

UNIVERSITI TEKNOLOGI MARA

**PREDICTION OF
DIABETIC RETINOPATHY USING
MACHINE LEARNING**

MUHAMMAD ADIB BIN Y HAMDAN

**BACHELOR OF INFORMATION SYSTEMS
(HONS.) BUSINESS COMPUTING**

27 JANUARY 2025

Universiti Teknologi MARA

**PREDICTION OF DIABETIC
RETINOPATHY USING MACHINE
LEARNING**

Muhammad Adib Bin Y Hamdan

**Thesis submitted in fulfilment of the requirements
for Bachelor of Information Systems (Hons.)
Intelligent Systems Engineering Faculty of
Computer and Mathematical sciences**

27 JANUARY 2025

SUPERVISOR APPROVAL

PREDICTION OF DIABETIC RETINOPATHY USING MACHINE LEARNING

By

MUHAMMAD ADIB BIN Y HAMDAN

2023385635

This paper was completed under the supervision of the project director Dr. Azliza Binti Mohd Ali. It was submitted to the Faculty of Computer and Mathematical Sciences and was admitted while partially meeting the requirements of the Bachelor of Information Systems (Hons) Intelligent Systems Engineering.

Approved by,



.....
DR. AZLIZA MOHD ALI

Project Supervisor

27 January, 2025

STUDENT DECLARATION

I certify that this paper and the project it refers to are the product of my own work, and any ideas or references to others' works, whether published or not, are fully recognized according to the standard reference practice of the discipline.



MUHAMMAD ADIB BIN Y HAMDAN

2023385635

27 January, 2025

ACKNOWLEDGEMENT

With profound praise and gratitude to Allah for His omnipotence and boundless grace, I am pleased to have completed my thesis within the specified timeframe. I extend my heartfelt thanks to everyone who assisted me in this endeavor. Without their support, this project would not have been possible.

First and foremost, I am grateful to Dr. Azliza Mohd Ali, my committed, intelligent, and patient supervisor. Her ongoing counsel, support, affirmation, encouragement, and inspiration were critical in completing this endeavour. I am deeply grateful to her for her unfailing cooperation and for graciously donating her important time to help me.

I'd also like to thank my CSP600 and CSP650 lecturers, Madam Siti Nur Kamaliah Kamarudin, Dr. Farah Aqilah and Dr. Azliza Mohd Ali. Their intelligent counsel and critical analysis were extremely useful throughout the project's completion.

I am also grateful to all of my helpful lecturers, who have led and supported me during the completion of this thesis. Special thanks to my beloved family, especially my father and mother, for their constant support and motivation, always standing by me as my biggest supporters.

Lastly, I would like to thank my friends for their encouragement and support, which allowed me to finish this project. Sufi, Akmal, Hyqal, and Fayyadh in particular deserve special recognition. I would like to express my gratitude to everyone who helped and supported this thesis. God bless you all for your generosity.

ABSTRACT

Diabetic Retinopathy (DR) is a serious consequence of diabetes that may lead to blindness if not identified and addressed promptly. Although early detection is crucial, conventional screening measures are frequently postponed because of asymptomatic initial phases and healthcare inequities. This study seeks to tackle these problems by creating a machine learning-driven predictive model for diabetic retinopathy to enhance early identification and intervention. The dataset for this study was supplied by Dr. Azimah in conjunction with the Department of Ophthalmology at Hospital Al-Sultan Abdullah encompassing patient demographics, clinical information, and health indicators. To ensure the integrity of the dataset, data pre-processing techniques were performed, such as handling of missing values, normalization, and encoding. Several machine learning models, such as Support Vector Machine (SVM), K-Nearest Neighbours(KNN), Logistic Regression and Random Forest were trained and evaluated. All assessment metrics to assess optimal model performance included are accuracy, precision, recall, F1-score and ROC-AUC. The Random Forest model has surpassed its prequels with 79.49% accuracy and 83.60% ROC-AUC to become the most powerful and recommendable prediction model for diabetic retinopathy. I have developed a prototype system that provides an accessible user-friendly platform for healthcare professionals to enter patient data and receive real-time prediction outcomes. This research illustrates the capability of machine learning to improve diabetic retinopathy detection which provides a scalable solution for clinical environments. Future endeavors will concentrate on enhancing the model efficacy by incorporating supplementary risk factors and extending its applicability to more extensive datasets for comprehensive clinical validation.

TABLE OF CONTENTS

CONTENT	PAGE
SUPERVISOR APPROVAL	iii
STUDENT DECLARATION	iv
ACKNOWLEDGEMENT	v
ABSTRACT	vi
LIST OF FIGURES	x
LIST OF TABLES	xi
CHAPTER 1: INTRODUCTION	1
1.1 Background of Study	1
1.2 Problem Statement	2
1.3 Research Question	3
1.4 Research Objective	3
1.5 Research Scope	4
1.6 Research Significance	4
1.7 Expected output	5
1.8 Summary	5
CHAPTER 2: LITERATURE REVIEW	6
2.1 Overview of Diabetic Retinopathy	6
2.1.1 Causes and Symptoms of Diabetic Retinopathy	7
2.1.2 Stages of Diabetic Retinopathy	8
2.2 Machine Learning in Healthcare	10
2.2.1 Support Vector Machine	11

2.2.2	Random Forest	11
2.2.3	Naïve Bayes	12
2.2.4	Artificial Neural Network	15
2.2.5	Deep Learning	16
2.2.6	Ensemble Method	16
2.3	Prediction of Diabetic Retinopathy	17
2.4	Related Work using Same Domain but Different Technique	18
2.5	Related Work using Same Technique but Different Domain	19
2.6	Summary	20
CHAPTER 3: METHODOLOGY		21
3.1	Research Design	21
3.2	Preliminary Study	25
3.3	Knowledge Acquisition	26
3.4	Data Collection	26
3.5	Data Pre-processing	27
3.6	Model Development	29
3.7	Model Evaluation	33
3.8	Prototype Development	35
3.9	Prototype Testing	38
3.10	Documentation	38
3.11	Summary	39
CHAPTER 4: RESULT AND FINDINGS		40
4.1	Overview	40
4.2	Experiment 1: Cross-Validation	41
4.2.1	Result of Evaluation	41

4.2.2 Confusion matrix:	42
4.2.3 ROC Curve	47
4.3 Experiment 2: Upsampled Data	48
4.3.1 Result of Evaluation	49
4.3.2 Confusion matrix	51
4.3.3 ROC Curve	57
4.4 Experiment 3: Mostly.ai	58
4.4.1 Model Performance Comparison	58
4.4.2 Confusion Matrix for 988 rows:	60
4.4.3 ROC Curve	66
4.5 Model Evaluation	67
4.6 System Evaluation	68
4.7 User Interface	71
4.8 Summary	77
 CHAPTER 5: SUMMARY	 78
5.1 Research Objectives Accomplishment	78
5.1.1 Objective 1	78
5.1.2 Objective 2	79
5.1.3 Objective 3	79
5.2 Research Strengths and Limitations	79
5.3 Recommendation and Future Works	80
 REFERENCE	 81
 APPENDIX	 85

LIST OF FIGURES

FIGURE	CONTENT	PAGE
1.1	Expected Output	5
2.1	Difference of Retina	7
2.2	Example of Retina for Every Stage	9
3.1	Research Methodology Overview	20
3.2	Dataset of Diabetic Retinopathy in Excel Format	26
3.3	Bar Chart of Missing Values Before Handling	27
3.4	System Architecture	30
3.5	User Interface (UI) for Predicting Diabetic Retinopathy	32
3.6	Confusion Matrix for Random Forest	34
3.7	ROC Curve for Random Forest	34
4.1	Confusion Matrix Logistic Regression Using Cross Validation	42
4.2	Confusion Matrix SVM Using Cross Validation	43
4.3	Confusion Matrix Random Forest Using Cross Validation	44
4.4	Confusion Matrix KNN Using Cross Validation	44
4.5	Confusion Matrix XGBoost Using Cross Validation	45
4.6	Confusion Matrix AdaBoost Using Cross Validation	46
4.7	ROC Curve for All Models Using Cross-Validation	47
4.8	Confusion Matrix Logistic Regression Using Upsampled Data	51
4.9	Confusion Matrix SVM Using Upsampled Data	52
4.10	Confusion Matrix Random Forest Using Upsampled Data	53
4.11	Confusion Matrix KNN Using Upsampled Data	54
4.12	Confusion Matrix XGBoost Using Upsampled Data	55

LIST OF TABLES

TABLE	CONTENT	PAGE
2.1	Stages of Diabetic Retinopathy	9
3.1	Research Design	21
3.2	Missing Data and Features	27
3.3	Summarized Parameters During Training Process	29
3.4	Model Evaluation Results for Cross-Validation	33
4.1	Model Performance Metrics	41
4.2	Comparison of Original Results and Upsampling Results	49

CHAPTER 1

INTRODUCTION

This chapter explains background of the study, problem statement, research questions, the objectives of the research, and the significance of the research. The goal of this research is to propose and develop a system for early detection and intervention of diabetic retinopathy using machine learning techniques.

1.1 Background of Study

Diabetic Retinopathy (DR) is a common and potentially blinding consequence of diabetes mellitus. It is an inflammatory condition affecting the retina, the light-sensitive tissue in the back of the eye. Degenerative disorders such as diabetic retinal disease, which is mostly brought on by consistently high blood sugar, are characterized by vascular damage to the retina. (Ling-Ping, 2021). Diabetes is a long-term medical condition that results in improper metabolism, which raises blood sugar levels. Regrettably, this condition can harm the eyes as well as other organs. Diabetes mellitus (DR) is a widely recognized consequence of the disease that typically causes blindness and visual impairments in working-age persons worldwide (Markan, 2020). From 1995 to 2021, the cumulative prevalence of diabetes among adults in Malaysia was 14.39%, implying that one in every seven Malaysians could be affected by the ailment. When compared to persons between the ages of 20 and 29, those over 60 had the highest prevalence of diabetes, accounting for around 33.46% of total cases. Over the last ten years, the prevalence of diabetes mellitus (DM) in Malaysia has risen significantly, increasing the incidence of DR (Akhtar, 2022). As a result, rapid DR diagnosis and treatment are critical for minimizing the risk of visual impairment in diabetics (Markan, 2020). In the medical field, making predictions would enable personnel to determine

the appropriate course of action for a patient based on their current state of health. Medical tasks like doing surgery and patient data analysis have been made easier by machine learning (ML). Predictive modeling using machine learning algorithms may be able to predict the presence or absence of DR by looking at the risk factors of DM patients (Gupta, 2018). Machine learning (ML) has become more and more popular as it uses big data to make predictions. It also has the potential to completely transform healthcare delivery by making it more focused and accurate when integrated (Cichosz, 2015). Vijayarani and Dhayanand used techniques like support vector machines and neural networks for their research on kidney disease prediction. In this study, the accuracy for both algorithms are 76.32% and 87.70%, respectively. SVM is a top machine-learning technique for both linear and nonlinear classification (Dhayanand, 2015). As it is most useful when analyzing training sets with noisy data, the KNN algorithm yields highly accurate and efficient results for Diabetes Mellitus prediction. According to Saxena (2014), the accuracy values achieved by the KNN model range from 57.0% to 70.0% (Saxena, 2014). This paper describes the application of risk-based machine learning algorithms to predict diabetic retinopathy. The developed machine learning system will assess a diabetic patient's likelihood of having DR based on their medical history and demographic data (Chikara, 2018).

1.2 Problem Statement

A number of interrelated factors can influence the delay in the treatment of Diabetic Retinopathy (DR). DR development often lacks conspicuous signs in its early stages, which leads a majority of the population to delay medical help until vision related issues become evident, thus highlighting the need for regular checkups, especially among diabetics (Sarkar and Pawar, 2023). Traditional DR analysis is manual, so it is chronic, dull, slow, and prone to mistakes that hinge on specialist straining, which is difficult to bear in developing economies (Hafeez et al., 2023). Also, misdiagnosis of DR is

common due to inadequate facilities and resources in lower-level rural areas, leading to a lack of healthcare equity (Smith et al., 2023). Most of these screening techniques still so depend on manual DR readings and fundus X-rays that estimates become unreliable and require trained personnel to interpret them (Zhang et al., 2024). Besides, poor integration of diabetes and eye services services creates further gaps in opportunities for early intervention and screening, which worsen the timeliness of actions necessary to ameliorate the patient's condition (Kaur et al., 2023). All overcoming these barriers is needed to accomplish the objectives of the studies, such as figuring out how to slow down the progression of DR, and design programs concentrating on improving DR detection and intervention using machine learning technologies (Lee et al., 2023).

1.3 Research Question

- 1) What are the key factors contributing to the development or presence of diabetic retinopathy (DR)?
- 2) How can machine learning techniques be effectively applied to design a predictive model for diabetic retinopathy (DR)?
- 3) What methodologies and algorithms are most suitable for developing a prototype for predicting diabetic retinopathy (DR) using a Machine Learning model?

1.4 Research Objective

- 1) To identify factors contributing to the development or presence of diabetic retinopathy (DR)
- 2) To design a machine learning model that can predict diabetic retinopathy (DR), enabling early intervention and detection methods
- 3) To develop a prototype for predicting diabetic retinopathy (DR) using a Machine Learning model

1.5 Research Scope

The scope of this project covers the collection and analysis of a dataset from the Department of Ophthalmology Hospital Al Sultan Abdullah. The dataset will include a variety of demographic variables like the patient's name, age, and marital status, alongside key medical information like their weight, height, and Body Mass Index (BMI), alongside the presence of stroke, asthma, hypertension, Diabetic Ketoacidosis (DKA), nephropathy, neuropathy, ischemic heart disease (IHD), anemia, dyslipidemia, and even the levels of glycated hemoglobin (HbA1c). The machine learning algorithms selected for this study which are Support Vector Machine (SVM), K-nearest Neighbors (KNN), Logistic Regression, Random Forest, XGBoost, and AdaBoost, will be implemented to derive predictive models against the specified dataset. The final product will be user friendly for General Practitioners (GP) or any other medical practitioners to enhance their ability to identify diabetic retinopathy and make appropriate clinical decisions, thereby facilitating effective interventions.

1.6 Research Significance

This study's contribution is the design of a system that can help identify diabetic retinopathy (DR) at advanced stages, which, if not treated timely, can lead to irreparable damage to the eyesight. The approach employs machine learning techniques to focus on diabetes patients' risk characteristics and provides individualized treatment using predictive analytics. Afterwards, patients who have diabetic retinopathy receive prompt diagnosis and treatment, which not only enhances the clinical outcome but also the overall health status of patients. Moreover, such a system is capable of transforming diabetes management towards more proactive prevention by detecting susceptible persons and adjusting their care to lessen the impact of vision impairment due to diabetic retinopathy. Besides that, this is also aligned to SDG 3: Good Health and Well Being, which broadens the scope of healthcare practice to preventable cases of vision impairment due to diabetic retinopathy.

1.7 Expected output

In broad terms, the outcome of the project will be a prototype that can be effectively utilized by medical professionals whereby a patient's demographic and medical information can be entered into the system. The data will then be analyzed through machine learning algorithms to detect if the patient has diabetic retinopathy (DR) or not. Upon completion of the analysis, the prototype will indicate the risk factors the patient may have in relation to developing DR. This further empowers healthcare providers to make reasoned decisions concerning prompt actions and the right treatment required which will lead to favorable results while minimizing cases of vision loss.

The figure consists of two side-by-side screenshots of a web application interface. The left screenshot shows a form titled "Please enter the following information:" with various input fields for patient data. The right screenshot shows the "Result:" section with "Diabetic Retinopathy: Absent".

Form Fields (Left Screenshot):

- Welcome to the prediction page!
- Please enter the following information:
- Age:
- Gender: ☒ Male ☐ Female
- Body Mass Index (BMI):
- Systolic Blood Pressure:
- Diastolic Blood Pressure:
- Duration of Diabetes Mellitus:
- Hypertension: ☒ Present ☐ Absent
- Nephropathy: ☒ Present ☐ Absent
- Neuropathy: ☒ Present ☐ Absent
- Diabetic Foot: ☒ Present ☐ Absent
- Stroke: ☒ Present ☐ Absent
- Ischemic Heart Disease: ☒ Present ☐ Absent
- Anemia: ☒ Present ☐ Absent
- Asthma: ☒ Present ☐ Absent
- HbA1c:
- Diabetic Ketoacidosis (DKA): ☒ Present ☐ Absent
- Dyslipidemia: ☒ Present ☐ Absent
-

Result Section (Right Screenshot):

- Result:
- Diabetic Retinopathy:
- Absent

Figure 1.1 Expected Output

1.8 Summary

This chapter has explained the research background and the problem in the research domain. It also highlights the objective and significance of this research to be proposed. The research questions and the scope of the research for solving the problem are also included in this chapter. The next chapter will discuss the literature reviews on the domain and previous research.

CHAPTER 2

LITERATURE REVIEW

This chapter explains the overview of Diabetic Retinopathy, causes and symptoms of Diabetic Retinopathy, stages of Diabetic Retinopathy and current assessment method. This chapter also reviews the previous solution that related to detection and prediction of Diabetic Retinopathy which will help to grasp the general concept, terminologies, and techniques in this study.

2.1 Overview of Diabetic Retinopathy

Diabetic Retinopathy (DR) is an eye ailment that occurs due to diabetes mellitus and is marked by the damage caused to the retinal blood vessels owing to elevated blood sugar levels (Emon et al., 2021). Unlike many other disorders, diabetic retinopathy progresses through various stages. At the onset, there might be slight non-proliferative changes, which can eventually develop into proliferative DR, a condition marked by growth of new abnormal blood vessels on the retina (Emon et al., 2021). As DR is one of the commonest causes for blindness amongst adults globally, it is one health condition that needs to be diagnosed early. If it is detected early, there is anti-VEGF therapy, vitrectomy, and laser treatment that greatly aids in preventing or slowing DR and preserving vision (Abas et al., 2024).

The early stages of diabetic retinopathy (DR) often go unnoticed owing to their symptomless nature, which highlights the need for periodic screening, especially in patients suffering from diabetes (Sarkar and Pawar, 2023). Continuous monitoring is essential as it allows discovering DR at a stage when intervention can do the most good. Also, using sophisticated predictive models that employ machine learning greatly enhances primary diagnosis accuracy, which in turn leads to improved clinical outcomes (Emon et al., 2021). With

the help of algorithms which scan thousands of data points, patients with DR can be seen and treated at the correct time by spotting trends and risk factors associated with the condition. It has been shown in recent studies that the standard screening of DR can be enhanced through the application of machine learning algorithms. The innovation proposed by Yuedong Zhao et al. (2022) concerning integration of such systems in diabetes retinopathy screening could alleviate the burden of blindness in such patients and increase the accuracy of the diagnosis (Yuedong Zhao et al., 2022).

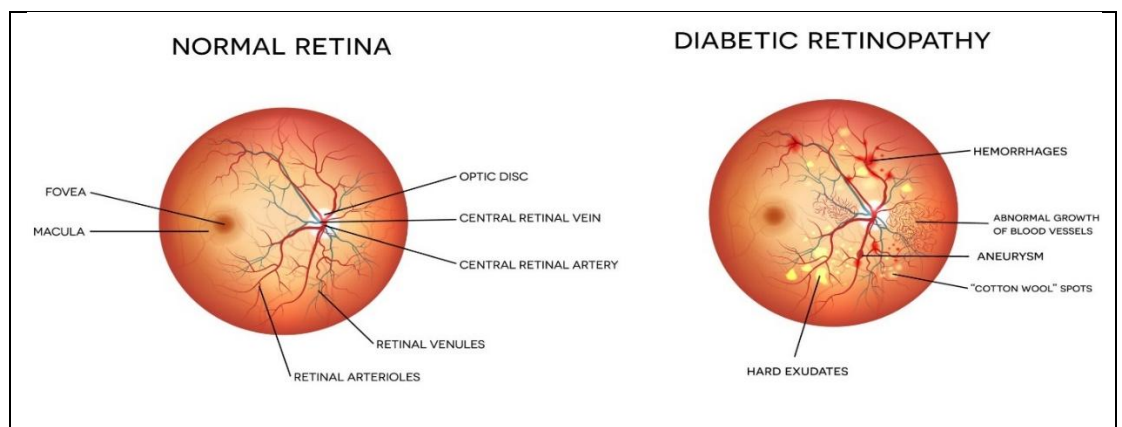


Figure 2.1 Visualization of a healthy retina and an unhealthy retina

(Source: Neoretina Eyecare Institute, 2022)

2.1.1 Causes and Symptoms of Diabetic Retinopathy

Diabetic Retinopathy (DR) occurs due to long-term elevated glucose levels that result in damage to the blood vessels located at the retina, causing them to leak and get blocked (Emon et al., 2021). This may ultimately result in the expansion of retinal tissues and the abnormal growth of neo-vasculature, which can be visually obstructive (Sarkar and Pawar, 2023). As reported by Nor Azamen et al. (2023), smoking, high blood pressure, high cholesterol, and diabetes which is poorly controlled are some of the leading risk factors for DR. The presence of family history and even pregnancy can further predispose one to it (Fatma Hilal Yagin et al., 2023). People's choices regarding their exercise and dietary habits also play a crucial role in the onset and progression of DR (Chua et al., 2021).

Once Diabetic Retinopathy (DR) worsens, symptoms such as blurry vision, floaters, dark spots, color blindness, and in extreme cases, complete vision loss, can be experienced. These symptoms are rarely visible in the early stages of the disease, but mark a notable decline in vision. Vision loss is highly prevalent in the advanced stages and can greatly affect the daily life of patients (Emon et al., 2021). To diagnose DR, fluorescein angiography, which involves imaging the retina, Optical Coherence Tomography (OCT) scanning for abnormalities, and dilated eye examinations are required. A thorough eye examination is usually required to diagnose DR (Sarkar and Pawar, 2023). Current strategies to manage DR focus on stopping the progress of the disease and safeguarding eyesight. This can be achieved by controlling blood pressure, cholesterol, and blood sugar, in addition to regular screenings and medical treatment at the onset of symptoms (Nor Azamen et al., 2023). New innovative therapies, including anti-VEGF injections and laser treatment, have proven to slow down the progress of DR and enhance vision (Brown et al., 2022).

2.1.2 Stages of Diabetic Retinopathy

Diabetic retinopathy has four broad categories which are proliferative, mild, moderately, and severely non-proliferative. In the first stage, the non-proliferative phase, small blood vessels within the retina expand in a balloon-like fashion. The second stage, moderate non-proliferative retinopathy, is characterized by occlusions in particular blood vessels within the eye. In the third stage known as severe non-proliferative retinopathy, further blood vessels become occluded. These leave deficit areas in the retina with insufficient blood flow (Shoaib et al., 2024). Without sufficient blood flow, the retina cannot sustain the growth of new blood vessels (Yagin et al., 2023). The term proliferative retinopathy refers to the fourth and final stage. It is characterized by the development of weak and abnormal new blood vessels in the retina, which can rupture and lead to blind spots, severe vision impairment, and even cognitive loss. There is often an overlap of the diabetic type and other serious eye diseases which Diabetic Macular Edema (DME), which occurs in nearly

50% of patients with diabetes retinopathy. DME develops when fluid seeps from blood vessels within the retina, resulting in the swelling of the macula, which is an area of the retina required for perceiving sharp images (Shoaib et al., 2024). Prevention of the disease progression and preservation of vision requires compelling regular screening and therapy (Rather and Malhotra, 2023).

Table 2.1 Stages of Diabetic Retinopathy

Stage	Description	Effects
Stage 1: Mild Non-Proliferative	Small blood vessels in the retina swell.	Early retinal changes.
Stage 2: Moderate Non-Proliferative	Some blood vessels in the retina become blocked.	Reduced blood flow in the retina.
Stage 3: Severe Non-Proliferative	Many more blood vessels become blocked, leading to poor blood flow.	Significant retinal damage risk.
Stage 4: Proliferative	New, weak, and irregular blood vessels form in the retina, which can leak.	Severe vision problems and potential blindness.

(Source: Sarkar and Pawar, 2023)

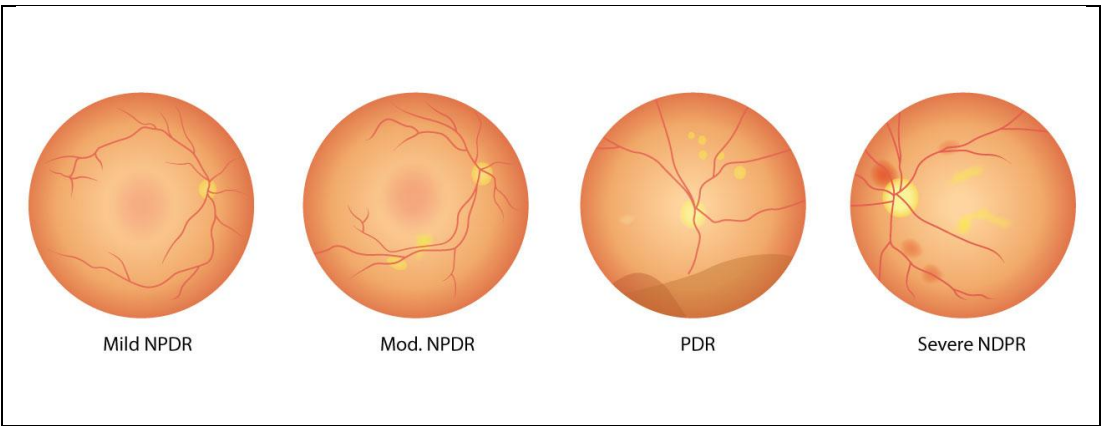


Figure 2.2 Example of Retina for every stages

(Source: Snyder's Eye Care, 2025)

2.2 Machine Learning in Healthcare

One of the key components of healthcare technology nowadays is machine learning, which enhances diagnosis precision, treatment procedures, and patient results using sophisticated data analytics. An ML algorithm can identify certain relations and patterns which specialized human practitioners may miss. A predictive model in machine learning is designed to analyze users' Electronic Health Records (EHRs) to predict epidemic events, immediate patient health deterioration, and to suggest custom-tailored treatment methods (Rather and Malhotra, 2023). Impressive results have been obtained using Convolutional Neural Networks (CNNs) for medical imaging. These models assist medical practitioners in making faster decisions regarding the presence of abnormalities in radiological images by helping them surpass human-level precision (Vyas et al., 2023).

There is a growing use of machine learning within healthcare systems to improve operational efficiency. Consider, for instance, how ML algorithms can be used to process medical claims and control patient flow in a hospital's emergency department, this automated workflow improves efficiency at the hospital's administrative level (Sarkar and Pawar, 2023). In addition, natural language processing (NLP) and ML in combination can analyze unstructured clinical notes and extract information further, helping improve clinical decisions and expand our knowledge of patients' clinical histories (Liu et al., 2021). Yet, as much progress as has been made, there are significant obstacles that need to be addressed in order for ML to be completely assimilated into healthcare, such as bias in training data, obtaining regulatory clearance, and safeguarding privacy. Nonetheless, these advancements are anticipated to make the future healthcare delivery systems more efficient and personal (Wang et al., 2019).

2.2.1 Support Vector Machine

Support Vector Machines (SVM), a form of supervised learning model, is helpful for regression and classification tasks. SVM has a defined proficient domain which is high dimensional spaces. The SVM algorithm finds the optimal hyperplane for dividing data into discrete classes by maximizing the margin between the closest data points of different classes (Cortes and Vapnik, 1995). Several studies have proved SVM's success in image analysis in the medical field. For example, Kaur et al. (2021) used SVM to classify retinal images into different stages of diabetes retinopathy with an accuracy of 90.55%. They used some pre-processing techniques as well and showed how SVM with some modifications can be used to detect diseased eyes. SVM is particularly well-suited for medical problems because such problems require accuracy due to minor changes in visual features. In addition, its ability to process non-linear data with the use of kernel functions increases SVM's selection for complex data distributions (Kaur et al., 2021). Healthcare is an area where SVM application versatility has been propelled by recent researches. For instance, Kumari et al., (2020) have used SVM for recognition of MRI images for diagnosing brain cancer and put forward the possibility of its application in medical imaging. The flexibility and performance of SVM serves to strengthen its relevancy in tackling difficulties had with medical diagnosis and imaging tasks.

2.2.2 Random Forest

Random forests, according to Zhao et al. (2022) and Biau and Scornet (2016), are a supervised classification learner that uses ensemble learning. During training, the model builds multiple decision trees and merges them together in order to classify as thoroughly as possible without overfitting. Rather than making a decision on their own, multiple models team up to provide an answer which is more accurate employing the "wisdom of the crowd" approach (Biau and Scornet, 2016). Moreover, Khan et al (2020) and Sarkar and Pawar (2023) highlight that Random Forests are most effective for datasets with a large

number of interacting variables as well as for higher dimensional data. Zhao Yuedong et al. used Random Forest and achieved 87% accuracy in prediction of diabetic retinopathy. The investigator proved the model's usefulness in medical imaging by retaking the photograph of the retina and demonstrating that it enhanced the features that could be extracted. In the medical field, there is great dependency on Random Forest feature significance scores in which the diagnosis of patients is done with high accuracy (Khan et al, 2020). Because of its precision and simplicity of application, it is frequently used in many fields, including health care. Apart from diabetic retinopathy, Zhang et al. (2024) and Sarkar and Pawar (2023) mention that Random Forest has also been applied in disease prognosis, identification of biomarkers, and predicting patient results.

2.2.3 Naïve Bayes

Naïve Bayes is an algorithm based on probabilities that assumes independence among features as conditions which execute Bayes theorem. This allows flexibility for its use in big data. Its ease of use as well as good results has made it appealing in predicting diseases particularly would in medical diagnosis. Maniruzzaman et al. (2021) used Naive Bayes to determine the severity levels of diabetic retinopathy through demographic and clinical data of the patients. The model achieved an accuracy of 80%. In medical applications, the model calculates posterior probabilities for each class and yields the one with the highest probability as a result, hence the name “naive.” This model has an additional feature of estimating probability distributions to deal with missing data which adds to its efficiency when dealing with incomplete records. Naive Bayes as the name suggests, does struggle with highly correlated features due to relying on the independence assumption. More recently, research leading towards improving Naive Bayes included feature selection techniques, hybrid models, and ensemble ones, increasing the predictive power of the model in classifying diseases (Sarkar and Pawar, 2023). As a result the efficiency and power of the algorithm makes it more accessible in predictive analytics,

exercising applications in healthcare such as screening for diseases or assessing patients' risk (Emon et al.,2021).

2.2.4 Logistic Regression

Due to its effectiveness in managing structured medical data, Logistic Regression (LR) has been one of the most employed methods for the prediction of Diabetic Retinopathy (DR). Smith and others (2023) have shown how LR can be reliable in clinical data. For example, one of the models trained on the patient's medical history had an accuracy and AUC of 78.21% and 81.61% respectively. According to the model's coefficients, the most important predictors of DR were found to be the levels of HbA1c and systolic blood pressure. Moreover, in a preceding study, Lee and others (2022) improved the performance of LR by varying the parameters in a stepwise regression model to reduce overfitting. Additionally, mainly in Lee's study (2022), combining LR with other machine learning methods, especially Random Forest and XGBoost, showed better sensitivity and specificity in the prediction of early stages of DR. Although deep learning models have better performance for image classification, LR is still an important approach, especially in risk stratification and primary screenings of patients, where computational power and IT infrastructure are inadequate.

2.2.5 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is one of the most popular learning algorithms which classifies and predicts new data points by voting from the k-nearest data points. KNN is utilized in DR prediction because it is favorable in dealing with nonlinear relationships within medical datasets. The model's effectiveness is highly dependent on the choice of k, where smaller k values tend to overfit while large k values can oversimplify important patterns. Studies like these by Ahmed et al. (2023) showed that using a k value of 5 was optimal achieving an accuracy of 74.5 percent while predicting DR based on clinical parameters. Also, feature scaling and dimensionality reduction using Principal Component

Analysis (PCA) has shown to improve KNN's performance by reducing the impact of irrelevant factors (Ahmed, 2023). Although KNN is easy to implement, it has some disadvantages such as higher computational costs when dealing with larger datasets since all training instances need to be stored for every prediction. Patel et al (2022) solved issues regarding KNNs low efficiency by implementing methods of approximate nearest neighbor search and found significant improvements in KNNs efficiency without losing accuracy. KNN is inefficient for larger datasets due to computational costs and need to have all training instances stored and distance calculations done for each prediction. KNN brings significant performance even in its weaknesses and is often used as a core component of the ensemble models combined with more complex algorithms (Patel, 2022)

2.2.6 AdaBoost

AdaBoost or Adaptive Boosting is an ensemble learning technique that enhances classification tasks by integrating several weaker learners, often decision stumps, into a single strong learner. The algorithm trains the models in steps, increasing the weight of instances that were incorrectly classified during the previous iteration, thus enabling the model to concentrate on harder examples (Freund & Schapire, 1997). Patel et al. (2023) achieved a DR prediction accuracy of 82.3% while employing AdaBoost on a clinical dataset. On the flip side, the accuracy of prediction models trained with AdaBoost decreases in the presence of noisy data and outliers because such instances tend to receive most of the weights when computing the classifier. In the light of this disadvantage, AdaBoost remains useful because of its impressive ability in improving weak classifiers as well as the medical decision-making processes.

2.2.7 XGBoost

XGBoost (Extreme Gradient Boosting) is an upgraded version of gradient boosting which utilizes regularization and parallel processing to optimize the prediction performance. While XGBoost follows the sequential building of decision trees just like AdaBoost, it adds L1 and L2 regularization for overfitting control making it ideal for big datasets (Chen & Guestrin, 2016). According to Zhao et al. (2022), XGBoost surpassed all other ML models during DR prediction with an AUC of 0.803 due to its skillful missing value treatment alongside computational optimizations. This ability to scale and process highly unbalanced datasets makes XGBoost favorable in DR screening tasks where accuracy and efficiency are highly needed.

2.2.8 Artificial Neural Network

Artificial Neural Networks (ANN), these networks are designed after the neural architectures found in animals' brain and are exceptional in prediction, categorization, and pattern recognition. In an artificial network, there is a collection of neurons or nodes organized in layers. Each neuron is connected to the other neurons in the following layer, where each connection is assigned a weight that will determine the strength of the signal when it passes through. ANNs are being increasingly used in medical diagnosis because they can be trained to learn from intricate datasets and recognize patterns that are subtle (Liu et al., 2021; Sarkar and Pawar, 2023). For example, Liu et al. used ANN for predictive analyses of diabetic retinopathy through retinal images. The study revealed high automation of medical images will lead to excellent accuracy, which provided promise to the utilization of ANN in the field of medicine. In addition, to achieve better performance, ANNs have also been combined with other forms of machine learning. One example is the development of ANN using genetic algorithms to enhance the performance of the system's network structure and training process (Vyas et al., 2023). ANNs, due to their flexibility and strength, are powerful and effective tools in multi-

disciplines like healthcare, patient monitoring, treatment planning, and disease diagnosis (Vyas et al., 2023; Sarkar and Pawar, 2023).

2.2.9 Deep Learning

Deep learning revolves around deep neural networks which enable learning through big data. As with traditional machine learning algorithms, deep learning models especially excel at recognizing patterns and complex representations in data that more simplistic models fail to understand. For example, Deep Convolutional Neural Networks (CNNs) have outperformed their peers in detecting various medical conditions radiologically and have shown outstanding image classification capabilities (Khan et al., 2023). In the same vein, Vyas et al. (2023) demonstrated that deep learning algorithms could be trained to recognize pneumonia from chest X-rays better than some radiologists. This not only puts such abilities at par with human experts, but poses certain perspectives where they could enhance or replace human efforts in particular domains. The ability to learn from unstructured data in the form of text and images makes deep learning versatile. This expands its scope of application to a variety of fields including, healthcare, medical imaging and natural language processing (Khan et al., 2023). The increasing availability of vast quantities of data as well as the constant improvements in computational speed have resulted in a deeper, more widespread adoption of deep learning techniques in real life, especially in medicine (Wang et al., 2024).

2.2.10 Ensemble Method

Ensemble methods combine different machine learning models to improve accuracy, and utilize bagging, boosting, and stacking techniques. These combine classifiers with differing levels of complexity to lower the variance and bias, which proves useful in challenging medical datasets, including those used for DR prediction. Random Forest, which is a bagging method, increases accuracy by training multiple decision trees on random subsets of the data and averaging the predictions to decrease overfitting (Breiman, 2001). The

boosting methods include AdaBoost and XGBoost which focus on improving the performance of weak models by training them sequentially through greater emphasis on previously misclassified cases (Freund & Schapire, 1997). Stacking has demonstrated greater accuracy than single models through the use of meta-models which consolidate and optimize predictions from base models (Liu et al., 2023). This model achieved a 7% increase in accuracy compared to single models while assisting in DR detection. While these models using ensemble methods increase classification accuracy and generalization to a high degree, the time needed to train these complex models is an important consideration for large scale Automated DR screening programs.

2.3 Prediction of Diabetic Retinopathy

Different machine learning forms have to be applied in predicting the risk of diabetic retinopathy (DR) which happens to be the most common cause of vision impairment among diabetes patients. Emon et al (2023) used multiple classifiers such as Bagging, Naive Bayes, Decision Tree, Logistic Regression, SGD, J48, SMO, and Random Forest to test their performance. From the primary results, it was evident that Logic Regression was the best performing classifier because of its favorable true positive rate and ROC value (Emon et al. 2023). Abas et al. (2024) remarked that Random Forest, KNN, and other algorithms as ensembling techniques with Deep Learning Decision Tree, Logistic Regression, SVM, and XGBoost were equally important. Also noted were the LightBGN ensemble model and XGBoost which increased accuracy rates, indicative of the usefulness of these modern models for predictive tasks (Abas et al., 2024). Zhu et al. (2023) and other researchers also showed that LightGBM and Gradient Boosting proved useful for DR because of their high prediction accuracy. Further, Sarkar et al (2023) showed that the ensemble approach strategy is more efficient in applying deep learning to machine learning to enhance predictive analytics accuracy through the comportsment of novel ensemble multi-classification techniques of machine learning with set

classifiers with set accuracy of 99.6% and 99.8 % in different datasets. With other methods like, Su et al., (2022) created a model that predicts diabetic retinopathy (DR) risks using plasma metabolites through the application of partial least squares regression and liquid chromatography-mass spectrometry (LC-MS). The model achieved an AUC of 0.770, while being able to identify distinctive predictors which included histidine, citrulline, phenylalanine, methionine, tyrosine, C3, and C24 (Su et al., 2022).

2.4 Related Work using Same Domain but Different Technique

This part focuses on the other applications and the added benefits that come with using Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbour (KNN) in medical diagnosis. To elaborate, Random Forest stands out because it is one of the dominant ensemble learning problems in medical detection. It is prolific in the construction of a myriad of decision trees during the training phase and improves the predictive performance by aggregating the results. Random Forest has been used successfully in analyzing patient records to determine the attendance of diabetic retinopathy (DR) and it's detection is amply facilitated with the algorithms ability to process high-dimensional and incomplete datasets in a reliable manner (Shanthi et al., 2021; Sarkar and Pawar, 2023).

Another sophisticated approach, that can be used for the categorization purpose in medical image analysis, is Support Vector Machine (SVM). SVM selects the best hyperplane which maximally separates data points of different categories. For instance, SVM can help classify a medical image by recognizing minute details and patterns related to specific medical conditions like MRI or retinal images. The advanced diagnostic capability of SVM stems from image data analysis. It enables the detection of small differences within images, making machine learning algorithms essential for classifying medical images. Additionally, SVM is unparalleled in medical diagnostics due to its

non-restrictive assumption of data structure; it is able to work efficiently with both linear and non-linear data (Gupta et al., 2020; Kaur et al., 2021).

Lastly, K-Nearest Neighbours (KNN), while being simple in nature, is a method which is effective in predicting the chances of diabetic retinopathy. KNN independently classifies new records in relation to the major class of its K nearest neighbors. KNN predicts the chances of developing diabetic retinopathy for a patient by examining the patient's records such as his or her blood pressure, glucose levels, and other health parameters. Its intuitive nature makes it particularly valuable for medical practitioners. When datasets are chosen carefully and KNN is set properly, KNN enables the early detection and intervention of diabetic retinopathy (Roy et al., 2021; Emon et al., 2021).

2.5 Related Work using Same Technique but Different Domain

This section describes some of the different medical fields that Artificial Neural Networks (ANNS) is used in which efforts are made to achieve their purpose at hand. This research demonstrates that ANN technology allows for the automated detection of Alzheimer's disease from brain images with a high level of accuracy, classification of skin cancer through dermoscopic images, and pneumonia from X-ray chest images. ANNs are capable of 'seeing' features on chest X-rays suggestive of pneumonia and, therefore, aids in its early diagnosis and management. After extensive coined X-ray imaging, models of ANN rapidly and accurately identify pneumonia. Such tools are helpful to radiologist and provide a speedy, noninvasive way to diagnose the patients (Emon et al., 2021; Diaz Redondo et al., 2022).

The use of ANN includes the interpretation of dermoscopic images to differentiate between malignant and benign lesions in skin cancers. Neural networks assist in the diagnosis of skin cancer by creating complicated structures in an image. For example, Sharma and Deep (2020) developed a technique for skin cancer image classification that is based on DE-ANN and

attained good results in classification of images. These studies underscore the fact that ANNs can serve as valid non-invasive diagnostic procedures for skin cancer (Sharma and Deep, 2020).

The use of ANNs now extends to the automatic diagnosis of Alzheimer's disease from data obtained from brain imaging studies like MRI scans. These ANN-based methodologies have demonstrated certain pathomorphological changes in Alzheimer's disease, for example, atrophy of the cerebral cortex and reduced size of the hippocampus, which are plausible predisposed alterations of the anatomy in AD. All these models present an advantageous approach for early diagnosis, which is the most pertinent in managing and slowing the progression of the disease. The ability of ANNs to provide rapid, accurate analysis of neuroimaging sets them apart as exceptional tools for MRI in dementia with Alzheimer's disease compared to other forms of dementia with lower sensitivity (Liu et al., 2021; Sharma et al., 2022).

2.6 Summary

This chapter reviews Diabetic Retinopathy (DR), its progression, and the significance of early detection. It explores machine learning (ML) models, including Logistic Regression, SVM, KNN, Naïve Bayes, Random Forest, AdaBoost, XGBoost, and CNNs, used for DR prediction and diagnosis. Ensemble methods like bagging, boosting, and stacking improve classification accuracy, while deep learning models enhance retinal image analysis. Studies highlight ML's effectiveness in handling large datasets, extracting features, and automating medical diagnoses. Feature selection and data preprocessing techniques further optimize predictive performance. The chapter also reviews ML applications in other diseases, demonstrating its adaptability. Recent advancements focus on improving interpretability, handling imbalanced datasets, and integrating hybrid models. The growing use of AI in healthcare aims to enhance screening, automate detection, and improve patient outcomes.

CHAPTER 3

METHODOLOGY

This chapter describes the methodology of the research that will be employed for completing this study. The phases in the methodology are constructed to fulfill the research objectives, which are to identify the factors contributing to the development or presence of diabetic retinopathy (DR), to design a predictive model for DR using machine learning techniques, and to develop a prototype for predicting DR using machine learning algorithms such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest. Further details of each phase in the methodology is elaborated in this chapter.

3.1 Research Design

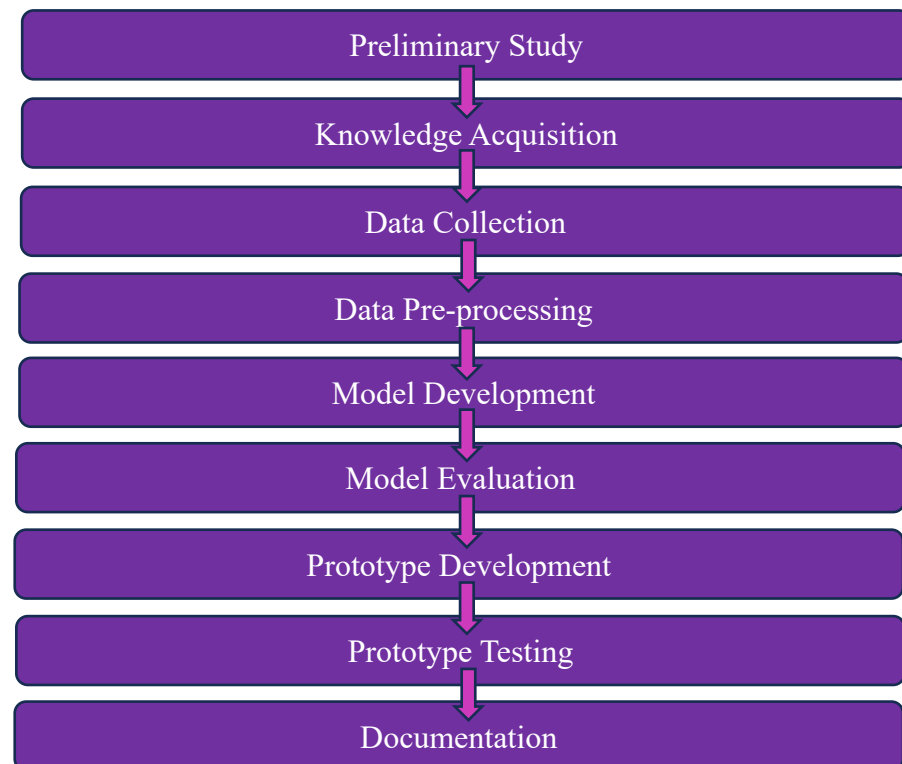


Figure 3.1 Research Methodology Overview

Figure 1.0 illustrates the overview of the research methodology for predicting diabetic retinopathy using machine learning, and it contains nine phases overall for completing the research. The research begins with performing the preliminary study, then continues with knowledge acquisition, data collection, data pre-processing, model development, model evaluation, prototype development, prototype testing, and ends with the documentation phase. Table 3.0 displays the research design for this research in which the activities, sources, and deliverables for each phase are briefly explained.

Table 3.1 Research Design

Objective	Phases	Activities	Sources	Deliverables
To identify the factors contributing to the development or presence of diabetic retinopathy (DR).	Preliminary Study	<ul style="list-style-type: none"> - Understanding the research background - Identifying the key factors contributing to the development or presence of diabetic retinopathy (DR) - Identifying the domain's problem - Identifying suitable methods for DR prediction - Searching and analyzing previous research related to DR 	ResearchGate, ACM Digital Library, Wikipedia, Scopus	<ul style="list-style-type: none"> - Background of study - Research objectives - Scope of study - Research significance
	Knowledge Acquisition	<ul style="list-style-type: none"> - Knowledge gathering and acquisition - Reviewing the literature of the domain and standard 		<ul style="list-style-type: none"> - Literature Review analysis with related work

		techniques used for the problem - Literature review analysis with related work		
To design a machine learning model to predict diabetic retinopathy (DR).	Data Collection	Use secondary data from the ophthalmology department	Ophthalmology Specialist, Department of Ophthalmology Hospital Al Sultan Abdullah	Dataset of factors that contributing to DR.
	Data Pre-processing	<ul style="list-style-type: none"> - Drop unnecessary columns like ID, Name, First Visit, Height, Weight, and Marital Status. - Handle missing values - Scale numerical features - Encode categorical features for compatibility 	Raw dataset of factors that contributing to DR.	Clean Dataset of factors that contributing to DR.
	Model Development	<ul style="list-style-type: none"> - Set model parameters (train-test split, selected features, evaluation metrics). - Train machine learning models: Logistic Regression, Random Forest, SVM, KNN, AdaBoost, and XGBoost. - Optimize model 	Dataset, Machine Learning libraries	Trained and Optimized Machine Learning Models

		performance using GridSearchCV for hyperparameter tuning. - Integrate trained models with a scalable system architecture.		
	Model Evaluation	- Evaluate the model's performance using metrics like accuracy, precision, recall, F1 score, ROC-AUC	Machine Learning libraries, Evaluation Metrics	Evaluation Results
To develop a prototype for predicting diabetic retinopathy (DR) using a Machine Learning model.	Prototype Development	- Code the machine learning model - Develop the user interface (UI) for the prediction prototype - Integrate the machine learning model with the UI	Python, Machine Learning libraries	Prediction Prototype of Diabetic Retinopathy
	Prototype Testing	- Test the system by using original dataset	Original dataset from Hospital Al Sultan Abdullah	Prototype Testing Results
	Documentation	- Write a full report of the project	Research Findings, Prototype Documentation	Full FYP Report

3.2 Preliminary Study

The initial study stage is an exhaustive conceptual exploration that will set the baseline toward a more detailed study focusing on how machine learning can be used to forecast diabetic retinopathy (DR). This stage starts with a contextual analysis, exploring the nature and implications of DR and searching for particular problems in the field that need solving. It involves formulating suitable methods for DR prediction by examining the available machine learning approaches and their application. The search and analysis of the existing DR studies is particularly important at this stage because it assists in pinpointing the existent study deficiencies and the evolving trends in the specialty. An extensive multifaceted dataset is obtained through the careful selection of credible sources such as ResearchGate, ACM Digital Library, Wikipedia, and Scopus. These stages aid in constructing a comprehensive study background, delineating the problem and purpose of the study, narrowing the aim and setting clear research objectives, and outlining pertinent research questions.

In particular, the results of this phase include a detailed introduction to the study by the context, problem statement, aim, scope, significance, and comprehensive literature review. This phase also includes the search for and analysis of literature relevant to the inquiry, which is fundamental to the preliminary investigation defined above. This piece forms the basis of a literature review that would enhance the understanding of Diabetic Retinopathy (DR) as well as its risk factors. Additionally, it is important to evaluate the machine learning techniques that have been implemented by other scholars to estimate the occurrence of DR. The study of the literature suggests that the Logistic Regression model is very popular in the scientific community and in particular, greatly overshadows other models in accuracy with regards to the prediction of diabetic retinopathy.

3.3 Knowledge Acquisition

The Knowledge Acquisition phase assists in predicting diabetic retinopathy (DR) through machine learning, and provides an outline for later work. As such, it consists of collecting and evaluating DR's etiological, symptomatic, and pathological information alongside other case studies. Collection of DR predictive models demands a thorough review of provided subject literature. This process, at its core, includes previously described articles from scholarly journals and reputable academic repositories such as ResearchGate, ACM Digital Library, or Scopus. This technique is meant to allow scholars to pinpoint gaps left by predecessors, as well as analyze recently advanced sophisticated predictive models. A summary and conclusion of the literature evaluation serves the intent of providing knowledge on new technologies and techniques, and formulating well-defined issue statements, study aims, and hypotheses. This phase generates a detailed literature review that directs research during subsequent stages, an accurately phrased issue description, a comprehensive investigational back drop, and specific research objectives.

3.4 Data Collection

The next stage in this study is the data collection. For the purposes of this research, the dataset was acquired from Dr. Azimah in association with the Hospital Al-Sultan Abdullah's Department of Ophthalmology. It is made up of organized medical data like patient demographics, clinical data, and DR diagnostic information. The dataset comprises 389 patient records and has both the independent variables as well as the dependent variable which is the diagnosis of DR. These features involve vital health parameters such as patient's ID, name, date of first visit, age, sex, blood glucose levels, blood pressure and other relevant clinical phenomena.

The target variable is binary, with the following categories:

- i. 0 - Non-Diabetic Retinopathy (Non-DR): Patients without signs of DR.
- ii. 1 - Diabetic Retinopathy (DR): Patients diagnosed with DR.

ID	Name	FirstVisit	Age	Gender	Ethnicity	MaritalStatus	weight	height	BMI	SystemicDiabetes	DurationDiabetes	Hypertension	Cholesterol	Neuropathy	Neuropathy	DiabeticFoot	Stroke	AMD	Asthenia	Asthma	Osteoporosis	IMR	Diabetic_Retinas		
CTC0005706	ZURIANI CHE NOOR	07/25/2018	51	2	1	2	76.7	153.01	33	131	95	2	1	0	0	0	0	0	1	0	0	1	7.7	0	
CTC0017652	ABD RAHIM WAHID	2/1/2017	64	1	1	2	82	165	29	127	67	23	1	0	1	0	0	0	0	1	0	0	1	8.1	1
1810548653	MUSTAPA KADIM	8/6/2018	60	1	1	2	113.2	188	40	147	71	12	1	0	1	0	0	1	0	0	0	0	1	8.87	0
CTC0009285	SHARAH AWALDEEN	12/4/2017	52	2	1	2	78.4	155.5	30	133	68	2	1	0	0	0	0	0	1	0	0	0	1	6.9	0
CTC0003981	GAHARAH AIP PORNUSAMPI	06/15/2017	72	2	9	2	75	161	29	135	73	21	1	0	0	0	0	0	1	0	0	0	0	8.87	1
CTC0016124	SABLI OTHMAN	11/1/2018	59	1	1	2	83	171.3	31	143	83	1	1	0	0	1	0	0	0	0	0	0	1	9.8	0
CTC0034047	ZULKARNAIN DIN	12/25/2018	56	1	1	2	85.5	172	29	135	75	12	1	0	0	0	0	0	1	0	0	0	0	10.1	0
1810506485	PAULINE KIM CHEE LAM	11/3/2014	51	2	2	2	74.3	162	28	143	89	15	1	0	0	0	0	0	0	0	0	0	1	11.3	0
CTC0009142	HABINAH BINTI IBRAHIM	11/4/2015	61	2	1	2	59.1	151	26	148	86	14	1	0	0	0	0	0	0	0	0	0	1	10.7	1
CTC0002068	NAGIE IBRAHIM	01/16/2014	55	1	1	2	61	161	29	146	123	7	1	0	1	0	0	0	0	1	0	0	1	7.4	1
1810511502	RANGGAWANGI	05/18/2014	64	2	9	2	86.65	154.99	36	137	73	2	1	0	0	0	0	0	0	0	0	0	1	6.7	0
CTC0015294	ASHAH NORDIN	01/18/2017	64	2	1	2	83.1	161	29	108	58	10	1	0	0	0	0	0	0	1	0	0	1	9	0
CTC0009884	MADNAN ALI	12/16/2014	62	2	1	2	86.3	161.01	33	145	98	12	1	0	0	0	0	0	0	0	0	0	1	10.6	1
CTC0012991	AINI SALMAH	01/18/2018	51	2	1	2	66.4	156.01	27	132	73	20	1	0	0	0	0	0	0	0	0	0	1	9.5	0
CTC0010176	LAU MEE NGIEW	06/14/2014	60	2	2	2	59.9	154.31	25	124	78	12	1	0	0	1	0	0	0	0	0	0	1	8.87	0
CTC0004185	RAGALI SULAIMAN	05/13/2018	55	1	1	2	91.9	166.5	33	134	85	7	1	0	0	0	0	0	0	0	0	0	1	7.4	0
CTC0013565	NORHAYATI SULAIMAN	9/7/2018	58	2	1	2	69.1	157	28	137	78	10	1	0	0	0	0	0	0	0	0	0	1	12	0
CTC0009262	FATIMAH MO LIM	01/18/2018	64	2	1	2	58	140	30	143	68	11	1	0	0	1	0	0	0	0	0	0	1	13.1	1
1810505325	SAFIYAH NADIR	7/2/2013	71	2	1	3	56.2	157.51	23	121	58	20	1	0	0	0	0	0	0	1	0	0	1	5.9	0
CTC0008284	HAMIR HAMZAH	10/26/2015	52	1	1	2	70.5	165	26	150	68	8	1	0	1	0	0	0	0	1	1	0	1	8.1	1
CTC0008571	NORHANI SHA ABU	01/20/2015	73	2	1	2	65.5	143.01	23	121	71	12	1	0	0	0	0	0	1	0	0	0	1	9	0
CTC0003449	SHUKRI CHE YEN	08/06/2014	42	1	1	2	92.95	179.2	31	172	100	11	1	0	0	1	0	0	0	0	0	0	1	10.4	1
CTC0000362	ZAHIDA RAMLI	1/6/2015	63	2	1	2	63.7	154.99	27	130	65	5	1	0	0	0	0	0	0	1	0	0	1	7	0
1810506675	AZIZ YOUSIF	05/27/2015	59	1	1	2	92	175.01	30	106	58	29	1	0	1	1	1	1	0	1	1	0	0	8.67	1

Figure 3.2 Dataset of Diabetic Retinopathy in excel format

3.5 Data Pre-processing

Prior to the training and testing phases, the data set of this study underwent significant data pre-processing. This step was crucial towards ensuring that the dataset was accurate, uniform and was capable of operating at the peak of model efficiency. Several procedures were tailored, which included the handling of missing values, normalization, and encoding. The collection contains data from several sources concerning diabetic retinopathy, including some patient demographics, physical measurements, and health issues. The base dataset consisted of 388 records which included different attributes, many of which required alterations to be suitable for machine learning processes.

The First Visit temporal variable was omitted from consideration, and columns with no relevance such as ID, Name, and FirstVisit were eliminated. The ID and Name did not have the ability to predict diabetic retinopathy, therefore, they were excluded. The FirstVisit being a temporal column did not possess direct relevance to the predictive features of diabetic retinopathy, and thus was not included in the consideration. Consequently, eliminating these columns results in a more focused dataset, minimising non-functional dimensions and improving the efficacy of the machine learning models.

A preliminary analysis of the dataset revealed a mixture of numerical and categorical features, with several missing values distributed across

different columns. The following table summarizes the dataset before pre-processing:

Table 3.2: The table highlights the missing data and standardization of features.

Feature Name	Data Type	Missing Values	Unique Values
Height	Numerical	4	102
Weight	Numerical	3	273
BMI	Numerical	3	30
Diabetic_Retinopathy	Target (Binary)	0	2

The first step of the process was handling missing values on both numerical and categorical columns. The missing numerical values such as Height and BMI were filled through median imputation which did not alter the data distribution. Figure 3.3 shows the visualization of missing values before pre-processing was done:

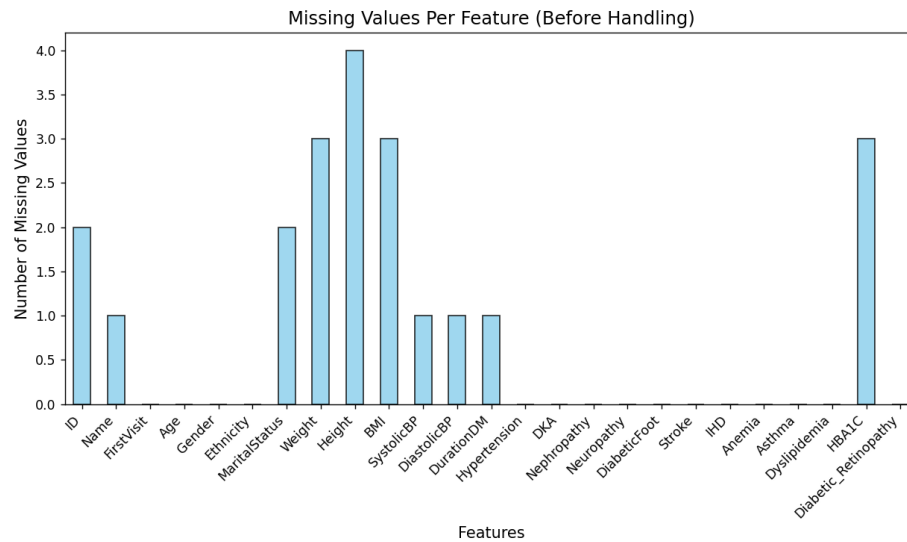


Figure 3.3 The bar chart of missing values before handling

All the numerical characteristics were normalized using the Min-Max method, which converts numbers to a range of values between 0 and 1. This stage facilitates faster convergence in the training stage while ensuring that no single

feature exerts too much influence on the model. Unprocessed values for key numerical attributes before and after standardization are shown below:

Various categorical variables like Gender were transformed into numerical formats using the Label Encoding technique. This replaced each category by an integer. The data had become suitable for machine learning algorithms because the category variable was now model-friendly. Diabetic_Retinopathy is the target variable which indicates whether the patient was Positive or Negative to the signs of diabetic retinopathy was having. Within such a framework for binary classification, it is expected that the goal variable is operationalized consistently across the sample. The last step after all pre-processing steps were conducted was saving the dataset in an Excel file for model training and evaluation. The revised dataset was capable of being machine learning ready because it was normalized, clean, and well-structured for accurate pattern and correlation identification for the prediction of diabetic retinopathy.

3.6 Model Development

The next step is developing machine learning models. The aim is to create a classification model for the prediction of diabetic retinopathy with the aid of machine learning and other data-based techniques. The models that were selected for this study are Logistic Regression, Random Forest, Support Vector Machines, K-Nearest Neighbors, as well as ensemble techniques like AdaBoost and XGBoost. The procedures followed for modeling and evaluating these models are provided in more detail later in the section.

a) Prepare the Model Parameters

The very first step of a model development process always starts with setting model parameters for training. Different form of machine learning algorithms were used including Logistic Regression, Support Vector

Machines, Random Forest, KNN, and ensemble techniques like AdaBoost and XGBoost. These models were trained and tested against both the standard settings and the optimized settings.

Table 3.3 The summarized of parameters used during the training process

Parameters	Value
Train-Test Split Ratio	80:20 (Stratified)
Features	Demographics, Physical Measurements
Target Variable	Diabetic Retinopathy (Binary)
Number of Features	Varies
Evaluation Metrics	Accuracy, ROC-AUC, Precision, Recall, F1
Optimization Algorithm	GridSearchCV for hyperparameter tuning

b) System Architecture

The system architecture is the computing paradigm describing the operations of the system. Figure 3.4 presents one design for the system architecture in this study.

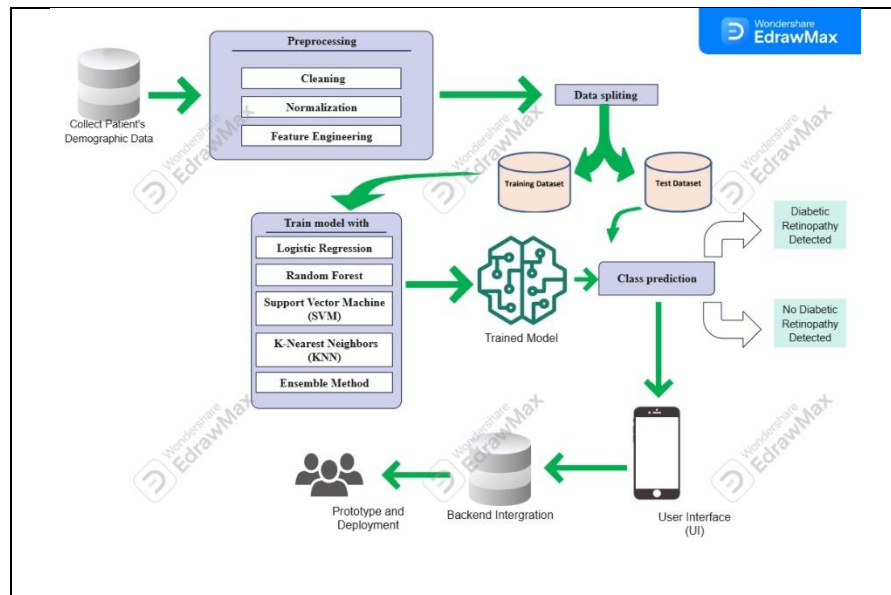


Figure 3.4 System Architecture

The procedures and techniques involved in the data architecture resemble the major components of the Diabetic Retinopathy Prediction project. Moreover, it was stratified into training and testing sets at an 80/20 split. The analysis of the diabetic retinopathy prediction was performed using multiple different models like Logistic Regression, Random Forest, SVM, KNN, ensemble models with step-wise XGBoost and AdaBoost machine learning models. Frontend for the healthcare providers allows seamless interaction with the system for data input, while predictions are provided on the spot, enabling instant access to system a vast amount of data.

c) Experiment Evaluation

The models were analyzed from three different perspectives to compare their performance and reliability. First, the models were validated through cross-validation where the data set is split into folds for training and testing. Primary performance metrics, precision, and recall, F1 score, and ROC AUC were calculated for each fold and overall and imposed over the IUC population. In addition, a class imbalance management strategy through up-sampling scheme where the class representatives are increased by artificially reproducing copies within the class was implemented. These tests on created model-trained tests allowed for greater creativity towards how the model bounds are defined. The model results were further enhanced for evaluation against the conditions where synthetic data provided by Mostly.ai was used. This allowed for further exploration of how adaptable and effective the model is when trained using various sources. Results from all three strategies of evaluation were then analyzed for the most effective.

d) Design User Interface (UI)

This stage includes the design of the user interface (UI) of the Diabetic Retinopathy Prediction System. The role of user interface is crucial in ensuring user and system interaction as it allows healthcare personnel or other users to input patient data and receive predictions for analysis.

The system is capable of collecting users data, such as age, gender, ethnicity, marital status, weight, height, BMI, and blood pressure, in a user-friendly manner as they engage with the interface. The inputs are essential for the system to generate precise predictions. The interface design contains buttons that allow the user to choose categories, which enhances user interaction with the system. The prediction systems design is user-friendly by making the system easy to use and interact with as shown in Figure 3.5. The layout is simple, clean, and professional, but still places importance on user accessibility and efficiency.

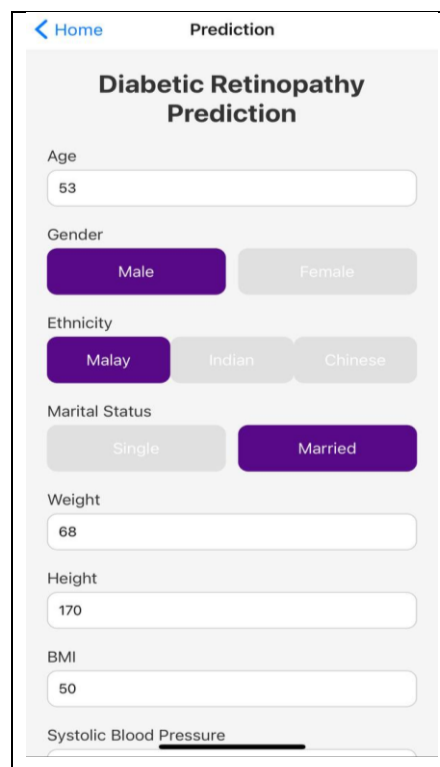
The image shows a mobile application interface for predicting Diabetic Retinopathy. At the top, there is a navigation bar with a blue back arrow and the text 'Home', and a title 'Prediction'. Below this is a header section with the title 'Diabetic Retinopathy Prediction'. The form contains several input fields and buttons: 'Age' with a text input field containing '53'; 'Gender' with two buttons, 'Male' (purple) and 'Female' (grey); 'Ethnicity' with three buttons, 'Malay' (purple), 'Indian' (grey), and 'Chinese' (grey); 'Marital Status' with two buttons, 'Single' (grey) and 'Married' (purple); 'Weight' with a text input field containing '68'; 'Height' with a text input field containing '170'; 'BMI' with a text input field containing '50'; and 'Systolic Blood Pressure' with a text input field. The interface is clean and professional, with a light grey background and purple accents for selected options.

Figure 3.5 User Interface (UI) for predict Diabetic Retinopathy

3.7 Model Evaluation

Evaluation of diabetic retinopathy classification model performance was done using numerous metrics which were accuracy, ROC AUC, precision, recall and F1 score. The multifaceted metrics are put in place to ensure the model is robust. Accuracy in particular measures the overall correctness of the prediction made, while ROC AUC measures the model's discriminative ability between positive and negative cases. Furthermore, precision measures the number of true positive predictions out of all positive predictions made, while recall measures the proportion of actual positive cases that were correctly identified. F1 score measures the balance between precision and recall.

Table 3.4 The result of model evaluation for 80/20 ratio using cross-validation

Model	Accuracy	ROC-AUC	Precision	Recall	F1-Score
Logistic Regression	78.21%	81.61%	0.76	0.75	0.75
Support Vector Machine	73.08%	79.00%	0.71	0.72	0.71
Random Forest	79.49%	85.65%	0.8	0.78	0.79
K-Nearest Neighbors	66.67%	73.64%	0.67	0.56	0.61
AdaBoost	71.79%	72.49%	0.67	0.75	0.71
XGBoost	69.23%	77.25%	0.68	0.64	0.66

When evaluating the models, Random Forest on the 80/20 ratios, ratio random sampled with replacement, was the most accurate at classifying cases of diabetic retinopathy with 79.49%. This model also achieved an ROC-AUC of 85.65%, confirming strong capability to classify cases of diabetic retinopathy. Logistic Regression on the other hand had an accuracy rate of

78.21%. This model, in addition, had a ROC-AUC of 81.61%, which shows that the model is quite reliable. Finally, with Support Vector Machine, they achieved an accuracy of 73.08% while ROC AUC was at 79.00%, showing moderate effectiveness. Of the few, the worst performing was XGBoost, this model obtained 77.25%, whereas AdaBoost achieved a percentage of 72.49%. k-NN was the least accurate, with only 66.67% correct classification while achieving ROC AUC of 73.64%, therefore, lower capacity to treat this dataset in comparison to other models.

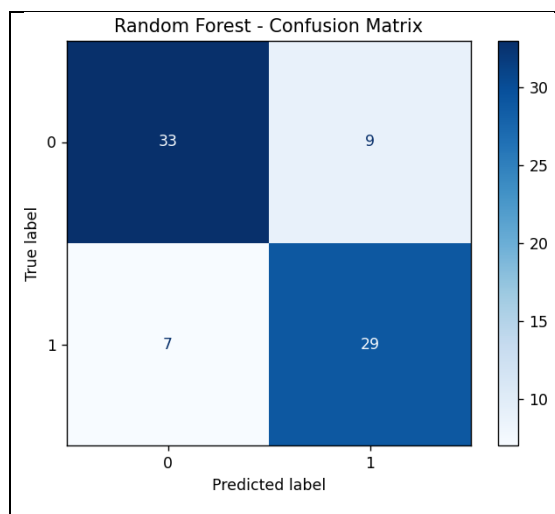


Figure 3.6 The confusion matrix for Random Forest

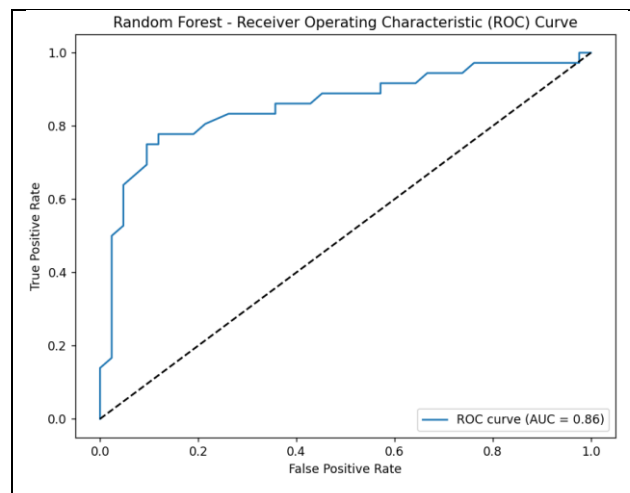


Figure 3.7 The ROC Curve for Random Forest

Were these matrices provided to achieve a deeper elucidation regarding the distribution of each method's true positives, true negatives, false positives, and false negatives? The matrices rendered crucial support in determining precisely where a model did not perform well in the case of errors. Moreover, all models ROC curves were also plotted to show how each model represented class separation visually. Random Forest achieved the highest distinction in class separation visually. Random Forest achieved the highest distinction in class discrimination which was also attested by its high ROC and AUC scores.

Through these assessments, Random Forest was selected as the optimal model for the project. Its accurate results and its capability to classify the presence of diabetic retinopathy cases versus non cases supports the classification task the best. This analysis asserts that Random Forest is exceptionally good and trusted in automatic diabetic retinopathy detection.

3.8 Prototype Development

A Diabetic Retinopathy Prediction Prototype was designed to be as effective, easy to use, and reliable as possible in addressing a significant challenge in the health sector. The objective of this investigation was to design a mobile phone application that has the potential to indicate a lack of diabetic retinopathy and offer relevant guidance to medical personnel.

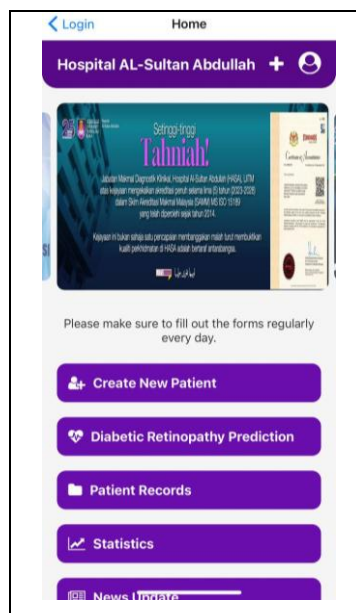


Figure 3.8 The main page for mobile apps

The prototype was built with a Flask API running on PythonAnywhere which interfaces with an SQLite database to store patients' documents and health data. A Logistic Regression model is developed to anticipate the likelihood of a patient having diabetic retinopathy, which drives the initial analytics. The application was built using React Native, which increases usability across devices while also improving user experience design.

a) Encoding Process

Figure 3.11 display the process of encoding that use for prototype which are executed in Visual Studio Code. The program is coded using python language.

A screenshot of the Visual Studio Code editor showing a Python file named 'app.py'. The code defines a Flask prediction route. It includes a list of required fields for the prediction model, checks for missing fields in the incoming JSON data, and prints the received data for debugging. The code is as follows:

```
80 | # Prediction route
81 | @app.route('/predict', methods=['POST'])
82 | def predict():
83 |     try:
84 |         data = request.json
85 |
86 |         # Ensure the correct number and order of fields
87 |         required_fields = [
88 |             'Age', 'Gender', 'Ethnicity', 'MaritalStatus', 'Weight', 'Height', 'BMI',
89 |             'SystolicBP', 'DiastolicBP', 'DurationDM', 'Hypertension', 'DKA',
90 |             'Nephropathy', 'Neuropathy', 'DiabeticFoot', 'Stroke', 'IHD', 'Anemia',
91 |             'Asthma', 'Dyslipidemia', 'HBA1C'
92 |         ]
93 |
94 |         # Check for missing fields
95 |         for field in required_fields:
96 |             if field not in data:
97 |                 return jsonify({'error': f'Missing field: {field}'}), 400
98 |
99 |         # Debug: Print received data and prepared features
100 |         print('Received data:', data)
101 |         features = np.array([data[field] for field in required_fields])
```

Figure 3.9 Encoding process using Visual Studio Code

b) Prediction process

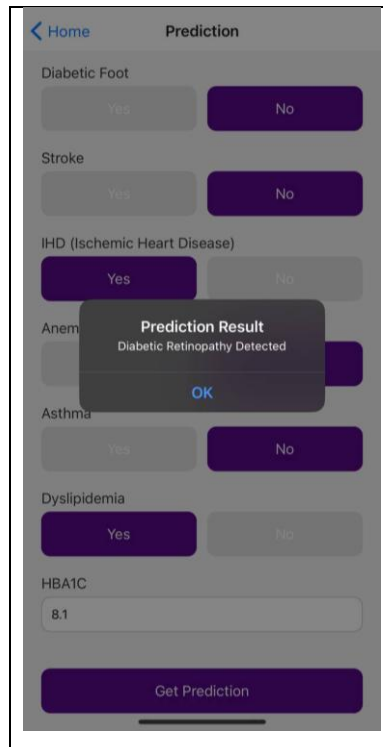


Figure 3.10 Pops up message appears for making a prediction

A pre-trained Random Forest model which is part of the Flask API is used for the prediction in the prototype. When the mobile application receives input through user submission like age, BMI, blood pressure, HBA1c levels and other health variables, those inputs are sent to the backend for processing. This approach takes advantage of the trained model by giving the model data and making the device perform computations to estimate the chances of the patient having diabetic tardive retinopathy. This information is then communicated back to the user in a simple and intuitive manner, like providing a diagnosis in the application's interface, which aids in early and proactive diagnosis and disease management.

3.9 Prototype Testing

The testing will focus on using the original data set from the Hospital Al Sultan Abdullah for testing the prototype of the Diabetic Retinopathy prediction system. This allows one to verify whether the system is functional and correct, and if it can actually be utilized in a clinical setting. Actual data is used for testing which enhances the credibility and effectiveness of the equipment. This information extracted from testing procedures should be used to further improve the performance of the system and the quality of service provided to users. The proposed comprehensively detailed test reports that define the capability and effectiveness of the system in real life are essential in this phase.

3.10 Documentation

The document stage consists of writing a report that explains all requirements and contains the R and D details of the development process of the diabetic retinopathy (DR) prediction system. This includes chronicling the study's background, problem statement, objectives, data collection and processing steps, model building and assessment, and system integration and deployment. Each section reviews the course of action and the challenges encountered. It includes code fragments, system design architecture, and the user interface designs of the study. In addition, manuals are prepared for users that would assist medical practitioners in utilizing the system. In the end, thorough documentation is produced that encapsulates the whole project for understanding and further research or applications.

3.11 Summary

As detailed in chapter 3, machine learning aids in DR prediction through a multi-faceted approach that can be divided into nine steps. These steps include preliminary study, knowledge acquisition, data collection, data preprocessing, model development, model evaluation, prototype development, prototype testing, and documentation. The process begins with identifying the problem, followed by literature reviews to gain some insights, then structured collection, and finally pre-processing to make the information viable for the machine learning models to utilize. The models developed and tested were Logistic Regression, Random Forest, SVM, and KNN. Among these the most performing algorithm was the Random Forest algorithm. Prototype development comprises a number of different stages and all of them will have to be tested and reviewed by experts so that it results in a practical application. This is required so that the goal of the research is met while still ensuring that the prediction of the disease is done as early on as possible.

CHAPTER 4

RESULT AND FINDINGS

This chapter explains the results and findings of the research after the methodology has been implemented. This chapter includes the images pre-processing results, experiments results, model evaluation, and system evaluation.

4.1 Overview

This research considers multiple machine learning models, such as Logistic Regression, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbours (KNN), XGBoost, and AdaBoost, to predict diabetic retinopathy. To assess each model's ability to generalize, an 80/20 train-test split was used. Random Forest had the best model performance after several evaluation measures like accuracy, ROC-AUC, precision, recall, F1-score were taken into account. Cross-validation was used to test models on multiple data subsets to look into model performance deeper. This procedure reduced overfitting and improved the accuracy of the prediction of the model's performance on unknown data.

Aside from cross-validation, experiments through Mostly.ai's synthetic data generation, along with an attempt to address class imbalance through upsampling, were performed to assess model performance. To address possible class discrepancy within the dataset's minor class, upsampling auxiliary data points was implemented for the underrepresented group. This model was designed to help the algorithm learn to predict both classes successfully. Also, realistic synthesized augmentations were produced through the synthetic data generation tool Mostly.ai in order to allow the models to train on a more comprehensive and heterogenous dataset. The combination of

these approaches increases the accuracy and the durability of the models, allowing for more reliable predictions of diabetic retinopathy.

4.2 Experiment 1: Cross-Validation

4.2.1 Result of Evaluation

Table 4.1 Model Performance Metrics

Model	Accuracy	ROC-AUC	Precision	Recall	F1-Score
Logistic Regression	78.21%	81.61%	0.76	0.75	0.75
Support Vector Machine	73.08%	79.00%	0.71	0.72	0.71
Random Forest	79.49%	85.65%	0.8	0.78	0.79
K-Nearest Neighbors	66.67%	73.64%	0.67	0.56	0.61
XGBoost	71.79%	72.49%	0.67	0.75	0.71
AdaBoost	69.23%	77.25%	0.68	0.64	0.66

Table 4.1 displays results for various machine learning models created to identify diabetic retinopathy using cross-validation techniques. Out of all tested algorithms, Random Forest was able to achieve 79.49% accuracy and 85.65% ROC-AUC score, making it the best performing model by far. Logistic Regression showed high level of competency in this binary classification task with accuracy of 78.21% and ROC-AUC of 81.61%. Other models, including SVM and K-Nearest Neighbors, did not perform so well at these tasks, with SVM having 79.00% ROC-AUC and K-Nearest Neighbors lagging behind at 73.64%. This difference might indicate an inability to perform well on unbalanced datasets. On the other hand, XGBoost did relatively well with a score of 77.25% while AdaBoost did significantly worse at only 72.49%. Both

precision and recall highlight the next level of detail in the composite measures and this measure of performance is known as the trade-off between false positives and false negatives. In this case, the ensemble methods as well as logistic regression have shown to be quite effective.

4.2.2 Confusion matrix:

a) Logistic Regression

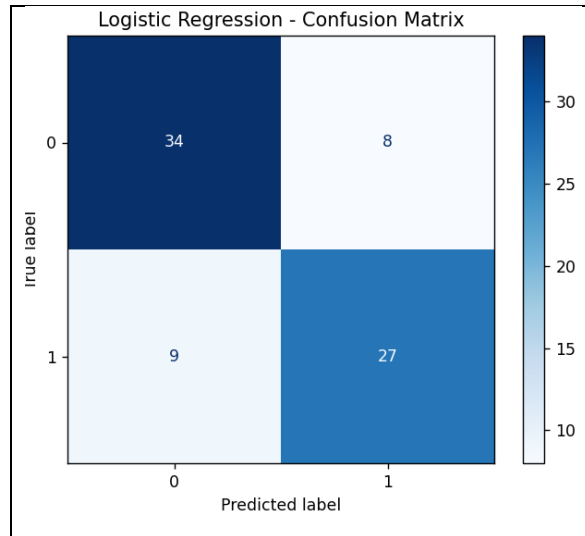


Figure 4.1 Confusion matrix logistic regression using cross validation

Confusion matrix from Logistic Regression illustrates capability regarding diabetic retinopathy cases. Out of total prediction, it identified 34 instances correctly as non-DR, whereas 27 are DR patients and correctly identified those as True Positives. At the same time, this model incorrectly identified 8 non-DR patients as DR, known as false positive, and 9 DR cases as non-DR known as False Negatives. These findings give a fairly reasonable balance between sensitivity, or true positive rate, and specificity, or true negative rate, showing that the model is useful; however, there is still room for improvement in minimizing misclassification.

b) Support Vector Machine

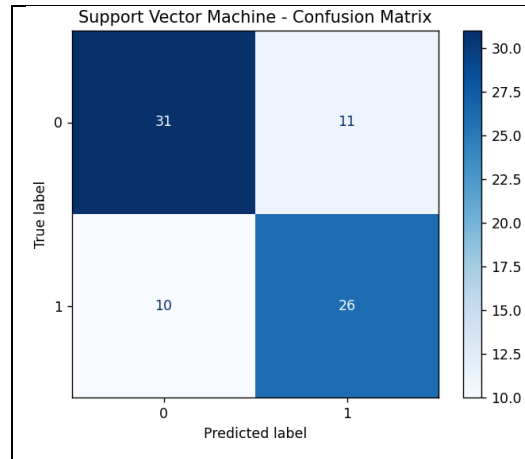


Figure 4.2 Confusion matrix SVM using cross validation

The confusion matrix of the SVM model has projected the classification capability for the cases of diabetic retinopathy. It rightly classified 31 non-DR as true negatives and 26 DR as true positives. Whereas, 10 non-DR patients who were misclassified as DR are the false positives, and 11 DR cases that are taken as non-DR are the false negatives. The larger false negative rate suggested that there was some limitation in recognizing DR instances reliably, which may affect its clinical usefulness while the model performed moderately.

c) Random Forest

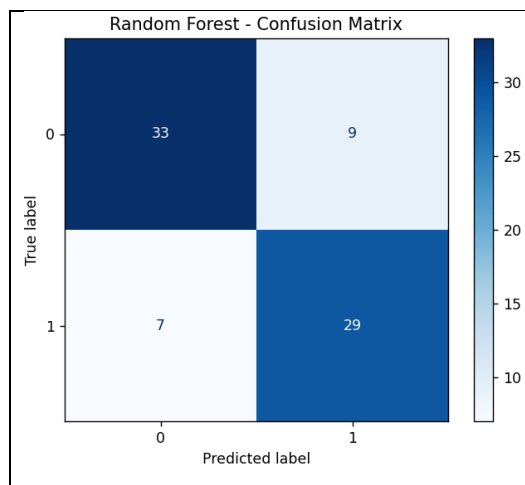


Figure 4.3 Confusion matrix Random Forest using cross-validation

The confusion matrix of the Random Forest model reflects a good performance of the model in the prediction of diabetic retinopathy. It has rightly predicted 29 non-DR cases as true negatives and 33 DR cases as true positives. However, it has wrongly identified 9 false positive predictions where non-DR cases were predicted as DR and 7 false negatives where DR cases were predicted as non-DR. This model presents a high degree of effectiveness and accuracy in view of a relative low number of misclassifications, and therefore, it can be reliably used for early detection of diabetic retinopathy.

d) K-Nearest Neighbors

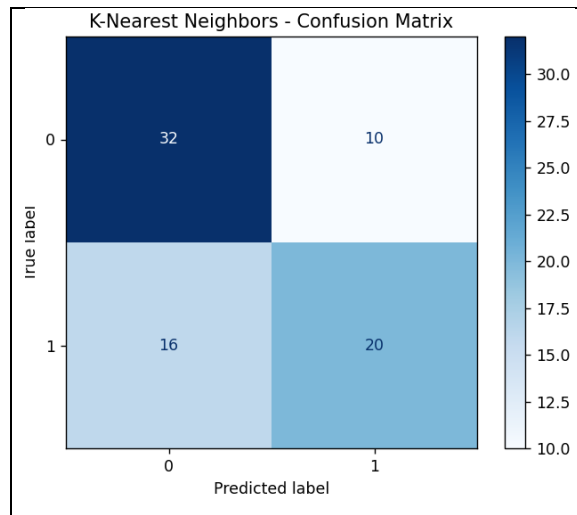


Figure 4.4 Confusion matrix for KNN using cross validation

The confusion matrix shows that the model KNN should give a very moderate performance when it comes to the prediction of Diabetic Retinopathy. There are 20 non-DR and 32 DR correctly predicted, while it has misidentified 10 as non-DR patients suffering from DR and 16 DR patients as non-DR. The larger share of false negatives indicates an inability to identify cases of DR more precisely, which places certain restraints on sensitivity and makes the model less reliable compared to other models for early identification of diabetic retinopathy.

e) **XGBoost**

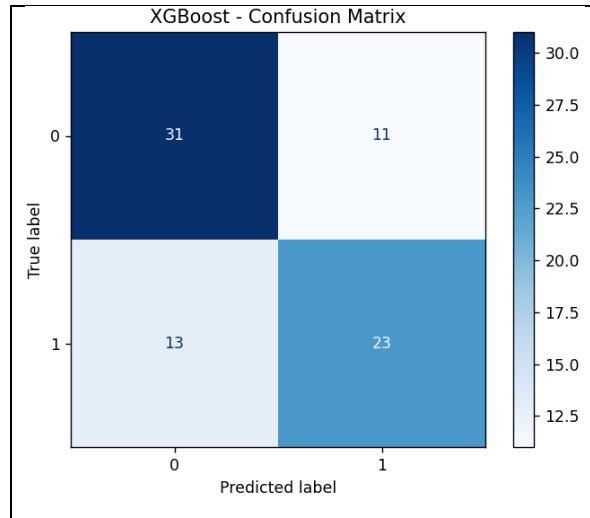


Figure 4.5 Confusion matrix XGBoost using cross validation

The confusion matrix of the XGBoost model represents the performance of the model in the classification of diabetic retinopathy. It rightly classified 31 non-DR cases as actual negatives, while 23 DR cases were actual positives. It misclassified 11 non-DR instances as DR false positives and 13 DR cases as non-DR false negatives. Although XGBoost had decent accuracy, the presence of both false positives and false negatives shows that some more tweaking might achieve a better sensitivity and precision, which is required to be dependable in clinical usage.

f) AdaBoost

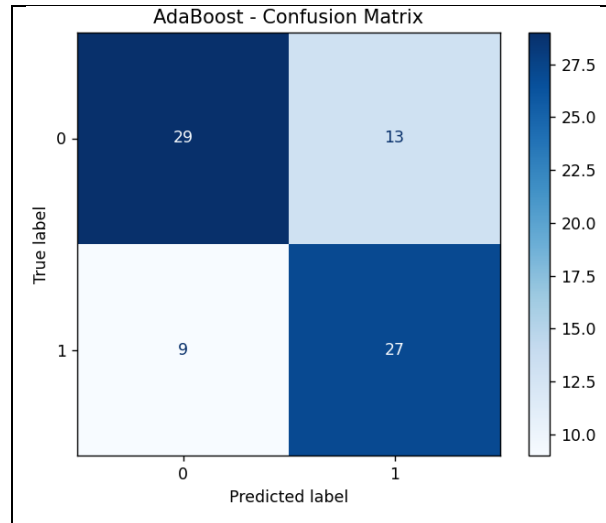


Figure 4.6 Confusion matrix AdaBoost using cross validation

The confusion matrix of the AdaBoost model is a representation of its ability to classify diabetic retinopathy instances. It correctly identified 27 non-DR and 29 DR, while it misclassified 13 as DR who are actually non-DR and 9 vice versa. That means AdaBoost is going well with a decent balance of true positives and true negatives; the false positive rate is still a concern, hence showing room for potential growth regarding precision.

4.2.3 ROC Curve

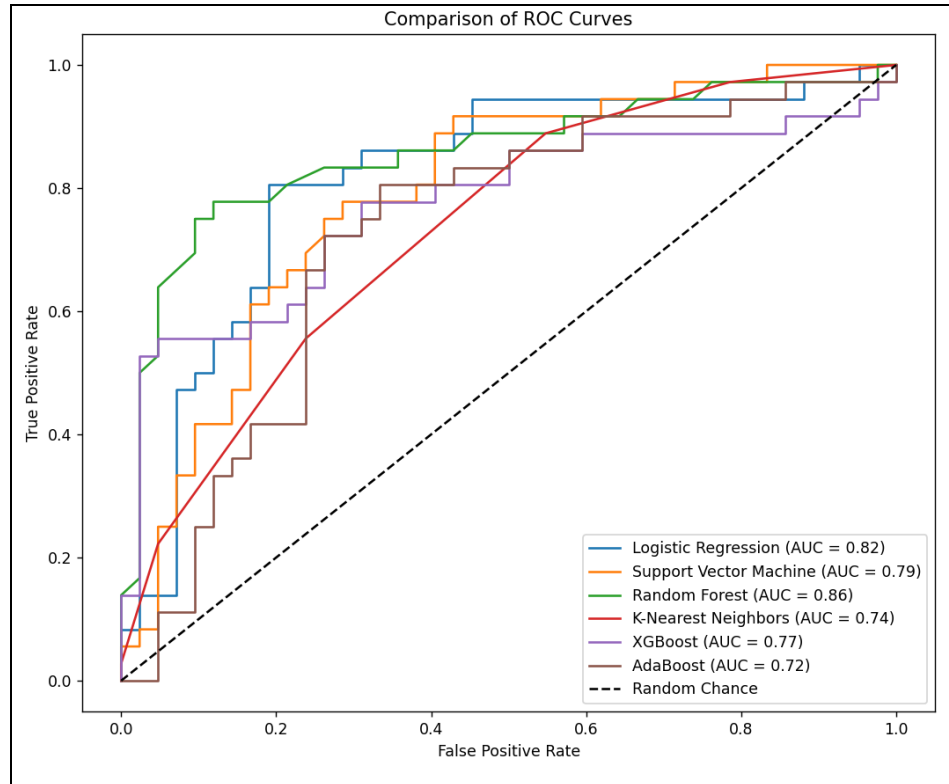


Figure 4.7 ROC curve for all model by using Cross-Validation

The ROC curve comparison in Figure 4.7 represents the performance of various machine learning models that are cross-validated in predicting diabetic retinopathy. The Random Forest model, with the best AUC score of 0.86, has the highest capability in distinguishing between positive (DR) and negative (non-DR) situations. Logistic Regression followed with an AUC of 0.82, showing its strong classification capability. Support Vector Machine achieved an AUC of 0.79, indicating moderate performance. With regard to AUC values, XGBoost showed 0.77, the AdaBoost model showed a bit worse, with 0.72, and, finally, K-Nearest Neighbors got an AUC value of about 0.74, thus showing its low suitability to the data under consideration. At the same time, it's worth noting that the curve corresponding to the Random Forest model keeps the closest to the top-left corner, meaning excellent trade-off sensitivity and specificity make this model top ranking among those compared within this study.

4.3 Experiment 2: Upsampled Data

Class imbalance is a prevalent issue in machine learning classification tasks particularly in medical datasets like diabetic retinopathy, where the minority class (DR patients) frequently has fewer samples than the majority class. To overcome this issue, the Synthetic Minority Oversampling Technique (SMOTE) was used. SMOTE generates synthetic samples by interpolating between existing minority-class samples, thereby balancing the dataset while keeping its structure. This strategy optimizes the training process by allowing models to learn from both classes equally.

By using SMOTE, the number of DR and Non-DR samples was equalized, mitigating the risk of models being biased toward the majority class. This balanced dataset would allow the models to improve their ability to detect DR cases accurately which leading to better recall and overall F1-scores. The impact of upsampling is reflected in the improved evaluation metrics across all models, as detailed in the subsequent comparison table.

4.3.1 Result of accuracy, ROC-AUC, precision, recall, and F1-score

Table 4.2 Comparison Original result and Upsampling result

Model	Accuracy (Cross- Validation)	Accuracy (Upsampled Data)	Precision (Cross- Validation)	Precision (Upsampled Data)	Recall (Cross- Validation)	Recall (Upsampled Data)	F1-Score (Cross- Validation)	F1-Score (Upsampled Data)
Logistic Regression	78.21%	71.00%	0.76	0.72	0.75	0.72	0.75	0.71
Support Vector Machine	71.79%	70.00%	0.71	0.70	0.67	0.70	0.69	0.70
Random Forest	79.49%	79.00%	0.80	0.79	0.78	0.79	0.79	0.79
K-Nearest Neighbors	65.38%	74.00%	0.65	0.74	0.58	0.73	0.61	0.73
XGBoost	74.36%	76.00%	0.74	0.76	0.69	0.76	0.71	0.76
AdaBoost	69.23%	73.00%	0.67	0.73	0.67	0.72	0.67	0.72

The table compares the results between the models that was trained using the original dataset and the ones trained using the upsampled dataset. It is interesting to note that most models reverted back to using the original datasets SMOTE upsampled strategy with class imbalance tackle because most of them significantly improved in regards to recall and F1 scores. This shows that the upsampling technique employed was indeed effective in improving the models ability to increase true positives relative to the number of false negatives for diabetic retinopathy instances. For example, K-Nearest Neighbours (KNN) achieved exceptional results and led the rest of the class with an accuracy rise of 8.62 percent to a whopping 74.00% and an increase in recall from 58% to 73%. Distance based models are very much favored by balanced datasets.

The Random Forest ensemble model, with its accuracy and consistency of 79.00%, remained at the top position. This model has led in performance together with attaining balanced precision, recall and F1 scores. With the application of upsampling, both Logistic Regression and XGBoost improved their recall and average balanced performance, while the Support Vector Machine (SVM) achieved only a moderate degree of success. The AdaBoost model, on the other hand, performed better with the upsampled dataset because of the consistent improvement seen in accuracy, balanced precision and recall. The results in total provide an understanding of class imbalance issues with regards to the models as the sensitivity to imbalance data significantly determines the performance of the model.

4.3.2 Confusion matrix

a) Logistic Regression

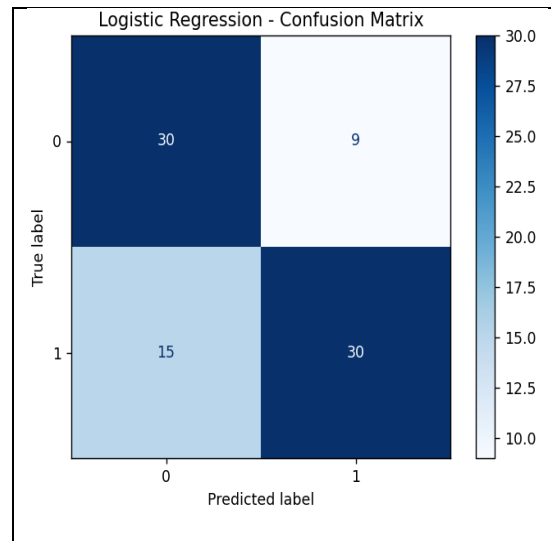


Figure 4.8 Confusion matrix Logistic Regression using Upsampled Data

The confusion matrix of the Logistic Regression model, trained with upsampled data, reflects the balancing of the dataset on the classification performance. It correctly classified 30 non-DR cases as true negatives and 30 DR cases as true positives, showing that it was balanced in its capability to recognize both classes. However, it misclassified 9 non-DR cases as DR and 15 DR patients as non-DR. While the upsampling strategy increases the recall by increasing the true positive rate, the much higher number of false negatives suggests that the model is still far from correctly identifying all the cases of DR; hence, it has room for more sensitivity.

b) Support Vector Machine

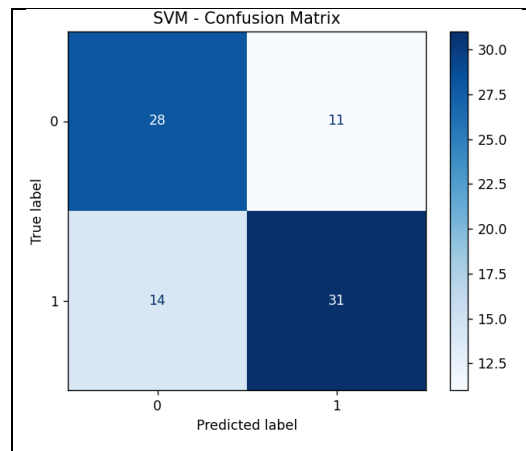


Figure 4.9 Confusion matrix for SVM using Upsampled Data

The confusion matrix presents the performance of Support Vector Machine modeling in classification tasks using upsampled data. The model has a great balance in identifying both classes: it correctly predicts 31 cases as DR positives and 28 Non-DRs as true negatives, while at the same time misclassifying 14 DR subjects as Non-DR (false negatives) and 11 Non-DR subjects as DR (false positives). Although the SVM model has a high capacity after upsampling, this is manifested by the comparably balanced true positives and true negatives, while the presence of false negatives suggests that further optimization may be necessary to improve memory and sensitivity for the exact recognition of DR situations.

c) Random Forest

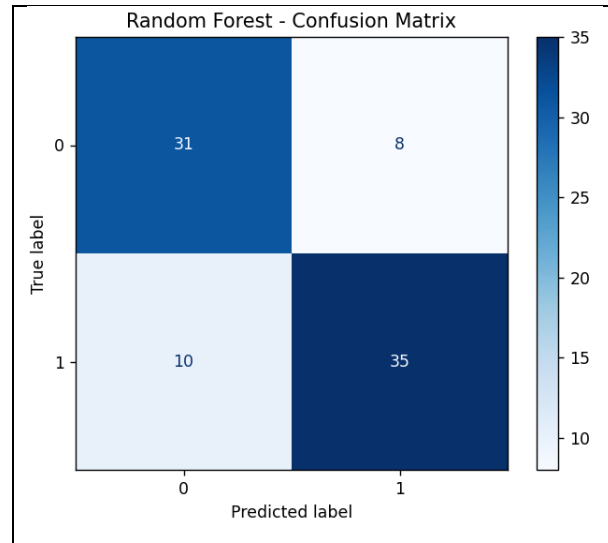


Figure 4.10 Confusion matrix for Random Forest using Upsampled Data

The Random Forest model on the upsampled data performed well in classifying Non-DR/DR. The model was able to accurately detect 31 Non-DR as true negatives and 35 DR cases as true positives. This shows that the model is able to effectively identify both classes. On the other hand, 10 DR cases were falsely detected as Non-DR – thus, were classified under ‘DR’ (false negatives). Some Non-DR cases, eight in total, were also classified as DR which is another false positive. The degree of these small misclassifications reveals how well the model balanced the precision and recall measures and solved the class imbalance problem simultaneously. Because of these reasons, the Random Forest model is the most appropriate model for applying the diabetic retinopathy test in this specific experiment.

d) K-Nearest Neighbors

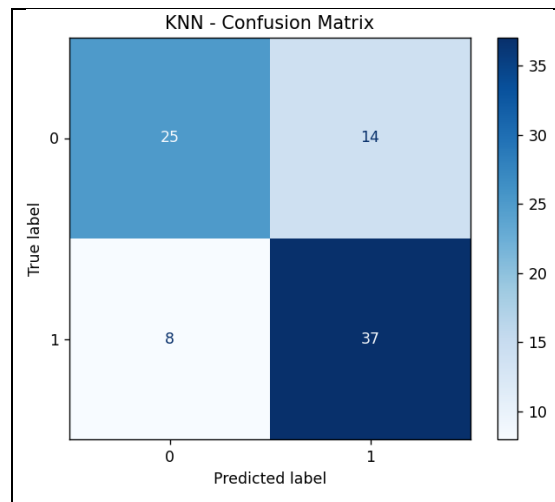


Figure 4.11 Confusion matrix for KNN using Upsampled Data

The confusion matrix representing the K-Nearest Neighbours (KNN) classification model performance with an upsampled test set indicates the algorithm was able to correctly classify 37 DR cases (True Positives) and 25 Non-DR cases (True Negatives). However, 8 DR cases have been misclassified as Non-DR (False Negatives) and 14 Non-DR cases have been misclassified as DR (False Positives). Considering the low number of false negatives resulting from upsampling, one could conclude that the sensitivity of the model to detecting DR instances has increased. Along with this improvement in sensitivity, there seems to be an increase in the number of false positives which do suggest that the model has become too liberal, and tends to make accurate diagnoses at the expense of some accuracy. Indeed, the KNN model with a bias class correction funnel oversensitivity feature usage but also oversensitivity restriction and bias avoidance enabled so DR cases did not decrease too severely without control.

e) **XGBoost**

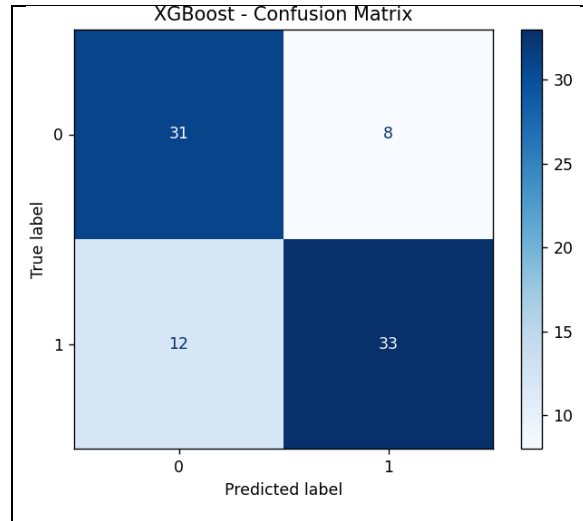


Figure 4.12 Confusion matrix for XGBoost using Upsampled Data

XGBoost model performance classification confusion matrix. The algorithm achieved a true positive rate of 33 in correctly classifying DR cases. In addition, 31 Non-DR cases were also successfully classified as correct true negatives. However, out of 12 DR cases, 12 were incorrectly assigned as Non-DR, which are considered false negatives, and out of the 8 Non-DR cases, 8 were marked as DR, thus being regarded as false positives. The absence of many false negatives such as these indicates that the model has adequate sensitivity in recognising DR cases. Of note, the presence of false positives, on the other hand, suggests that the model does perform reasonably well in recognising DR cases, but tends to overpredict to some extent, which greatly diminishes specificity. Under these particulars, the XGBoost model demonstrates satisfactory performance with regard to the detection of DR cases but it equally puts an emphasis on renouncing detail.

f) AdaBoost

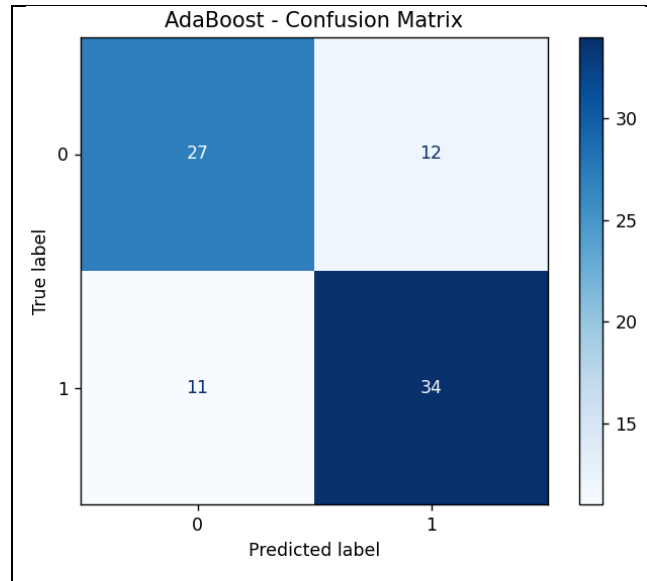


Figure 4.13 Confusion matrix for AdaBoost using Upsampled Data

The confusion matrix depicts how well the AdaBoost model performed on the upsampled data for the classification of diabetic retinopathy. The model correctly classified 27 Non-DR cases as true negatives and 34 DR cases as true positives. Nevertheless, the model incorrectly categorised 12 Non-DR cases as DR (false positives) and 11 DR cases as Non-DR (false negatives). The model does maintain a good precision and recall score; however, the greater of the two false positive cases suggests there is an overconfidence in predicting DR cases. In conclusion, AdaBoost performed superbly on the balanced dataset and remains, with the accepted accuracy, a dependable model for Non-DR and DR detection because of the good sensitivity and specificity.

4.3.3 ROC Curve

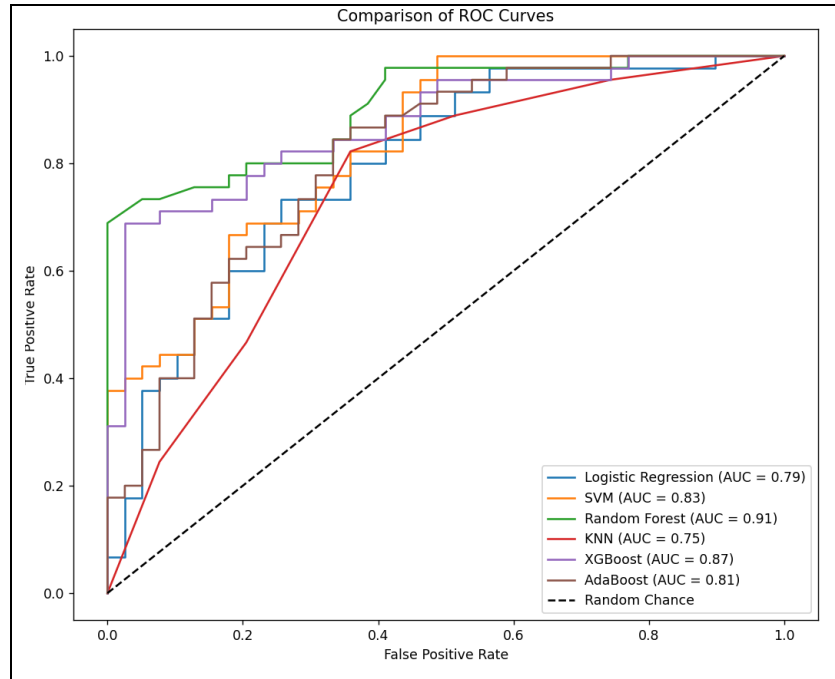


Figure 4.14 ROC Curve for all model by using Upsampled data

The ROC curve analysis for models developed with upsampled data indicates the models varying abilities to discern the presence or absence of diabetic retinopathy (DR) cases. The AUC score that the Random Forest model achieved is the highest of all at 0.91, which clearly means this model uses the best combination of false positive rate as measured through 1-specificity and true positive rate measured through sensitivity. XGBoost comes in second with an AUC of 0.87, also showing good performance with a balanced dataset. SVM comes third but with an AUC score of 0.83 which is also a considerable improvement over the original dataset. The regression models that performed the least were AdaBoost and conventional regression at 0.81 and 0.79 AUC score respectively. While KNN failed to demonstrate any good performance with an AUC of 0.75 which is the worst of all. In general, the upsampling strategy was helpful to underlying models and boosted their performance as seen in the case of Random Forest, which continues to be the best model for this problem as well.

4.4 Experiment 3: Mostly.ai

To address the issues of factoring data availability and privacy when creating synthetic diabetic retinopathy data, Mostly. AI was used. The platform uses privacy compliant approaches to create realistic synthetic datasets through statistical feature simulation. The synthetic data flawlessly replicated the relationships and structure within the original dataset, thus perfect for training machine learning models without violating privacy. The synthetic dataset was balanced with regard to DR and non-DR cases, thereby making it comparable to the upsampling procedure of the DR dataset. This approach solved the problems related to the initial imbalanced dataset and made model training unbiased. The useful data along with the generated data was verified for fidelity and alignment with the original data structure before being incorporated for model training and testing.

4.4.1 Model Performance Comparison

Table 4.3 Comparison performance with different of synthetic datasets

Model	Accuracy (888 rows)	ROC- AUC (888 rows)	Accuracy (988 rows)	ROC- AUC (988 rows)	Accuracy (1188 rows)	ROC- AUC (1188 rows)
Logistic Regression	55.62%	0.5690	58.08%	0.6102	56.72%	0.5778
Support Vector Machine	53.93%	0.5518	52.53%	0.5494	54.62%	0.5602
Random Forest	62.36%	0.5927	59.60%	0.5745	53.78%	0.5659
K-Nearest Neighbors	52.25%	0.5401	51.52%	0.5374	54.62%	0.5441
AdaBoost	56.18%	0.5845	58.59%	0.5850	57.98%	0.5742
XGBoost	55.62%	0.5842	56.06%	0.5936	57.56%	0.5813

In the Table 4.3, the results for the six models, namely, Logistic Regression, SVM, Random Forest, KNN, AdaBoost, and XGBoost are benchmarked with three synthetic datasets of sizes 888, 988, and 1088 rows. In the case of the 888 row dataset, Random Forest is able to achieve the maximum accuracy score first at 62.36%. However, due to overfitting or issues with generalizing on the synthetic data, its accuracy drops on the larger datasets. Conversely, XGBoost and AdaBoost consistently perform well across all dataset sizes. In addition, XGBoost demonstrated the lowest overfitting as compared to other models, which is also presented by the highest ROC-AUC value of 0.5936 and accuracy of 57.56% on the largest dataset.

In terms of the 988-row dataset, both SVM and Logistic Regression outperform others with moderate gains. It is apparent that logistic regression outperforms by reaching 58.08% accuracy and 0.6102 ROC-AUC, before declining slightly. However, K-Nearest Neighbours (KNN) display consistently low accuracy and ROC-AUC across all datasets, thereby showcasing the challenge posed by artificial data. In addition, the findings also show that Random Forest has a tough time generalising with increasing dataset sizes, while models such as XGBoost and AdaBoost do consistently well with fake data. From all considerations, the 988-row dataset is the most optimal for training and testing models. It strikes the greatest balance between ROC-AUC and accuracy - ensuring forecasted models that are precise and applicable on a large scale.

4.4.2 Confusion Matrix for 988 rows:

a) Logistic Regression

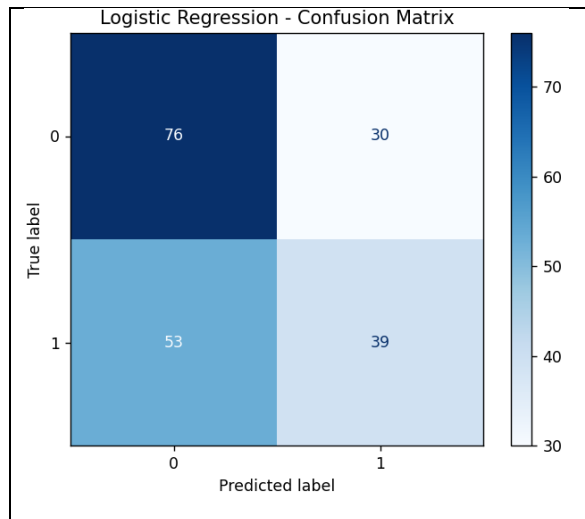


Figure 4.15 Confusion matrix for Logistic Regression using Mostly.ai

The fallout of this confusion matrix for the Logistic Regression classifier on the synthetic dataset of 988 rows shows that the model correctly flagged 76 true negatives (non-diabetic retinopathy cases) and 39 true positives (diabetic retinopathy cases). On the other hand, the classifier incorrectly flagged 30 cases as false positives whereas 53 cases were flagged as false negatives. The greater rate of false negatives reveals that the model is having difficulties detecting the presence of diabetic retinopathy cases while relatively better at distinguishing its absence. These do not suffice conclusively and suggest that optimisation or the inclusion of a considerably more competent model is needed to enhance the sensitivity in diabetic retinopathy identification.

b) Support Vector Machine

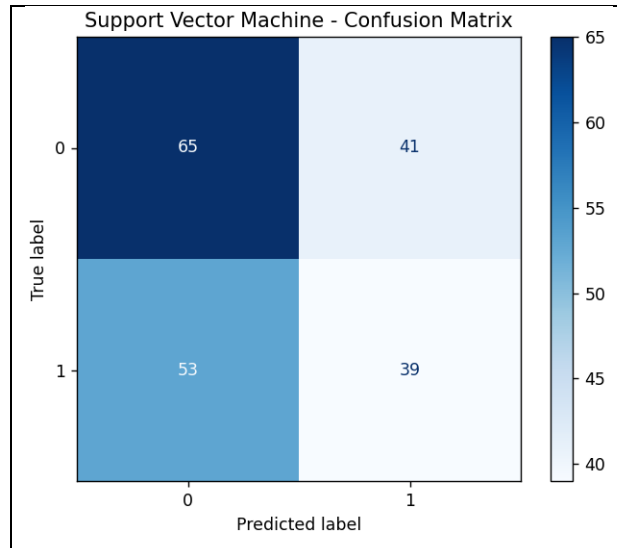


Figure 4.16 Confusion matrix for SVM using Mostly.ai

The classification of the SVM model on the synthetic dataset consisting of 988 rows illustrates that the model accurately identified 65 non-diabetic retinopathy cases as true negatives, along with 39 cases of diabetic retinopathy identified truly as positives. In addition, there were false positives and negatives of 41 and 53 respectively. The high count of false negatives suggests that the model has a very low sensitivity and is likely missing a substantial amount of actual instances of diabetic retinopathy which the SVM model is configured to detect. This indicates a need for further tuning of the model and improvement of features to achieve better and more reliable detection of diabetic retinopathy cases.

c) Random Forest

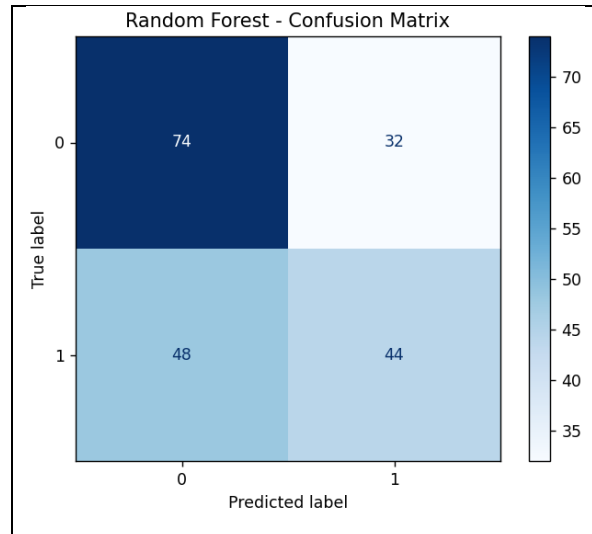


Figure 4.17 Confusion matrix for Random Forest using Mostly.ai

The confusion matrix of the Random Forest model on synthetic data with 988 rows indicates there are 74 true negatives and 44 true positives-that is, 74 correct instances of non-diabetic retinopathy and 44 correct instances of diabetic retinopathy are predicted by the model. Further, it has produced 32 false positives and 48 false negatives. The model performs decently in case identification, but the occurrence of false negatives indicates a great challenge in the complete detection of diabetic retinopathy cases, which may limit its clinical application. The sensitivity of the model and overall performance may further improve by better feature selection or tuning of hyperparameters.

d) K-Nearest Neighbors

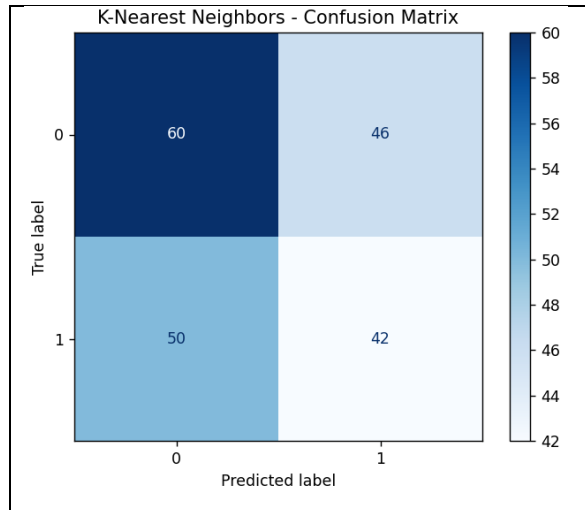


Figure 4.18 Confusion matrix for KNN using Mostly.ai

From the confusion matrix obtained on the synthetic data using the K-Nearest Neighbours model with 988 rows, it can be noticed that the true negatives and true positives from the exact predictions of the model are 60 and 42, respectively. Whereas, the number of false positives and false negatives generated from the model amounts to 46 and 50, respectively, showing an equal amount of misclassifications. Thus, it means that, in this context, KNN failed to reach a good balance between sensitivity and specificity, most likely owing to its vulnerability to feature scaling and noise in the dataset. Performance gains may be achieved after optimization by feature engineering or tuning parameters.

e) **XGBoost**

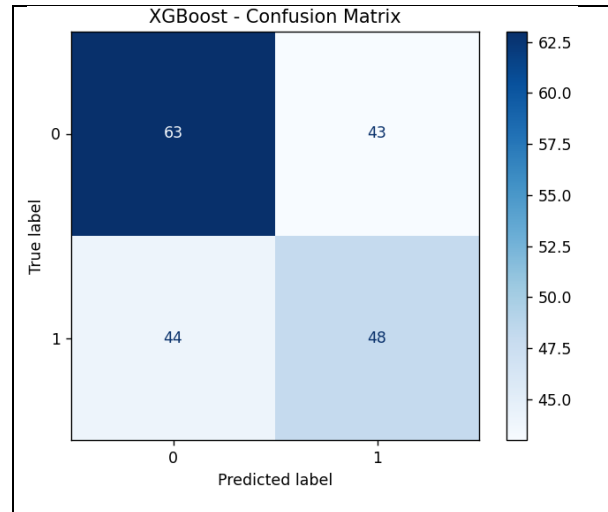


Figure 4.19 Confusion matrix for XGBoost using Mostly.ai

In the XGBoost model using a synthetic dataset, with 988 rows, the confusion matrix presents 63 true negatives and 48 true positives. The model performed correctly in these two respects, while in contrast, the model developed 43 false positives and 44 false negatives, relatively balanced compared with some of the models discussed in the text. According to Balance, XGBoost has demonstrated the capability of finding complex patterns in the data while showing good sensitivity and specificity. Its robust performance demonstrates its effectiveness for the present forecasting exercise but could be further improved with more tuning.

f) AdaBoost

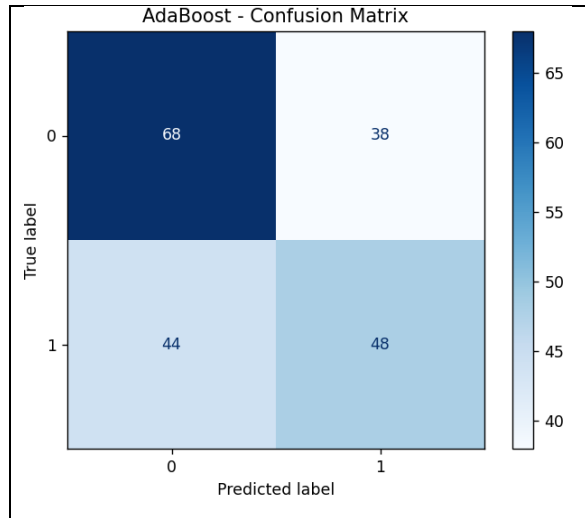


Figure 4.20 Confusion matrix for AdaBoost using Mostly.ai

The confusion matrix for the AdaBoost model with the synthetic dataset of 988 rows predicts 68 true negatives and 48 true positives, showing that the model has appropriately predicted these occurrences. Still, there are 38 false positives and 44 false negatives to show misclassifications in both categories. This finding suggests that AdaBoost, though balanced in its abilities to recognize positive and negative cases, would be better with more tuning or tweaks in performance to manage inaccurate predictions. It shows intermediate sensitivity and specificity for this activity.

4.4.3 ROC Curve

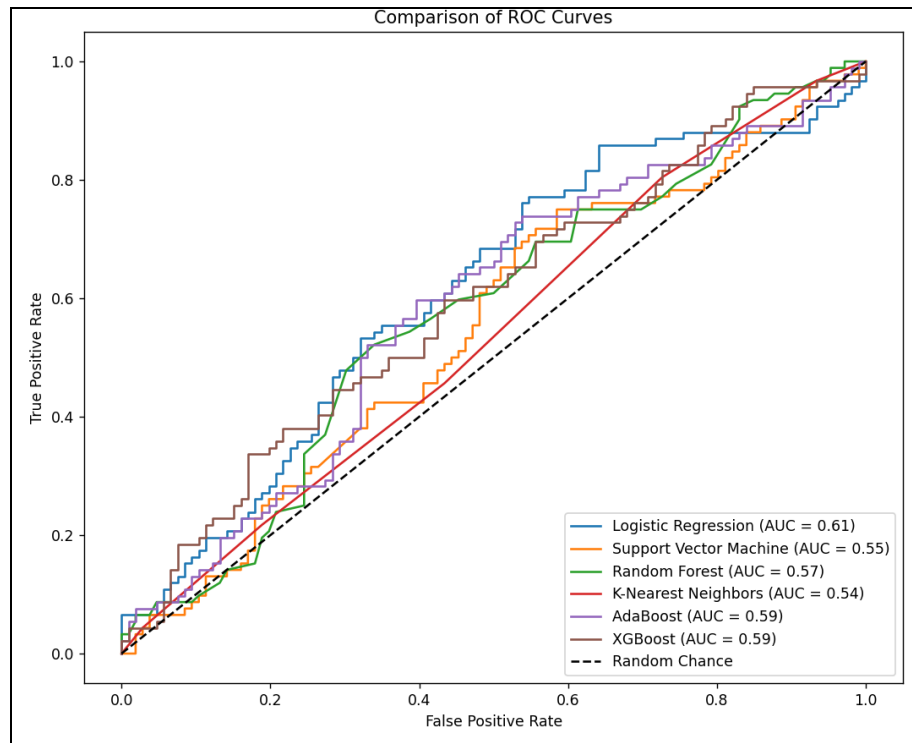


Figure 4.21 ROC Curve for all model by using Mostly.ai

The evaluation of the comparative ROC curves aims to understand the discrimination power of different models on a synthetic dataset with 988 rows created for prediction of diabetic retinopathy. The highest AUC of 0.61 was achieved by Logistic Regression, indicating its highest ability to discriminate between the positive and negative cases. Both AdaBoost and XGBoost recorded AUC of 0.59, suggesting that they performed equally well over the entire dataset. Random Forest achieved an AUC of only 0.57 and did not perform well as expected from accuracy in the earlier metrics. Support Vector Machine (AUC = 0.55) and K-Nearest Neighbours (AUC = 0.54) performed worse than the former two models and were also considered to have low discrimination accuracy well above random guessing. Out of all these models, it is clear that with this particular synthetic dataset, Logistic Regression is the best suited model as it is proven to be the most effective tool in identifying diabetic retinopathy cases.

4.5 Model Evaluation

The effectiveness of the machine learning models was determined through the analysis of comparison for three experiments, which are cross validation, upsampling, and production of synthetic data with the help of Mostly.AI. Random forests achieved the highest accuracy and ROC-AUC on both cross validation and upsampled datasets, demonstrating XGboost and AdaBoost's reliable and ROC-AUC accuracy. Additionally, Random Forest showed the best ROC-AUC and accuracy among K Nearest Neighbor, Logistic Regression, and Support Vector Machine (SVM) outperformed moderate when compared to synthetic datasets, but upsampling provided the best results for both of them. The remaining two models, K Nearest Neighbors, consistently recorded underscored scores, describing the model's performance in complex and high dimensional data. KNN scored the worst compared to the other models.

Table 4.4 Overall performance for every model

Model	Cross-Validation Accuracy	Upsampled Data Accuracy	Synthetic Data Accuracy
Logistic Regression	78.21%	71.00%	55.62% - 58.08%
Support Vector Machine	71.79%	70.00%	52.53% - 54.62%
Random Forest	79.49%	79.00%	62.36% - 53.78%
K-Nearest Neighbors	65.38%	74.00%	51.52% - 54.62%
AdaBoost	69.23%	73.00%	56.18% - 57.98%
XGBoost	74.36%	76.00%	55.62% - 57.56%

Table 4.4 shows what is the summary overall accuracy of the models considering cross validation, upsampling, and creation of synthetic datasets. Random Forest achieved the highest accuracy during cross validation (79.49%) and upsampling (79.00%), which proves that he learns from original

samples and balanced samples. However, his accuracy on synthetic data sets was slightly lower, meaning that he was affected by the data generation procedures. XGBoost and AdaBoost performed well and consistently in all experiments including upsampling for which XGBoost scored an accuracy of 76.00%. Although all both models Xgboost and adaboost declined with the synthetic data, they still retained their competitiveness and good ROC-AUC scores.

Both the Logistic Regression and Support Vector Machine produced average results. After applying upsampling, the performance improved, but these algorithms trained on more complex synthetic data seemed to indicate that they were not very adaptable. K-Nearest Neighbours always had the lowest score out of all the tested algorithms. Even though its performance improved after upsampling, KNN still had the lowest accuracy on synthetic data, which is probably due to the sensitivity of KNN to complex high dimensional data. Thus, for the analysis of Diabetic Retinopathy dataset, Random Forest with upsampling was the most appropriate model and technique.

4.6 System Evaluation

The performance of the Diabetic Retinopathy Prediction System was tested on an independent test dataset, which was a subset of the original dataset of 30 patients at the Department of Ophthalmology, Hospital Al-Sultan Abdullah, for correct classification of the patients as having or not having DR. This dataset has not been used during training and validation, hence, a fair evaluation of the performance of the system can be assured. It was now matched with the ground truth labels of the test dataset. Actual testing had started where errors and misclassifications are recorded, and results were analyzed for precision and recall. Tabulated summary evaluation result in terms of precision and recall.

Table 4.5 Result of prototype testing by using original dataset

Patient	Patient ID	Age	Gender	Actual DR	Predicted DR	Result
1	1810018440	54	Female	No DR	No DR	Classified
2	CTC0008434	60	Male	DR	No DR	Misclassified
3	CTC0004997	64	Male	No DR	DR	Misclassified
4	CTC0003492	53	Female	DR	No DR	Misclassified
5	CTC0000630	61	Male	DR	DR	Classified
6	CTC0000319	61	Male	No DR	DR	Misclassified
7	CTC0008375	58	Male	DR	DR	Classified
8	CTC0023299	63	Male	No DR	No DR	Classified
9	CTC0008338	57	Female	No DR	No DR	Classified
10	CTC0001828	59	Female	No DR	No DR	Classified
11	1810023341	49	Female	No DR	No DR	Classified
12	1810141309	51	Male	No DR	DR	Misclassified
13	1810020644	65	Female	No DR	No DR	Classified
14	CTC0009689	62	Male	DR	DR	Classified
15	CTC0007113	53	Male	No DR	No DR	Classified
16	1810110437	58	Male	No DR	DR	Misclassified

17	CTC0005546	52	Female	No DR	No DR	Classified
18	CTC0011446	65	Female	DR	No DR	Misclassified
19	1810016675	51	Male	No DR	DR	Misclassified
20	CTC0013964	28	Male	DR	DR	Classified
22	CTC0007707	55	Male	DR	DR	Classified
23	CTC0009258	70	Female	DR	No DR	Misclassified
24	CTC0007561	57	Female	No DR	No DR	Classified
25	1810036541	57	Female	DR	No DR	Misclassified
26	CTC0006762	61	Male	DR	DR	Classified
27	1810137745	57	Female	DR	DR	Classified
28	CTC0005613	64	Female	DR	No DR	Misclassified
29	1810030163	73	Male	No DR	No DR	Classified
30	CTC0009209	63	Female	No DR	No DR	Classified

The evaluation results show that from the cohort of 30 patients, the system correctly classified 18, hence yielding a classification accuracy of 60%. Nine patients with "No DR" and nine patients with "DR" were correctly predicted. A total of 12 cases were misclassified, hence giving a misclassification rate of 40%. The misclassifications mostly occurred when the system predicted "No DR" for patients with "DR" or contrariwise. These results are indicative that although the system presents a modest accuracy, further optimization is necessary to improve the ability of the system in distinguishing between diabetic retinopathy and nondiabetic retinopathy cases.

4.7 User Interface

The UI of the system, developed on React Native, offers a responsive and user-friendly experience to healthcare professionals. The features include appointment management, add and tracking of patient records, and diabetic retinopathy prediction, news update from official website of hospital and gallery of the hospital. The UI seamlessly integrates with a Flask API hosted on PythonAnywhere, ensuring secure and efficient data exchange. Features like interactive charts for statistical insights, form-based patient data entry, and predictive analytics enhance the usability and effectiveness of the application.

i) Homepage

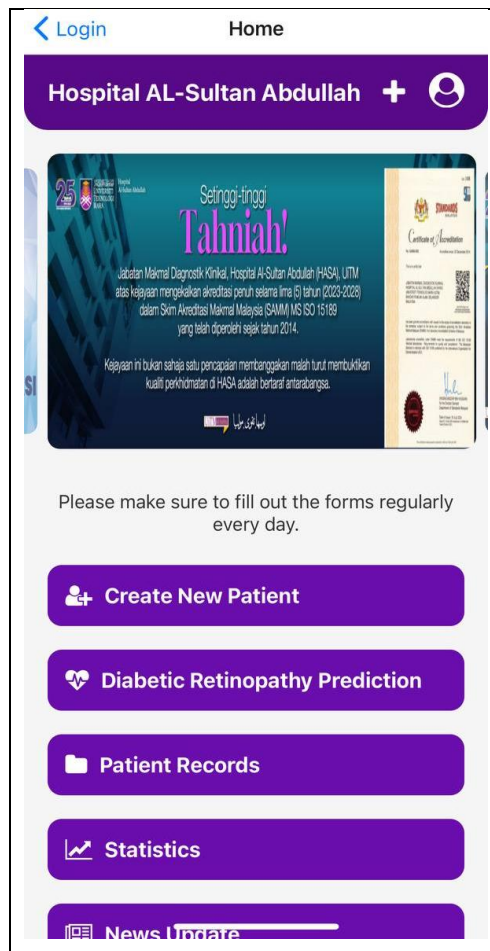


Figure 4.22 The homepage for mobile apps

The homepage in Figure 4.22 of the application is the main navigating center through which a user can jump to any core functionality of the application in no time. It contains a highlighted carousel to be used for announcing and updating information. Immediately after the carousel, it should remind the users to fill in the necessary forms on a regular basis. Major functionalities are reached by oversized, well-separated buttons like new patient record creation, diabetic retinopathy prediction, view of patients' records, statistical data view, checking for news updates, and view galleries. Structure-wise, clean and well-organized, the theme is very comfortable to work with, allowing health professionals to handle their tasks in a hassle-free manner.

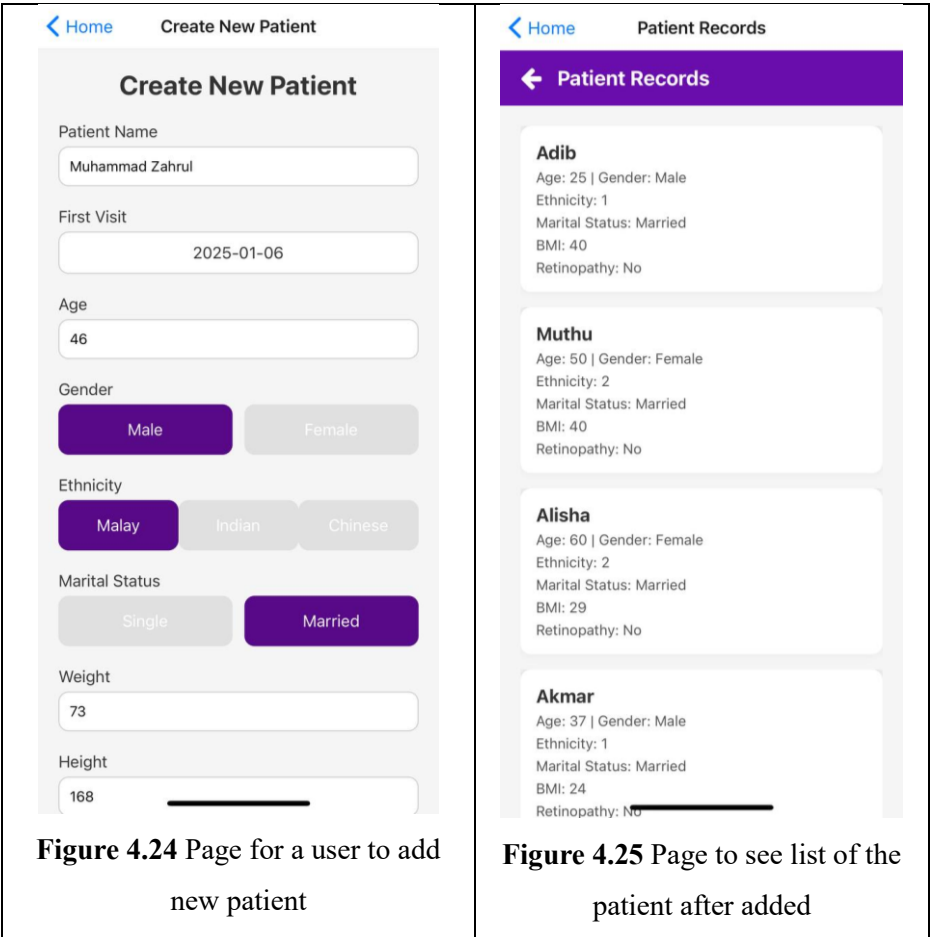
ii) Prediction Page

The screenshot shows a mobile application interface for a 'Diabetic Retinopathy Prediction' page. At the top, there is a navigation bar with a back arrow and the text 'Home', and a title 'Prediction'. Below this, the main heading is 'Diabetic Retinopathy Prediction'. The form contains several input fields and buttons: 'Age' with the value '60', 'Gender' with 'Male' selected, 'Ethnicity' with 'Malay' selected, 'BMI' with the value '33', 'Systolic Blood Pressure' with the value '127', 'Diastolic Blood Pressure' with the value '65', 'Duration of Diabetes Mellitus' with the value '12', and 'Hypertension' with 'Yes' selected. The 'Yes' button is highlighted in purple.

Figure 4.23 Page for user to make prediction

The Prediction page in Figure 4.23 provides the health expert with an online system for estimating the possibility of a diabetic retinopathy problem based on information that includes every aspect of the patients' information: age, gender, ethnicity, along with BMI, systolic and diastolic blood pressure, duration of diabetes mellitus, hypertension, DKA, nephropathy, neuropathy, diabetic foot, stroke, IHD, anemia, asthma, dyslipidemia, and HBA1C level. It provides such an ample amount of information that it then systemizes inputs and creates a prediction toward early detection and personalized care planning. This form is quite intuitive and thus efficient to navigate.

iii) Add and Record of Patient Page



Patient Records

Adib
Age: 25 | Gender: Male
Ethnicity: 1
Marital Status: Married
BMI: 40
Retinopathy: No

Muthu
Age: 50 | Gender: Female
Ethnicity: 2
Marital Status: Married
BMI: 40
Retinopathy: No

Alisha
Age: 60 | Gender: Female
Ethnicity: 2
Marital Status: Married
BMI: 29
Retinopathy: No

Akmar
Age: 37 | Gender: Male
Ethnicity: 1
Marital Status: Married
BMI: 24
Retinopathy: No

Figure 3.1.3 illustrate the "Create New Patient" interface, which allows healthcare practitioners to enter detailed patient information such as demographics and health characteristics for customized diabetic retinopathy estimates. Figure 3.1.4 shows the "Patient Records" feature, which provides an organized overview of patient profiles and medical information, hence optimizing data management for effective clinical decision-making.

iv) Visualization Page

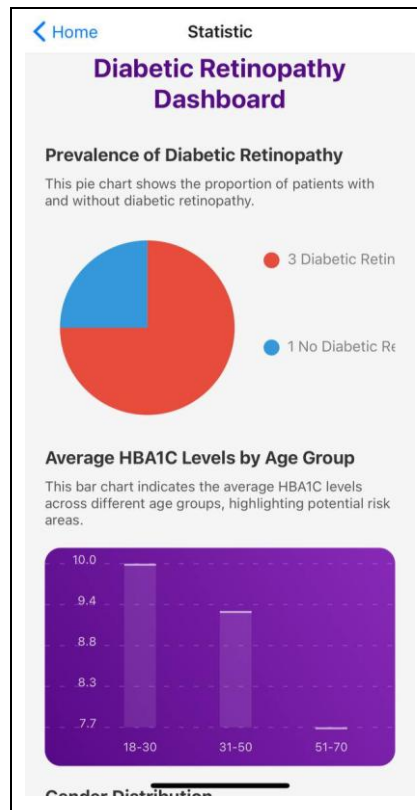


Figure 4.26 The dashboard by demographic patient data

Figure 3.1.5 illustrate the Diabetic Retinopathy Dashboard which gives a visual information about crucial patient statistics. The pie chart illustrates the prevalence of diabetic retinopathy among individuals distinguishing those who diagnosed from those who do not have the ailment. Furthermore, the bar chart depicts average HBA1C levels across age groups, emphasizing high-risk categories. These visualizations help healthcare personnel quickly spot trends and prioritize care for at-risk patients, thereby improving data-driven decision making.

v) News and Gallery Page

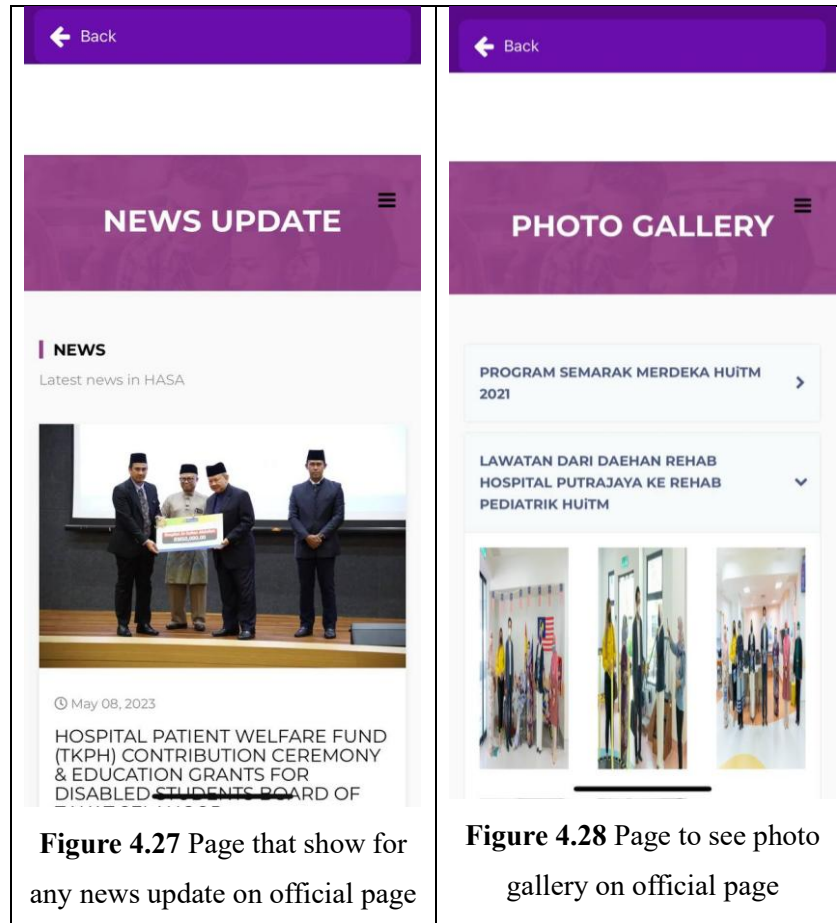


Figure 4.27 Page that show for any news update on official page

Figure 4.28 Page to see photo gallery on official page

News and Galleries provide access to the news about very recent events, including photo highlights concerning the hospital. The news page presents information on the most recent happenings, events, or announcements, including welfare fund contribution or grants. To this respect, it allows the user not to miss something important. The Galleries page displays photos of important programs and events at the hospital, including visits and celebrations that have occurred at the hospital for which photo documentation should be viewed by users. Both pages will be integrated into the system very well to give exciting and educative experiences to the users.

vi) **Create and Records of Appointment**

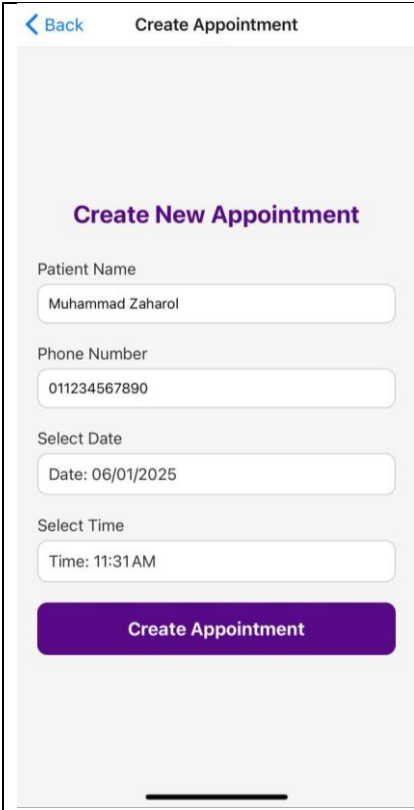
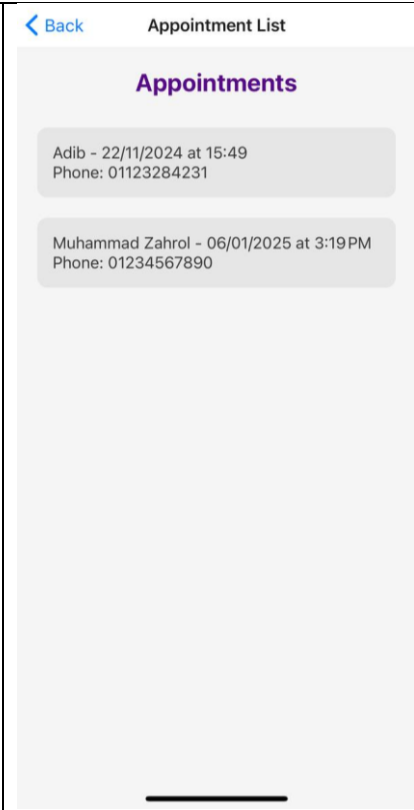
 <p>The 'Create Appointment' screen features a 'Back' button at the top left. The main heading is 'Create New Appointment'. Below this, there are four input fields: 'Patient Name' (containing 'Muhammad Zaharol'), 'Phone Number' (containing '011234567890'), 'Select Date' (containing 'Date: 06/01/2025'), and 'Select Time' (containing 'Time: 11:31AM'). A prominent purple 'Create Appointment' button is located at the bottom of the form area.</p>	 <p>The 'Appointment List' screen has a 'Back' button at the top left. The heading is 'Appointments'. It displays a list of two appointments in grey boxes. The first appointment is for 'Adib' on '22/11/2024 at 15:49' with phone '01123284231'. The second appointment is for 'Muhammad Zahrol' on '06/01/2025 at 3:19 PM' with phone '01234567890'.</p>
<p>Figure 4.29 Page to create new appointment</p>	<p>Figure 4.30 Page to check appointment list</p>

Figure 3.16 illustrates the "Create New Appointment" functionality, enabling users to arrange appointments by inputting details including the patient's name, phone number, day, and time. This optimised format guarantees effective appointment scheduling for healthcare practitioners. Figure 3.1.7 illustrates the "Appointment List" interface, which organises and presents planned appointments together with details such the patient's name, date, time, and contact information. Together, these features enhance the system's usability by simplifying appointment management and improving patient care coordination.

4.8 Summary

In Chapter 4, the development and classification for the Diabetic Retinopathy (DR) prediction model is done through different machine learning algorithms such as Logistic Regression, SVM, Random Forest, KNN, XGBoost, and AdaBoost. It tackles issues like the class imbalance problem with SMOTE and synthetic data generation which was enhanced by Mostly.ai. Random Forest had the best results consistently when the models were defined with the accuracy, ROC-AUC, precision, recall and F1 scores, especially for balanced datasets. While ensemble models like XGBoost and AdaBoost did quite well, KNN did not do well with high dimensional data. Another aspect of this chapter is the assessment of the prototype system that was deployed on live data, and its performance was commendable but there were certain accuracy issues there as well which require thorough examination. The system responsive React Native front end provided to the healthcare professionals for managing patient DR records and predictions, alongside visualizations ensure the ease of use and efficiency from the backend Flask API.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

This chapter summarized all the results of this research that had been obtained and provide recommendations for the future reference. The accomplishment of the research's objectives, strengths of the research, limitations of the research, and recommendations for future works are included and explained in this chapter.

5.1 Research Objectives Accomplishment

This study met its primary goal of predicting diabetic retinopathy by using machine learning techniques. Multiple research initiatives met major objectives such as identifying factors contributing to diabetic retinopathy, designing and evaluating machine learning models, and creating a functional prototype for healthcare practitioners. The initiative produced significant outcomes, illustrating the ability of machine learning to improve early detection and intervention in clinical settings.

5.1.1 Objective 1: Identify factors contributing to the development of diabetic retinopathy (DR)

Through the analysis of a dataset provided by the Department of Ophthalmology of hospital Al-Sultan Abdullah, the study effectively determined some important contributors that are responsible for the development of DR. This dataset had demographic, clinical, and health-related variables involving age, gender, ethnicity, marital status, height, weight, BMI, nephropathy, neuropathy, diabetic foot, stroke, IHD, anemia, asthma, dyslipidemia, and HbA1c. In order to maintain the integrity and robustness of the data, hoisting missing values, normalization, and encoding were among the many preprocessing techniques that were implemented. These factors were

used to comprehend their impact on the development of DR and aided in model building.

5.1.2 Objective 2: Design a machine learning model to predict diabetic retinopathy for early detection and intervention

The predictive ability of various machine learning models which included Random Forest, Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Extreme Gradient Boost, and AdaBoost, was computed. The Random Forest algorithm achieved the highest success rate of 79.49% and ROC-AUC score of 85.65%. With these accuracy assessments, the model can be relied upon for the diagnosis of diabetic retinopathy (DR) and for early screening. The results of this model demonstrate the advantages of using machine learning to solve the problem of DR diagnosis in the clinical environment.

5.1.3 Objective 3: Develop a prototype for predicting diabetic retinopathy using a machine learning model

The final goal of producing an intuitive prototype for DR prediction was achieved by integrating the Random Forest model into a web-based system using Flask. The system allows healthcare practitioners to enter patient data and receive real-time projections which support early diagnosis and assist with clinical decision-making.

5.2 Research Strengths and Limitations

This study revealed strengths in employing data preprocessing and a user-deployed framework into primary machine learning models such as Random Forest which made accurate predictions (79.49%) and ROC-AUC (85.65%) for Diabetic Retinopathy (DR). These approaches resulted in an accurate model and the gap between research and clinical applications were met which positively impacts the Sustainable Development Goal 3 (SDG3) in providing better health care. However, the study had its limitations, such as using a small and homogenous sample that might impair the

cross-population applicability of the model. The model was limited in the amount of nuanced predictors it accounted for due to absence of retinal image data and genetic information, in addition to not yet being validated in a real-world clinical setting. Moreover, these factors alongside the lack of explainability features for the model could decrease the professional's trust in the model, all posing further challenges to resolve in the future.

5.3 Recommendation and Future Works

Further research should center on the incorporation of larger, more diverse datasets to improve the generalizability of the models when predicting diabetic retinopathy (DR). Furthermore, extracting health records data in real-time from EHRs (electronic health records) systems as well as from wearable devices would allow for continuous monitoring and enhance early detection. Model performance can be boosted and data imbalance can be tackled by techniques such as synthetic data generation using Mostly.Ai or GANs. The development of explainable AI (XAI) capabilities would further assist healthcare specialists interpret these predictions accurately and enhance adoption of the systems. Ultimately, testing the system in real-world clinical setting as well as performing longitudinal studies to monitor the long-term impacts will guarantee that the system is scalable and effective in enhancing early intervention and reduction of DR associated vision impairment.

REFERENCE

- Abas, M. Z., Li, K., Hairi, N. N., Choo, W. Y., & Wan, K. S. (2024). Machine learning based predictive model of Type 2 diabetes complications using Malaysian National Diabetes Registry: A study protocol. *Deleted Journal*, 13(1). <https://doi.org/10.1177/22799036241231786>
- Alfian, G., Syafrudin, M., Fitriyani, N. L., Anshari, M., Stasa, P., Svub, J., & Rhee, J. (2020). Deep Neural Network for Predicting Diabetic Retinopathy from Risk Factors. *Mathematics*, 8(9), 1620. <https://doi.org/10.3390/math8091620>
- Azamen, N. T. N. N., Ali, A. M., & Aziz, N. a. A. (2023). Prediction of Diabetic Retinopathy Based on Risk Factors Using Machine Learning Algorithms. <https://doi.org/10.1109/aidas60501.2023.10284646>
- Balamurugan, N. M., Maithili, K., Babu, T. K. S. R., & Adimoolam, M. (2022). Stage-Wise categorization and Prediction of diabetic retinopathy using ensemble learning and 2D-CNN. *Intelligent Automation & Soft Computing*, 36(1), 499–514. <https://doi.org/10.32604/iasc.2023.031661>
- Dutta, A., Hasan, M. K., Ahmad, M., Awal, M. A., Islam, M. A., Masud, M., & Meshref, H. (2022b). Early Prediction of Diabetes Using an Ensemble of Machine Learning Models. *International Journal of Environmental Research and Public Health*, 19(19), 12378. <https://doi.org/10.3390/ijerph191912378>
- Emon, M. U., Zannat, R., Khatun, T., Rahman, M., Keya, M. S., & Ohidujjaman, N. (2021). Performance Analysis of Diabetic Retinopathy Prediction using Machine Learning Models. <https://doi.org/10.1109/icict50816.2021.9358612>

- Husain, A., & Malhotra, D. (2023). Development of intelligent framework for early prediction of diabetic retinopathy. In *Lecture notes in electrical engineering* (pp. 491–503). https://doi.org/10.1007/978-981-99-0601-7_38
- Jebaseeli, T. J., Durai, C. a. D., Alelyani, S., & Alsaqer, M. S. (2021). Prediction of diabetic retinopathy using machine learning techniques. *Journal of Engineering Research*. <https://doi.org/10.36909/jer.13947>
- Liu, L., Wang, M., Li, G., & Wang, Q. (2022). Construction of Predictive Model for Type 2 Diabetic Retinopathy Based on Extreme Learning Machine. *Diabetes, Metabolic Syndrome and Obesity, Volume 15*, 2607–2617. <https://doi.org/10.2147/dmso.s374767>
- Naramala, V. R., Kumar, B. A., Rao, V. S., Mishra, A., Hannan, S. A., El-Ebiary, Y. A., & Manikandan, R. (2023). Enhancing Diabetic Retinopathy Detection Through Machine Learning with Restricted Boltzmann Machines. *International Journal of Advanced Computer Science and Applications*, 14(9). <https://doi.org/10.14569/ijacsa.2023.0140961>
- Odeh, I., Alkasassbeh, M., & Alauthman, M. (2021). Diabetic Retinopathy Detection using Ensemble Machine Learning. <https://doi.org/10.1109/icit52682.2021.9491645>
- Sarkar, P., & Pawar, S. (2023). Novel machine learning model for fast and accurate diabetes prediction. *Research Square (Research Square)*. <https://doi.org/10.21203/rs.3.rs-3500371/v1>
- Saxena, R., Sharma, S. K., Gupta, M., & Sampada, G. C. (2022). A Comprehensive Review of Various Diabetic Prediction Models: A Literature Survey. *Journal*

of Healthcare Engineering, 2022, 1–15.

<https://doi.org/10.1155/2022/8100697>

Shoaib, M. R., Emara, H. M., Zhao, J., El-Shafai, W., Soliman, N. F., Mubarak, A.

S., Omer, O. A., El-Samie, F. E. A., & Esmail, H. (2023). Deep learning innovations in diagnosing diabetic retinopathy: The potential of transfer learning and the DiaCNN model. *Computers in Biology and Medicine*, 169, 107834. <https://doi.org/10.1016/j.compbiomed.2023.107834>

Vyas, A., Raman, S., Sen, S., Ramasamy, K., Rajalakshmi, R., Mohan, V., &

Raman, R. (2023). Machine Learning-Based Diagnosis and Ranking of Risk Factors for Diabetic Retinopathy in Population-Based Studies from South India. *Diagnostics*, 13(12), 2084.

<https://doi.org/10.3390/diagnostics13122084>

Wagai, G., Firdous, S., & Sharma, K. (2022). A survey on diabetes risk prediction using machine learning approaches. *Journal of Family Medicine and Primary Care*, 11(11), 6929. https://doi.org/10.4103/jfmpe.jfmpe_502_22

Wang, Z., Chen, S., Liu, T., & Yao, B. (2024). Multi-Branching Temporal Convolutional Network With Tensor Data Completion for Diabetic Retinopathy Prediction. *IEEE Journal of Biomedical and Health Informatics*, 28(3), 1704–1715. <https://doi.org/10.1109/jbhi.2024.3351949>

Wright, D. M., Chakravarthy, U., Das, R., Graham, K. W., Naskas, T. T., Perais, J., Kee, F., Peto, T., & Hogg, R. E. (2023). Identifying the severity of diabetic retinopathy by visual function measures using both traditional statistical methods and interpretable machine learning: a cross-sectional study.

Diabetologia, 66(12), 2250–2260. <https://doi.org/10.1007/s00125-023-06005-3>

Yagin, F. H., Yasar, S., Gormez, Y., Yagin, B., Pinar, A., Alkhateeb, A., & Ardigò, L. P. (2023). Explainable artificial intelligence paves the way in precision diagnostics and biomarker discovery for the subclass of diabetic retinopathy in Type 2 diabetics. *Metabolites*, 13(12), 1204.

<https://doi.org/10.3390/metabo13121204>

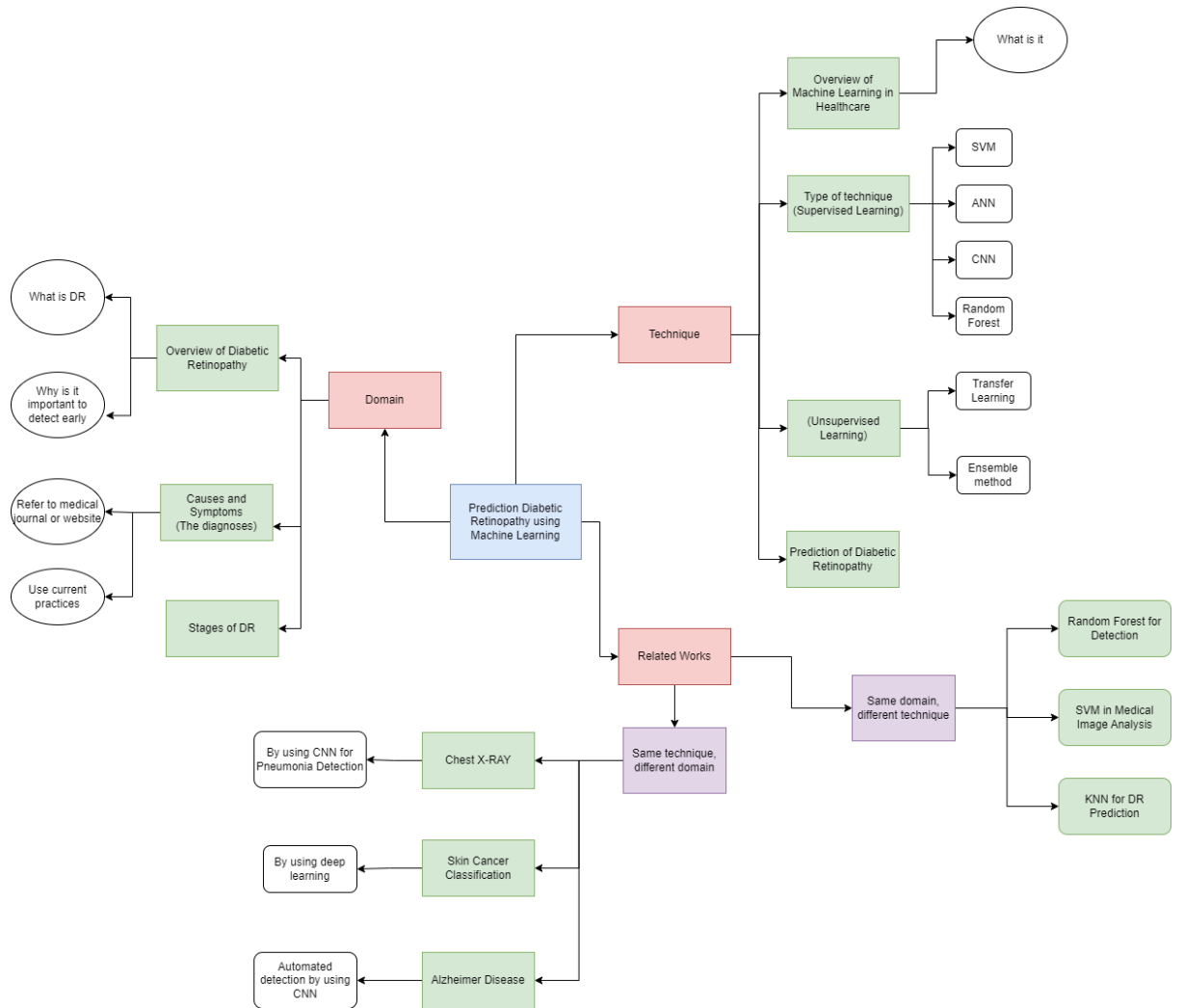
Zhao, Y., Li, X., Li, S., Dong, M., Yu, H., Zhang, M., Chen, W., Li, P., Yu, Q., Liu, X., & Gao, Z. (2022). Using machine learning techniques to develop risk prediction models for the risk of incident Diabetic retinopathy among patients with Type 2 diabetes mellitus: a cohort study. *Frontiers in Endocrinology*,

13. <https://doi.org/10.3389/fendo.2022.876559>

Zhu, C., Zhu, J., Wang, L., Xiong, S., Zou, Y., Huang, J., Xie, H., Zhang, W., Wu, H., & Liu, Y. (2023). Development and validation of a risk prediction model for diabetic retinopathy in type 2 diabetic patients. *Scientific Reports*, 13(1).

<https://doi.org/10.1038/s41598-023-31463-5>

APPENDIX



FINAL YEAR PROJECT GANTT CHART
USING MACHINE LEARNING: PREDICTION OF DIABETIC RETINOPATHY USING MACHINE LEARNING

SEMESTER	SEM 5														SEM 6													
PHASE / WEEK	W 1	W 2	W 3	W 4	W 5	W 6	W 7	W 8	W 9	W 10	W 11	W 12	W 13	W 14	W 1	W 2	W 3	W 4	W 5	W 6	W 7	W 8	W 9	W 10	W 11	W 12	W 13	W 14
PHASE 1 Preliminary Study																												
PHASE 2 Knowledge Acquisition																												
PHASE 3 Data Collection																												
PHASE 4 Data Pre-processing																												
PHASE 5 Model Development																												
PHASE 6 Model Evaluation																												
PHASE 7 System Development																												
PHASE 8 System Testing																												
PHASE 9 Documentation																												

Authors	Title	Problem Statement	Objectives	Techniques	Findings	Remarks	Dataset
Mohamad Zulfikrie Abas et al., 2024	Machine learning based predictive model of Type 2 diabetes complications using Malaysian National Diabetes Registry: A study protocol	Increasing prevalence of diabetes in Malaysia necessitates accurate prediction models for complications.	Develop ML-based predictive models for Type 2 diabetes complications using MNDR data.	Logistic random forest, regression, XGBoost, k-NN, decision tree, , SVM, LightGBM. Data cleaning, imputation, k-fold cross-validation, hyperparameter tuning, performance evaluation.	Develops precise predictive models with ROC-AUC, recall, precision, F1 score, accuracy, and early intervention and resource allocation in mind.	High potential for improving diabetes management, requires further validation.	Clinical audit datasets from the Malaysian National Diabetes Registry (2011-2021).
Chengjun Zhu et al., 2023	Development and validation of a risk prediction model for diabetic retinopathy in type 2 diabetic patients	One of the main complications of diabetes that can result in blindness is diabetic retinopathy (DR). It's necessary to predict risks accurately.	Develop and validate a risk prediction model for DR in type 2 diabetes patients.	Logistic regression model, meta-analyses, electronic patient-reported outcome questionnaire, ROC curve analysis.	Achieved high prediction accuracy with AUC of 0.912, sensitivity of 86.7%, and specificity of 86.7%.	High accuracy, further validation needed with larger sample sizes.	Meta-analyses with 15,654 cases and 12 risk factors.

Minhaz Uddin Emon et al., 2024	Performance of Diabetic Retinopathy Prediction using Machine Learning Models	Diabetic retinopathy is a complication of diabetes that can lead to blindness if not diagnosed early.	Predict diabetic retinopathy using various machine learning models and identify the best-performing algorithm.	Naive Bayes, SMO, logistic regression, SGD, bagging, J48, decision tree, random forest.	Logistic regression performed the best with 75% accuracy and an ROC value of 83%.	Logistic regression provides the highest accuracy and ROC value among the tested models.	Diabetes Retinopathy Debrecen dataset from UCI, 1151 iterations with 19 attributes.
Nor Tasha Nadira Nor et al., 2023	Prediction of Diabetic Retinopathy Based on Risk Factors using Machine Learning Algorithms	Diabetic Retinopathy (DR) can lead to permanent blindness if not treated early. Early detection using risk factors is essential.	Predict DR using ML algorithms based on risk factors and develop a web-based application for healthcare use.	Logistic regression, SVM, k-nearest neighbors (KNN)	With a random state value of 11, logistic regression produced the highest accuracy (83.78%) and specificity (83.78%).	Logistic regression provides the highest accuracy among tested models. Further improvement with more data is needed.	Data from Ophthalmology Clinic, UiTM Sg Buloh Medical Centre (361 instances of Type 2 Diabetes patients).
Roshan Vasu Muddalur	A Comparative Study of Filters and Deep	Diabetic retinopathy (DR) is difficult to diagnose early and	Compare the outcomes of various deep learning models	Deep learning models: InceptionNetV3,	InceptionNetV3 with Gaussian filter achieved the highest	Gaussian filter significantly enhances retinal	Grayscale DR dataset with 3664 retinal

u et al., 2023	Learning Models to predict Diabetic Retinopathy	must be classified according to severity in order to prevent blindness.	with different filters to improve DR diagnosis.	EfficientNet, DenseNet121, MobileNetV2; Filters: Gaussian, Greyscale, Gabor.	accuracy (96%) for DR severity classification.	image quality, aiding in better feature extraction for DR diagnosis.	images from Kaggle.
Roshan Vasu Muddalur u et al., 2022	A Comparative Study of Filters and Deep Learning Models to predict Diabetic Retinopathy	Diabetic retinopathy (DR) is difficult to diagnose early and must be classified according to severity in order to prevent blindness.	Compare the outcomes of various deep learning models with different filters to improve DR diagnosis.	Deep learning models: InceptionNetV3, EfficientNet, DenseNet121, MobileNetV2; Filters: Gaussian, Greyscale, Gabor.	InceptionNetV3 with Gaussian filter achieved the highest accuracy (96%) for DR severity classification.	Gaussian filter significantly enhances retinal image quality, aiding in better feature extraction for DR diagnosis.	Grayscale DR dataset with 3664 retinal images from Kaggle.
Shimoo Firdous et al., 2022	A Survey on Diabetes Risk Prediction Using Machine Learning Approaches	Diabetes is a chronic condition that can lead to severe complications. Accurate early prediction is essential.	Examine and evaluate machine learning algorithms that are used to early-stage diabetes prediction.	Logistic regression, random forest, SVM, Naive Bayes, k-NN, decision tree, and MLP are examples of machine learning algorithms.	With the highest accuracy of 96.54%, SVM was followed by random forest and k-NN, both of which had accuracies of approximately 96%.	Accurate prediction models improve early detection and management of diabetes.	Various datasets including Pima Indian Diabetes Dataset (PIDD) and others.

Minhaz Uddin Emon et al., 2021	Performance Analysis of Diabetic Retinopathy Prediction using Machine Learning Models	Diabetic retinopathy (DR) is a major complication of diabetes and a leading cause of blindness. Accurate prediction models are needed.	Predict DR with machine learning models and select the best performing algorithm.	Naive Bayes, SMO, logistic regression, SGD, bagging, J48, decision tree, random forest.	Logistic regression demonstrated the best performance, with an accuracy of 75% and a ROC value of 83%.	Logistic regression is the most effective among the tested models for DR prediction.	Diabetes Retinopathy Debrecen dataset from UCI with 1151 instances and 19 attributes.
Lei Liu et al., 2022	Construction of Predictive Model for Type 2 Diabetic Retinopathy Based on Extreme Learning Machine	Diabetic retinopathy (DR) is the major cause of blindness among type 2 diabetics. Accurate prediction models are crucial.	Create an extreme learning machine (ELM) DR prediction model and compare it against SVM, ANN, KNN, and RF models.	ELM, SVM, KNN, RF, ANN; measures include ACC, sensitivity, specificity, accuracy, NPV, training time, and AUC.	The ELM model outperformed the others in terms of ACC (84.45%), accuracy (83.93%), specificity (93.16%), AUC (88.34%), and training time (1.24s). SVM had the highest NPV and sensitivity.	ELM is effective for DR prediction with high accuracy and efficiency.	Information from the computerised medical records of the 1093 patients at Anhui Medical University's Lu'an Hospital in China.

Ganjar Alfian et al., 2020	Deep Neural Network for Predicting Diabetic Retinopathy from Risk Factors	Diabetic retinopathy (DR) can cause vision impairment. Accurate prediction using risk factors is essential.	Develop a deep neural network (DNN) combined with recursive feature elimination (RFE) for early DR prediction.	DNN, SVM-RFE, data normalization, grid search for hyperparameter optimization, stratified 10-fold cross-validation.	The proposed model achieved 82.033% accuracy, outperforming other models. Key risk factors include DM duration, FBS, HDL, Age, and A1c.	Integration of DNN with RFE improves DR prediction accuracy.	Publicly available dataset from Khodadadi et al. (133 diabetic patients).
Israa Odeh et al., 2021	Diabetic Retinopathy Detection using Ensemble Machine Learning	Diabetic retinopathy (DR) can cause blindness if not detected early. DR identification requires expertise, which can be costly and time-consuming.	Develop an ensemble-based model for automatic DR detection to improve diagnosis accuracy.	Ensemble learning combining Random Forest, Neural Network, and SVM; feature selection with InfoGainEval and WrapperSubsetEval.	Ensemble model achieved 75.1% accuracy on the original dataset, 70.7% on InfoGainEval top 5 features.	Ensemble learning improves prediction accuracy, reducing complexity and processing time.	MESSIDOR dataset with 1151 fundus images
Aishwariya Dutta et al., 2022	Early Prediction of Diabetes Using an Ensemble of Machine Learning Models	Diabetes leads to severe complications; early prediction is crucial.	Develop an ensemble of ML models for early diabetes prediction using a newly labeled dataset from Bangladesh.	Selecting features, K-fold cross-validation, missing value imputation, grid search hyperparameter	The ensemble model, which consists of DT, RF, XGB, and LGB, had the best accuracy (73.5%) and AUC (0.832).	Ensemble model improves prediction performance; new dataset contributes to	New dataset from Bangladesh Demographic and Health Survey (2011,

				optimisation, Decision Tree, Random Forest, LightGBM, and XGBoost.		developing robust models.	2017-2018), 4751 diabetes cases, 2814 non-diabetes cases.
Dr. Venkateswara Rao Naramala et al., 2023	Enhancing Diabetic Retinopathy Detection Through Machine Learning with Restricted Boltzmann Machines	Diagnosing diabetic retinopathy (DR) requires a lot of work and time. Though difficult, early detection is essential.	Improve DR diagnostic accuracy using Restricted Boltzmann Machines (RBM) and U-network model for optic segmentation.	RBM, Squirrel Search Algorithm (SSA) for hyperparameter optimisation, U-network for optic segmentation, and RIM-ONE DL dataset for evaluation.	Achieved 99.2% accuracy on RIM-ONE DL dataset, demonstrating robust performance and potential clinical relevance.	High accuracy and efficiency in DR detection, contributing significantly to the medical field.	RIM-ONE DL dataset for training and validation.
Jing-Yang Su et al., 2022	Establishment of Metabolite Prediction Model for the Risk of Diabetic Retinopathy in Chinese Type 2	Diabetic retinopathy (DR) is a leading cause of blindness in type 2 diabetes patients. Accurate prediction models are essential.	Using plasma metabolites, create a DR risk prediction model and assess its efficacy.	For metabolite measurement, partial least squares regression, logistic regression, and ROC curve analysis, use liquid	<ol style="list-style-type: none"> 1. histidine 2. citrulline 3. phenylalanine 4. methionine 5. tyrosine 6. C3 7. C24 	Effective model for DR risk detection in Chinese diabetic patients, further validation needed.	743 hospitalized patients from a tertiary hospital in China, divided into

	Diabetic Population			chromatography-mass spectrometry (LC-MS).			DR and non-DR groups.
Jing-Yang Su et al., 2022	Establishment of Metabolite Prediction Model for the Risk of Diabetic Retinopathy in Chinese Type 2 Diabetic Population	Diabetic retinopathy (DR) is a leading cause of blindness in type 2 diabetes patients. Accurate prediction models are essential.	Develop a DR risk prediction model using plasma metabolites and evaluate its performance.	Liquid Chromatography-Mass Spectrometry (LC-MS) for metabolite measurement - Partial least squares regression - Logistic regression - ROC curve analysis	- An AUC of 0.770 was obtained by a DR risk prediction model that included seven metabolites (tyrosine, histidine, citrulline, phenylalanine, methionine, C3, and C24).	Effective model for DR risk detection in Chinese diabetic patients, further validation needed.	A Chinese tertiary hospital admitted 743 patients, who were split into DR and non-DR groups.
Abhishek Vyas et al., 2023	Machine Learning-Based Diagnosis and Ranking of Risk Factors for Diabetic Retinopathy in Population-Based Studies from	One of the main causes of blindness in diabetics is diabetic retinopathy (DR). Reliable forecasting models are essential.	Develop and rank risk factors for DR using various machine learning models.	<ol style="list-style-type: none"> 1. K-Nearest Neighbor 2. Decision Tree 3. SVM 4. Logistic Regression 5. Naive Bayes 	KNN and Decision Tree achieved the highest AUC (0.79 t-test, 0.77 SHAP). Key risk factors: systolic blood pressure, glycosylated hemoglobin, diabetes duration. KNN	High accuracy and robust risk factor ranking, validates using t-test and SHAP methods.	Data from 4 extensive population-based studies conducted in India between 2001 and 2010; 3133 non-DR cases

	South India			6. Ensemble models	achieved 82.6% accuracy (t-test), 78.3% (SHAP).		and 857 DR cases.
Zekai Wang et al., 2024	Tensor data completion and a Multi-Branching Temporal Convolutional Network for Predicting Diabetic Retinopathy	DR is a leading cause of vision loss, and early detection is needed due to low screening compliance and high costs.	Develop a MB-TCN with Tensor Data Completion for DR prediction using EHR data.	MB-TCN, Tensor Data Completion, CP decomposition, residual blocks, dilated causal convolution, missing value masks.	Achieves AUROC of 0.949 and AUPRC of 0.793. Handles imbalanced data and missing values, capturing temporal correlations.	High accuracy and efficiency for DR prediction, applicable to other health monitoring scenarios.	2018 Cerner Health Facts® data warehouse, 414,199 diabetic patients with 3% DR positive rate.
Yuedong Zhao et al., 2022	Using Machine Learning Techniques to Develop Risk Prediction Models for the Risk of Incident Diabetic Retinopathy: A Cohort Study of Patients with	One of the main causes of blindness in people with type 2 diabetes is diabetic retinopathy, or DR. We require precise prediction models.	Diabetic retinopathy, or DR, is one of the primary causes of blindness in individuals with type 2 diabetes. We need accurate prediction models.	Random Forest, XGBoost, Logistic Regression, SVM, K-Nearest Neighbor; GridSearchCV for hyperparameter optimization; fivefold cross-validation.	XGBoost achieved highest performance with AUC 0.803, accuracy 88.9%, sensitivity 74.0%, specificity 81.1%. Key risk factors: HbA1c, diabetes duration, age, FBG, SUA.	High accuracy and early risk identification (up to 2.895 years before diagnosis). Needs multi-center prospective validation.	Retrospective cohort study, 7943 patients from Dalian Medical University Affiliated Central Hospital, 1692

	Type 2 Diabetes Mellitus						diagnosed with DR.
--	--------------------------	--	--	--	--	--	--------------------