimental results demonstrated the effectiveness of the proposed approach. The model achieved a validation accuracy of 85.01% and a quadratic weighted kappa score of 0.9134, indicating excellent agreement between predicted and true labels. The model showed particularly strong performance in detecting Severe and Proliferative DR, which are the most clinically critical stages. Most classification errors occurred between neighboring severity levels, especially Mild and Moderate DR, due to the visual similarity between adjacent stages.

In addition to model training and evaluation, a web application was developed to demonstrate the practical use of the proposed system. The application allows users to upload retinal fundus images and obtain automatic diabetic retinopathy predictions, including the predicted class, confidence score, risk level, and visual attention map.

This implementation shows how the trained model can be integrated into an interactive tool that supports preliminary screening and helps make the system more accessible.

Overall, this work confirms the potential of deep learning, particularly EfficientNet-B3, for diabetic retinopathy classification from retinal fundus images. The proposed system can serve as a decision-support tool for preliminary screening, awareness, and educational purposes. However, it should not be considered a replacement for professional medical diagnosis, since final clinical decisions must always be confirmed by qualified ophthalmologists.

Future improvements may include testing the model on independent external datasets, improving the detection of early and intermediate DR stages, adding more explainability techniques, and enhancing the web application with patient history management, report generation, database storage, and real clinical workflow integration. These improvements could make the system more robust, reliable, and suitable for practical medical environments.

References

[1] D. S. A. L. P. F. F. L. a. K. R. Fong, "Diabetic retinopathy," *Diabetes Care,* vol. 26, no. suppl_1, pp. s99-s102, 2003.

[2] R. K. B. E. K. M. S. E. a. C. K. J. Klein, «The Wisconsin epidemiologic study of diabetic retinopathy: XVII. The 14-year incidence and progression,» *Ophthalmology,* vol. 105, n° %110, pp. 1801-1815, 1998.

[3] C. P. F. F. L. K. R. E. e. a. Wilkinson, «Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales,» *Ophthalmology,* vol. 110, n° %19, pp. 1677-1682, 2003.

[4] J. a. W. T. Y. Ding, «Current epidemiology of diabetic retinopathy and diabetic macular edema,» *Current Diabetes Reports,* vol. 12, n° %14, pp. 346-354, 2012.

[5] T. D. R. S. R. Group, «Photocoagulation treatment of proliferative diabetic retinopathy,» *American Journal of Ophthalmology,* vol. 81, n° %14, pp. 383-407, 1976.

[6] I. D. Federation, IDF Diabetes Atlas, Brussels: International Diabetes Federation, 2021.

[7] J. W. Y. Yau, «Global prevalence and major risk factors of diabetic retinopathy,» *Diabetes Care,* vol. 35, n° %13, pp. 556-564, 2012.

[8] R. P. Finger, «Inter-observer agreement for diabetic retinopathy grading,» *British Journal of Ophthalmology,* vol. 93, n° %15, pp. 613-616, 2009.

[9] V. Gulshan, «Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,» *JAMA,* vol. 316, n° %122, pp. 2402-2410, 2016.

[10] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[11] Y. B. Y. a. H. G. LeCun, «Deep learning,» *Nature,* vol. 521, n° %17553, pp. 436-444, 2015.

[12] A. S. I. a. H. G. E. Krizhevsky, «ImageNet classification with deep convolutional neural networks,» chez *Advances in Neural Information Processing Systems (NeurIPS)*, 1097-1105, 2012.

[13] D. E. H. G. E. a. W. R. J. Rumelhart, «Learning representations by back-propagating errors,» *Nature,* vol. 323, n° %16088, pp. 533-536, 1986.

[14] S. J. a. Y. Q. Pan, «A survey on transfer learning,» *IEEE Transactions on Knowledge and Data Engineering,* vol. 22, n° %110, pp. 1345-1359, 2010.

[15] K. Z. X. R. S. a. S. J. He, «Deep residual learning for image recognition,» chez *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] A. Canipek, «Eyepacs, Aptos, Messidor Diabetic Retinopathy Dataset,» 2024. [En ligne]. Available: https://www.kaggle.com/datasets/ascanipek/eyepacs-aptos-messidor-diabetic-retinopathy.

[17] A. Buslaev, «Albumentations: fast and flexible image augmentations,» *Information,* vol. 11, n° %12, p. 125, 2020.

[18] M. a. L. Q. V. Tan, «EfficientNet: Rethinking model scaling for convolutional neural networks,» chez *International Conference on Machine Learning (ICML)*, 2019.