IMPROVING DIABETIC RETINOPATHY GRADING ACCURACY WITH A CNN-BASED ENSEMBLE METHOD

PROPOSAL TESIS

By:
Puji Kuswanti
2311601617

PROGRAM STUDI MAGISTER ILMU KOMPUTER

FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS BUDI LUHUR

JAKARTA

GENAP 2025/2026

IMPROVING DIABETIC RETINOPATHY GRADING ACCURACY WITH A CNN-BASED ENSEMBLE METHOD

PROPOSAL TESIS

Diajukan untuk memenuhi salah satu persyaratan memperoleh gelar Magister Ilmu Komputer (M.Kom)

By:
Puji Kuswanti
2311601617

PROGRAM STUDI MAGISTER ILMU KOMPUTER

FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS BUDI LUHUR

JAKARTA

GENAP 2025/2026

PROGRAM STUDI MAGISTER ILMU KOMPUTER
FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS BUDI LUHUR

LEMBAR PENGESAHAN

Nama	:
Nomor Induk Mahasiswa	:
Program Studi	: Magister Ilmu Komputer
Bidang Peminatan	:
Jenjang Studi	: Strata 2
Judul	:

Laporan Proposal Tesis ini telah disetujui, disahkan dan direkam secara elektronik sehingga tidak memerlukan tanda tangan tim penguji.

Jakarta, Sabtu 13 Maret 2021

Tim Penguji:

Ketua	:
Anggota	:
Pembimbing	:
Ketua Program Studi	:

ABSTRAK

Retinopati diabetik (DR) merupakan salah satu komplikasi serius dari penyakit diabetes yang dapat menyebabkan kebutaan permanen jika tidak ditangani sejak dini. Tingkat keparahan DR diklasifikasikan dalam lima kategori, dan klasifikasi yang akurat sangat penting untuk pengambilan keputusan klinis. Saat ini, metode berbasis deep learning seperti convolutional neural networks (CNN) banyak digunakan untuk otomatisasi proses ini, namun performanya sering kali tidak konsisten jika hanya menggunakan satu model. Masalah utama dalam penelitian ini adalah bagaimana meningkatkan akurasi dan keandalan klasifikasi retinopati diabetik dengan menggabungkan beberapa model CNN menggunakan pendekatan ensemble. Tujuan dari penelitian ini adalah untuk mengembangkan model stacking ensemble yang mengintegrasikan tiga arsitektur CNN pralatih, yaitu DenseNet50, InceptionResNetV2, dan EfficientNetV2, dengan memanfaatkan meta-classifier seperti Multilayer Perceptron (MLP), Random Forest, dan XGBoost. Model ini diharapkan dapat mempelajari cara optimal untuk menggabungkan prediksi dari masing-masing CNN guna meningkatkan akurasi klasifikasi. Metode yang digunakan mencakup praproses citra fundus dari dua dataset publik (APTOS 2019 dan DRTiD), pelatihan masing-masing model CNN, ekstraksi vektor probabilitas softmax, penggabungan fitur, serta pelatihan meta-classifier. Evaluasi dilakukan menggunakan metrik akurasi, F1-score, Kappa Cohen, dan AUC. Hasil yang diharapkan dari penelitian ini adalah terciptanya prototipe sistem klasifikasi retinopati diabetik yang lebih akurat, andal, dan dapat diimplementasikan untuk mendukung deteksi dini di bidang kesehatan, khususnya oftalmologi.

Kata Kunci: Retinopati Diabetik, Deep Learning, Stacking Ensemble, CNN, Meta-Classifier

ABSTRACT

Diabetic retinopathy (DR) is a serious complication of diabetes that can lead to permanent blindness if not detected and treated early. The severity of DR is classified into five categories, and accurate grading plays a crucial role in clinical decision-making. Currently, deep learning methods such as convolutional neural networks (CNNs) are widely used to automate this process. However, their performance often lacks consistency when relying on a single model. The main problem addressed in this research is how to improve the accuracy and reliability of diabetic retinopathy classification by combining multiple CNN models through an ensemble approach. The objective of this study is to develop a stacking ensemble model that integrates three pre-trained CNN architectures: DenseNet50, InceptionResNetV2, and EfficientNetV2, using meta-classifiers such as Multilayer Perceptron (MLP), Random Forest, and XGBoost. This ensemble model is expected to learn how to optimally combine the predictions of individual CNNs to improve classification performance. The method involves preprocessing fundus images from two public datasets (APTOS 2019 and DRTiD), training each CNN individually, extracting softmax probability vectors, concatenating the features, and training the meta-classifiers. The ensemble models will be evaluated using metrics such as accuracy, F1-score, Cohen's Kappa, and AUC. The expected outcome of this study is the development of a prototype system for diabetic retinopathy grading that is more accurate and reliable, which could support early screening and diagnosis in the healthcare domain, particularly in ophthalmology.

Keywords: Diabetic Retinopathy, Deep Learning, Stacking Ensemble, CNN, Meta-Classifier,

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to those who have supported and guided me throughout the completion of this research proposal.

My deepest appreciation goes to Dr. Ir. Achmad Solicihin, S.Kom., M.T.I., Dean of the Faculty of Information Technology, Universitas Budi Luhur, who introduced me to the field of machine learning in a truly engaging and inspiring way. His teaching sparked my interest, and I have enjoyed exploring this field ever since.

I would also like to thank Dr. Rusdah, S.Kom., M.Kom., Head of the Master of Computer Science Program at Universitas Budi Luhur, for her constant encouragement, kind support, and for always believing in me. Her words of motivation have helped me stay focused and confident throughout the proposal process.

Special thanks to my supervisor, Luhur Bayuaji, S.T., M.Eng., Ph.D., for his continuous support, constructive feedback, and insightful direction. I truly appreciate the time he has generously spared to supervise us, even with his demanding schedule.

To all who have contributed in various ways, I am truly grateful.

Lists of Tables

Table of Figures

TABLE OF CONTENTS

CHAPTER I
INTRODUCTION

1.1 Research Background

Diabetes mellitus (DM) has become more and more common in the world's population and it is now one of the world's most common diseases found among people. Report by the World Health Organization (WHO), in the late 20th century there are only about 200 million people in the world suffering from diabetes and the number has increased significantly to approximately 800 million in second decade of 21th century. Changes of lifestyle is one of the highest contributing factors to the increase of diabetes whish mostly due poor dietary habits and reduced physical activity. DM is a condition when the body has insufficient insulin production or ineffective insulin action, usually results in persistent hyperglycaemia and a range of metabolic complications (Soelistijo, 2021). The new dynamic of the era, where people prefer instant food, frequent consumption of fast food, high-sugar beverages, and inactive behaviors have contributed significantly to the increase of DM cases, particularly among young adults (Atika and Susilawati, 2022; Hermawan et al., 2021). In many parts of the world, DM poses a serious public health concern. The International Diabetes Federation (IDF) reported that about 500 million adults were living with diabetes in 2021, and this number is projected to rise to 600 million by 2030 and 800 million by 2045. Diabetes cases in Indonesia are expected to reach slightly over 20 million by 2025 and estimated to increase by 40% by 2050 (IDF, 2025).

DM cases have been seen increasingly common not only in older people, but also in younger generations, leading to further complications following the rise prevalence of diabetes. These include both macrovascular and microvascular complications. Microvascular issues related to damage to small blood vessels while macrovascular issues involved damage to large blood vessels. In this case, microvascular complications such as diabetic retinopathy, diabetic nephropathy and diabetic neuropathy are more specific to diabetes. Among the others, diabetic retinopathy (DR) is one of most common microvascular complications of diabetes, which cause patients to suffer sight loss if left untreated. Prolonged high level of blood sugar can do damage to the small blood vessels in the retina which can lead to DR. Diabetic patients usually suffer blurriness and distortion of vision in the early stage of DR. Many diabetes patients are unaware that they have DR due unnoticeable symptoms in the early stages. Other than that, some are also lacking in information of diabetes-related complications which end up in gradual loss of visual acuity (Purnama, 2023; Soelistijo, 2021). In more advanced stages, DR can become vision-threatening, a condition referred to as VTDR which includes severe no-proliferative DR, proliferative DR and diabetic macular edema (Trisera et al., 2024). Reaching this stage, patients have their eyes suffer extensive retinal haemorrhages, venous abnormalities and microvascular damage.

Diagnosing DR is not a simple task that can be done by general practitioner, as it typically involves particular retinal imaging techniques, such as fluorescein angiography, optical coherence tomography (OCT), and B-scan ultrasonography (Soelistijo, 2021). DR is usually diagnosed by analysing retinal fundus images, which are two-dimensional colour photographs of internal surface of the eye. These images allow physicians to examine the blood vessels, the macula and the optic disc for signs of damage. Only trained experts or ophthalmologists are able to accurately diagnose DR from the fundus images. However, the analysis is subject to variability in interpretation among clinicians,

particularly when early signs are subtle (Srinivasan et al., 2023). In smaller clinics or public health care centre, the shortage of trained eye care specialists added challenge for early diagnosis and intervention.

In small cities or in rural urbans doctors often lacked both the tools and training to identify early signs of DR from fundus images. Many efforts were made to resolve this issue, with research focusing on the development of more accessible and efficient methods for early detection. With the advancement of technology, particularly in artificial intelligence (AI), automatic DR detection has more possibilities. Development of image-processing algorithms further enhances the potential to detect signs of DR by using fundus images.

Nowadays, AI and deep learning have advanced significantly, along with one of their key architectures, the convolutional neural networks (CNNs). These architectures have transformed the field of medical image analysis. Research shows that CNN-based models have demonstrated state-of-the-art performance in DR detection and grading by learning complex features from retinal images (Taifa et al., 2024a). The CNN models can classify DR severity level into standardized grading scales based on the International Clinical Disease Severity Scale (ICDS) for diabetic retinopathy, providing a scalable and automated solution to assist an accurate diagnosis. Incorporating AI into the CNN models, such as Inception-V3, Inception-V4, and VGGNet Ensemble, has achieved high accuracy rates of 80% to 90% in identifying the severity of DR across datasets (Yusran et al., 2022). A number of AI systems, including AlexNet and variants of InceptionV3, have been adapted for detecting advanced DR. Some offline tools, such as ADVEN-i and Selena+, have also been developed to support DR screening without the need for continuous internet access (Tahir et al., 2025).

These tools offer promising options for mass screening and for expanding access to eye care services, especially in underserved areas. Deep learning as part of the AI technology developed in recent years, mainly convolutional neural networks (CNNs) has a strong potential for identifying fine vascular changes in retinal images due DR. Many other methods are also being explored, such as capsule networks and graph-based models. Other than the CNNs architecture, there are also no-code platforms that make it possible for healthcare practitioners without coding knowledge to use these systems by using simple, intuitive interfaces (Budi Susilo et al., 2025).

Many researchers have designed sophisticated algorithms to grade the severity of DR by using fundus images. The most common architecture used is convolutional neural networks (CNN) with its various models such as EfficientNet, MobileNet, ResNet, Inception and DenseNet121. A review on diabetic retinopathy grading reported the accuracy of different models, especially ResNet, DenseNet and Inception, have a range of 70% to 90% (Saraswathy, 2023). This range shows that the accuracy of each model varies based on models and dataset used. An ensemble of CNN architectures InceptionV3, Xception, ResNet-50, MobileNetV2, and DenseNet-201 results in accuracy about 90% by using majority voting (Deepa et al., 2022) while a combination of other deep learning models such as EfficientNet, EfficientNetV2, LCNet, MobileNetV3, TinyNet, and FBNetV3 using ensemble stacking techniques results in testing accuracy of 84.17% (Handoyo and Kusuma, 2022).

By using APTOS 2019 dataset, ensemble of the ResNet-50, MobileNet, and EfficientNet architectures by weighted voting results in accuracy of 93.3 % (Desiani et al., 2024). The accuracy shows that ensemble method gives higher accuracy than it's individual base models. The ensemble methods accuracies vary approximately from 80%

to 90% despite using similar dataset. This means that there is still room for improvement in the classification performance of diabetic retinopathy using ensemble CNN models. Though some reports stated accuracy above 90%, the variation in outcomes across different studies suggests that there are factors affecting the results, such as model selection, ensemble method, pre-processing techniques, and dataset diversity. Other than that, most existing research focus on accuracy, and rarely addresses issues such as model generalizability, robustness or computational efficiency. This opens opportunities for further exploration, particularly in optimizing ensemble combinations, refining fusion techniques, and evaluating performance across diverse datasets to achieve more reliable and scalable DR grading systems.

This research is to address the opportunity for further exploration to improve the accuracy of diabetic retinopathy grading by developing a CNN-based ensemble method. Taking advantage of the strength of multiple CNN architectures and integrating them through and optimized ensemble strategy, this study is pursuing the goal of enhancing classification performance on fundus images. Other than to achieve higher accuracy, the objective of this research is also to ensure better model reliability and generalizability, contributing to the development of automated diabetic retinopathy diagnosis systems.

1.2 Statement of Problem

1.2.1 Problem Identification

Based on the background described earlier, the problems addressed in this study are as follows:

1. DR is a diabetes complication that can lead to blindness when left untreated. Accurate detection of DR, especially its severity levels, is critical to prevent irreversible vision loss.
2. Manual grading of DR from retinal fundus images by ophthalmologists is time-consuming, prone to inter-observer variability, and not scalable for mass screening, especially in resource-limited settings where eye care specialists are not available.
3. Deep learning models, particularly Convolutional Neural Networks (CNNs), have shown promising performance in automating DR detection and classification. However, most studies rely on single CNN architectures (e.g., ResNet, DenseNet, EfficientNet), which may lack generalizability, especially in multi-class classification tasks with imbalanced data.
4. Although CNN-based models have been widely used for DR classification, their performance remains inconsistent, with accuracy varying across different model architectures and datasets, giving room for improved ensemble approaches to enhance reliability and generalization.

1.2.2 Scope and Limitation

To ensure the research remains focused and feasible, the scope of this study is limited to the following:

1. The study exclusively utilizes retinal fundus images for both training and evaluation.
2. The research is limited to deep learning models, specifically the CNN-based architectures DenseNet201, EfficientNetV2, and InceptionResNetV2, and their ensemble.
3. The classification is restricted to the five severity levels of diabetic retinopathy (0–4), based on the International Clinical Diabetic Retinopathy Disease Severity Scale, supporting the hypothesis on severity grading accuracy.

4. The study uses the APTOS (Asia Pacific Tele-Ophthalmology Society) Blindness Detection dataset and DRTiD (Diabetic Retinopathy Two-field image Dataset), both are publicly available and clinically labeled benchmark. The dataset is split into training and testing sets to evaluate generalization and performance.

1.2.3 Research Questions

In line with the background, identified problems, and defined limitations, the research question formulated is:

Can a CNN-based ensemble model, combining DenseNet201, EfficientNetV2, and InceptionResNetV2, improve the accuracy of diabetic retinopathy severity grading compared to individual CNN models?

1.3 Research Objectives and Significance

1.3.1 Research Objectives

The primary objective of this research is to determine whether a CNN-based stacking ensemble, integrating DenseNet50, InceptionResNetV2, and EfficientNetV2, can improve the accuracy of diabetic retinopathy severity grading compared to individual CNN models. To achieve this, the study will design and implement an ensemble framework combining the three CNN architectures, explore and compare the effectiveness of different meta-classifiers, namely Multilayer Perceptron (MLP), Random Forest, and XGBoost, within the framework, and develop a functional prototype for automated diabetic retinopathy grading using the proposed approach.

1.3.2 Significance

This research is significant in both technical and clinical contexts. First, it proposes a novel stacking ensemble framework that integrates three high-performing CNN architectures, DenseNet50, InceptionResNetV2, and EfficientNetV2, with the goal of improving the accuracy and consistency of diabetic retinopathy severity grading, a key requirement for reliable AI-assisted diagnosis. Second, by evaluating and comparing multiple meta-classifiers (MLP, Random Forest, and XGBoost), the study provides valuable insights into optimal strategies for integrating CNN outputs in medical image analysis, offering guidance for future ensemble learning research in healthcare applications. Finally, the development of a functional prototype for automated DR grading demonstrates the practical applicability of the proposed method, with the potential to assist clinicians in early detection and grading of diabetic retinopathy, particularly in resource-limited or remote settings where specialist expertise is scarce.

1.4 Research Outline

This research report is organized into four chapters as follows:

Chapter I – Introduction

Provides the background, research problems, scope, objectives and significance of the study, research methodology overview, and research outline.

Chapter II – Theoretical Framework and Literature Review

Explains the relevant theories and references used in the research, serving as the foundation for addressing the research problems.

Chapter III – Research Methodology and Design

Describes the workflow, stages, and design implemented to support and achieve the research objectives.

Chapter IV – Conclusion

Presents conclusions of the research proposal.

Chapter II
THEORETICAL AND CONCEPTUAL FRAMEWORK

2.1 Theoretical Background

2.1.1 Diabetic Retinopathy (DR)

Diabetic retinopathy (DR) is a diabetes complication mostly caused by damage in blood vessels. It is a form of microangiopathy, a disease involving capillaries in which the capillary walls become thick and weak that they bleed and slowing the blood flow. The main sign of DR is the microvascular occlusion and vascular leakage, where there is blockage of closure of tiny blood vessels in the retina and leakage of blood or fluid out of retinal blood vessels into the surrounding of retinal tissue. It is a chronic microvascular complication due to prolonged hyperglycemia (Purnama, 2023). If this condition progresses over time without appropriate medication, it poses a high risk of visual impairment and eventual blindness.

At the early stage, eyes with DR suffer Non-Proliferative Diabetic Retinopathy (NPDR) and at a later stage, the Proliferative Diabetic Retinopathy (PDR). They are the two main stages of DR based on the condition of the blood vessels in the retina. NPDR is the early phase of the disease while PDR occurs at an advanced stage. Mostly in NPDR stage, it can be seen in the fundus image the presence of microaneurysms, intraretinal haemorrhages, and hard exudates, indicating mild to moderate vascular damage. PDR is the more advanced stage of DR, in which there are new blood vessels formed in the retina which is commonly named neovascularization. In this stage, there is a formation of fragile, abnormal new blood vessels that are prone to rupture, leading to vitreous haemorrhage and an increased risk of retinal detachment (Nneji et al., 2022). Figure 2.1 (a) shows the normal retina condition while the fundus image on the right (b) depicts a retina with diabetic retinopathy symptoms. Clear differences can be seen from the images, showing changes when diabetic patients suffer from DR. Haemorrhages, aneurysm, cotton wool spots, hard exudates and abnormal growth of blood vessels formed in the retina with severe DR. These characteristics are important features in differentiating healthy and DR retina images.

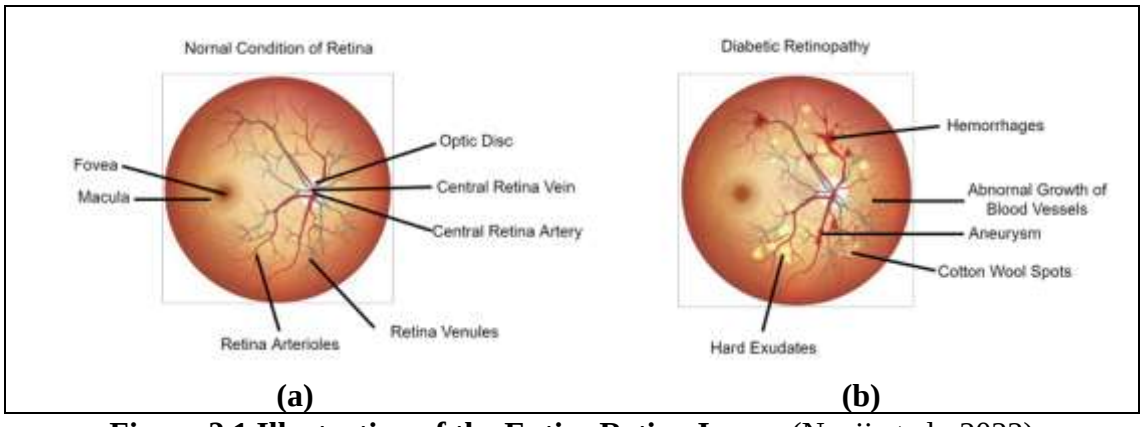


Figure 2.1 Illustration of the Entire Retina Image (Nneji et al., 2022)

At early stage, people with diabetes are often unaware of having DR due to the unnoticeable symptoms. They may start noticing slight symptoms such as blurred vision and the appearance of dark spots or floater as it gets worse. In the long run, they might experience difficulties seeing at night or even sudden loss of vision (Purnama, 2023).

Patients who developed DR and did not receive appropriate treatment will suffer deterioration of their retina and gradually lead to blindness. It is stated that about 5.0% of global blindness cases are caused by diabetic retinopathy, making it the fourth leading cause of blindness worldwide, following cataracts, glaucoma, and macular degeneration.

DR can be classified into several stages based on severity of the impairment. International Clinical Diabetic Retinopathy (ICDR) Severity Scale is the standard used to grade severity of DR, which was developed to simplify the more complex Early Treatment Diabetic Retinopathy Study (ETDRS) scale. There are five stages stated in ICDR, which was based on the presence and extent of specific retinal abnormalities (Kandhasamy et al., 2020; Yang et al., 2022):

a. No Apparent Retinopathy (normal)

   There are signs of damage retina or abnormalities in blood vessels. The patients' vision remains unaffected.

b. Mild Non-Proliferative Diabetic Retinopathy (Mild NPDR)

   When eyes started to have microaneurysms, they are in the early stage of DR which severity is mild. At this stage small red dots in fundus image are usually seen clustered around the macula. Patients with diabetes often ignore the slight symptoms they feel, as they start to feel that their vision is impaired. They do not get proper health care as early detection typically relies on routine retinal screening using fundus photography or optical coherence tomography (OCT) analysed by experts.

c. Moderate Non-Proliferative Diabetic Retinopathy (Moderate NPDR)

   As the symptoms progress, the retina will not only show microaneurysms but also haemorrhages. Hard exudates and cotton wool spots are visible in the retina which indicate wider leakage and damage of blood vessels. These findings indicate more widespread vascular leakage and retinal ischemia which happens when capillaries are blocked or damaged. More visible symptoms are present at this stage where patients may report blurred vision and dark spots especially at night.

d. Severe Non-Proliferative Diabetic Retinopathy (Severe NPDR)

   The critical state of DR is the severe stage which is the most advanced NPDR. This phase is characterized by the presence of widespread haemorrhages, bleeding in at least two regions in the retina, or intraretinal microvascular abnormalities in the retina. At this stage, patients begin to experience more visual symptoms. Their vision started weakening, there are dark spots, increased floaters, and reduced visual acuity. The retina suffers inadequate blood supply due to blockage which becomes extensive and widespread. The risk of progression to the proliferative stage increases significantly once the patients reach severe NPDR.

e. Proliferative Diabetic Retinopathy (PDR)

   The last stage of DR is PDR in which case the retina has severely damaged. New thin blood vessels are formed as a result of blockage in the blood circulation. The symptom is called neovascularization where the abnormal vessels are prone to break causing severe haemorrhage, retinal detachment, and potentially vision loss.

Figure 2.2 shows the progression of DR severity from normal to PDR. It can be seen clearly the development of abnormalities in the fundus images as the severity advanced. From the images, it is clearly visible that there are more spots and blood leakage in the retina as it deteriorates. These changes in the fundus images are important in the feature extraction as they characterized the severity of DR.

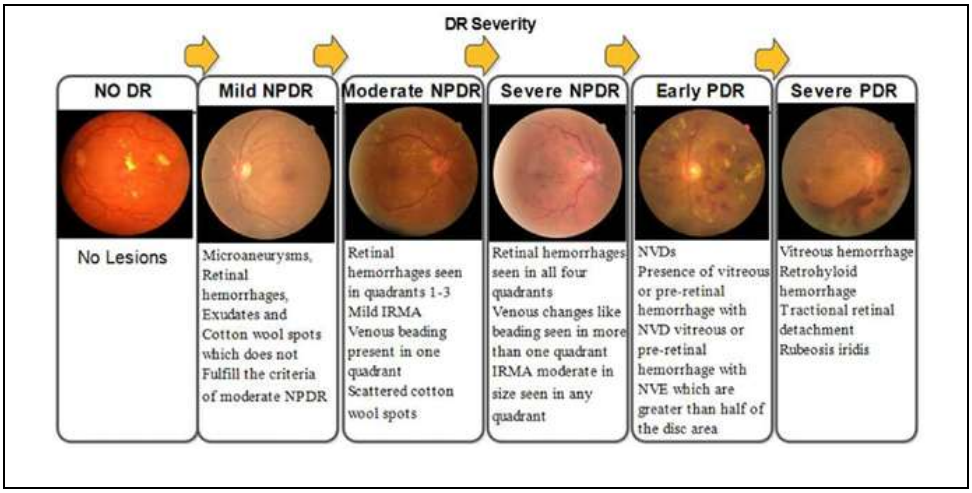


Figure 2.2 Retina Images for Five DR stages and their features (Usman et al., 2023)

2.1.2 Medical Image Analysis

Retinal fundus image is a photograph of the inner back portion of the eyeball visible through the pupil, showing the interior surface at the back of the eye opposite the lens. Referring to Figure 2.1 (a), fundus image consists of retina, optic disc, macula, fovea and retinal blood vessels. Ophthalmologists use fundus image to examine the health of eyes, including the sign of DR. In this case, the fundus images capture visible features such as microaneurysms, haemorrhages, hard exudates, and neovascularization, which are key indicators of DR progression (Lim et al., 2020; Usman et al., 2023). Compared to invasive methods like fluorescein angiography, fundus imaging offers a more patient-friendly approach while still providing valuable clinical information (Oh et al., 2021).

Computer vision technology has rapidly improved over decades, motivating the study of fundus image in medical research. Their analysis has expanded with the advancements in deep learning providing reliable computer-aided diagnosis (CAD) in medical practice with promising outcomes. CNNs as one of the deep learning architectures, were used in many medical studies and have proven to give high accuracy in detecting and classifying DR severity. In some cases, the results were reported to be more accurate than manual assessment by physical observation (Usman et al., 2023; Vardhan et al., 2024). Advancement of fundus camera give a better view of the surface behind the eyeball with ultra-wide-field (UWF) fundus photography. More area in the retina has become more visible and more peripheral lesions can be observed which was often missed in standard images and improving detection sensitivity (Oh et al., 2021).

Manual grading of DR severity from fundus images by ophthalmologists poses several challenges. Among experts, there can be different opinion about the severity level based on fundus images observation. Ophthalmologists with their own expertise and training view fundus image differently, therefore inter-observer variability occurs. In this case, diagnosis are influenced by a clinician's experience, fatigue, or interpretation differences, particularly in borderline cases or poor-quality images (Lim et al., 2020; Taifa et al., 2024a). These manual diagnoses are often inconsistent and are time-consuming, limiting scalability in population-wide screening efforts (Lin and Wu, 2023; Zhang et al., 2025). Other than that, there are more factors affecting the reliable identification of DR severity. Some of the factors are image quality such as illumination artifacts, motion blur, and occlusions that can hinder the identification of DR-specific

features. There can be mistakes in recognising some retinal structures due to similar pixel intensities. These might complicate accurate visual interpretation (Oh et al., 2021; Usman et al., 2023a). To resolve these issues, automated systems based on CNNs can consistently process large datasets, eliminate subjectivity, and reduce the burden on ophthalmologists. These systems also show promise in supporting DR grading by learning to identify subtle patterns in fundus images that may be difficult for human eyes to distinguish (Vardhan et al., 2024; Zhang et al., 2025).

2.1.3 Deep Learning in Medical Imaging

The growth of artificial intelligence technology has stimulated the evolution of machine learning into a more complex approach known as deep learning. This approach learns features and patterns on its own with large data and computing ability. To be able to do so, deep learning employs artificial neural networks with multiple layers to learn hierarchical data representations. Deep learning is able to learn complex pattern automatically and does not rely on handcrafted features. This is one of the reasons deep learning gained popularity as effective tool in analysing unstructured data such as images, audio, and text (Fuad et al., 2021; Xue et al., 2024). The growing technology in computing, availability of large datasets and the development of sophisticated computer architectures have been the driving force behind expansion of deep learning.

Structures of a neural network in deep learning are built to learn and extract three-dimensional features such as edges, textures and shapes of unstructured data. CNNs, as one of the deep learning architectures, is one of the most popular in medical research. It suited for medical imaging application due to its strength in feature extraction. Fundus images has complex structures which makes it difficult for manual feature engineering. Feature selections are critical for identifying changes in retinal fundus images. Their hierarchical structure makes CNNs effective in diabetic retinopathy (DR) detection, where subtle lesions like microaneurysms, haemorrhages, and exudates must be identified with precision (Usman et al., 2023a). CNNs models are designed to extract features automatically from the datasets removing the dependency of manual feature extraction. This ability is highly beneficial in medical research, where accurate identification of lesion impacts directly the diagnosis and prescribed treatment (Zhang et al., 2025).

A significant advancement in computer vision has motivated researcher to build a large database to develop stronger architecture as deep learning needs more samples to train. CNNs has advanced rapidly with the use of transfer learning from the databases built by the collaboration of researchers around the world. This technique involves using models pretrained on large general-purpose datasets (e.g., ImageNet) and fine-tuning them on medical datasets. Studies in medical field are usually hindered by confidentiality and experts' analysis issues, which affects the availability of labelled datasets. Transfer learning addresses the challenge of limited labelled medical images and has been shown to significantly enhance performance in DR classification (Alyoubi et al., 2021; Atwany et al., 2022). Started with handcrafted feature selection, CNNs have evolved rapidly and now there are many pretrained architectures such as ResNet, InceptionV3, DenseNet, and EfficientNet. They have been widely used in DR grading studies as are potentially give high accuracy results. These models, when combined with preprocessing techniques like contrast enhancement, data augmentation, and normalization, demonstrate high accuracy and generalization capabilities (Oh et al., 2021; Pamungkas et al., 2025a).

Hybrid deep learning system have been explored extensively to further enhance the performance of CNN architectures. Individual models are combined with other object

detection methods, such as YOLOv3 to perform both classification and lesion localization, improving interpretability and diagnostic utility (Vardhan et al., 2024). This proven to improve the accuracy of DR classification. Other than that, ensemble learning approaches have also been adopted. Ensemble learning combined multiple models to improve prediction stability and reduce bias. Johnson et al., (2022) proposed an ensemble of DenseNet121 and InceptionResNetV2 for feature extraction, followed by a multilayer perceptron classifier. This model achieved an F1-score above 90% in multiclass DR grading. Similarly, Pamungkas et al. (2025) demonstrated that class balancing with SMOTE increased EfficientNet-B4 performance from 80.61% to 97.78%, highlighting the importance of data handling techniques in clinical settings. In addition to architectural innovations, hyperparameter optimization techniques such as Bayesian optimization are employed to fine-tune model settings like learning rate, batch size, and dropout rates. All these innovative methods have shown a significant improvement in terms of classification accuracy and robustness, especially when they are applied to imbalanced and noisy data (Gupta and Khadka, 2025).

CNNs architecture and their deep learning models, offer potential framework to be used in grading DR severity as they are powerful and accurate in terms of feature extraction and classification. Ensemble strategy that combines several CNNs models with their ability to automatically extract and learn complex features have revolutionized computer-aided diagnosis in ophthalmology. The adoption of these models in this study is thus well-justified, given their proven performance in previous research and strong applicability in real-world clinical settings.

2.1.4 Convolutional Neural Networks

Regular neural networks have only one layer without hidden layers. They are a type of multilayer perceptron which typically refer to fully connected networks. In fully connected networks, each neuron in one layer connects to all neurons in the next layer. Data overfit might happen in this case, as the neural network learns the training samples too well and doesn't generalize well. Unlike regular neural networks, CNNs are different due to the existence of hidden layers. They use a special kind of layer called a convolutional layer which is the main part of a CNN. In image analysis, the first layer of CNNs often detects the edges. As the network goes deeper, layers start to identify more complex features from the training samples. These feature maps are then combined at the end of the network, which helps in making the final decision about the image. Along with convolutional layers, CNNs also include other types of layers like pooling and characterization layers.

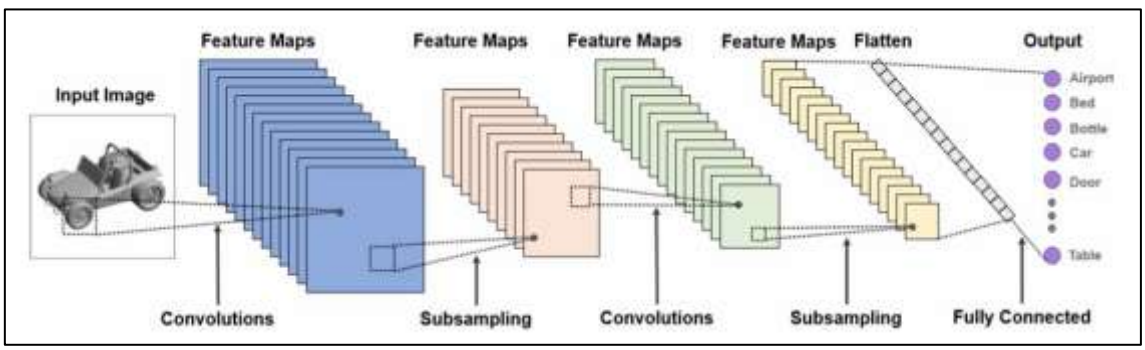


Figure 2. 3 The Standard Convolutional Neural Network (CNN) Architecture
(Alzahrani et al., 2024)

The standard CNN architecture consists of multiple layers as shown in Figure 2.3. Convolutional layers consist of several layers of feature extraction. In these layers image features are filtered and extracted to create feature maps. The outputs of the previous layers are the input for the next layers creating several features and generate the different feature maps. These are the output of the convolutional layer. After being convolved, features are collected as subsamples (Fuad et al., 2021). At pooling or subsampling, the spatial dimensions (width and height) of the images are reduced to lower the network's computational costs and prevents overfitting. This layer does not affect the depth dimension of the feature maps, so the number of feature maps passed to the next layer remains the same (Alzahrani et al., 2024). There are various types of pooling operations, but the most commonly used are max pooling and average pooling. The output of the pooling layer is the generated pooled feature maps. Fully Connected (FC) layers are used after a series of convolutional and pooling layers. In this layer, each neuron is fully connected to all the activation functions in the previous layer. The purpose of FC layers is to convert the 2D pooled feature maps into a flat 1D feature vector for further processing. In classification and recognition tasks, the last FC layer is usually connected to a classifier like softmax, which produces the final output, such as a class label, based on the input image (Alzahrani et al., 2024; Pamungkas et al., 2025a).

CNNs are known for their end-to-end learning, allowing the model to learn relevant features directly from raw input without the need for manual feature engineering. This is particularly important in DR, where subtle differences in fundus images can signal critical changes in disease severity. To enhance performance and robustness, this study employs three well-established CNN architectures: DenseNet201, EfficientNetV2, and InceptionResNetV2. Each model offers distinct advantages and complements the others in an ensemble setup.

1. DenseNet121

Densely Connected Convolutional Network (DenseNet) is one of the CNNs based models that has several architectures such as DenseNet-169, DenseNet201, DenseNet-264 and DenseNet-121. They differ mainly in depth or number of layers and complexity. These networks mostly have unique connectivity pattern where each layer receives input from all previous layers creating a highly compact architecture. Features are being reuse in the training; therefore, fewer parameters are needed than other models. One of the most commonly used is DenseNet-121 as it is considered more efficient and well balanced than other models. DenseNet-121 is able to learn complex pattern deeply, but still efficient to train and deploy on common hardware (Chilukoti et al., 2024).

The output from DenseNet-121 has a dense connectivity allowing it to learn more robust and detailed features using fewer parameters than other CNNs. Hence, it is ideal for deep feature extraction in DR classification. Equation (2.1) expresses mathematically, for a given layer l, the output is defined as:

$$x_l = H_l([x_0, x_1, \dots, x_{l-1}]) \tag{2.1}$$

where

- $[x_0, x_1, \dots, x_{l-1}]$ is the concatenation of all previous layers' feature maps,
- H_1 is a composite function of batch normalization, ReLU, and convolution operations.

Structure of DenseNet121 consists of an initial convolution and pooling layer. There are four dense blocks following the pooling layer where each separated by transition layers. Down sampling is carried in the transition layers. Multiple convolutional layers are compiled in one dense block and the number of layers increases progressively in each block. The transition layers reduce the dimensionality of the feature maps using a 1×1 convolution followed by a 2×2 average pooling layer. At the end, a global average pooling layer aggregates the learned features before passing them to the fully connected classification layer.

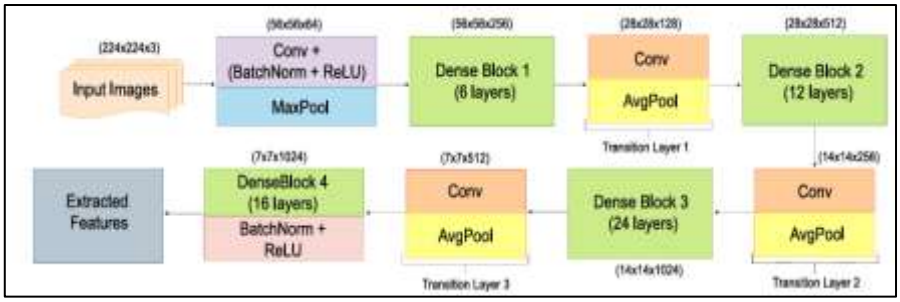


Figure 2. 4 Structure of DenseNet121 model (Taifa et al., 2024a)

Fig.2.4 shows the structure of the DenseNet121 model, that is used in this work. It has four dense blocks to improve the depiction of features. Also, it has three transition layers, each of which has a 1x1 convolution, batch normalization, rectified linear unit(ReLU) Activation to minimize the number of feature mappings, and average pooling to reduce the spatial dimensions' sample size. This study skipped the top (output) layer of the DenseNet121 model for utilizing it as a feature extractor. Finally, it will extract the features from images.

In the context of diabetic retinopathy (DR) classification, DenseNet121 has demonstrated excellent performance. DenseNet121 achieved 97.30% accuracy on the APTOS 2019 dataset when combined with preprocessing techniques such as CLAHE and resizing (Mohanty et al., 2023). Improving quality of the fundus image by integrating ESRGAN-based enhancement in which retinal fine features like microaneurysms and haemorrhages can be restored, increase the accuracy to reach 98.7% (Alwakid et al., 2023). Besides being used in single analysis, DenseNet121 is also frequently used in ensemble learning settings. It is combined with models like InceptionResNetV2 and EfficientNet, to boost classification robustness and generalizability (Chilukoti et al., 2024; Zhang et al., 2025).

DenseNet121 is appropriate to be used in the medical image analysis due to its ability to capture both low-level and high-level features from fundus images. Features used in the training are collected to create a dense connection. Main features extracted from fundus images such as microaneurysms, exudates, and hemorrhages are preserved and utilized throughout the network, enhancing diagnostic accuracy.

2. EfficientNetV2

EfficientNet was developed by Google AI in 2019. It gets the name from its ability to classify objects with fewer parameters; hence faster in training. EfficientNet has achieved high accuracy in many studies involving classification. This model is not only improving its structure, but also optimizing its training to analyse objects which makes it

appropriate for medical image analysis applications such as diabetic retinopathy (DR) grading(Pamungkas et al., 2025).

Combination of Mobile Inverted Bottleneck Convolution (MBConv) and Fused-MBConv blocks is shown in Figure 2.5. These shows that EfficientNetV2 combines the two layers to balance accuracy and training speed. Reduction of the number of parameters is done by MBConv blocks by applying depthwise separable convolutions, while Fused-MBConv blocks merge certain layers for better efficiency in early network stages. Additionally, EfficientNetV2 supports both standard and large kernel sizes, enabling it to capture both fine-grained and global image features effectively, an essential characteristic when analyzing complex structures in retinal fundus images.

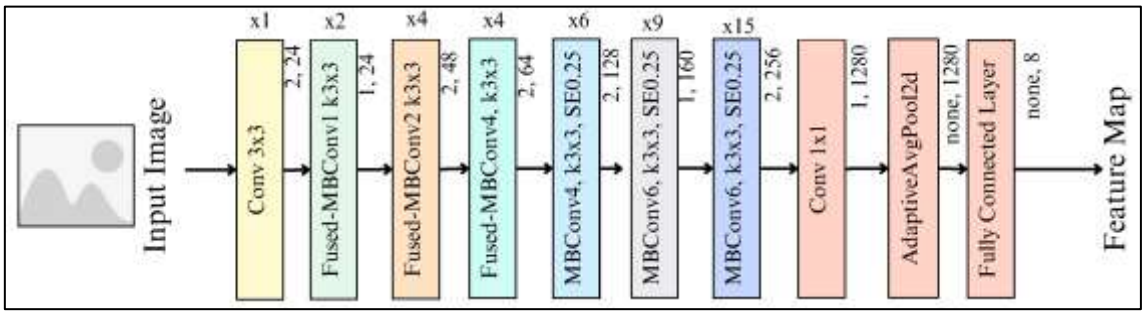


Figure 2. 5 Structure of EfficientNetV2 Model (Aldakhil et al., 2024)

A significant improvement of EfficientNetV2 is the compound scaling method. This enables the depth, width and input resolutions to be scaled uniformly, increasing the model's capacity without overfitting or resource waste. Equation (2.2) shows the mathematical expression of the scaling.

$$d = \alpha^{\phi}, \quad w = \beta^{\phi}, \quad r = \gamma^{\phi} \tag{2.2}$$

Equation (2.2) is subject to the constraint:

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \tag{2.3}$$

Where:

- d is the depth of the network,
- w is the width (number of channels),
- r is the resolution (input image size),
- ϕ is a user-defined coefficient to control scaling,
- α, β, and γ are constants found through grid search.

Having the compound scaling method, EfficientNetV2 has the capacity to learn complex features that makes it ideal for classifying the severity levels of DR based on fine features such as microaneurysms, exudates, and hemorrhages in fundus retina images that are sometimes less visible to the human eyes. The model works in fewer parameter enables it to achieve higher classification accuracy(Pamungkas et al., 2025; Xue et al., 2024). Other than that, EfficientNetV2 is highly compatible transfer learning boosting its performance when trained on medical image datasets such as APTOS and EyePACS. EfficientNet is still accurate even when training data is limited. Pamungkas et al. (2025), reviewed that EfficientNetV2-B4 outperformed several other CNN architectures in

grading DR severity, achieving accuracy above 97% when combined with SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance.

3. InceptionResNetV2

InceptionRestNetV2 is a hybrid CNN architecture that combines Inception module from Google's Inception and residual connection from Microsoft's ResNet. This combination gives InceptionResNetV2 the strength of both base models. Inception modules learn features at multiple scales in parallel making it good for special feature extraction while residual connections allow very deep networks to train more easily by avoiding vanishing gradients (Sushith et al., 2025; Taifa et al., 2024). InceptionResnetV2 merges these benefits into a deep, efficient and high performing CNN.

The Inception module employs multiple parallel convolutional filters (e.g., 1×1, 3×3, 5×5) to capture fine and coarse features simultaneously. This multi-branch architecture helps detect diverse lesion sizes such as microaneurysms and haemorrhages in retinal fundus images (Zhang et al., 2025). InceptionResNetV2 is mathematically expressed in equation (2.4).

$$y = F(x) + x \tag{2.4}$$

where:

- x is the input to the residual block,
- $F(x)$ is the output of the stacked convolutional layers within the block,
- y is the final output of the block after adding the shortcut connection.

The structure of InceptionResNetV2 allows the model to learn residual mappings, stabilizing training and improving convergence in deeper layers. The InceptionResNetV2 architecture has several components which are stem layer, Inception-Resnet blocks, reducation blocks and global average pooling. Stem layers consists of a series of convolutions and max-polling layers in which initial feature extraction is carried out. There are three type of Inception-Resnet blocks, each with distinct convolutional filter paths and residual connections. Reduction of the spatial dimensions while increasing dept is done in the reduction blocks followed by global average pooling. Final path is the dense and softmax layers for classification.

This architecture has been successfully applied in diabetic retinopathy classification. InceptionResNetV2 in a hybrid ensemble with DenseNet and MobileNetV2 achieved about 97.27% accuracy across severity levels (Taifa et al., 2024). Similarly, interpretability and robustness of InceptionResNetV2 when combined with transformer-based models is shown, highlighting its ability to extract detailed features while maintaining generalization (Zhang et al., 2025).

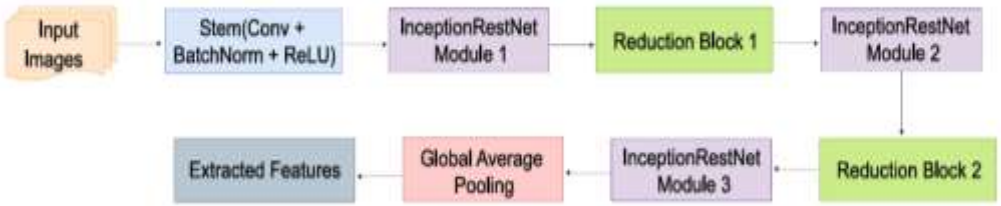


Figure 2. 6 Structure of InceptionResNetV2 model

InceptionResNetV2 has the ability to capture a broad range of features at various levels of abstraction. This makes it very useful for feature extraction, including in medical image analysis. Fig. 2.6 shows the structure of the InceptionRes NetV2 model that is used in this study. It starts with a stem made up of ReLU, Batch Normalization, and Convolution. Three InceptionResNet modules make up the remaining portion. Several parallel convolutional layers make up these modules. This enables the network to concurrently capture characteristics at different scales and levels of complexity. Additionally, it features reduction blocks, which are usually made up of pooling layers and 1 × 1 convolutions. By using these blocks, the number of feature maps can be increased while the spatial dimensions are decreased. Finally, by skipping the top (output) layer of the model, we will get the extracted feature. Given its multi-scale receptive field and residual stabilization, InceptionResNetV2 is highly effective for analyzing medical images with complex and hierarchical features, making it a strong backbone in ensemble DR classification models.

2.1.5 Ensemble Method

Ensemble learning combines predictions from multiple models to produce a better-performing final model. The objective of ensemble methods is to enhance the robustness of an estimator by integrating forecasts of multiple learning algorithms. In ensemble learning, predictions from multiple base learners are combined to improve predictive performance beyond that of single model. There are various methods to combine the several models after each being trained independently. Single models of CNN-based architectures, either homogeneous or heterogeneous can be ensembled by majority voting, weighted averaging or stacking. These methods help mitigate individual model biases and variances to produce more reliable predictions (Akram et al., 2025).

The diversity of the constituent models highly affects the prediction performance of an ensemble. When several CNN architectures such as ResNet-50, MobileNet, EfficientNet, and custom CNNs are integrated in an ensemble, their individual performance affects the final prediction. In this case, each model learns different aspects or levels of abstraction from the input data due to their unique depth, layer configurations, and parameterization (Desiani et al., 2024). Model diversity can be enhanced further by training them with wide-ranging data augmentations, initial weights, and learning schedules.

There are many different ensemble learning strategies, specifically ways to combine multiple models to improve prediction performance. Several of them are:

- Majority Voting
  In majority voting, the final prediction is determined by using the most common class label predicted by the base models. Prediction from each model will be taken into account and the majority will decide the decision. This method is simple; however, it has weakness, the final prediction will be accurate only when model accuracies are comparable.
- Weighted Voting
  Different from majority voting, weighted voting gives its final predictions based on a weight assigned to each model. This means that each model's past performance is important and given more weight in making decision. This technique is useful when certain models consistently perform better.

- Stacking
  Stacking method uses a meta-learner, where outputs from multiple base models are fed into the meta-learner to determine the final classification. There are many types of meta learners that can be used in stacking method, such as logistic regression, linear regression, decision trees, random forest, boosting models, shallow neural networks and support vector machines.
- Branch-Based Ensemble (Hierarchical CNNs)
  CNN architectures can be divided into multiple branches at different depth levels, capturing features at various spatial resolutions. These branches are then ensembled to enhance feature richness and flexibility.

In this research, ensemble techniques chosen for classifying DR is stacking ensemble. This is due to its ability to learn how to optimally combine predictions from multiple base models, rather than simply aggregating them through fixed rules like majority or weighted voting. Stacking ensemble uses a meta-learner that learns from the softmax output of each base CNN model. This enables the system to recognize and adapt to patterns in model perf performance, leading to more accurate and robust predictions (Ergun and Ilhan, 2023). Besides, the differences in the fundus image features such as microaneurysms or haemorrhages are sometimes subtle that makes it difficult to distinguish. These are often better captured when combining multiple models with different strengths (e.g., DenseNet50, EfficientNetV2, InceptionResNetV2). A stacking ensemble allows the meta-classifier to intelligently leverage these complementary strengths, outperforming traditional ensemble techniques that do not involve learning.

2.1.6 Stacking Ensemble Meta-Classifier Design

Stacking ensemble learning strategy is chosen to improve predictive accuracy and robustness in medical classification tasks. Stacking (or stacked generalization) is a two-level model architecture where multiple base classifiers are trained in parallel, in this case, DenseNet121, EfficientNetV2 and InceptionRestnet50. These base learners were chosen due to their proven performance in various ensemble learning studies for medical prediction tasks, including stress level detection (Fonda et al., 2024), diabetes classification (Ali et al., 2022) and emergency heart disease readmission prediction (Ghasemieh et al., 2023).

Empirical studies confirm the benefits of ensemble approaches for DR detection and grading. When multiple CNN models were integrated into an ensemble, they were successfully classify DR stages with an AUC score of 0.88. This means that the ensemble is able to differentiate the severity level with an accuracy of 88% (Hendrawan et al., 2024). Similarly, CNN-based stacking ensemble with self-adjusting classifier weights reached a higher accuracy of 98.35% (Nadda et al., 2025). An ensemble of EfficientNet achieved 95% accuracy and 97% recall which is higher than thirteen established CNN architectures for DR classification (Shelke et al., 2025). These results highlight the capacity of stacking ensembles to integrate multiple CNNs and enhance performance across all DR severity levels.

Stacking ensembles has shown significant benefits in medical image analysis. In terms of diabetic retinopathy, they can improve early-stage prediction which is critical for further medication to prevent vision loss. Stacking ensemble combine models that are individually more sensitive to subtle lesions such as microaneurysms or early haemorrhages and determine if the retina shows any sign of DR (Deepa et al., 2022). This

CAD is valuable in large-scale screening in health care facilities which are lacking in resources and experts as the results are accurate (Hendrawan et al., 2024; Shelke et al., 2025).

Stacking is well suited for CNNs-based ensemble particularly in the context of deep learning and computer vision. This is due to the ability of different architectures to extract information from various dimension of images. For example, DenseNet is strong at feature extraction, reusing them in the final determination and gradient flow, InceptionResNetV2 combines multi-scale feature extraction with residual learning, and EfficientNetV2 balances scaling of depth, width, and resolution for efficiency (Arora et al., 2024; Tummala et al., 2023). By integrating these architectures in a stacking framework, it is possible to take advantage of their complementary strengths for more accurate diabetic retinopathy grading.

Three diverse and widely accepted base classifiers were selected to be tested in this research.

1. Random Forest (RF)
   RF uses bagging with decision tress as base models. It is selected to enhance classification performance as a number of decision trees are combined to create RF. Final decision is determined by using the majority voting method. By using RF, more randomness is incorporated on decision tree predictors to obtain more diverse classifier.
2. Extreme Gradient Boosting (XGBoost)
   XGBoost is one of boosting techniques that handles feature extraction and organizing them efficiently. XGBoost created parallel tree boosting effectively and determine the final decision based on the gradient-boosted decision tree.
3. Multilayer Perceptron (MLP)
   MLP is a shallow neural network classifier used to determine outputs of CNNs that are complex. It can model non-linear relationships between base models. MLP consists of at least an input layer, a hidden layer and an output layer.

This research is planned to use XGBoost due to its consistent performance in enhancing ensemble by effectively capturing higher-order interactions from the base learners' predictions. XGBoost is employed as a meta-learner in a stacking ensemble is able to outperform traditional classifiers in predicting emergency readmissions for heart disease patients (Ghasemieh et al., 2023). Another research by Fonda et al. (2024) used several classifiers such as RF, MLP, and SVM with XGBoost to classify stress level and achieved an accuracy of 95%. This shows that XGBoost has the potential to be implemented in ensemble of CNN-based models to classify DR.

2.1.7 Performance Metrics

Evaluation is an important part of building any automatic system, in this case classification of DR severity level. Performance metrics are crucial to evaluate the reliability of deep learning models used for DR classification from fundus retinal images. In medical image analysis, firm evaluation is important to ensure that the model meets clinical accuracy requirements and safety. Performance metrics provide a quantitative assessment basis to measure a model's predictive capabilities. Other than that, they can facilitate model selection, hyperparameter tuning, and comparative analysis between algorithms (Schlosser et al., 2024).

In classification tasks for DR detection, performance metrics are derived from the confusion matrix. The results of the computation include false negatives (FN), false positives (FP), true positives (TP) and true negatives (TN) which then arranged in a matrix that can be used to evaluate the system's performance. Accuracy, precision, recall, and specificity are several metrics commonly used in the performance evaluation (Akhtar et al., 2025; Bilal et al., 2022; Blair et al., 2023; Sebti et al., 2022). Other than those metrics, F1-score (Akhtar et al., 2025; Avolio et al., 2023) and ROC-AUC are used as additional evaluation (Dai et al., 2021). The existence of data imbalance makes measuring only accuracy for the overall proportion of correct predictions may not provide adequate insight. Therefore, other evaluation metric is needed. Precision will give better understanding, as it evaluates the proportion of true positive predictions among all positive predictions. Other than that, recall enhance evaluation as it measures the proportion of actual positives correctly identified, are often prioritized in clinical applications (Müller et al., 2022).

Commonly found in medical image, datasets imbalance usually poses challenges. Similarly, in diabetic retinopathy classification where class imbalance and subtle pathological features are present, combining metrics such as the F1-score and ROC-AUC enhances the robustness of model evaluation. The F1-score is important when the value FN is high as it gives the harmonic mean of precision and recall. Area under the curve analysis shows the model's discrimination abilities across all edges (Terven et al., 2025; Vujović, 2021) which can be used to determine if the model can distinguish different classes accurately.

Confusion-matrix metrics, ROC-based evaluations, and statistical tests are standard metric performance which are crucial to ensure reproducibility and fair model (comparison Rainio et al., 2024; Schlosser et al., 2024). However, interpretability and transparency in model evaluation has gained more attention beyond the standard quantitative evaluation. Metrics designed for saliency-based explanation methods assess the alignment between model focus and relevant image regions, supporting explainable AI in clinical settings (Fresz et al., 2024). With the high risk of medical analysis, improper metric selection or biased implementations can threaten the credibility and utility of diagnostic models. Müller et al. (2022) emphasize the need for careful metric application, especially in studies reporting inflated performance.

Performance evaluation metrics should be determined appropriately and adequately as they are vital in diabetic retinopathy (DR) classification, especially when using deep learning models such as convolutional neural networks (CNNs) and their ensembles. DR grading is typically a multi-class and ordinal classification problem; therefore, it is essential to assess both the overall accuracy of predictions and the model's ability to correctly identify each class, particularly in imbalanced datasets where misclassification of minority classes can have serious clinical consequences. To revolve these challenges, evaluation metrics such as accuracy, sensitivity, specificity, F1-score, Cohen's Kappa, and Area Under the ROC Curve (AUC) are included in the performance evaluation. These metrics provide a comprehensive understanding of classification performance across all classes, making them especially suitable for medical image analysis.

1. Confusion Matrix-Based Metrics

The confusion matrix is constructed based on the calculation of TP, TN, FN and TN. It allows for detailed performance analysis in classification problems, especially multi-class settings. Key metrics derived from it include:

- Accuracy:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (2.5)$$

- Sensitivity (Recall):

$$Sensitivity = \frac{TP}{TP+FN} \times 100\% \quad (2.6)$$

- Specificity:

$$Specificity = \frac{TNN}{TN+FP} \times 100\% \quad (2.7)$$

- F1-Score:

$$F1 - Score = \frac{2TP}{2TP+FP+FN} \times 100\% \quad (2.8)$$

- Cohen's Kappa:

$$\kappa = \frac{2(TP \cdot TN - FN \cdot FP)}{(TP+FP)(FP+TN)+(TP+FN)(FN+TN)} \times 100\% \quad (2.9)$$

The performance of a classification model can be thoroughly evaluated using several metrics derived from the confusion matrix. Accuracy (Eq. 2.5) is a measure of correctly classified instances. This is the most common measure used to explain the performance of classification task. While sensitivity or recall (Eq. 2.6) emphasizes the model's ability to correctly detect positive cases. In the other hand, specificity (Eq. 2.7) highlights the ability to identify negative cases accurately. Common problem for datasets with imbalanced classes is resolved by calculating the F1-Score (Eq. 2.8) which is a combination of precision and recall. Cohen's Kappa (Eq. 2.9) is added to account for the agreement level between predicted and actual classifications beyond chance, making it a more robust measure of performance. These metrics were applied in various studies. Desiani et al. (2024) used them to compare individual CNN models and a weighted ensemble combining EfficientNet, ResNet-50, and MobileNet. The ensemble achieved 93.3% accuracy, 93.42% F1-score, and Cohen's Kappa of 0.866, demonstrating substantial improvements over individual models.

2. ROC Curve and AUC

The ROC (Receiver Operating Characteristic) curve is a graphical analysis used to visualize the performance of a classification model across various decision thresholds. The graph plots the True Positive Rate (TPR) against the False Positive Rate (FPR):

- True Positive Rate (TPR):

$$TPR = \frac{TP}{TP+FN} \quad (2.10)$$

- False Positive Rate (FPR)

$$FPR = \frac{FP}{FP+TN} \quad (2.11)$$

The AUC can be mathematically defined as:

$$AUC = \frac{1}{|S^+||S^-|} \sum_{s^+ \in S^+} \sum_{s^- \in S^-} \mathrm{II}(s^+ > s^-) \quad (2.12)$$

Where:

S^+ and S^- are the sets of scores for positive and negative samples, respectively.

$II(s^+ > s^-)$ is an indicator function equal to 1 if the positive score is higher than the negative, and 0 otherwise.

True Positive Rate (Eq. 2.10) and the False Positive Rate (Eq. 2.11) are plotted to create ROC graph. This plotting helps to analyse the trade-off between sensitivity and false positive. The AUC is shown in a number between 0 and 1, it represents the probability that the model assigns a higher score to a randomly chosen positive instance than to a negative one (Eq. 2.12). A higher value of AUC that is closer to 1 shows a stronger discriminative ability to classify objects. Sufficient evaluation metrics is crucial for guaranteeing the reliable performance analysis in classifying DR severity level by using CNN ensemble models. Standardized metrics including accuracy, F1-score, and Cohen's Kappa potentially show detailed insights into classification balance and consistency, while AUC provides a threshold-independent assessment.

2.2 Previous Studies

As one of the major complications of diabetes, DR poses as leading cause of vision loss and blindness if it is undetected and left untreated at an early stage. The rising number of cases of individuals diagnosed with diabetes globally, there is an urgent demand for fast, accurate, and reliable early detection systems. Advancement in AI technology, including in deep learning methods, open new opportunities to develop systems to assist in the automated diagnosis and classification of DR through retinal image analysis.

This potential results in numerous studies have been conducted to design automated system and to evaluate the effectiveness of various AI algorithms in detecting DR. Table 2.1 has been constructed to show the literature review of research in DR classification with different CNN architectures. The table is expected to offer an overview of the research methods, objectives, and outcomes which focus on accuracies from relevant sources.

A wide range of deep learning models for classification of DR severity levels has been seen growing recently, particularly convolutional neural networks (CNNs). Among the models, ResNet-based architectures are one of the most commonly used due to their residual learning capabilities and robustness in training. A study shows that ResNet-50 combined with Random Forest achieved 96% accuracy on the Messidor-2 dataset (Yaqoob et al., 2021), while a revised ResNet-50 version reached a training accuracy of 83.95% on a large Kaggle dataset (Lin and Wu, 2023). More advanced versions, such as ResNet152, also demonstrated strong performance with 94.40% accuracy (Usman et al., 2023).

Other than family of ResNet, DenseNet is also common in DR classification. These variants, known for efficient feature reuse, reducing the parameter used in the system. DenseNet121 and DenseNet201 achieved high accuracy across multiple datasets, including 98.36% accuracy reported for DenseNet121 on Kaggle data (Taifa et al., 2024) and 97.68% when enhanced with Bayesian uncertainty quantification (Akram et al., 2025). Another architecture comparable to RestNet and DenseNet is the InceptionV3. It consistently performed well, showing 95.9% precision and 97% accuracy across both EyePACS and IDRiD datasets (B P et al., 2023; Sushith et al., 2025). Hybrid models incorporating InceptionResNetV2, MobileNetV2, and DenseNet also achieved performance over 97%, highlighting the strength of ensemble-based approaches (Taifa et al., 2024).

Beyond the standard image pre-processing, several studies also introduced advanced techniques such as Principal Component Analysis (PCA), edge enhancement, and adaptive histogram equalization to improve image quality before training in the hope of better prediction performance. Further, ordinal evaluation metrics like Quadratic Weighted Kappa (QWK) were emphasized in studies by Chilukoti et al. (2024) and Zhang et al. (2025) to better reflect the severity-aware performance of DR grading models. Studies reviewed shows that CNN architectures such as ResNet50/152, DenseNet121/201, and InceptionV3 remain leading in the CNN models that are highly effective for DR detection. Models that combined pretrained CNN frameworks with techniques like Bayesian inference, ensemble learning, or ordinal loss evaluation tend to outperform single-model individual results. Reviewed studies also show that DenseNet121 and InceptionV3 are the most commonly models that give high accuracy across different datasets and evaluation metrics.

Classification of DR has increasingly focused on the integration of ensemble deep learning methods to improve diagnostic accuracy and robustness beyond the individual model. A common approach among this research is the use of multiple pre-trained convolutional neural networks (CNNs), such as ResNet, EfficientNet, MobileNet, Inception, and VGG architectures in an ensemble. The final prediction is derived by combining the models by stacking, weighted voting, or feature-level fusion. These ensemble strategies aim to exploit the strengths of different models and anticipates individual weaknesses, which in the end enhancing performance across multiclass DR grading tasks.

Fundus image dataset such as APTOS 2019, EyePACS and IDRiD are available publicly and often used as the primary dataset. APTOS 2019 is the most commonly used primary dataset in many studies it provides standardize image with reliable annotation. Fundus images in the public datasets usually are varies in terms of size, resolution and lighting, therefore before training, the images have to go through pre-processing to ensure they are appropriate for the training. There are many pre-processing techniques like image resizing, normalization, contrast enhancement, and augmentation which are frequently applied to improve image quality and images uniformity (Pamungkas et al., 2025; Saxena et al., 2024). In the case of imbalance dataset, Pamungkas et al. (2025) applied SMOTE oversampling to address the problem which significantly increasing the classification accuracy from 80.61% to 97.78%.

Saxena et al. (2024) introduce new approach by incorporating AI in the system build to classify DR severity level. DRNET ensemble, which integrates ResNet50, AlexNet, and GoogleNet, were proposed and not only delivers high classification performance but also provides interpretable visual outputs and uncertainty scores, enhancing the model's clinical trustworthiness. Another study by Handoyo and Kautsar (2022) who applied a stacking ensemble technique using EfficientNet and MobileNetV3 as base models, used ANNs and SVMs meta-learners has proven to achieve 84.17% accuracy. This study emphasises the advantage of combining deep features with traditional classifiers.

Another study by Deepa and Venkatesan (2022) suggested an ensemble with a multi-stage patch-based combining InceptionV3 and Xception networks, which achieved 96.2% accuracy. The detection of subtle lesions is possible with the usage of localized image patches to address a critical challenge in DR diagnosis. Desiani et al. (2024) proposed a weighted voting ensemble using ResNet-50, MobileNet, and EfficientNet, reaching 93.3% accuracy and 93.42% F1-score. Each classifier's contribution is adjusted dynamically to improve the decision consistency.

Table 2. 1 Literature Review on Diabetic Retinopathy Classification

Author, Title, Year	Method	Dataset	Results
AbdelMaksoud et al, (2020) Diabetic Retinopathy Grading System Based on Transfer Learning.	Customized EfficientNet B0	IDRiD	Accuracy: 86%DSC: 78.45% (train), 65% (validation)
Yaqoob, et al, (2021) ResNet Based Deep Features and RandomForest Classifier for Diabetic Retinopathy Detection. (2021)	ResNet-50, VGG-19, Inception-v3, MobileNet, Xception, VGG16	Eyepacs Messidor2	Accuracy of 96% on the Messidor-2 dataset Accuracy 75.09% on the Eyepacs
Kandimalla, et al,(2022) Screening and Staging of Diabetic Retinopathy Using Convolution Neural Networks.	ResNet-50	Kaggle DR	Accuracy: 74.2%Compared: AlexNet (57%), E-Net (66%)
Handoyo, et al, (2022) Severity Classification of Diabetic Retinopathy Using Ensemble Stacking Method.	Ensemble Stacking (EfficientNet, MobileNetV3, etc.) with meta-learners (ANN, SVM)	APTOS	Accuracy: 84.17%, F1-score: 70.16% (test set)
Aishwarya, et al, (2023). *Diabetic retinopathy using Inception V3 model*.	Inception V3	eyePACs	Accuracy 97%
Usman, et al, (2023) *Diabetic retinopathy detection using principal component analysis multi-label feature extraction and classification*	Transfer learning with ResNet50, ResNet152, and SqueezeNet1	RMU, Pakistan	Accuracy of 93.67% for ResNet 50 Accuracy of 91.94% for Squeezenet1 Accuracy of 94.40% for ResNet 152
Lin et al, (2023) *Development of revised ResNet-50 for diabetic retinopathy detection*.	ResNet-50	Kaggle dataset	Train accuracy: 0.8395 and Test accuracy: 0.7432
Malik, et al, (2024) *Inception-v3 vs. DenseNet for automated detecon of diabetic retinopathy*.	DenseNet, Inception V3	APTOS	Accuracy : Densenet : 89.2%, Inception : 95.9% Precision : Densenet : 89.6% Inception :95.9% Recall : densenet 89.3% Inception : 95.7%
Sathiya, et al, (2024) *Detection and classification of diabetic retinopathy using Inception V3 and Xception architectures*	CNN with Inception V3 and Xception	IDRiD	Accuracy : 97%
Taifa, I. et al, (2024) *A hybrid approach with customized machine learning classifiers and multiple feature extractors for enhancing diabetic retinopathy detection*.	DenseNet121, MobileNetV2 and InceptionResNetV2	Kaggle data	Accuracy DenseNet121 : 98.36%

Author, Title, Year	Method	Dataset	Results
			MobileNetV2 : 97.82% InceptionResNetv2 : 97.27%
Chilukoti et al, (2024) *A reliable diabetic retinopathy grading via transfer learning and ensemble learning with quadratic weighted kappa metric*.	EfficientNet-B3, VGG, ResNet	Eyepacs APTOS Messidor-2	QWK : Eyepacs Chilukoti et al. 0.901 APTOS 0.967 Messidor-2 0.944
Saxena et al, (2024). Deep learning ensemble framework for multiclass diabetic retinopathy classification.	DRNET: Ensemble of ResNet50, AlexNet, GoogleNet	APTOS	Accuracy : Boosting ensemble 90.2 % Stacking ensemble 88.1% Bagging ensemble 97.3%
Desiani, et al, (2024). Weighted Voting Ensemble for Enhanced DR Classification Using CNN.	Ensemble: ResNet-50, MobileNet, EfficientNet with weighted voting	APTOS + EyePACS	Accuracy: 93.3%, F1: 93.42%, Kappa: 0.866
Sundar, et al, (2024) Classification of Diabetic Retinopathy Disease Levels by Extracting Spectral Features Using Wavelet CNN.	Wavelet CNN with SVM, XGBoost, and Random Forest classifiers	EyePACS	Best result with SVM:Accuracy: 98.3%F1-score: 0.983AUC: 0.976–0.978
Zhang, et al, (2025). *Interpretable deep learning for diabetic retinopathy: A comparative study of CNN, ViT, and hybrid architectures*.	CNNs (ResNet-50), ViTs (Vision Transformer and SwinV2-Tiny), and hybrid models (Convolutional Vision Trans former, LeViT-256, and CvT-13)	eyePacs APTOS	The best-performing model (CvT-13) achieved a Quadratic Weighted Kappa (QWK) score of 0.84 and an AUC of 0.93 on the test set.
Akram, et al, (2025) *Uncertainty-aware diabetic retinopathy detection using deep learning enhanced by Bayesian approaches*.	CNN DenseNet with Bayesian approach	APTOS DDR	Bayesian-augmented DenseNet-121 test accuracy: 97.68%
Pamungkas, et al, (2025). Enhancing Diabetic Retinopathy Classification in Fundus Images using CNN Architectures and Oversampling Technique.	CNN (EfficientNet-B4, ResNet-50, DenseNet-201, Xception, Inception-ResNet-v2) with SMOTE	APTOS	Accuracy improved from 80.61% to 97.78% with SMOTE

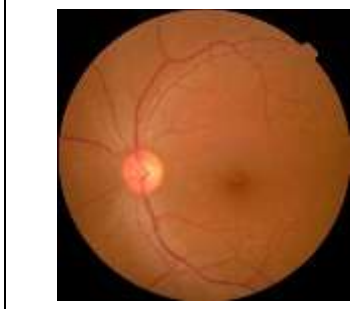
2.3 Research Object

The research object of this study is fundus images taken from the public datasets. It is a two-dimensional picture showing the interior surface of the eye that capture important anatomical features such as the retina, optic disc, macula, and retinal vasculature. Fundus images are usually obtained using colour fundus photography (CFP), an image gathering technique that is widely used in ophthalmology. As a non-invasive procedure, fundus photography is accessible and effective in providing fundus images. They are used as a primary tool for the screening and diagnosis of diabetic retinopathy (DR) based on the visible markers of DR (Gupta and Khadka, 2025).

Retinas suffering from DR gradually show changes. Fundus images demonstrate the subtle changes in the nervous system and surrounding conditions. The important features such as microaneurysms, haemorrhages, exudates, spots, and neovascularization are visible in the fundus images. The existence of these markers are used to determine the severity stage of DR based on International Clinical Diabetic Retinopathy Disease Severity Scale (ICDRS) and the Early Treatment Diabetic Retinopathy Study (ETDRS) (Chilukoti et al., 2024; Zhang et al., 2025).

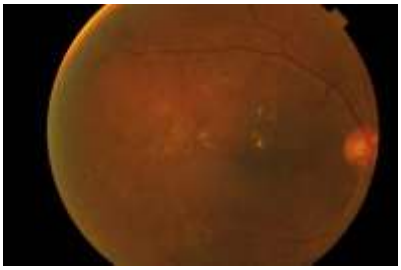
Computer vision advancement combined with deep learning, give fundus images a significant opportunity in developing automated DR detection and classification systems. CNNs have shown outstanding performance in learning complex patterns in fundus images and identifying subtle lesions that may not be easily visible to the human eye (Pamungkas et al., 2025; Usman et al., 2023). These models effectively extract spatial features like texture, color, and vascular structure directly from raw image pixels, eliminating the need for handcrafted features.

Recent research has demonstrated the effectiveness of using CNN-based architectures like DenseNet, ResNet, and EfficientNet on fundus image datasets such as APTOS 2019 and EyePACS. These models achieved high classification accuracy across multiple DR severity levels (Chilukoti et al., 2024; Pamungkas et al., 2025). To improve the classification performance, pre-processing techniques are given more attention. Preprocessing techniques such as contrast enhancement (CLAHE), image normalization, and cropping are often applied before training to improve the quality of input images and enhance lesion visibility (Atwany et al., 2022; Oh et al., 2021). The pre-processing is important as fundus images provide information necessary for accurate diabetic retinopathy classification. Their ability to visualize lesion-specific patterns, combined with the power of deep learning, makes them an ideal modality for both clinical screening and automated DR grading systems. Figure 2.7 shows the changes in the retina due to increasing severity of diabetic retinopathy.



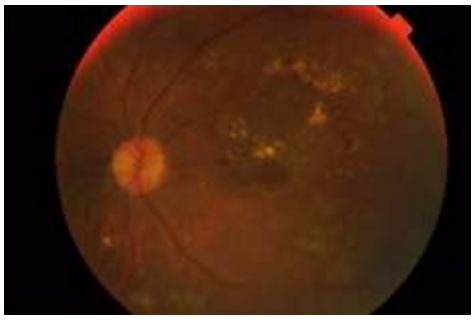
(a) No DR

(b) Mild DR



(c) Moderate DR

(d) Severe DR



(e) PDR

Figure 2. 7 Sample of Original Images from each class from the APTOS 2019 dataset

2.4 Conceptual Framework

A conceptual framework is constructed to support the building of an ensemble-based deep learning model in classifying DR severity level using fundus images. In this research, two datasets will be used, the APTOS 2019 and DRTiD. Both are annotated public datasets where APTOS 2019 is the most commonly used images and DRTiD is the newer datasets collected in 2023. Figure 2.8 shows that the framework starts with the datasets as the input, both categorized based on ICDR grading scale ranging from Level 0 (No DR) to Level 4 (Proliferative DR). The next step is enhancing image quality to highlight important features such as microaneurysms and hemorrhages. This pre-processing uses several techniques including CLAHE (Contrast Limited Adaptive Histogram Equalization), Gaussian filtering for noise reduction, and green channel extraction for vascular detail enhancement.

After pre-processing, the data is split into training and testing datasets and futher processed in parallel through three state-of-the-art Convolutional Neural Network (CNN) architectures: DenseNet201, InceptionResNetV2, and EfficientNetB2. These models are selected for each individual strength in deep feature extraction, multi-scale lesion recognition, and computational efficiency. Feature maps are produced by each model, which is then fused by using softmax method to collect a more diverse representations of retinal abnormalities. The combined feature output is passed to a meta-classifier, Random Forest, XGBoost and MLP which is responsible for the final classification of DR severity levels.

Multiple performance metrics are used to assess the effectiveness and clinical viability of the model. These include accuracy, F-1 score and confusion matrix for classification performance, along with ROC/AUC. They are used to improve predictive

accuracy of the model and also to ensure the model's relevance in clinical decision-making by reducing both false positives and false negatives in DR diagnosis.

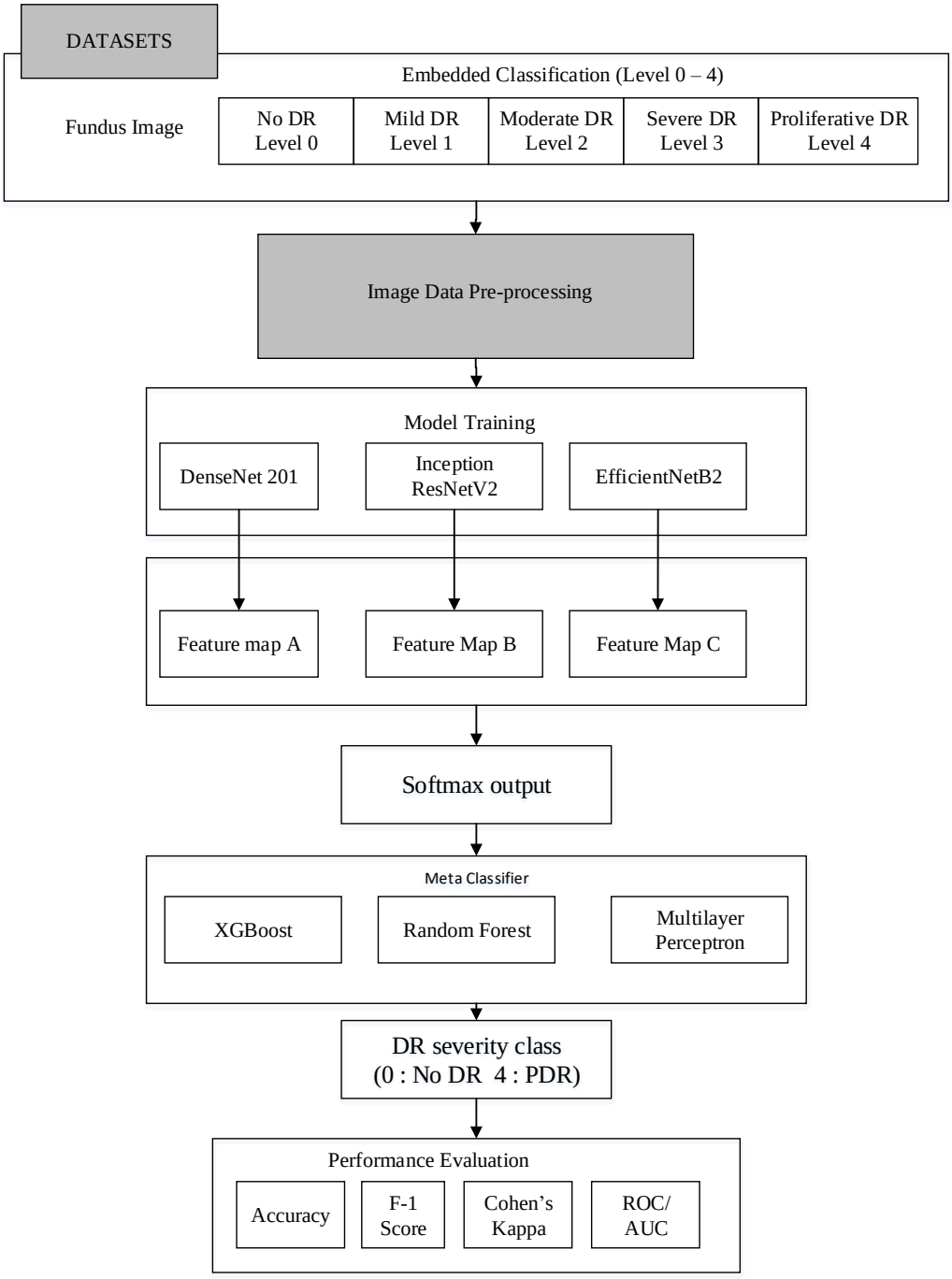


Figure 2. 8 Dataset Processing and Modelling Diagram

2.5 Statement of Hypothesis

Hypothesis formulated in this research is if a CNN-based ensemble model integrating DenseNet201, EfficientNetV2, and InceptionResNetV2 is used for diabetic retinopathy grading, then it will achieve significantly higher accuracy compared to each individual CNN model.

Chapter III
Research Design

3.1 Methodology

3.1.1 Datasets

Over decades, ophthalmologists from many countries collected fundus images from hospitals to create datasets that are available for public, particularly for research purpose including diabetic retinopathy detection and classification. They origin from different countries and have been collected and updated over decades. These image datasets have played a significant role in the advancement of deep learning-based diagnostic systems in medical field. The images are labelled and annotated by experts. The diversity of the datasets includes images of retinas without DR and those showing biomarkers of varying degrees of DR severity. Other than the severity level, quality, and size of these datasets are also varies making them essential for training and validating machine learning models.

As shown in Table 3.1, several standard retinal fundus image datasets that have been widely used in diabetic retinopathy research, are EyePACS, E-Ophtha, Messidor-2, IDRiD, APTOS, DDR, and DRTiD. There is variation on the datasets in terms of the number of sample images provided, label, and type of image. All datasets are intended to study diabetic retinopathy, including severity grading (0–4 levels), binary classification (DR and non-DR), and lesion detection such as microaneurysms, exudates, and haemorrhages. For example, APTOS 2019 and EyePACS follow five level severity grading scale based on the international standard, which can be used in parallel for standardized training and evaluation. On the other hand, IDRiD provide a fine lesion annotation, supporting not only classification but also segmentation and localization tasks.

Table 3. 1 Comparison of Standard Retinal Fundus Image Datasets for Diabetic Retinopathy

Dataset	Year Published	No. of Images
EyePACS	2009	88,702
E-Ophtha	2013	~200
Messidor-2	2016	1,748
IDRiD	2018	516
APTOS	2019	3,662
DDR (DeepDR)	2022	13,673
DRTiD	2023	3,100

The Asia Pacific Tele-Ophthalmology Society 2019 Blindness Detection (APTOS 2019 BD) dataset contains 3,662 samples collected from participants in rural India, organized by Aravind Eye Hospital. It is commonly used for training and testing models that classify the severity of the disease. This research chose APTOS 2019 Blindness Detection dataset from Kaggle as its primary datasets as it is a standardized benchmark in diabetic retinopathy research. The dataset consists of high-resolution retinal fundus images labelled according to the International Clinical Diabetic Retinopathy (ICDR) severity scale, which categorizes DR into five levels: 0 (No DR), 1 (Mild), 2 (Moderate), 3 (Severe), and 4 (Proliferative DR).

To improve reliability, this research also incorporates a more recent dataset collected in 2023 from the Shanghai Diabetic Eye Study called Diabetic Retinopathy Two-field image Dataset (DRTiD). They consist of 3,100 fundus images. This dataset provides a valuable comparison to APTOS in terms of size and relevance. The distribution of diabetic retinopathy severity levels in both APTOS and DRTiD is shown in Table 3.2.

Table 3. 2 Distribution of DR Severity Levels in the Training Set

Diabetic Retinopathy Stage	Label	Number of Samples	
		APTOS	DRTiD
Normal (No DR)	0	1805	1494
Mild DR	1	370	280
Moderate DR	2	999	812
Severe DR	3	193	398
Proliferative DR	4	295	116
Total		3662	3100

As illustrated in Figure 3.1, there is a significant imbalance in the dataset distribution. Approximately 50% of the samples are classified as “No DR,” while “Moderate DR” accounts for 27% and “Mild DR” for only 10%. The severe stages “Severe DR” (5%) and “Proliferative DR” (8%), are underrepresented. This imbalance may bias the model toward the majority class, limiting its ability to accurately detect advanced but clinically critical stages.

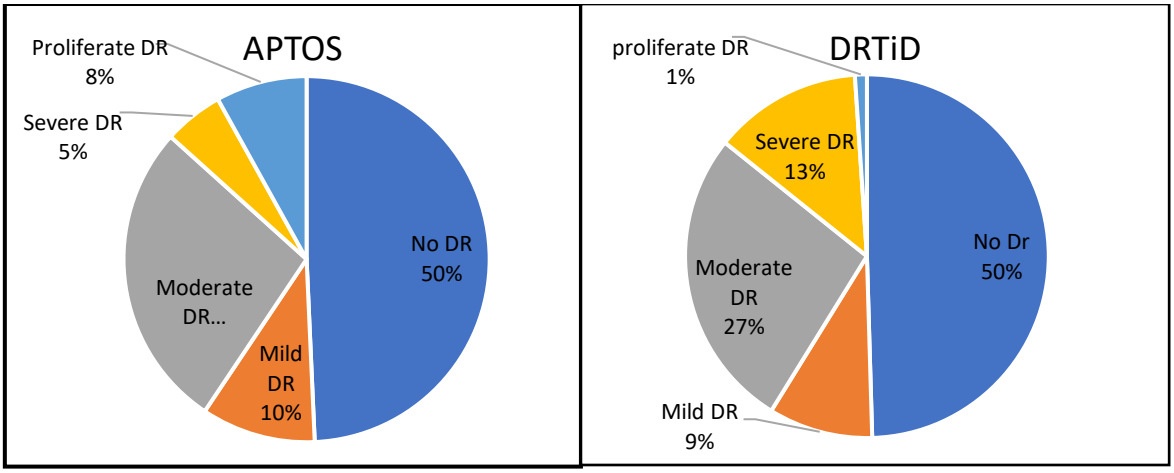


Figure 3. 1 Distribution of Level of Diabetic Retinopathy datasets

In total, 5,590 images (3,662 from APTOS and 3,100 from DRTiD) were used in this study. These datasets were split into training (80%) and testing (20%) sets, providing both sufficient data for model training and an unbiased portion for performance evaluation.

Figure 3.1 shows the imbalance in the dataset classes with the no DR images in the dominant part. Taking into account for the imbalance dataset is important to improve model reliability and accuracy in clinical application. There are several approaches that can be used to address the imbalance dataset such as data augmentation for the small number of samples, or oversampling methods like SMOTE. This class imbalance presents a common challenge in medical imaging classification tasks and highlights the need for

effective pre-processing, data augmentation, and suitable evaluation metrics that account for misclassification severity.

3.2 Analysis Technique, Design and Testing of model

There are two main part in this research, pre-processing of datasets and creating ensemble for the classification of DR based on the proposed framework in this research.

3.2.1 Data Preparation

Pre-processing of the retinal fundus images is one of the main phases in this research to create high quality datasets. APTOS 2019 and DRTiD datasets will be collected and checked for uniformity in size. All images will be resized to 224 x 224 pixels. Fundus images often suffer from variability in lighting, contrast, resolution, and noise due to differences in acquisition devices and conditions. The preprocessing pipeline is essential to standardize the data, enhance relevant features, and remove irrelevant artifacts, thus improving model accuracy and robustness. All images will be labeled according to the International Clinical Diabetic Retinopathy (ICDR) grading scale, assigning each image a class label ranging from 0 (No DR) to 4 (Proliferative DR). This ensures consistent ground truth labels across both datasets for model training and evaluation.

3.2.2 Base Model Development (CNNs)

Three pre-trained convolutional neural network (CNN) models: DenseNet50, InceptionResNetV2, and EfficientNetV2 will be prepared and loaded. In this case, each model will be initialized with pretrained weights from ImageNet to leverage transfer learning. To adapt these models for the diabetic retinopathy classification task, a two-phase fine-tuning process will be applied. In the first phase, the early layers of each CNN will be frozen, and only the top classification layers will be trained on the new dataset. In the second phase, all layers will be unfrozen and fine-tuned using a lower learning rate to further adapt the entire network to the domain-specific features of DR. Each model will use a softmax activation function in the output layer to generate class probabilities for the five DR severity levels. After training, each individual model will be evaluated on the validation or test set using key performance metrics including accuracy, F1-score, and Cohen's Kappa. These metrics provide a comprehensive view of each model's classification performance, accounting for both overall correctness and the balance between sensitivity and specificity across classes.

3.2.3 Feature Extraction from Softmax Output

For each input fundus image, the softmax probability vectors (of size 5 for DR grades 0–4) will be extracted from the outputs of all three base CNN models, DenseNet50, InceptionResNetV2, and EfficientNetV2. These probability vectors represent the model's confidence for each class and will be concatenated to form a 15-dimensional feature vector per image, combining the insights from all three models.

3.2.4 Meta-Classifier Design (Stacking Ensemble)

A new meta-dataset will be constructed using the 15-dimensional softmax vectors as input features and the original DR severity labels as targets. This dataset will then be used to train three different meta-classifiers: a Multilayer Perceptron (MLP), XGBoost,

and Random Forest. These models will learn how to best combine the outputs of the base CNNs to improve final prediction accuracy.

3.2.5 Evaluation of Ensemble Models

Each stacking configuration will be evaluated using 5-fold cross-validation or a holdout validation strategy. Performance will be assessed using multiple metrics, including accuracy, macro or weighted F1-score, Cohen's Kappa, and optionally Area Under the ROC Curve (AUC). These metrics will help measure not only the overall correctness but also how well the ensemble handles class imbalance.

3.2.6 Statistical Comparison

The performance of the three meta-classifier models will be compared directly using the evaluation metrics. Visual tools such as confusion matrices and ROC curves (per class or averaged) will be used to illustrate classification results, highlight strengths and weaknesses of each configuration, and support final model selection.

3.2.7 Selection of Best Model

The final stacking ensemble model will be selected based on which meta-classifier achieves the highest and most consistent performance across the datasets. The selection will be justified using both the evaluation metrics and the model's robustness across classes and data splits.

3.2.8 Prototype Development

A working prototype will be developed to demonstrate the practical application of the model. This prototype will be capable of accepting new fundus images as input, processing them through the three base CNNs, aggregating their predictions using the selected stacking ensemble model, and outputting the final diabetic retinopathy grade.

3.3 Research Process

The research begins with the data source stage, where publicly available and clinically relevant retinal image datasets are collected. These datasets used is APTOS 2019, contain 3,662 of labeled fundus photographs representing different stages of diabetic retinopathy (DR), including No DR, Mild, Moderate, Severe, and Proliferative DR. High-quality and diverse image data are essential to train deep learning models that generalize well across different populations and imaging conditions.

In the data selection and preprocessing stage, the raw retinal images undergo a series of transformations to improve model learning. Preprocessing techniques include green channel extraction, which enhances blood vessel visibility; CLAHE (Contrast Limited Adaptive Histogram Equalization), which improves image contrast; and noise reduction using Gaussian or Wiener filtering. Additionally, data augmentation techniques such as rotation, flipping, and brightness adjustment are applied to increase dataset variability and prevent overfitting. Class balancing is also considered, either by undersampling or oversampling techniques, to ensure the model does not favor dominant classes.

Following preprocessing, the next step is feature extraction, which is performed through multiple convolutional neural network (CNN) architectures. Pre-trained models denseNet 121, EfficientNetV2, and InceptionResNetV2 are employed to automatically learn deep features from the input images. Each CNN processes the image and outputs a

class probability vector, representing the likelihood of the image belonging to each DR severity level. These CNNs are fine-tuned using transfer learning to adapt their learned weights to the DR classification task.

In the model training and ensemble construction phase, each CNN model is trained individually using the prepared dataset. The performance of each model is evaluated on a validation set to obtain metrics like F1-score or Cohen's Kappa. The model evaluation and interpretation step are crucial to assess the effectiveness of both individual CNNs and the ensemble model. Evaluation metrics include accuracy, sensitivity, specificity, F1-score, Cohen's Kappa, and Area Under the ROC Curve (AUC). These metrics are derived from the confusion matrix and provide a detailed view of the model's classification performance, particularly in the presence of class imbalance. ROC and AUC analysis further help in understanding the model's discriminatory power across decision thresholds.

Finally, in the knowledge extraction stage, the insights obtained from the trained ensemble model are consolidated. This includes identifying the most informative CNN architecture, understanding misclassification patterns, and validating the model's applicability in real clinical scenarios. The ultimate goal is to produce a reliable and accurate DR grading system that can assist ophthalmologists in screening and diagnosing diabetic retinopathy more efficiently, thereby contributing to early detection and treatment planning. The results and insights can also inform future research or clinical tool development for other medical image classification problems.

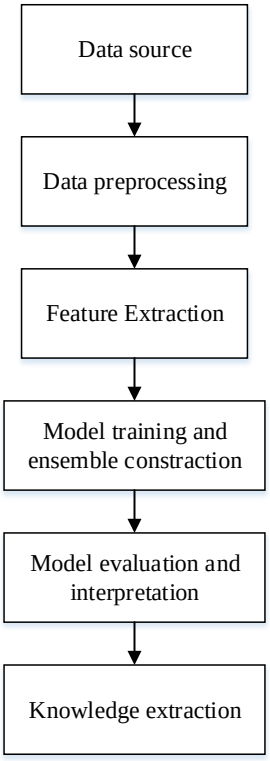


Figure 3. 2 Research Processing Model

Figure 3.2 illustrates the overall research processing model applied in this study. The workflow begins with the data source, followed by data preprocessing to enhance image quality and standardize inputs. Next, feature extraction is performed to identify relevant patterns within the retinal fundus images. These features are then used in model

training and ensemble construction, where multiple classifiers are integrated to improve prediction accuracy. The trained models undergo evaluation and interpretation to validate performance. Finally, the process concludes with knowledge extraction, where insights from the model results are derived to support clinical decision-making and further research.

3.4 Research Schedule

Table 3.3 presents the research time schedule across six months, outlining the key activities undertaken throughout the study. The process begins with the development of the research proposal, followed by an extensive literature review. Subsequently, research questions are developed to guide data collection, which takes place in the middle phase of the project. Data analysis is then conducted, leading to report writing. Finally, the research concludes with the submission of the completed study.

Table 3. 3 Research Time Table

Research activity	1st Month	2nd Month	3rd Month	4th Month	5th Month	6th Month
Develop research proposal						
Literature review						
Develop questions for data collection						
Data collection						
Data analysis						
Report writing						
Submission						

CHAPTER IV
CONCLUSION

This research proposal aims to improve the accuracy and reliability of diabetic retinopathy (DR) severity grading using a CNN-based ensemble learning approach. DR is a leading cause of vision loss worldwide, and early, accurate classification is essential for timely treatment. However, manual grading by specialists is labor-intensive and inconsistent, making it impractical for large-scale screening. While deep learning models, particularly Convolutional Neural Networks (CNNs), have demonstrated strong performance in DR classification, most studies rely on single architectures, which may struggle with generalization and class imbalance. To address these limitations, this study proposes an ensemble model that combines DenseNet201, EfficientNetV2, and InceptionResNetV2 using meta classifiers to enhance robustness and predictive accuracy.

The research will use retinal fundus images from the APTOS 2019 Blindness Detection dataset and DRTiD from Shanghai diabetic centre, focusing on classifying five DR severity levels based on the International Clinical Diabetic Retinopathy Disease Severity Scale. The proposed model will be optimized through transfer learning, preprocessing (e.g., green channel extraction and CLAHE), and fine-tuning techniques. Its performance will be evaluated against individual CNNs using metrics such as accuracy, F1-score, Cohen's Kappa, and AUC. This study not only contributes to improving DR detection but also explores the feasibility of deploying ensemble models in real-time, resource-limited clinical environments, supporting broader access to automated eye disease screening.

BIBLIOGRAPHY

Akhtar, S., Aftab, S., Ali, O., Ahmad, M., Khan, M.A., Abbas, S., Ghazal, T.M., 2025. A deep learning based model for diabetic retinopathy grading. Sci Rep 15. https://doi.org/10.1038/s41598-025-87171-9

Akram, M., Adnan, M., Ali, S.F., Ahmad, J., Yousef, A., Alshalali, T.A.N., Shaikh, Z.A., 2025. Uncertainty-aware diabetic retinopathy detection using deep learning enhanced by Bayesian approaches. Sci Rep 15. https://doi.org/10.1038/s41598-024-84478-x

Aldakhil, L.A., Alhasson, H.F., Alharbi, S.S., 2024. Attention-Based Deep Learning Approach for Breast Cancer Histopathological Image Multi-Classification. Diagnostics 14. https://doi.org/10.3390/diagnostics14131402

Ali, M., Haider, M.N., Lashari, S.A., Sharif, W., Khan, A., Ramli, D.A., 2022. Stacking Classifier with Random Forest functioning as a Meta Classifier for Diabetes Diseases Classification, in: Procedia Computer Science. Elsevier B.V., pp. 3453–3462. https://doi.org/10.1016/j.procs.2022.09.404

Alwakid, G., Gouda, W., Humayun, M., Jhanjhi, N.Z., 2023. Deep learning-enhanced diabetic retinopathy image classification. Digit Health 9. https://doi.org/10.1177/20552076231194942

Alyoubi, W.L., Abulkhair, M.F., Shalash, W.M., 2021. Diabetic retinopathy fundus image classification and lesions localization system using deep learning. Sensors 21. https://doi.org/10.3390/s21113704

Alzahrani, M., Usman, M., Jarraya, S.K., Anwar, S., Helmy, T., 2024. Deep models for multi-view 3D object recognition: a review. Artif Intell Rev 57. https://doi.org/10.1007/s10462-024-10941-w

Arora, L., Singh, S.K., Kumar, S., Gupta, H., Alhalabi, W., Arya, V., Bansal, S., Chui, K.T., Gupta, B.B., 2024. Ensemble deep learning and EfficientNet for accurate diagnosis of diabetic retinopathy. Sci Rep 14. https://doi.org/10.1038/s41598-024-81132-4

Atika, R., Susilawati, 2022. Gaya Hidup Sebagai Faktor Risiko Diabetes Melitus Tipe 2. Journal of Cahaya Mandalika 2. https://doi.org/https://doi.org/10.36312/jtm.v2i1.714

Atwany, M.Z., Sahyoun, A.H., Yaqub, M., 2022. Deep Learning Techniques for Diabetic Retinopathy Classification: A Survey. IEEE Access 10, 28642–28655. https://doi.org/10.1109/ACCESS.2022.3157632

Avolio, M., Fuduli, A., Vocaturo, E., Zumpano, E., 2023. Multiple Instance Learning for Diabetic Retinopathy Detection.

Bilal, A., Zhu, L., Deng, A., Lu, H., Wu, N., 2022. AI-Based Automatic Detection and Classification of Diabetic Retinopathy Using U-Net and Deep Learning. Symmetry (Basel) 14. https://doi.org/10.3390/sym14071427

Blair, J.P.M., Rodriguez, J.N., Lasagni Vitar, R.M., Stadelmann, M.A., Abreu-González, R., Donate, J., Ciller, C., Apostolopoulos, S., Bermudez, C., De Zanet, S., 2023. Development of LuxIA, a Cloud-Based AI Diabetic Retinopathy Screening Tool Using a Single Color Fundus Image. Transl Vis Sci Technol 12. https://doi.org/10.1167/tvst.12.11.38

B P, A., M S, C., K S, P., K A, S., 2023. Diabetic Retinopathy using Inception V3 Model. Int J Res Appl Sci Eng Technol 11, 4612–4616. https://doi.org/10.22214/ijraset.2023.51304

Budi Susilo, Y.K., Yuliana, D., Mahadi, M., Rahman, S.A., Ariffin, A.E., 2025. Artificial Intelligence for Early Detection and Prognosis Prediction of Diabetic Retinopathy. https://doi.org/10.1101/2025.03.29.25324873

Chilukoti, S.V., Shan, L., Tida, V.S., Maida, A.S., Hei, X., 2024. A reliable diabetic retinopathy grading via transfer learning and ensemble learning with quadratic weighted kappa metric. BMC Med Inform Decis Mak 24. https://doi.org/10.1186/s12911-024-02446-x

Dai, L., Wu, L., Li, H., Cai, C., Wu, Q., Kong, H., Liu, R., Wang, X., Hou, X., Liu, Y., Long, X., Wen, Y., Lu, L., Shen, Y., Chen, Y., Shen, D., Yang, X., Zou, H., Sheng, B., Jia, W., 2021. A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nature Communications 12. https://doi.org/10.1038/s41467-021-23458-5

Deepa, V., Kumar, C.S., Cherian, T., 2022. Ensemble of multi-stage deep convolutional neural networks for automated grading of diabetic retinopathy using image patches. Journal of King Saud University - Computer and Information Sciences 34, 6255–6265. https://doi.org/10.1016/j.jksuci.2021.05.009

Desiani, A., Primartha, R., Hanum, H., Dewi, S.R.P., Al-Filambany, M.G., Suedarmin, M., Suprihatin, B., 2024. Weighted Voting Ensemble Learning of CNN Architectures for Diabetic Retinopathy Classification. JURNAL INFOTEL 16. https://doi.org/10.20895/infotel.v16i1.999

Ergun, O.N., Ilhan, H.O., 2023. Advancing Diabetic Retinopathy Severity Classification Through Stacked Generalization in Ensemble Deep Learning Models. Traitement du Signal 40, 2495–2506. https://doi.org/10.18280/ts.400614

Fonda, H., Irawan, Y., Melyanti, R., Wahyuni, R., Muhaimin, A., 2024. A Comprehensive Stacking Ensemble Approach for Stress Level Classification in Higher Education. Journal of Applied Data Sciences 5, 1701–1714. https://doi.org/10.47738/jads.v5i4.388

Fresz, B., Lörcher, L., Huber, M., 2024. Classification Metrics for Image Explanations: Towards Building Reliable XAI-Evaluations, in: 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024. Association for Computing Machinery, Inc, pp. 1–19. https://doi.org/10.1145/3630106.3658537

Fuad, M.T.H., Fime, A.A., Sikder, D., Iftee, M.A.R., Rabbi, J., Al-Rakhami, M.S., Gumaei, A., Sen, O., Fuad, M., Islam, M.N., 2021. Recent advances in deep learning techniques for face recognition. IEEE Access 9, 99112–99142. https://doi.org/10.1109/ACCESS.2021.3096136

Ghasemieh, A., Lloyed, A., Bahrami, P., Vajar, P., Kashef, R., 2023. A novel machine learning model with Stacking Ensemble Learner for predicting emergency readmission of heart-disease patients. Decision Analytics Journal 7. https://doi.org/10.1016/j.dajour.2023.100242

Gupta, S., Khadka, A., 2025. Diabetic Retinopathy Detection through Multiclass Classification of Fundus Image Using Convolutional Neural Network. Pokhara Engineering College Journal (PECJ) 2.

Handoyo, A.T., Kusuma, G.P., 2022. Severity Classification of Diabetic Retinopathy Using Ensemble Stacking Method. Revue d'Intelligence Artificielle 36, 881–887. https://doi.org/10.18280/ria.360608

Hendrawan, K., Handayani, A.T., Andayani, A., Titiek, E., Gumelar, A.B., 2024. Classification of diabetic retinopathy using ensemble convolutional neural network architectures. Universa Medicina 43, 188–194. https://doi.org/10.18051/univmed.2024.v43.188-194

Hermawan, D., Wahyudi, T., Djamaludin, D., 2021. Pengaruh Gaya Hidup Terhadap Kejadian Diabetes Mellitus Pada Usia Produktif Di Wilayah Kerja Puskesmas Gading Rejo Kabupaten Pringsewu Tahun 2019 Lifestyle Influence against Incidence of Diabetes Mellitus in Productive Age in the Work Area of Gading Rejo Health Center, Pringsewu Regency in 2019, Jurnal Dunia Kesmas. Online.

Kandhasamy, J.P., Balamurali, S., Kadry, S., Ramasamy, L.K., 2020. Diagnosis of diabetic retinopathy using multi level set segmentation algorithm with feature extraction using SVM with selective features. Multimed Tools Appl 79, 10581–10596. https://doi.org/10.1007/s11042-019-7485-8

Lim, G., Bellemo, V., Xie, Y., Lee, X.Q., Yip, M.Y.T., Ting, D.S.W., 2020. Different fundus imaging modalities and technical factors in AI screening for diabetic retinopathy: a review. Eye and Vision. https://doi.org/10.1186/s40662-020-00182-7

Lin, C.L., Wu, K.C., 2023. Development of revised ResNet-50 for diabetic retinopathy detection. BMC Bioinformatics 24. https://doi.org/10.1186/s12859-023-05293-1

Mohanty, C., Mahapatra, S., Acharya, B., Kokkoras, F., Gerogiannis, V.C., Karamitsos, I., Kanavos, A., 2023. Using Deep Learning Architectures for Detection and Classification of Diabetic Retinopathy. Sensors 23. https://doi.org/10.3390/s23125726

Müller, D., Soto-Rey, I., Kramer, F., 2022. Towards a guideline for evaluation metrics in medical image segmentation. BMC Res Notes. https://doi.org/10.1186/s13104-022-06096-y

Nadda, R., Singh, J., Shrivastava, U., 2025. Automatic diabetic retinopathy detection using an ensemble learning approach and classifiers with self-adjusting weights. Soft comput. https://doi.org/10.1007/s00500-025-10773-y

Nneji, G.U., Cai, J., Deng, J., Monday, H.N., Hossin, M.A., Nahar, S., 2022. Identification of Diabetic Retinopathy Using Weighted Fusion Deep Learning Based on Dual-Channel Fundus Scans. Diagnostics 12. https://doi.org/10.3390/diagnostics12020540

Oh, K., Kang, H.M., Leem, D., Lee, H., Seo, K.Y., Yoon, S., 2021. Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images. Sci Rep 11. https://doi.org/10.1038/s41598-021-81539-3

Pamungkas, Y., Triandini, E., Yunanto, W., Thwe, Y., 2025. Enhancing Diabetic Retinopathy Classification in Fundus Images using CNN Architectures and Oversampling Technique. Journal of Robotics and Control (JRC) 6, 413–425. https://doi.org/10.18196/jrc.v6i1.25331

Purnama, R.F.N., 2023. Retinopati Diabetik : Manifestasi Klinis, Diagnosis, Tatalaksana dan Pencegahan. Lombok Medical Journal 2, 39–42. https://doi.org/10.29303/lmj.v2i1.2410

Rainio, O., Teuho, J., Klén, R., 2024. Evaluation metrics and statistical tests for machine learning. Sci Rep 14. https://doi.org/10.1038/s41598-024-56706-x

Saraswathy, C., 2023. Diabetic Retinopathy Detection and Multi Stage Classification using Deep Learning Models: A Quick Review. IJARCCE 12. https://doi.org/10.17148/ijarcce.2023.12675

Saxena, Mudit, Narra, P., Saxena, Mayank, Saxena, R., 2024. Deep learning ensemble framework for multiclass diabetic retinopathy classification. Telkomnika (Telecommunication Computing Electronics and Control) 22, 665–672. https://doi.org/10.12928/TELKOMNIKA.v22i3.25794

Schlosser, T., Friedrich, M., Meyer, T., Kowerko, D., Professorship, J., 2024. A Consolidated Overview of Evaluation and Performance Metrics for Machine Learning and Computer Vision. https://doi.org/10.13140/RG.2.2.14331.69928

Sebti, R., Zroug, S., Kahloul, L., Benharzallah, S., 2022. A Deep Learning Approach for the Diabetic Retinopathy Detection, in: Lecture Notes in Networks and Systems. Springer Science and Business Media Deutschland GmbH, pp. 459–469. https://doi.org/10.1007/978-3-030-94191-8_37

Shelke, N., Somkunwar, R., Pimpalkar, A., Maurya, S., Chhabria, S., 2025. Ensemble EfficientNet: a novel technique for identification, classification and prediction of diabetic retinopathy. Automatika 66, 543–558. https://doi.org/10.1080/00051144.2025.2514884

Soelistijo, S.A., 2021. PEDOMAN PENGELOLAAN DAN PENCEGAHAN DIABETES MELITUS TIPE 2 DEWASA DI INDONESIA-2021 PERKENI i Penerbit PB. PERKENI.

Srinivasan, S., Suresh, S., Chendilnathan, C., Prakash V, J., Sivaprasad, S., Rajalakshmi, R., Anjana, R.M., Malik, R.A., Kulothungan, V., Raman, R., Bhende, M., 2023. Inter-observer agreement in grading severity of diabetic retinopathy in wide-field fundus photographs. Eye (Basingstoke) 37, 1231–1235. https://doi.org/10.1038/s41433-022-02107-1

Sushith, M., Sathiya, A., Kalaipoonguzhali, V., Sathya, V., 2025. A hybrid deep learning framework for early detection of diabetic retinopathy using retinal fundus images. Sci Rep 15. https://doi.org/10.1038/s41598-025-99309-w

Tahir, H.N., Ullah, N., Tahir, M., Domnic, I.S., Prabhakar, R., Meerasa, S.S., AbdElneam, A.I., Tahir, S., Ali, Y., 2025. Artificial intelligence versus manual screening for the detection of diabetic retinopathy: a comparative systematic review and meta-analysis. Front Med (Lausanne). https://doi.org/10.3389/fmed.2025.1519768

Taifa, I.A., Setu, D.M., Islam, T., Dey, S.K., Rahman, T., 2024. A hybrid approach with customized machine learning classifiers and multiple feature extractors for enhancing diabetic retinopathy detection. Healthcare Analytics 5. https://doi.org/10.1016/j.health.2024.100346

Terven, J., Cordova-Esparza, D.M., Ramirez-Pedraza, A., Chavez-Urbiola, E.A., Romero-Gonzalez, J.A., 2025. Loss Functions and Metrics in Deep Learning. https://doi.org/10.1007/s10462-025-11198-7

Trisera, O., Himayani, R., Apriliana, E., Yusran|, M., Diabetik, R., Penglihatan, M., Yusran, M., 2024. Retinopati Diabetik yang Mengancam Penglihatan.

Tummala, S., Thadikemalla, V.S.G., Kadry, S., Sharaf, M., Rauf, H.T., 2023. EfficientNetV2 Based Ensemble Model for Quality Estimation of Diabetic Retinopathy Images from DeepDRiD. Diagnostics 13. https://doi.org/10.3390/diagnostics13040622

Usman, T.M., Saheed, Y.K., Ignace, D., Nsang, A., 2023. Diabetic retinopathy detection using principal component analysis multi-label feature extraction and classification. International Journal of Cognitive Computing in Engineering 4, 78–88. https://doi.org/10.1016/j.ijcce.2023.02.002

Vardhan, G.H., Jyoshna, M.V.S., Viswanath, P.K., Zubayr, S., Sravanth, V., 2024. Diabetic Retinopathy Detection Using InceptionResnet-V2 and Densenet121. Journal of Image Processing and Intelligent Remote Sensing 30–40. https://doi.org/10.55529/jipirs.42.30.40

Vujović, Ž., 2021. Classification Model Evaluation Metrics. International Journal of Advanced Computer Science and Applications 12, 599–606. https://doi.org/10.14569/IJACSA.2021.0120670

Xue, J., Wu, J., Bian, Y., Zhang, S., Du, Q., 2024. Classification of Diabetic Retinopathy Based on Efficient Computational Modeling. Applied Sciences (Switzerland) 14. https://doi.org/10.3390/app142311327

Yang, Z., Tan, T.E., Shao, Y., Wong, T.Y., Li, X., 2022. Classification of diabetic retinopathy: Past, present and future. Front Endocrinol (Lausanne). https://doi.org/10.3389/fendo.2022.1079217

Yaqoob, M.K., Ali, S.F., Bilal, M., Hanif, M.S., Al-Saggaf, U.M., 2021. Resnet based deep features and random forest classifier for diabetic retinopathy detection†. Sensors 21. https://doi.org/10.3390/s21113883

Yusran, M., Ilmu, B., Mata, K., Kedokteran, F., 2022. Kecerdasan Buatan dalam Diagnosis Retinopati Diabetik.

Zhang, W., Belcheva, V., Ermakova, T., 2025. Interpretable Deep Learning for Diabetic Retinopathy: A Comparative Study of CNN, ViT, and Hybrid Architectures. Computers 14. https://doi.org/10.3390/computers14050187