PREDICTIVE MODELING FOR EARLY DETECTION OF DIABETES USING MACHINE LEARNING TECHNIQUES

BY

MAHMUDUL HASAN SAGOR
ID: 213-15-4413

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

DR. SHEAK RASHED HAIDER NOORI
Professor & Head
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

DR. MD ZAHID HASAN
Associate Professor
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

July 2024

APPROVAL

This Project titled "Predictive Modeling For Early Detection Of Diabetes Using Machine Learning", submitted by Mahmudul Hasan Sagor to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 16-07-2024.

BOARD OF EXAMINERS

Dr. Sheak Rashed Haider Noori (SRH) — **Board Chairman**
Professor & Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Zahid Hasan (ZH) — **Internal Examiner 1**
Associate Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Abdus Sattar (AS) — **Internal Examiner 2**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Manowarul Islam (MI) — **External Examiner**
Associate Professor
Department of Computer Science and Engineering
Jagannath University

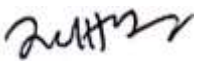
DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dr. Sheak Rashed Haider Noori, Professor & Head, Department of Computer Science and Engineering, Daffodil International University.** We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Dr. Sheak Rashed Haider Noori
Professor and Head
Department of CSE
Daffodil International University

Co-Supervised by:



Dr. Md. Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Submitted by:

Mahmudul Hasan Sagor
ID: 213-15-4413
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty for His divine blessing making it possible for us to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Dr. Sheak Rashed Haider Noori, Professor & Head,** Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to the **Head of the Department of CSE**, **Dr. Sheak Rashed Haider Noori** sir for his kind help in finishing our project and also to other faculty members and the staff of the Department of CSE, Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Diabetes is a prevalent chronic disease with significant health implications worldwide. It is usually prolonged in a patient for their entire vitality. Early detection and intervention are vital for successfully managing and preventing any complications. Diabetes can lead to complications if not recognized and diagnosed early enough. In this dissertation, I will be talking about how machine-learning methods are crucial for predictive modeling. Aimed at early detection of diabetes. These models will be based on different factors, including demographic, clinical, and lifestyle, among others, with large datasets being used to come up with them. Therefore, the author prefers using machine learning methods such as SVM, KNN, ANN, Naive Bayes, logistic regression, XGB Classifier, and Decision Tree. The results are evaluated using performance measures including recall, accuracy, precision, and the F-measure, which are computed from the confusion matrix. I designed a predictive model to identify whether a patient will develop diabetes, utilizing specific diagnosis measurements in the dataset. This project tries to find a way to improve healthcare outcomes by enabling early intervention and enhanced disease management.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES **PAGE NO**

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Introduction:

Diabetes is a long-term metabolic condition that afflicts millions of people all over the world. It is a serious condition in the body when our body cannot produce insulin on its own or if produced, it doesn't work effectively.. As a result, the level of sugar or glucose in the blood becomes higher than normal. [1]Diabetes is a disease that is caused by other disorders such as heart attacks, blindness, cardiovascular system disease (CVD), and kidney problems. According to Wikipedia, as of 2021, an estimated 537 million individuals worldwide have diabetes. Type 2 accounts for 90% of occurrences, and it affects 10.5% of the adult population. It is estimated that by 2045, approximately 783 million adults will have diabetes, a 46% increase. The prevalence is highest in low- and middle-income countries, with annual healthcare expenditures of $760 billion. Although there is no long-term treatment for diabetes, it can be controlled and prevented with accurate early prediction of the most challenging tasks is the prediction of diabetes. Prediction in, and prevention of, diabetes occupies prime importance in the domain of public health, and the development of accurate predictive models has been commendable using complex relationships in data with the application of ML techniques in this area.. Machine learning (ML) has various forms, including classification, regression, and clustering, each with its impact. My focus is on classification methods, which are used in medical domains, particularly in diagnosing diseases like diabetes. Common ML classification algorithms include SVM, KNN, ANN, Naive Bayes, logistic regression, and decision trees, which are used to identify diabetes patients early. These methods provide high accuracy and performance in predicting future activities or information from a given dataset. This paper is focused on the development of a diabetes prediction model, mainly to determine whether a patient is capable of having the disease, and explores techniques that enhance model accuracy statistically meaningful only when the segmentation problem gets more challenging and when less target training data is available. [2] There are mainly two types of diabetes: type 1 and type 2

Type-1 Diabetes

This is the case when the liver completely stops producing insulin. Insulin is a hormone that needs to be present for the body to absorb the amount of glucose in the blood flow and further use this glucose for the building of the body. On the other hand, the absence of insulin in the body will cause the blood sugar level to go up, thus leading to type 1 diabetes. It is usually present in children and adolescents. This disorder mostly happens because of genetic disorders. It is, for the most part, called juvenile disorder. Its usual symptoms comprise frequent urination. Weight loss, thirst increase, blurred vision, and nerve problems are the consequences of dieting. This can be the solution to the problem with insulin therapy.

Type-2 Diabetes

It is a metabolic condition that normally affects adults over the age of 40 and persists for a long time. It is evident in the conditions of high blood sugar and insulin resistance. The primary causes of obesity and lack of exercise are obesity and the fast pace of life. This unhealthy lifestyle can cause glucose to be deposited in the blood, which can be the reason for diabetes. [3]Type 2 diabetes has an impact on about 90% of people. Metformin is applied to the treatment of insulin resistance and thus it is possible to get rid of it.

Although no cure exists for diabetes, it can be controlled and prevented by the early prediction that is very accurate. Nowadays Computer Aided Diagnosis is a very fast-growing, dynamic area of research in the medical industry. Diabetes prediction is a difficult issue. Diabetes prediction and prevention are the most important public health issues and computational techniques have proved to be very useful by creating prediction models that are based on complex data relationships. Including Machine learning, Ensemble Learning, Deep Learning, Fuzzy logic, Ontology, and NPL.

1.2 Motivation:

The invention of ever-new methods in the field of medicine to advance the diagnostic process is much more than a simple quest for methods. It is about care; it helps you serve a patient and improve the health of society. In the case of chronic diseases, diabetes being one of them poses a great danger, and therefore detection needs to be done earlier on before any interventions are made. So, we're employing sophisticated computer systems to anticipate who will get diabetes before it's too

late. We may make better assumptions about who is at risk by looking at a variety of factors such as age, weight, family history, and lifestyle choices In artificial intelligence, machine learning is a revolution, which is at the heart of it and allows the model to learn from large sets of data so that it can predict diabetes. These supervised learning methods, well known for their efficiency in data processing, are the main technical instruments of this expedition. Through a scrupulous examination of many machine learning algorithms, not just accurate prediction of diabetes diseases will be achieved but also principal risk factors will be recognized. The study undertakes this ambitious comparative analysis to reveal the most effective strategies, and thus it shows us how to take the first step towards a revolutionary change in diabetes diagnostics and, at the end of the day, raise the level of healthcare worldwide.

1.3 Rational of the Study:

This work, "Predictive Modeling For Early Detection of Diabetes Using Machine Learning: Comparative Analysis of Supervised Learning Techniques," is necessitated by the need to solve some of the major predicaments hampering the prediction and management of diabetes. Diabetes is one of the key health challenges worldwide, and early, accurate prediction efforts could be of immense help in formulating effective treatment and prevention measures. Machine learning is a very powerful tool of artificial intelligence that can be used in attempting to enhance the model's ability by identifying diabetes risk based on patterns. It applies different techniques of supervised learning, which are well known for the efficiency of data analysis. It is not only to predict diabetes at this level; it is also required to understand what risk factors among the total are the most important. Comparing different machine learning algorithms, it was of interest to know which of the proposed methods would be better at predicting diabetes. The essence of this comparative analysis is, therefore, very related to the enhancement of current prediction tools and the development of more accurate and reliable systems. In the study, the strengths and weaknesses of each approach will be discussed constructively to give valuable insights into refining diabetes prediction methods. Ultimately, this would lead to enhancing the precision and efficiency of predictions in diabetes diagnostics. More specifically, by enriching the current knowledge on how machine learning can best fit in this respect, it aspires to deliver its contribution to academic knowledge and real-world healthcare, in which better prediction tools can result in palpable differences for patients and outcomes related to public health. An integral complement to the further

development of existing diagnosing tools as well as for fostering more accurate and trustworthy systems.

1.4 Research Questions:

Among the key questions that this research hopes to answer fall within the domain of "Early Prediction of Diabetes Using Machine Learning: Comparative Analysis of Supervised Learning Techniques." On its part, the guiding research to the full understanding of intricacies involved in leveraging the advanced machine learning techniques for enhanced diagnostic output in medicine has therefore posed some questions. Broadly, the following are propounded for the research questions:

1. Which machine learning algorithms (logistic regression, KNN, ANN, SVM, random forests, neural networks) best suit this research?

The question speaks to the core role of different algorithms in machine learning in improving the accuracy and efficiency of predicting diabetes. The research will, therefore, help in how these techniques optimize the process of such a prediction to enable an improved understanding of their effectiveness in medical diagnosis. It speaks to which model is more efficient and more reliable for this project.

2. How well can the various supervised learning models predict diabetes, and how do they contribute to the entire process?

The question points to the individual contributions of various supervised learning models in the processes used for diabetes prediction. It specifically aims to assess their performance about each other, which gives representation on how it can improve the overall diagnosis capability and opens further research into improvement through design studies of predictive modeling**.**

3. How do the sensitivity and specificity of diabetes prediction go in contrast across traditional methods and integrative approaches in feature selection under machine learning models?

It tries to address the question of how combining feature selection techniques with machine learning models affects sensitivity and specificity for diabetes prediction and compares their performance

against traditional methods. The focus is on measuring diagnostic accuracy using state-of-the-art techniques to reveal possible improvements and challenges.

4. What are the most important risk factors determined by machine learning models for the motive of early diabetes prediction, and how have these factors contributed to this prediction?

It identifies key risk factors that the machine learning models are using and describes their contribution to the prediction process. The aim of this study on explanation would be "Critical risk factors in early diabetes prediction," which gives insight into the embedded patterns.

5. What are the strengths and weaknesses of different supervised learning algorithms in diabetes prediction, and how could such knowledge be used in the construction of more robust and accurate predictive models?

This question encompasses an in-depth review of the strengths and limitations intrinsic to different supervised learning techniques. By identifying and understanding these factors, it was intended that the study would inform the development of predictive models by guiding future efforts toward the comparative analysis of different supervised learning techniques that contribute to the refinement of existing predictive methodologies and the advancement of state-of-the-art diabetes prediction.

1.5 Expected Output:

In our efforts to reform diabetes diagnosis, our work is set to provide significant improvements that could change the field of healthcare intervention. Here is a look at the projected outcomes:

1. The current research methodically creates an approach for predicting diabetes.
2. The objective is to reduce the number of early fatalities caused by diabetes.
3. This study's findings might help the medical community create reliable patient diagnosis and treatment prediction models.
4. By employing the most efficient algorithms, we can create trustworthy models to enhance patient outcomes by including the appropriate elements.
5. Increase public awareness of diabetes after prediction.
6. Development of Robust Predictive Approach
7. Reduction in Early Diabetes-Related departed.

8. Algorithms Optimization for Improved Patient Outcomes.

1.6 Project Management and Finance

Full-cycle project management and careful control of the financial activities are the two pillars on which the success and sustainability of any initiative rest. Project management incorporates detailed planning, organizing, and execution of activities from collected data to train the model and evaluate in the proposed study of 'Predictive Modeling for Early Diabetes Detection Using Machine Learning.' A properly structured project plan will be devised that turns a spotlight on all pivotal milestones, resource allocation, and establishing timelines for work to ensure a disciplined yet efficient workflow. Equally important is sound financial management, which will be the backbone of this project. This includes a proper allocation of resources to data collection and computational resources, collaborative research, and a well-planned and frugal financial strategy for optimal resource use. These combined management and financing models will be set to spearhead the researcher through the intricacies of the study by ensuring that funds allocated are matched against the set objectives of the research and seamlessly steer the progress of the process toward the intended outcomes.

Table 1.6.1: Project Cost Breakdown Table

Budget Item	**Description**	**Estimated Cost**	**Actual cost**
Hardware	Computational Resources	6000	5000
Training and Workshops	Professional development for members. (machine learning workshop)	7000	5500
Travel	Travel expenses for conferences and meetings	2000	1500
Publication Fees	Fees for publishing		

	the research in a journal.		
Software & tools	Expenses for necessary software	3000	2500
Various Expenses	Other expenses	1500	2000
Total		=19500	=16500

1.7 Report layout

The proposed report shall have a comprehensive layout that guides readers through the research process, its findings, and the implications of the research. Each chapter thus serves a purpose that makes the document coherent and deep.

Chapter 1: Introduction

The introductory chapter sets the scene for the research with a background on early diabetes detection, applications of machine learning models, and creating a felt need to do predictive modeling. It incorporates the research questions, objectives, and framework of the study.

Chapter 2: Background

The background chapter critically reviews the literature concerned with past research on the chosen subject. This part, elaborately explains the theoretical backgrounds, approaches, and technologies used in diabetes prediction and machine learning. This chapter provides the grounds for the present study and exposes existing knowledge gaps.

Chapter 3: Research Methodology

The methodology chapter is a brief outline of how the study is to be conducted, detailing insight into the research design, techniques of data collection, model architecture, and rationales upfront in the choice of methodologies for the study. The chapter will also go on to discuss some ethical considerations and limitations that may affect the study.

Chapter 4: Experimental Result

In this chapter, we will go through the results of the experiments conducted for the early diabetes prediction model on individuals using machine learning models. The performance of various predictive models under study will be presented with appropriate visual displays and statistical measures, all put together in support of the results.

.

Chapter 5: Impact on Society

The chapter on the impact on society elaborates on and discusses the broader implications of the findings of research in healthcare and medical diagnosis. It looks into how a refinement in diabetes prediction methodologies could affect patient care, treatment, and general public health. Consideration is also made for probable future technological improvements and their eventual, but consequent social impact.

Chapter 6: Summary, Conclusion, Recommendation, and Implications for Future Research

This last chapter is, in some sense, a synthesis of the whole report. In this, key findings of the study are summed up, conclusions from the research restated, and recommendations made in terms of practical application and further studies. The discussion encompasses future research implications, therefore providing a kind of roadmap for any future researcher who desires to build on the current study.

That means the information will be in a logical order, basically helping any reader through the research journey from the introduction to the broader implications and recommendations. One will be better placed to gain insight into the process of the research and its likely effect on the early prediction of diabetes with an academic as well as practical disposition using Machine Learning methods. This discussion therefore provides a roadmap for persons interested in expanding upon this study.

CHAPTER 2
Background

2.1 Preliminaries/Terminologies

This chapter sets up the preliminaries and clarifies the terminologies in use for a clear understanding of the key concepts and methodologies applied in this research on "Predictive Modeling for Early Detection of Diabetes Using Machine Learning.". Machine learning forms the prime backbone of artificial intelligence and is the main concept holding this study together. ML helps models learn patterns from data and predict without being explicitly programmed. This approach is, hence, very important in analyzing such huge data sets of health attributes for the efficient and accurate prediction of diabetes risk. Predictive modeling refers to statistical techniques and ML algorithms for predicting the occurrence of future events from historical data. In diabetes prediction, predictive modeling would involve developing models that can establish the probability of any person developing diabetes, given health data. Feature selection is the process of selecting pertinent variables from data that significantly reward the predictive power of the model. This step is, therefore, very important to improve model accuracy and reduce computational complexity. Training data is the data used to train the relationships and patterns within the data for the ML model. It consists of a wide array of health metrics and historical information about patients that the model needs to learn from in order to understand and predict diabetes risk. Testing data: Another set of data, usually containing a small portion of the whole dataset, is used to test model performance, more precisely, generalizability to new, unseen data. This is a very important assessment for the real-world effectiveness of the model that will be trained. It is in ML that models get trained based on labeled data, which means the input data is paired correspondingly with a comprehensive output. In this study, techniques of supervised learning are used to train the model to recognize the pattern associated with the onset of diabetes. Classificatory algorithms, such as logistic regression, decision trees, or support vector machines, are used for putting data into pre-defined classes. Nodes could be diabetic or nondiabetic. These algorithms may be chosen for handling complicated and nonlinear relationships between variables. These terminologies have been used throughout this research in efforts to draw

synergies between machine learning approaches and the early prediction of diabetes. Preliminaries about these terminologies are essentially only established for a better understanding of the intricacies of the study and the nuances of advanced techniques employed in diabetes prediction. It is this foundational knowledge that will guide how such exploration into machine learning may revolutionize diagnosis for the betterment of patient outcomes through early interventions in diabetes.

2.2 Related Works

Machine-learning approaches have been used in recent studies to identify diabetes patients based on their symptoms. For this study, I explore several papers related to predicting diabetes. The authors build several models to predict diabetes and its risk factors. They several techniques in their research. Authors Edgar et al [2] apply 11 different machine learning algorithms to four datasets for the prediction of diabetes with increased precision. They used k-fold cross-validation to evaluate performance across performance metrics: They referred to the Mendeley dataset, the (PIDD) dataset, the Diabetes Early Stage (DES) dataset, and the Vanderbilt dataset, with emphasis on the leading 2, top 3, and all attributes adopted. The best accuracy among the three classifiers was achieved by the random forest C, which had an accuracy of 99%. 67% for the Mendel dataset. The highest value for the random forest and logistic regression with 99. 68% The kernel SVM model shows the highest accuracy, 82% over the 4% for the PIDD dataset. The best performer is the random forest, accuracy of 98 %. DES dataset provides The neural network architecture of multi-layer perceptron had the highest accuracy rate (96. 52 %).

Ganie and Malik [3] suggested an ensemble learning-based framework for the early prediction of type II diabetes mellitus. The authors employed eight different machine learning algorithms that are based on ensemble methods and were applied using 10-fold cross-validation for the prediction of the disease. The authors gather raw data from different hospitals in Jammu and Kashmir, UT. The dataset had 1939 records and 11 parameters. The bagged decision tree was better than all other classification algorithms in terms of accuracy (99%). 41%), precision (99.13%), recall (95.83%), specificity (99.11%), and F1-score (99.15%).

Zhang et al. [4] The study in the paper applied ML/EL techniques to predict the risk of type 2 diabetes mellitus (T2DM), which include logistic regression (LR), classification and regression tree (CART), artificial neural networks (ANN), support vector machines (SVM), random forest (RF), and gradient boosting machine (GBM) models. The data was obtained from the Henan rural cohort, which consisted of 36652 cases and 10 parameters. The GBM obtained the highest accuracy (81.2%).

Massari et al. [5] The research compares different ML/Ontology techniques like Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, XGBoost. The authors used the Pima Indians Diabetes Database (PIDD), originally sourced from the National Institute of Diabetes and Digestive and Kidney Diseases. The data set consists of 768 instances and 8 attributes. The Naïve Bayes and Random Forest classifiers got an accuracy of 80% and The ontology classifier has the highest Precision of 81. 2% for both the test and the practice modes.

The authors of this paper, Sneha et al. [6], are looking for a way to use significant features, create a prediction algorithm using Machine Learning, and select the best classification to provide medically useful outcomes. They used SVM, NB, KNN, Random forest, and Decision tree framework to predict diabetes from Diabetes dataset where instances are 2500 and 15 attributes. The decision tree algorithm and Random forest are the ones giving the highest accuracy of 98. 20% and 98. 00%, respectively.

Ahmed et al. [7] suggested a model based on the fused machine learning approach for diabetes prediction. They created conceptual frameworks for SVM and ANN models. The output of these models will be the input membership function to the fuzzy model, and at last, fuzzy logic decides whether diabetes is positive or negative. If the results of ANN and the SVM model are positive, then the diabetes becomes positive. In case of positive results from ANN models and negative from SVM models, then it is negative diabetes 1. If the result is negative from both ANN and SVM, then it is diabetes negative 1. There are 520 instances and 17 attributes based on diabetic symptoms. The proposed fused machine learning model showed an accuracy of 94.87%. R. Rahman et al. [8] The authors gathered 520 instances with 16 features called attributes: age, sex, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring,

itching, irritability, delayed healing, partial paresis, muscle stiffness, and alopecia. Which are collected from the direct questionnaires of the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh. They have the best accuracy at 97%. 4% with Random Forest using the cross-validation technique.

Joti et al. [9] The authors suggested different classification algorithms that were examined to predict diabetes, such as J48, Naïve Bayes, CART, k-Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM). the biggest accuracy at 99% is then the Decision Tree. The diabetes dataset for the prediction consisted of 2000 cases with nine features. The study employed machine learning algorithms for the diagnosis of diabetes at an early stage, and the Decision Tree method had the highest accuracy.

2.3 Comparative Analysis and Summary

According to IDF projections, in 2045, one in eight adults, or almost 783 million people will be affected by diabetes—an increase of 46 per cent. More than 90 per cent of diabetes patients-those with type 2 diabetes-are caused by socioeconomic, demographic, environmental and genetic variables. In 2019, this was 8.4 million for Bangladesh subjects aged 20 to 79 years, and by 2045 is almost expected to quadruple to 15.0 million. Since it is an incurable disease, early prediction of Diabetes can save so many lives. Artificial Intelligence and machine learning techniques seem to work pretty well in this respect. Several researchers have built their models to predict this disease.

Table 2.3.1: Comparative Analysis Table

SL No	Author Name	Used Algorithm	Best Accuracy with Algorithm
1.	, Edgar et al. [1]	Random Forest Classifier, Logistic Regression XGBoost, Decision Tree Classifier, SVC (linear kernel, Kernel SVM, Naive Bayes, AdaBoost Classifier, KNN, Quadratic Discriminant Analysis, Multi-Layer Perceptron.	Kernel SVM =82.4%(PID) Random Forest=98.12%(DES) Random Forest=99.67%(Mendeley)
2.	Malik et al. [2]	Bagging, Boosting, and Voting technique	Bagged decision tree=99.41%
3.	Zhang et al. [3]	Logistic Regression,, CART, ANN, SVM, Random forest, and GBM.	GBM=81.2%
4.	Massari. et al. [4]	Logistic Regression, Decision Tree, Random Forest, SVC, XGBoost, NB	Random Forest=80% Naïve Bayes=80%
5.	Sneha. et al. [5]	VM, NB, KNN, Random Forest, Decision tree	Random Forest=98.2 %
6.	Ahmed. et al. [6]	SVM, ANN, Fuzzy Logic	94.84%
7.	Rahman al. [7]	Logistic Regression, KNN, Random Forest, SVM, NB	Random Forest=97.4%
8.	Joti et al. [8]	including J48, NB, CART, knn, Logistic Regression, Decision Tree, Random Forest, and SVC	Decision tree=99%

2.4 Scope of the Problem

We read some papers that address Diabetes. We found out that while some researchers were very successful in their studies, others did not match the learning objectives**.** We find out a few widely used techniques for predicting diabetes, including Machine learning, Deep learning, Ensemble techniques, and Fuzzy Logic. The existing research can improve the overall effectiveness of predicting Diabetes in the early stage. But there are also a few promising areas to explore. Developing interpretable AI models with the assistance of medical experts, automating diabetes analysis, Build real-time diabetes predictors, especially apps or sites. Few diabetic predictor apps and sites already exist but user-friendly sites or apps should be developed from a Bangladeshi perspective. And Develop ML models that consider Bangladeshi dietary patterns, genetic predispositions, and environmental factors that might influence diabetes risk.

2.5 Challenges

To realize an accurate, reliable, and practically applicable approach to the early prediction of diabetes using machine learning, several challenges need to be resolved. One of them is related to the quality and availability of the available data. The prediction of diabetes demands very rich datasets that contain highly informative data in terms of patients, their medical history, life habits, and genetic predispositions. The construction of such large, high-quality datasets is difficult to obtain for some reasons: concerns about privacy, fragmentation of data across a large number of healthcare systems, and variability in recording practice. Another challenge is within the feature selection and preprocessing stages. The diversity and complexity of factors leading to diabetes demand that only the most relevant features be identified and suitably preprocessed to avoid bias or noise in the model. This relates to the treatment of missing values, data normalization, and class balancing issues that may have a strong influence on model performance. Model interpretability is another critical challenge while machine learning models, especially complex ones like neural networks, attain high accuracy, their inner mechanisms of decision-making are pretty obscure. Gaining the trust of healthcare providers and patients requires that these models be interpretable and transparent. This means coming up with techniques for explaining model predictions and ensuring that the models conform to medical knowledge and reasoning.

The challenges persist in a clinical setting when machine learning models are deployed. Such machine learning models will have to interact seamlessly with electronic health record systems and make predictions in real-time to be integrated into the current healthcare workflows. In diverse and dynamic clinical environments, models also need to be highly robust and reliable, handling variations in patients' demographics and clinical practices. Lastly, there are the ethical considerations: how to make sure the privacy of patient data is maintained, get informed consent for the use of data, and iron out possible biases in the models. In this regard, machine learning models should always be engineered with a focus on fairness, accountability, and transparency to prevent some far-reaching perils and inequities in health outcomes that might result.

Hence, the way to overcome the challenge is to address it so that machine learning models for early diabetes prediction can be successfully applied and there are advanced medical diagnostics for better patient outcomes.

CHAPTER 3
Research Methodology

3.1 Research Subject and Instrumentation

This study focuses on the application of machine learning for early diabetes prediction. The objective is the successful development of an effective, quick predictive modeling system that can apply to any population at risk of developing the disease. These advanced machine-learning frameworks include logistic regression, decision trees, support vector machines, and random forests. Standard methods of data preprocessing and feature engineering are needed to improve the model's performance. Examples of techniques include normalization, handling missing values, and selecting relevant features for diabetes prediction. In the process of training and tuning, all of these machine learning models make use of different algorithms and their parameters as a function of the diabetes dataset. The performance of these predictive models is measured by accuracy, precision, recall, and F1-score, which portray the magnitude of the success in identifying those who are at risk. These metrics were derived after running a series of experiments on a validation set and another independent test set to show the effectiveness of the model in real-world, unseen cases. This research is envisioned to positively impact the development of optimal practices for the early detection of diabetes and to further improve the implementation of machine learning strategies in healthcare. It aims to [preview video] enable the potential premature intervention and better handling of diabetes with improved predictive accuracy, hence improving and enhancing the quality of life and health of the public at large.

Instrumentation:

This project was concerned with assessing the techniques of machine learning algorithms for predicting diabetes risk. All experimentation was done in a Jupyter Notebook environment using Python. Explore machine learning techniques. This involved the evaluation of various machine learning algorithms for their effectiveness in diabetes prediction, among which were the following: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Logistic Regression, Decision Tree, Random Forest, and XGBoost Classifier. These algorithms are chosen because of their quite different approaches to learning and their established performance for a wide range of classification tasks. In the implementation of models and analysis, Python libraries

were used. Where the original study was done using a highly advanced environment from Google Colab with a GPU, this project is done with Python libraries in a Jupyter Notebook, which runs reasonably on a local machine.

3.2 Data Collection Procedure

Datasets

This study was undertaken to collect full data about diabetic patients from Sylhet Diabetes Hospital, Sylhet, Bangladesh. The data used in the research are collected directly through questionnaires designed with much importance placed on extracting information about patient's medical history, lifestyle, and management of their diabetes. This approach instilled confidence in collecting high-quality data, which would make a valuable contribution to the existing knowledge of the diabetes landscape in this region. The questionnaires were designed with the help of medical professionals to capture all the views on diabetes management and lifestyle practices of patients. Both open- and closed-ended questions were included to elicit quantitative as well as qualitative responses. The mixed-method approach ensured more profound insight into each patient's condition and management strategies. Participants were selected using a random sampling technique to ensure the sample was representative of the larger patient population in the hospital. This entailed patients who had been drawn from different wards and clinics within the hospital, thereby ascertaining that they were drawn from background experiences varied by age, sex, and socio-economic status. And this dataset was collected from Kaggle.com for my experiment.

Process of Experiments

The need for the accurate prediction of diabetes has been very critical because it gives room for early intervention and effective management. This paper, therefore, focuses attention on the use of machine learning algorithms in predicting diabetes with a dataset from Kaggle. The dataset had 520 instances and 17 features that consisted of demographic and symptomatic variables. Support Vector Machine, K-Nearest Neighbors, Artificial Neural Networks, Logistic Regression, Decision Tree, Random Forest, and XGBoost Classifier were used and compared in this study.

Dataset Descriptions

The dataset contains the following features

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 520 entries, 0 to 519
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Age                 520 non-null    int64 
 1   Gender              520 non-null    object
 2   Polyuria            520 non-null    object
 3   Polydipsia          520 non-null    object
 4   sudden weight loss  520 non-null    object
 5   weakness            520 non-null    object
 6   Polyphagia          520 non-null    object
 7   Genital thrush      520 non-null    object
 8   visual blurring     520 non-null    object
 9   Itching             520 non-null    object
 10  Irritability        520 non-null    object
 11  delayed healing     520 non-null    object
 12  partial paresis     520 non-null    object
 13  muscle stiffness    520 non-null    object
 14  Alopecia            520 non-null    object
 15  Obesity             520 non-null    object
 16  class               520 non-null    object
```

Table 3.2.2: sample attribute with a value

Attribute Name	**Convention**	**Values**
Age	age	Numeric
Gender	gender	Object
Polyuria	polyuria	Object
Polydipsia	polydipsia	Object
sudden weight loss	Sudden_weight _loss	Object
Weakness	weakness	Object
Polyphagia	polyphagia	Object
Genital thrush	genital_thrush	Object
Visual Blurring	visual_blurring	Object
Itching	itching	Object
Irritability	irritability	Object
delayed healing	delayed_healing	Object
partial paresis	Partial_paresis	Object
Alopecia	alopecia	Object
Obesity	obesity	Object
Class	class	Numeric

Methodology

Data Collection and Preprocessing:

The dataset was drawn from Kaggle, one of the largest data science competition platforms. The initial preprocessing steps included handling missing values, encoding categorical variables, and normalizing numeric features so that all of them were on the same scale.

Feature Selection:

All 17 features were used in this analysis to capture all types of information provided by the dataset. During the model evaluation phase, feature importance analysis was used to determine which variables impact the predictions the most.

Machine learning algorithms

The following algorithms were used to develop predictive models involving the development of the Support Vector Machine, K-nearest neighbors, and Artificial Neural Networks.

SVM: It finds out the best-separating hyperplane of classes in the feature space. Thus, it is very effective in high-dimensional spaces and robust to overfitting.

KNN: K-nearest neighbor gives the majority class for the K-nearest neighbors of a data point. This algorithm has the advantage of simplicity in implementation and interpretability. However, it is computationally time-consuming when dealing with large datasets.

Artificial Neural Networks

A great deal of similarity exists between ANN models and the human brain, in which neurons are also interconnected through layers. They can manage hard patterns and nonlinear relations in data.

Logistic Regression

Logistic regression offers a way to estimate probabilities of a binary outcome from one or more predictor variables. This results in an easily interpretable model for binary classification in a very direct way.

Decision Trees

Decision Trees are a recursive data-splitting method on the feature values in the data to predict the target variable. It gives an easily interpretable and visualizable model but is prone to overfitting.

Random Forest

It is an ensemble learning method where many decision trees vote to give a more accurate and robust prediction. This can help reduce overfitting by averaging out the predictions across many trees.

XGBoost Classifier

XGBoost An advanced gradient boosting algorithm that ensembles decision trees sequentially so that each subsequent tree minimizes the previous tree's prediction error. It is highly efficient and often has better predictive accuracy than other algorithms.

Experiment and Results

Data Split:

It was then split into a training set containing 78% of the data and a test set containing 22%. This split allowed the models to have enough data to learn from while providing a good test set for evaluating their performances**.**

Model Training and Evaluation

Each trained algorithm is tested on the test set after training on the training dataset. Measured and compared models in performance include evaluation metrics using accuracy, precision, recall, and F1-score. Cross-validation was also done to guarantee the robustness of the results.

Performance Comparison

Preliminary results indicated ensemble methods to perform very well with high accuracy and high F1 scores, including Random Forest and XGBoost. Moreover, strong performance was observed in SVM and ANN. Simpler models like logistic regression and KNN gave good benchmarks for the comparison.

Conclusion:

In this work, various machine learning algorithms were tested for the ability to diagnose diabetes based on a complete set of features. Of these, ensemble methods were the top performers: Random Forest and XGBoost both showed a special ability to deal with complex datasets highly accurately. These results would be very useful in helping to create more sophisticated predictive models and inform strategies for early detection and management of diabetes.

Healthcare researchers have the potential to improve predictive analytics by using several machine learning techniques, allowing clinicians to conduct their duties efficiently and leading to better patient outcomes.

3.3 Statistical Analysis

This statistical analysis within this research forms an essential tool through which meaningful interpretations of results are derived. Barring the performance of different machine learning algorithms used in diabetes prediction, rigorous statistical methods quantify and attain assessment. Descriptive statistics in terms of mean and standard deviation summarize central tendencies followed by variability in the obtained data. What is more, the methods based on inferential statistics provide tests of hypotheses and confidence intervals for evidence about observed differences and similarities among the different methodologies. The essence of the statistical analysis is to establish the nature of the findings, and thus provide a quantitative basis through which one can be in a better position to compare the efficacies of algorithms such as Support Vector Machine, K-Nearest Neighbors, Artificial Neural Networks, Logistic

Regression, Decision Tree, Random Forest, and XGBoost in diabetes prediction. They not only validate the robustness of the experimental results but provide grounds for generalizing the findings to a broader population of patients.

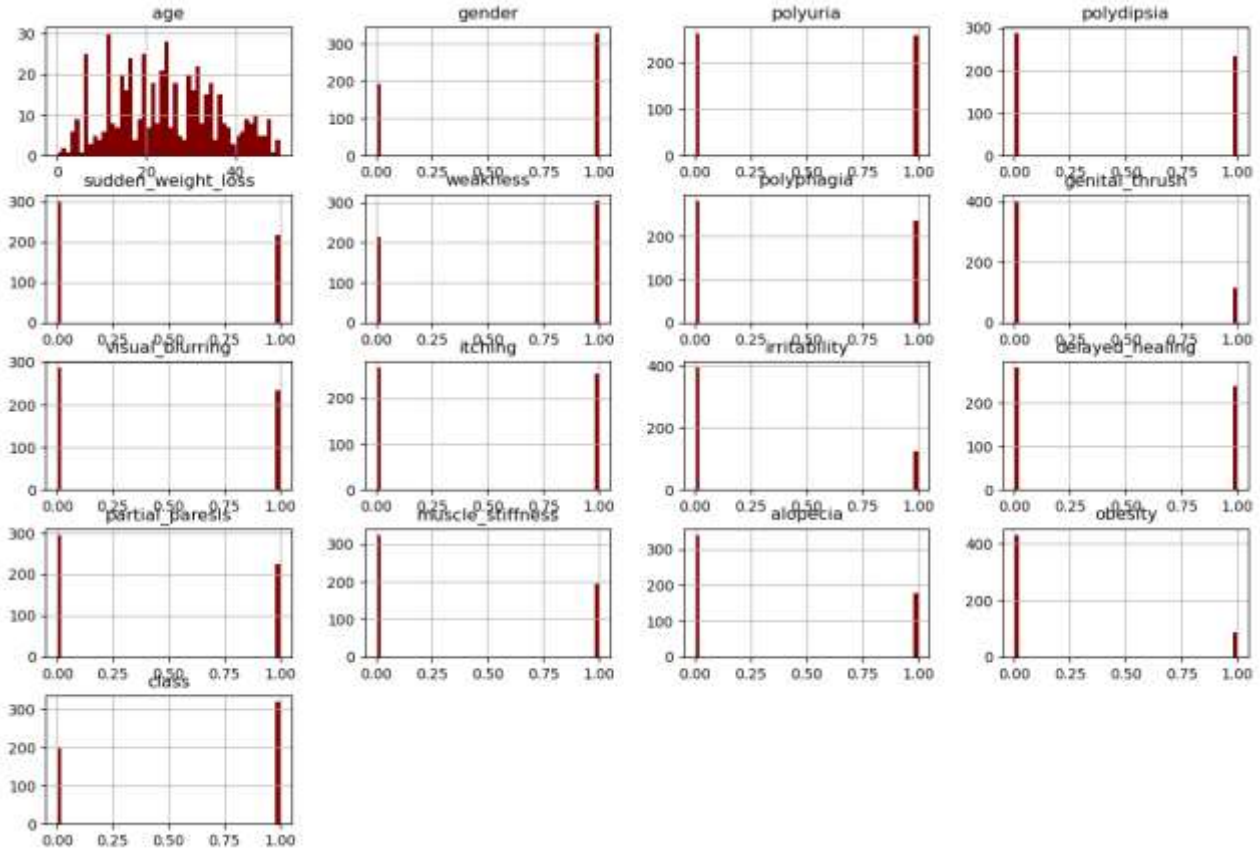


Figure 3.2.1: Histogram plot for every feature.

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing
count	520.000000	520	520	520	520	520	520	520	520	520	520	520
unique	NaN	2	2	2	2	2	2	2	2	2	2	2
top	NaN	Male	No	No	No	Yes	No	No	No	No	No	No
freq	NaN	328	262	287	303	305	283	404	287	267	394	281
mean	48.028846	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
std	12.151466	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	16.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25%	39.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	47.500000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	57.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max	90.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 3.2.2:Description of Features.

More than this, statistical insights will be important in understanding the full comparative study for any trends, patterns, or possible correlations in the differences between the numerous data points. Statistical rigor permits examination in a trustworthy and objective format of the different approaches that increase credibility and validity for research. Essentially, statistical analysis is the process of gaining key knowledge regarding empirical data to improve the current understanding of the efficacy of various approaches within this overly complicated domain of diabetes prediction.

3.4 Proposed Methodology

Bringing the prediction's whole modeling, the early stage of diabetes by machine learning was recognized to require steps that should be taken ruthlessly and methodically to have the correct classification be made reliable. Several very important research steps are data collection, model validation, and assessment, all carried out with extreme caution aimed at ensuring that the predictions are robust and accurate. In this regard, the proposed methodology acknowledges the variability in patients' data, the paucity of labeled data, and the generalization of the model; it includes detailed data preprocessing, feature selection, and validation to get optimized results. So, we followed a systematic process to enhance valuable insight into the field of diabetes prediction by giving useful recommendations to health professionals.

Frequency of Diabetes

While this had been our objective to develop a predictive model for diabetes the class imbalance that characterizes this dataset must be highlighted. The class distribution for this dataset is therefore as follows:

Class 1: This class consisted of about 320 labeled instances as 1, consequentially meaning that the outcome was positive or affected by diabetes.

Class 0: This class characterizes a negative outcome or no case of diabetes, and the dataset contains around 200 examples annotated as 1.

This becomes important during the development and evaluation of this model because class balance may impact the performance of the model. The rectification of class balance, therefore, is very critical to correct and reliable prediction. These techniques include resampling, creation of synthetic data, or adjustment of algorithms so class-balance effects are adjusted on the analysis applied.

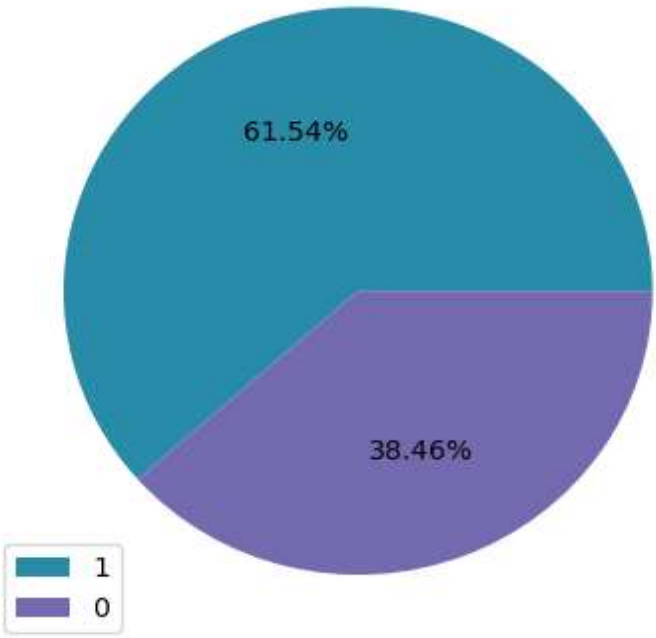


Figure 3.4.1:Distribuation of diabetic and non diabetic patient.

Proposed Methodology Steps

Data Collection: In this research, this dataset has been collected with utmost care from Sylhet Diabetes Medical Hospital using questionnaires. The dataset comprised 17 different features under the proper guidance of medical experts so that the dataset becomes accurate and reliable. Besides, it contains 520 instances that can contribute to developing a complete model for analysis.
A dataset such as this would therefore have a set of features: personal information, case history, habits in terms of lifestyle, and multiple clinical measurements relating to diabetes. This rich dataset, detailed and expert-verified, is a very sound basis for any training and validation of models toward the early prediction of diabetes.

Data Preprocessing: This stage cleans up the dataset, handles missing values in a dataset, normalizes data, or encodes categorical variables. This step ensures that the best input data is used for training machine learning models.

Feature Selection:

Feature selection is done to identify the most relevant features, which have predictive power about Diabetes. Variable Selection: Techniques confirmation of the hypothesis, Importance Ranking.

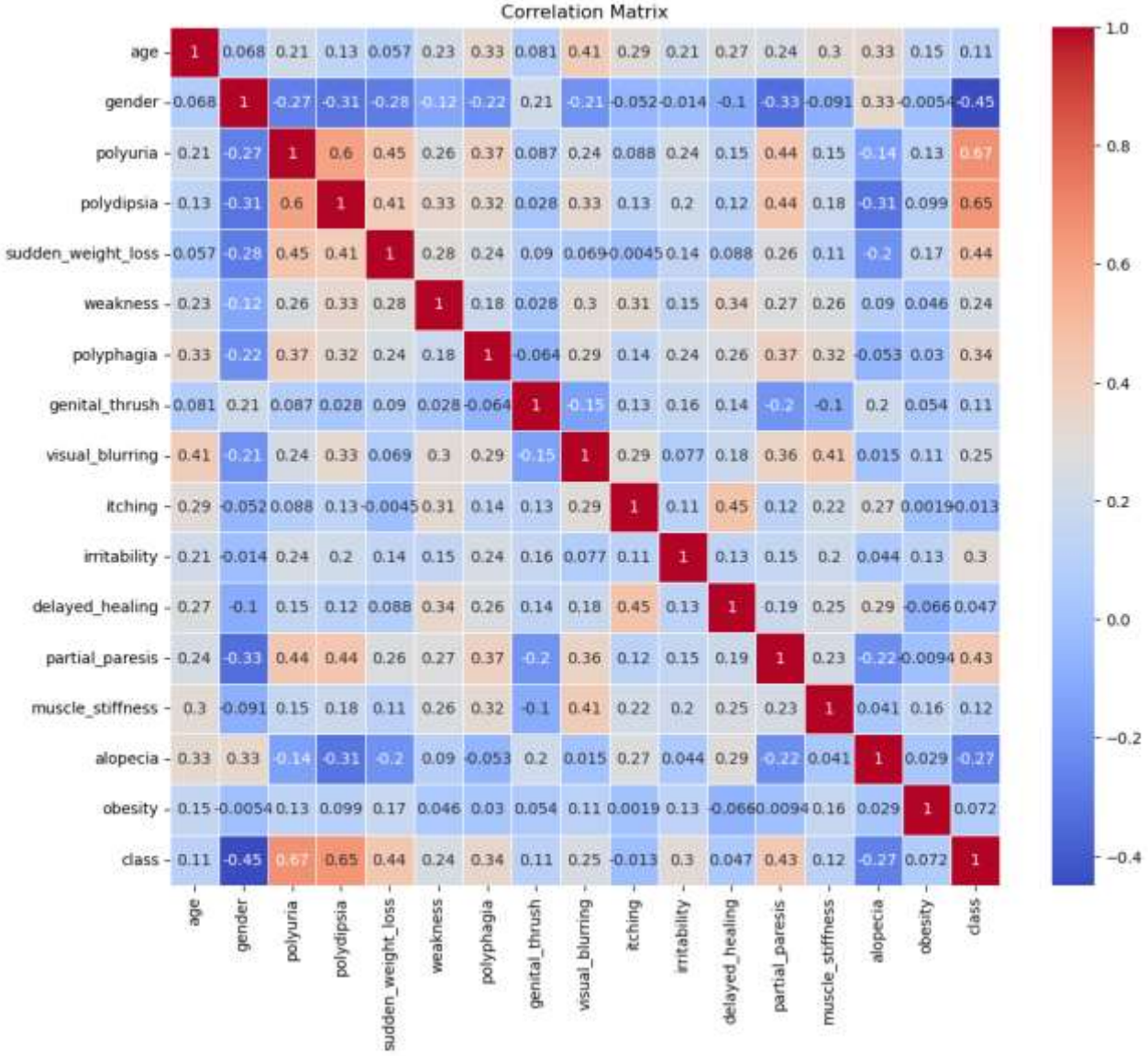


Figure 3.4.2 The Correlation between Features.

A correlation matrix is simply a table that displays a summary of the correlation coefficients of each pair of features in your dataset. A correlation coefficient is a value between -1 and 1 indicating both the strength and orientation of the linear relationship between variables. With the correlation matrix, we can gain meaningful insights about how the features are related to one another in our dataset. This is information that can be utilized in many ways, from data cleaning to feature selection and model building. In this figure shows the correlation between

each feature.

Model Selection:

Many machine learning algorithms will be sampled during this study, such as the Support Vector Machine, K-Nearest Neighbors, Artificial Neural Networks, Logistic Regression, Decision Tree, Random Forest, and XGBoost. Initial performance metrics are used in selecting the best model.

Model Training:

The chosen model gets trained on the labeled dataset by appropriate training techniques. Transfer learning can be applied when using pre-trained models, and the tuning of hyper-parameters was done to optimize the model's performance. This is followed by a split of data into training and test sets, wherein 22% of data will be used to test.

Model Evaluation:

The trained model is evaluated concerning a validation set consisting of at least one performance metric: accuracy, precision, recall, or F1-score. This test is important for fine-tuning your model and knowing where to improve in.

Testing Model

The model is then tested on another independent test set to ensure that it generalizes well to completely new data. Thus it would confirm that this model correctly predicts diabetes in practical scenarios and produces accurate and reliable results.

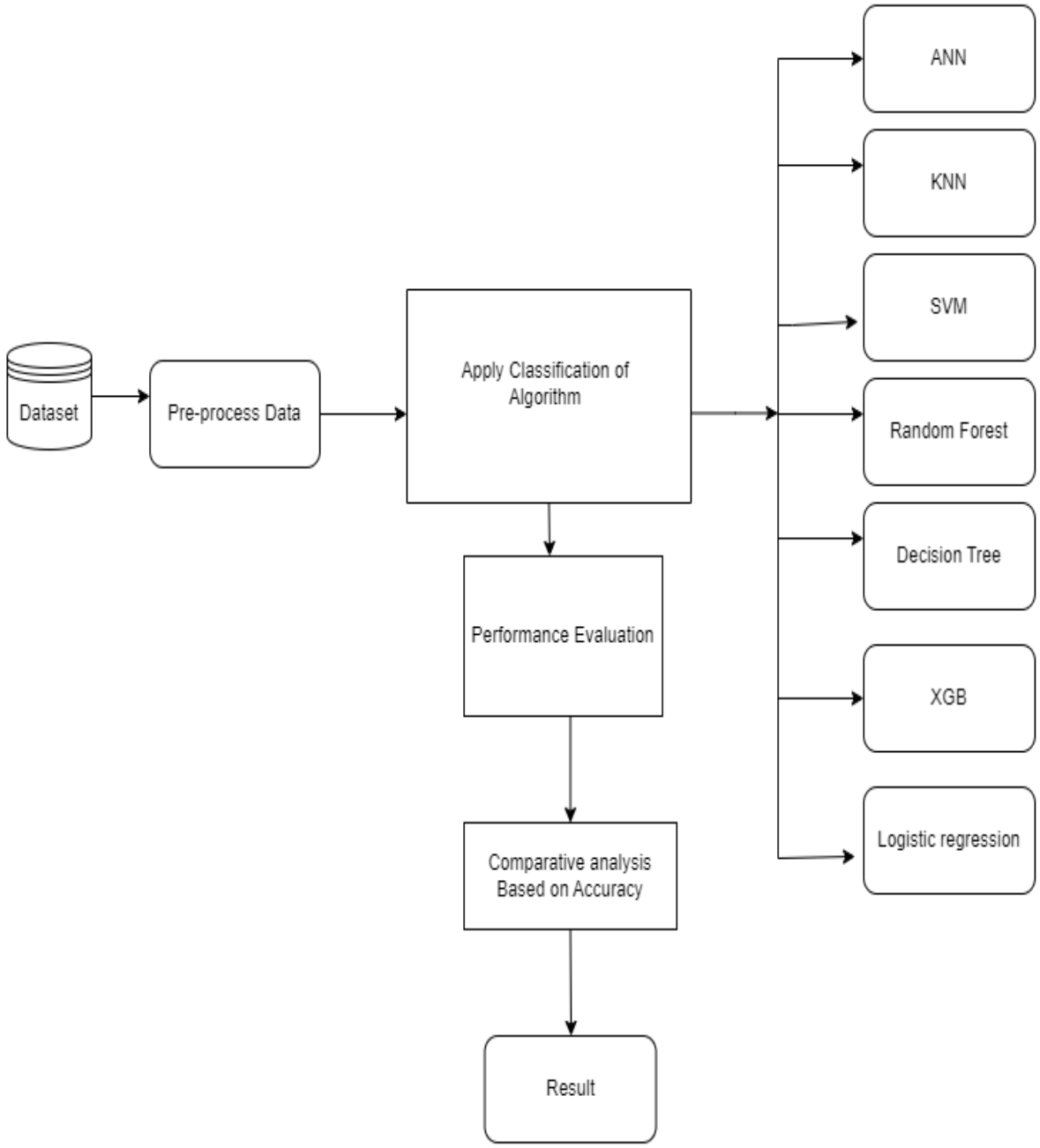


Figure 3.4.3: Proposed Model Diagram

Accuracy:

Accuracy measures the proportion of correctly classified cases from among the total cases. It is given by:

Accuracy=(TP+TN)/(TP+TN+FP+FN)----------(i)

Precision:

Precision is the ratio of correct identifications. This is useful in cases where the cost of false positives might be high. It is formally calculated as:

Precision=TP/(TP+FP)----------(ii)

Recall:

The recall represents the ratio of actual positives that go unidentified. It is of utmost importance in those cases where the cost of being a false negative is high. Recall is defined as:

Recall=TP/(TP+FN)----------(iii)

F-1 Score:

The F1-score represents the harmonic mean between precision and recall and, therefore, it gives a balanced measure of both. It is defined as:

F-1 Score=2*(Precision*Recall)/(Precision + Recall)----------(iv)

All these metrics are summarized in the confusion matrix, which enables a direct evaluation of values such as TP, FP, TN, and FN.

3.5 Implementation Requirements

Some of the key resources and tools that would be required in implementing the proposed methodology of predictive modeling for early prediction and diagnosis of diabetes using machine learning techniques are as follows:

Computing Resources: High-performance computing, like GPUs or TPUs, is an essential tenet in effectively training models for inference. These supportive resources will help fine-tune and run complex machine-learning models.

Machine-Learning-Libraries: Deep learning architectures-based ML libraries, such as TensorFlow, and Keras, among others, would be required in the development, training, and evaluation of the predictive models. These libraries will provide the required modules for the construction of complex algorithms and neural networks.

Data Processing Libraries: Other libraries like NumPy and pandas are applied for mathematical computations and data processing. Scikit-learn is also applied in the implementation of conventional machine-learning algorithms with evaluation metrics.

Data Preprocessing: Inventing a diverse dataset with good labeling and comprehensive health metrics is required for training and testing models. A wide variety of patient demographics and indicators of health should be involved in the dataset for the robust performance of models. Data preprocessing techniques are applied, which involve normalization, handling missing values, and feature selection to prepare a dataset that will further be used for model training.

Experimental Platform: Google Colab and Jupyter Notebook will be used for the experiments as a cloud-based platform that has enough computational power to carry out experiments. It also provides integrated tools for carrying out implementation processes and progress reporting, which is also useful for reporting the results of this implementation.

Reporting and Communication: The results of the projects are effectively communicated through text processing tools and documentation across Google Colab and Jupyter Notebook. This enables clear reporting on the methodologies applied, the experimental setup, and the results.

CHAPTER 4
Experiment and Result Discussion

4.1 Experimental Setup

Given this fact, an attempt was made to systemically design an experimental setup for this research on predictive modeling for the early detection of diabetes using machine learning so that a thorough model evaluation process is followed up with its performance assessment. Experiments were conducted both in Jupyter Notebook and Google Colab to make use of availed robust computational resources, which included even Tesla GPUs from the Google Colab environment, which very instrumental in efficiently training and running a model that was key to running due to the complexity of the tasks of machine learning.

The data used in this study was quite diverse, ranging from a wide variety of health metrics to patient demographic information, which is very important in robust predictive model building. Some data preprocessing activities include normalization to scale the features and treating missing values for completeness, along with feature selection aimed at choosing only the most relevant attributes that increase the probability of predicting diabetes. Therefore, such data preprocessing methods were put in place to ensure the availability of a good dataset for training these models and generally improve overall model performance.

In this paper, various machine learning algorithms were involved, including Logistic Regression, Decision Trees, Random Forests, SVM, KNN, ANN using TensorFlow and Keras, and the XGBoost Classifier. Each algorithm was implemented using Python machine learning libraries like TensorFlow, Keras, and sci-kit-learn so that all predictive methods will be covered comprehensively.
Individual model performance metrics were standard, consisting of accuracy, precision, recall, and the F1-score. It was computed through a set of validation and test experiments on an independent test set to show the effectiveness of the models in real-world and unseen cases.

During the experimental process, Google Colab and Jupyter Notebook provided an interactive and collaborative environment wherein to code, debug, and document. These platforms helped in clearly

reporting about the implementation process, progress, and results, with effectual communication of the project result.

In essence, the experimental setting was carefully devised to use modern computational resources with robust data preprocessing techniques and a suite of machine learning algorithms to build an accurate and reliable system of predictive modeling to be used in early diabetes detection.

4.2 Experimental Results & Analysis

Result of Experiment

The results for early diabetes detection predictive modeling using applied machine learning algorithms show different models with various levels of accuracy. For this study, the following models have been taken into consideration: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, Artificial Neural Network from TensorFlow and Keras, XGBoost, Naive Bayes; the performance was evaluated in terms of accuracy.

It also yields an accuracy of 0.9913 from random forest and XGBoost. Probably the most consistent takeaway from these results is that ensemble methods—random forest and XGBoost—have very good performance on this dataset. Their strength could be related to their ability to handle complex interactions between features and reduce overfitting by making aggregation of trees' predictions. On the other hand, the Decision Tree model also performed well, with an accuracy of 0.94, thus proving that it captured relevant patterns from the data. But unlike ensemble methods, it suffered from lagging a little behind, thereby suggesting that one single decision tree might prove to be less powerful as opposed to a forest of trees.

These models' accuracies were a little below the baseline, at 0.935 for Logistic Regression, 0.93 for SVM, and 0.93 for Naive Bayes. Of course, all of these models have the advantages of simplicity and relative efficiency, so they are suitable for the very first tests of models to compare against a baseline. Although their performances were relatively high, they were outperformed by more complex ensemble methods. The K-Nearest Neighbors model returned an accuracy of 0.91. Though it is a very simple, interpretable algorithm, it can be pretty sensitive to the choice of hyperparameters

and how the dataset is scaled. It worked fairly well in this case but didn't do quite as well as the best-performing models. The ANN returned the lowest accuracy of 0.85. Although ANN is a very powerful and flexible model, its performance is fully dependent on the chosen architecture and hyper-parameters together with the quality and size of data. The relatively lower accuracy may perhaps indicate that more tuning may be necessary, probably using larger datasets.

Table 4.2.1: Performance Summary of Model: Training vs. Test Accuracy for Early Prediction of Diabetes.

Architecture	Training Accuracy	Test Accuracy
SVM	94%	93.91%
Logistic-Regression	93.58%	93.51%
Naïve-Bayes	88.88%	93.04%
XGBoost	97.77%	99.13%
KNN	93.33%	91%
Random-Forest	100%	99.13%
Neural-Network	88.15%	85.22%
Decision-Tree	93.34%	94%

Table 4.2.2: Precision, Recall, F1-Score, and Support Classification Report.

SVM				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.95	0.89	0.92	47
Diabetes-Prediction	0.93	0.97	0.95	68
Accuracy			0.94	115
Macro Avg	0.94	0.93	0.94	115
Weighted Avg	0.94	0.94	0.94	115

Logistic-Regression				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.89	0.95	0.92	44
Diabetes-Prediction	0.97	0.93	0.95	71
Accuracy			0.94	115
Macro Avg	0.93	0.94	0.94	115
Weighted Avg	0.94	0.94	0.94	115
Naïve-Bayes				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.85	0.98	0.91	41
Diabetes-Prediction	0.99	0.91	0.94	74
Accuracy			0.93	115
Macro Avg	0.92	0.94	0.93	115
Weighted Avg	0.94	0.93	0.93	115

XGBoost				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.98	1.00	0.99	47
Diabetes-Prediction	1.00	0.99	0.99	68
Accuracy			0.99	115

Macro Avg	0.99	0.99	0.99	115
Weighted Avg	0.99	0.99	0.99	115

KNN				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	1.00	0.82	0.90	57
Diabetes-Prediction	0.85	1.00	0.92	58
Accuracy			0.91	115
Macro Avg	0.93	0.91	0.91	115
Weighted Avg	0.93	0.91	0.91	115
Random Forest				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	1.00	0.98	0.99	48
Diabetes-Prediction	0.99	1.00	0.99	67
Accuracy			0.99	115
Macro Avg	0.99	0.99	0.99	115
Weighted Avg	0.99	0.99	0.99	115

Neural Network				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.95	0.81	0.87	47
Diabetes-Prediction	0.88	0.97	0.92	68
Accuracy			0.90	115
Macro Avg	0.92	0.89	0.90	115
Weighted Avg	0.91	0.90	0.90	115

This table provides a comprehensive representation of the performance metrics of different machine-learning models that have been considered in this study for early diabetes detection. These include SVM, logistic regression, Naïve Bayes, KNN, random forest, XGBoost, decision tree, and neural network. For each, the precision, recall, F1-score, and support for the two classes, Normal and Diabetes Prediction, are computed.

The models with the highest accuracy scores will be 99% by Random Forest and XGBoost. Both models did amazingly well on precision and recall, showing very good F1 scores for both classes, thus proving these models for robustness and reliability in making accurate predictions on diabetes. These are then followed by an accuracy of 94% by both the SVM, logistic regression, and decision tree models, also with balanced results for both precision and recall. These models thus have very good performance overall in differentiating between normal and diabetic cases. The KNN model, although it had less recall for the Normal class, still had good overall performance with an accuracy of 91%. It suggests that KNN is relatively reliable but might benefit from further tuning or additional data to improve its recall in normal cases. The class Diabetes-Prediction shows an accuracy of 93% for Naïve Bayes with precision and recall quite high compared to the Normal, which is a relatively lower performance, indicating that it is especially good at picking out diabetic cases. The Neural Network model is the worst among the models, with an accuracy of 90%. Although the precision and recall of the Normal class within it are relatively high, when compared to those of all other models, these values are poor, thereby leaving scope for improvement in its predictive power. In summary, the Random Forest and XGBoost models rank at the top among all of the models,

evidencing the strongest predictive power for early diabetes detection. The remaining models, including SVM, Logistic Regression, and Decision Tree, exhibit relatively good performance and therefore could be considered a reliable alternative. However, some models, like KNN, Naïve Bayes, and Neural Network, may require further optimization concerning predictive accuracy and reliability.

Training and Validation Accuracy and loss of Machine Learning Model:

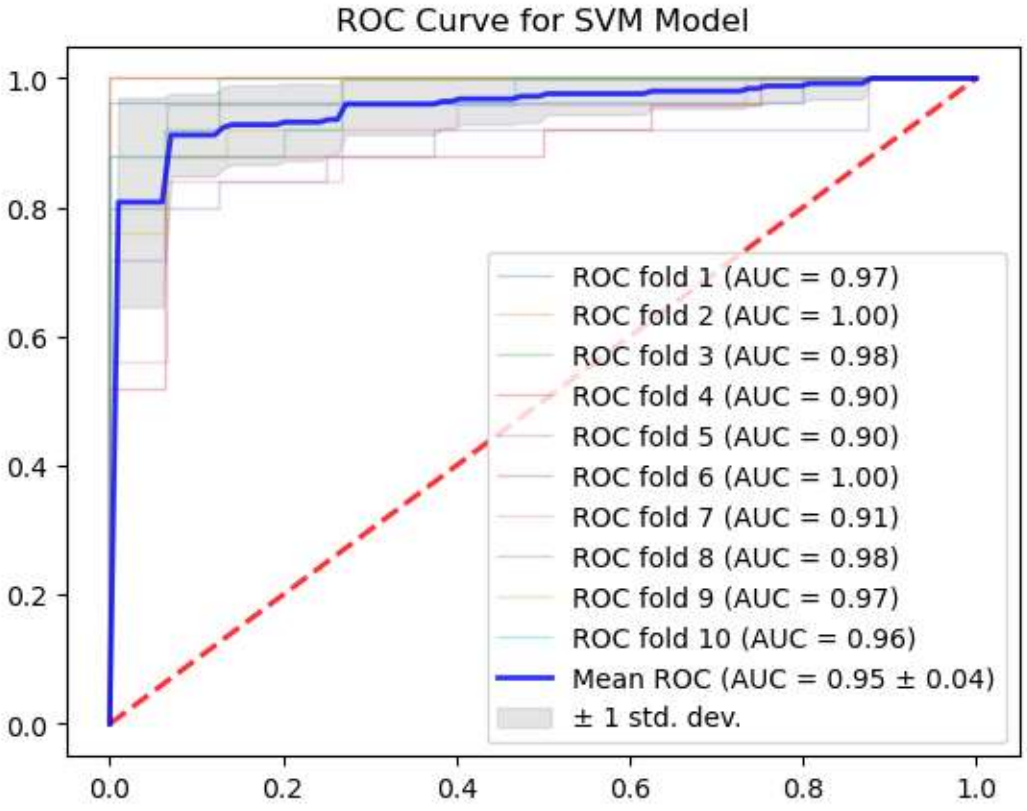


Figure 4.2.1 ROC curve of our proposed SVM model.

The figure shows The ROC curve in the image shows 10 folds of the SVM model, which means the data was split into 10 parts, The model was trained on 9 parts and tested on the remaining 1 part, 10 times. Each fold is represented by a different colored line in the graph and an AUC value. AUC stands for Area Under the Curve, and it summarizes the performance of the model between 0 (worst) and 1 (best).

The fact that all the curves are close to the top left corner indicates that the model performs well across all folds. The mean ROC AUC is 0.95 which is also a good value. Overall, the ROC curve suggests that the SVM model is good at distinguishing between positive and negative cases.

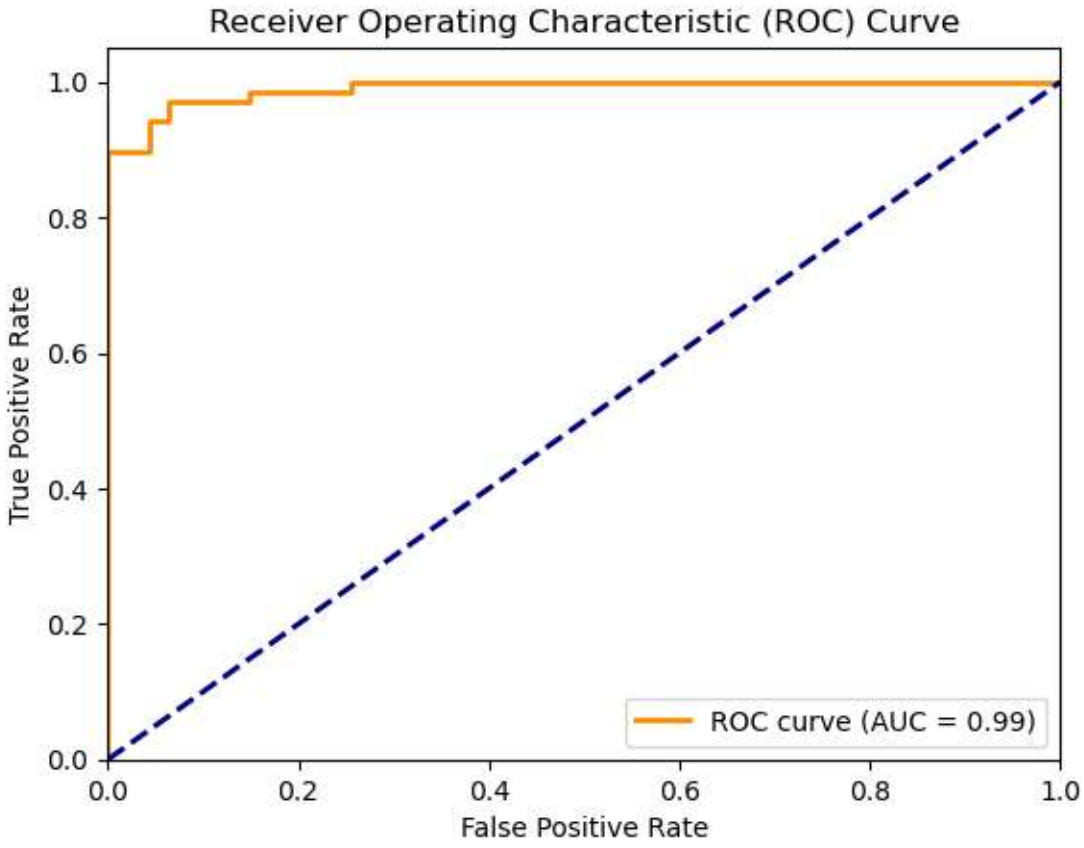


Figure 4.2.2 ROC curve of our proposed Logistic Regression model.

The figure is shown. The x-axis gives the number of iterations that the model goes through, and the y-axis gives the loss and the accuracy. The ROC curve can be defined as a graph showing the performance of a classification model at all classification thresholds. Linear Regression ROC curve is almost the same as SVM ROC curve representation.

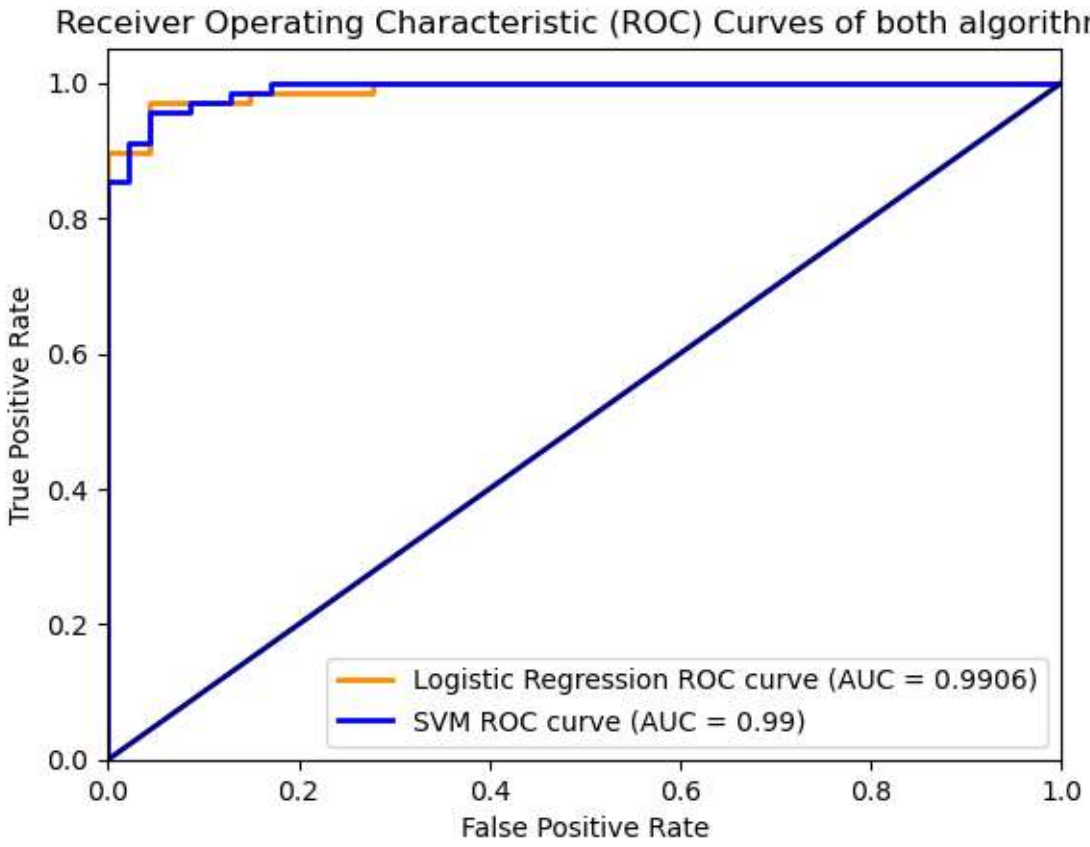


Figure 4.2.3 ROC curve of our proposed Linear Regression and SVM model.

The two ROC curves of the figure, one corresponding to Logistic Regression and the other to SVM, lie pretty close to the upper-left corner. That means essentially that both models perform quite well. The AUC is a number that may be used to measure the performance of a model. It is mentioned in the legend of each curve. The higher the AUC is, the better the performance. Using the corresponding line graph in this image, the AUC for a Logistic Regression model is 0.9906 versus 0.99 for an SVM model. Therefore, by this line graph, a logistic regression model is on average doing a bit better at classifying the positive and negative instances within this dataset.

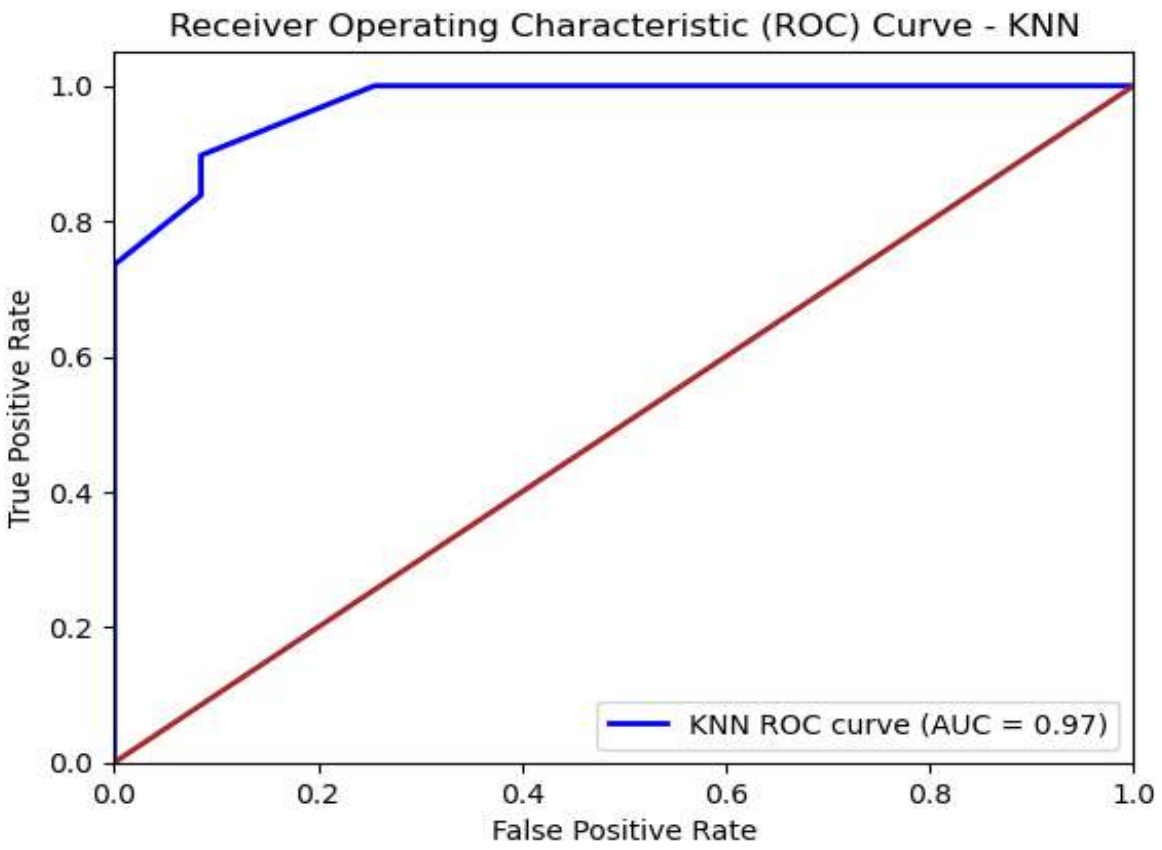


Figure 4.2.4 ROC curve of our proposed KNN model.

The image shows that the ROC curve corresponds to a KNN model with K = 5. More precisely, it implies what is: in K-nearest neighbor, a new data point is classified by looking at the k-nearest data points in the training data set. In this case, k is set to 5. The model then makes a classification based on a majority vote of its neighbors. A perfect classifier would have an ROC curve that goes straight up the left side of the graph, and then levels off at the top. This means that the model can perfectly classify all of the positive instances and none of the negative instances. Nevertheless, the random classifier would form an ROC curve, arcing from the lower left to upper right on the graph. This explains a model not improved over random guessing.

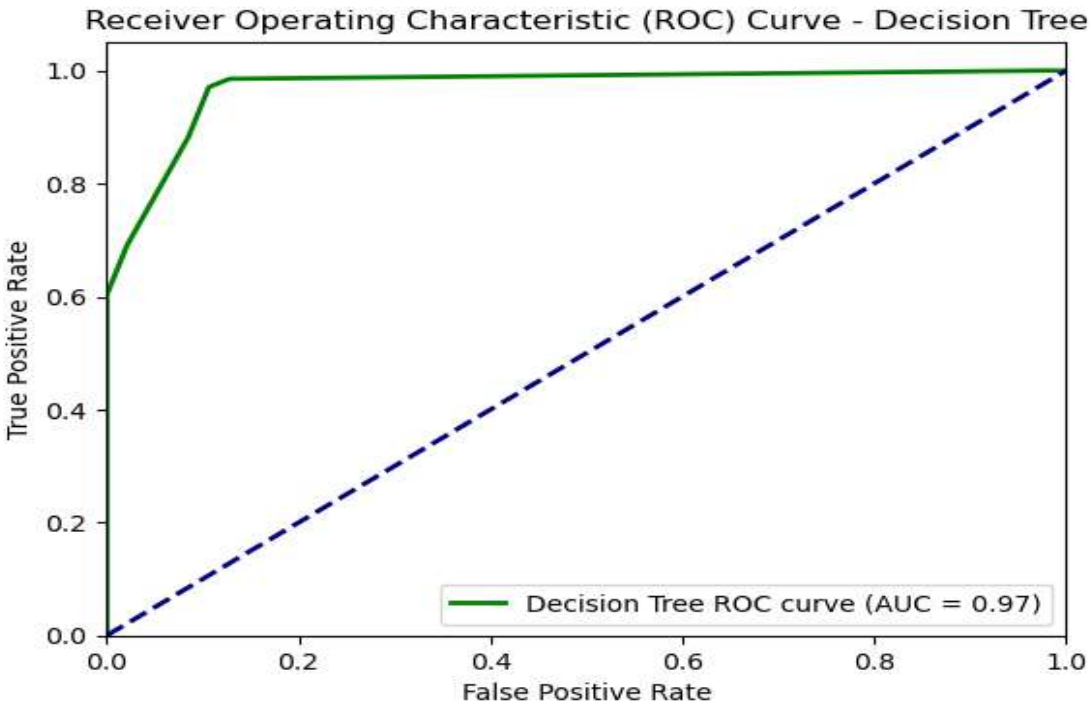


Figure 4.2.5 ROC curve of our proposed Decision Tree model.

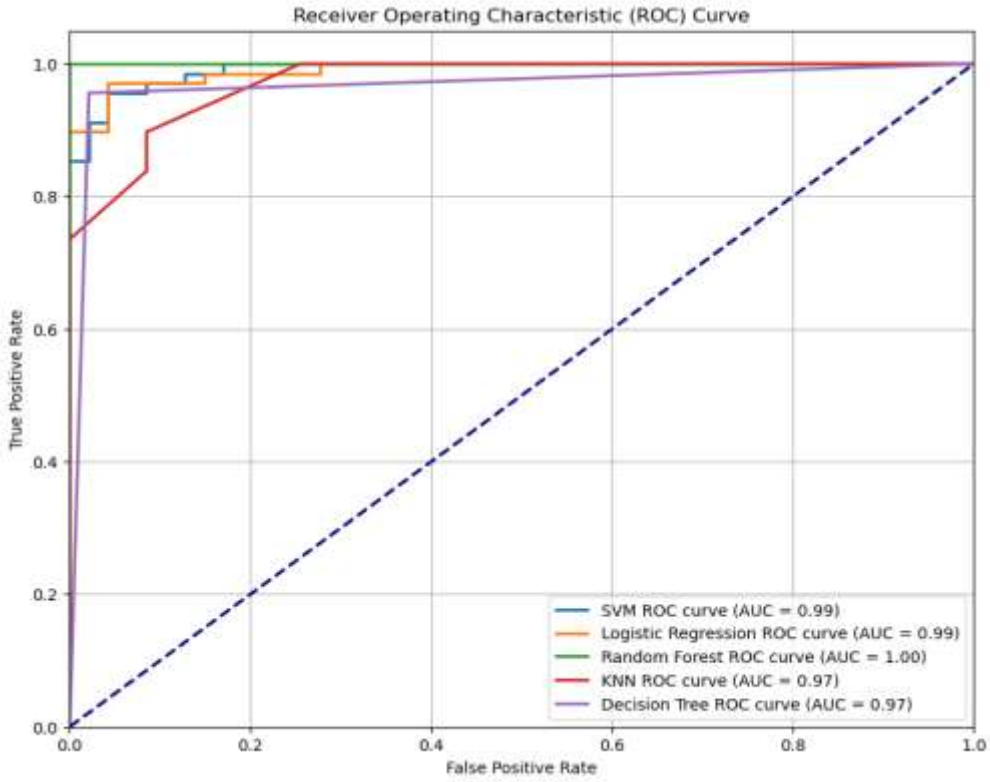


Figure 4.2.6 ROC curve of our proposed SVM,LR,RF,KNN,Decision Tree model.

Training and Validation Accuracy and Loss of Neural Network:

In the plot depicted here, showing the training performance of a machine learning model in 30 epochs, two critical parameters may be identified: loss and accuracy.

First of all, the trends in loss are analyzed: both the Training Loss and the Validation Loss drop very fast at the beginning, which already indicates that effective learning has taken place. Then, they stabilize, thus giving the signal that learning by the model has stabilized. A small gap between Train and Validation Loss may suggest some overfitting; regularization techniques or early stopping may be applied in this respect by the researcher .Accuracy Trends: Both Training Accuracy and Validation Accuracy are high at the beginning but stay constant thereafter. The model is working more accurately on training data; however, the gap that comes in with validation accuracy suggests 'keeping a close eye'. It simply means that this disparity highlights whether a model generalizes well or further training results in diminishing returns. Considering these trends is important in tuning hyperparameters to improve the performance of models. One should be adjusting the hyperparameters, trying different architectures, and implementing early stopping to ensure high performance without allowing overfitting.

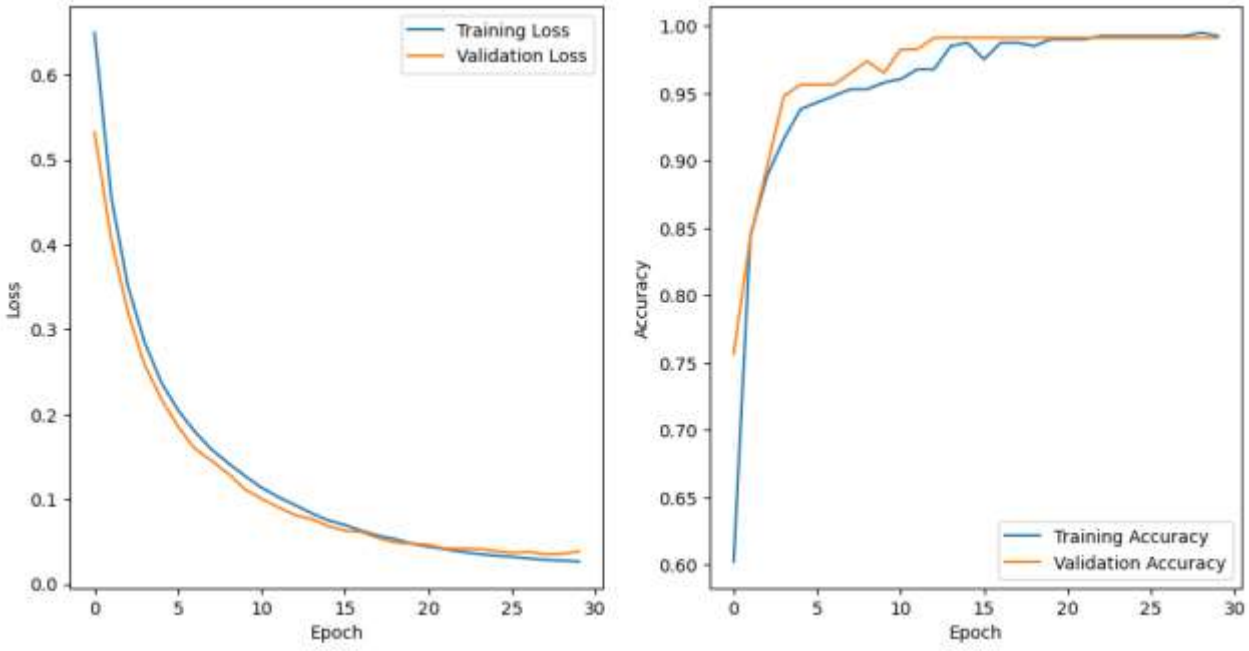


Figure 4.2.7: Training and validation accuracy and loss over the epochs (Neural Network)

Confusion Matrix of All Model after Experiment

	Normal	Diabetics Prediction
Normal	**42**	**5**
Diabetics-Prediction	**2**	**66**

Figure 4.2.8: CM after SVM

	Normal	Diabetics Prediction
Normal	**42**	**5**
Diabetics-Prediction	**2**	**66**

Figure 4.2.9: CM after Logistic Regression

	Normal	Diabetics Prediction
Normal	**40**	**7**
Diabetics-Prediction	**1**	**67**

Figure 4.2.10: CM after Naïve Bayes Model

	Normal	Diabetics Prediction
Normal	**47**	**0**
Diabetics-Prediction	**1**	**67**

Figure 4.2.11: CM after XGB Classifier

	Normal	Diabetics Prediction
Normal	**47**	**0**
Diabetics-Prediction	**1**	**67**

Figure 4.2.12: CM after Random Forest.

	Normal	Diabetics Prediction
Normal	**47**	**0**
Diabetics-Prediction	**10**	**58**

Figure 4.2.7: CM after KNN

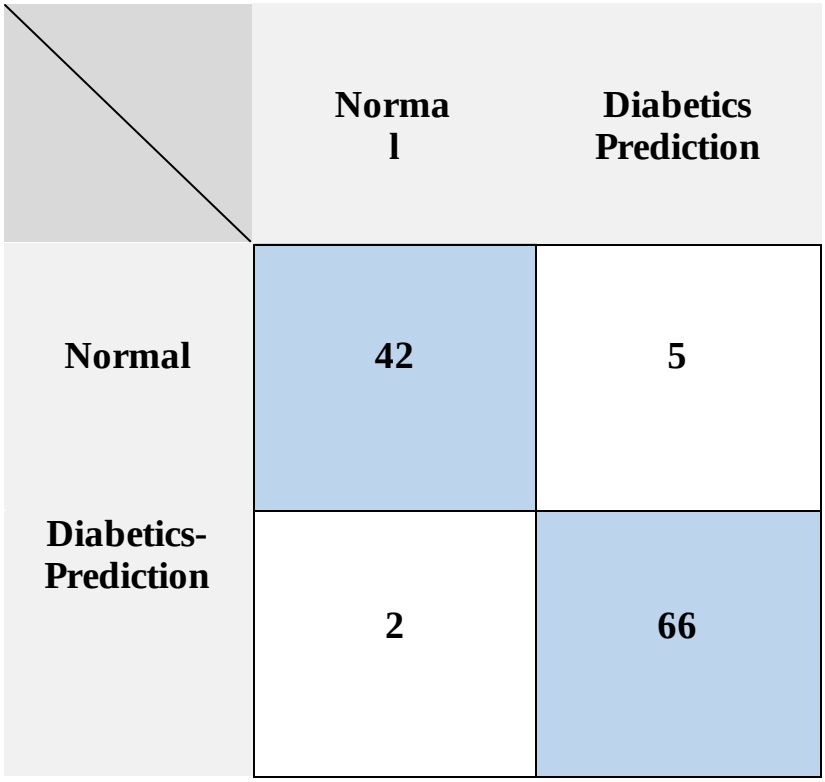


Figure 4.2.13: CM after Decision Tree

The experimental results indicate that the ensemble methods give the highest accuracy in early diabetes detection, especially Random Forest and XGBoost. The findings demonstrate that indeed, advanced machine learning techniques could yield robust and reliable predictive models. Future work would then be fine-tuning these models and probably introducing more features or data augmentation techniques to improve model performance even more.

4.3 Discussion

The parameters evaluated in the experimental results and analysis include the performance of various machine learning algorithms with models that involve predicting the early detection of diabetes. These models in this study represent the following: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, Artificial Neural Network, XGBoost, and Naive Bayes. Each of these models was trained and quantity tested with the same

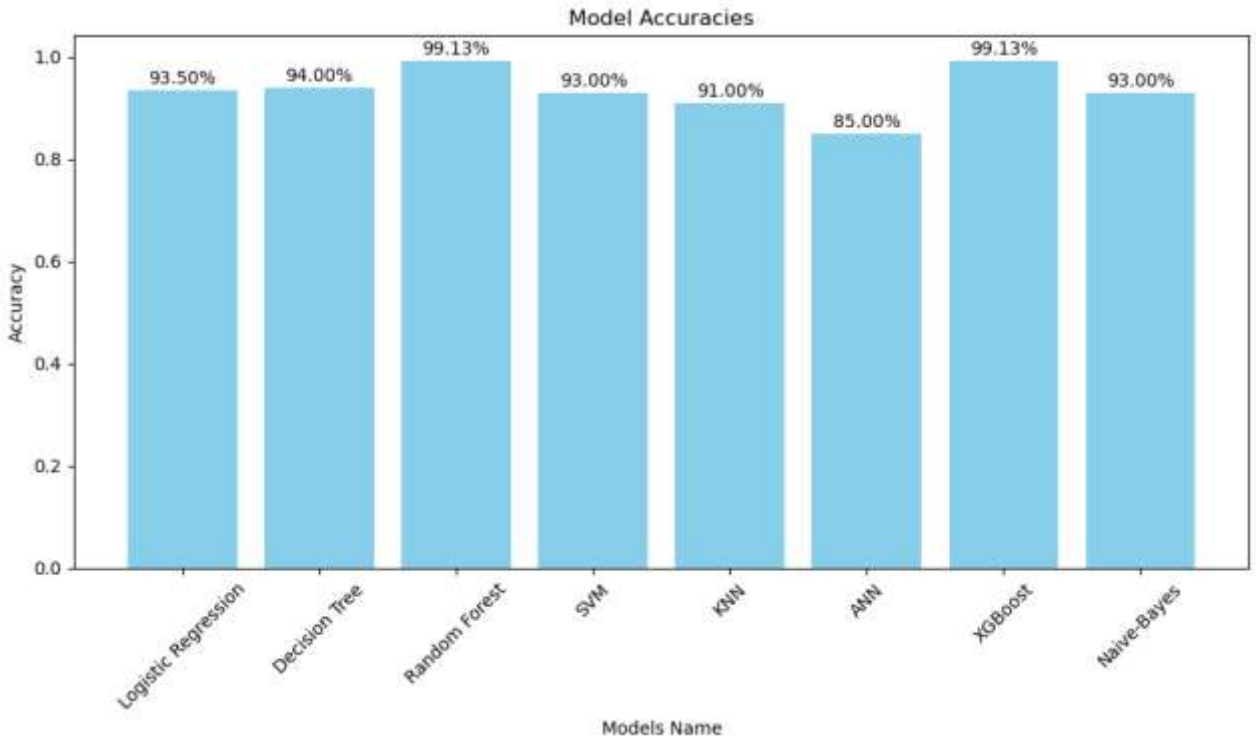


Figure 4.3.1: Accuracy comparison

dataset for a fair comparison purpose. For measuring the performance of each model, accuracy in each was determined. It emerged that the Random Forest and XGBoost had the highest accuracy, approximately 99.13%. This high goodness-of-fit value can be attributed to their ensemble nature of combining predictions from several base learners to get a generalized model that is robust to misfits. The Decision Tree model also performed excellently with an accuracy of 94%, a little outperforming the Logistic Regression model with an accuracy of 93.5%.

In the same vein, SVM and Naive Bayes models achieved test accuracies of 93% and 93.04%, respectively, applied to this case to show their prowess in handling classification tasks efficiently. The KNN model itself had an accuracy of probably slightly less than 91% because of sensitivity to the choice of k-value and the nature of the dataset. While the ANN model had complex potential with high capability to capture intricate patterns in the data, it returned an accuracy of 85.22%, thus calling for further tuning or optimization. After evaluating all algorithms we got the highest accuracy of 99.13% and the highest training accuracy of 100% of Random Forest so I chose this algorithm for my proposed algorithm.

Proposed Model Structure

In this research, the proposed model is Random Forest for its reliability and its highest accuracy. RF is explained as a process that develops many decision trees with the help of referring to each one for decision-making. Normally, n number of data points are selected from the dataset, and when combined together, a stable decision is generated. In case of more guesses, the average of all predictions is considered. The RF technique is used to solve both the classification and regression problems. As shown in This process produces quite a large number of trees in the forest. The more the number of trees the healthier the forest is, the findings are then produced with high accuracy. The RF architecture comprises multiple trees; each tree provides a certain option. All options are then averaged to assess the most recent prediction. Important components of a Random Forest model structure include many decision trees independently created from each other, where every decision tree is trained on a different random subset of the dataset; during training, every tree makes its predictions independently. The final prediction is usually aggregated from all the predictions made by the individual trees in the Random Forest model using majority voting for classification tasks. It is an ensemble technique that uses the power of many models to make good predictions and also be robust. The application of a random forest model has several crucial implications for this research work.

It handles class imbalance in the dataset with techniques such as bootstrap sampling and random feature selection, which ensure that the built model will then be unbiased and accurate. Furthermore, Random Forest is able to capture complex non-linear interactions between features, which is crucial for predicting diabetes correctly from 17 features that describe a very diverse set. It further gives feature importance scores, which provide the relative importance of all features used and identify the most important predictors of diabetes. This information is very useful to health experts. Thus, in this study, besides improving the accuracy and reliability of the diabetes prediction model based on the Random Forest algorithm, explained variance will also enable a better understanding of the underlying factors of diabetes that play a vital role in diagnosis and treatment planning.

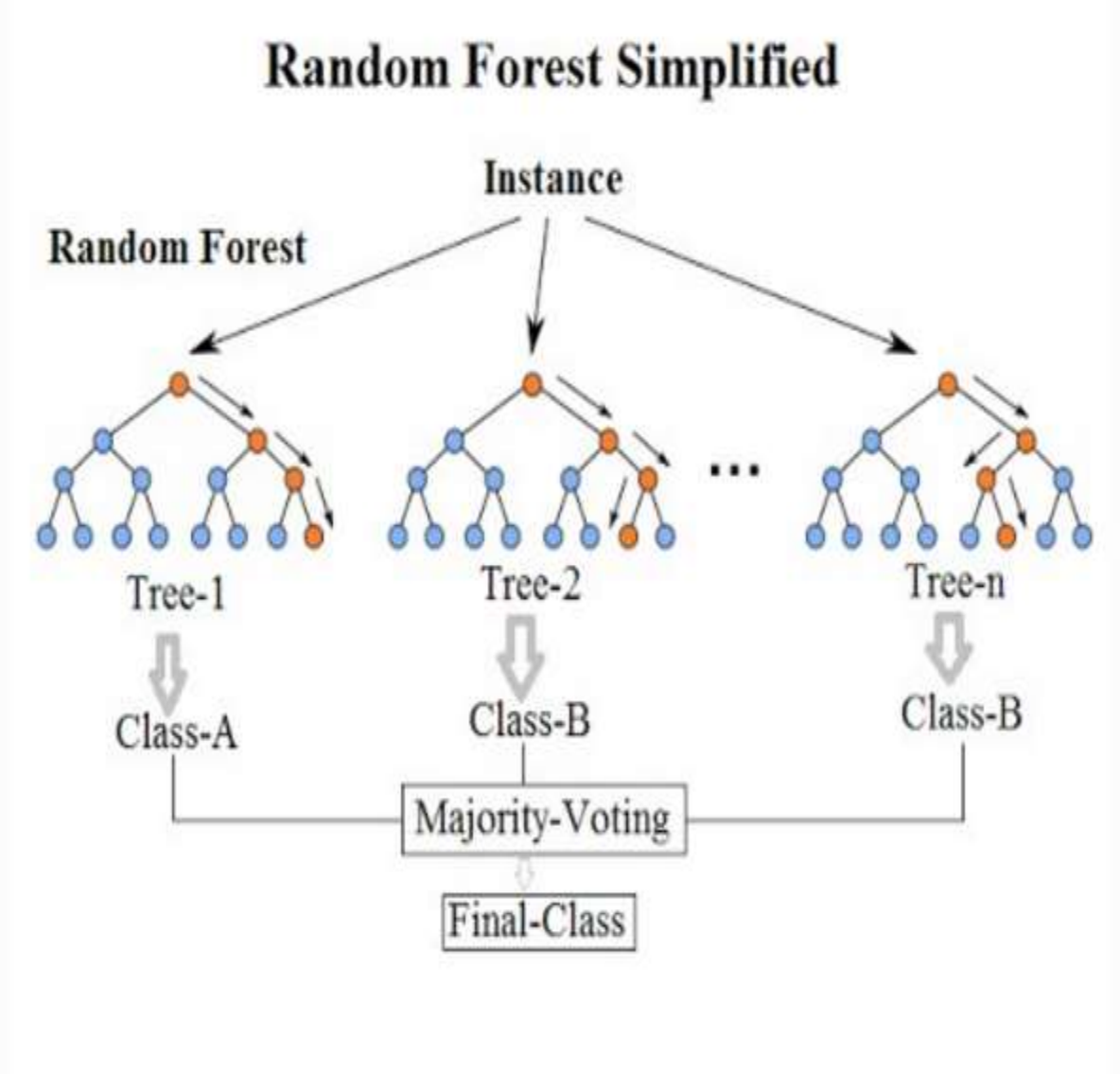


Figure 4.3.2: Random Forest Structure.

CHAPTER 5
Impact on Society

5.1 Impact on Environment

The potential of machine learning in indicative modeling for the early detection of diabetes is immeasurably huge for society, especially on matters of health. Diabetes is just one of those serious chronic conditions that affect millions across the globe and may present serious health complications if it continues to run unmanaged. Early detection may, therefore, play a very important role in mitigating its impact by allowing timely intervention and treatment that eventually works toward improving the outcome and hence the quality of life of patients.

In terms of benefits to society, the most important would be the reduced burden on healthcare systems. Early identification of those at risk of developing the condition could have healthcare providers take proper prevention measures that will drastically reduce the incidents of serious complications such as cardiovascular diseases, kidney failures, and neuropathy. That could also help reduce cases of admission to hospitals, lower healthcare expenditure, and bring about more effective deployment of medical resources.

Moreover, increased access to machine learning-based diagnostic facilities can reduce health disparities in underserved and remote areas. Traditional diagnostic procedures require special instruments and trained human resources, which are generally not available in these areas. In contrast, applying the machine learning models within a mobile health application or a telemedicine platform can be more cost-effective for scaling early diagnosis of diabetes. This democratization can ensure access to critical diagnostic services for the vast majority of people, no matter the geographical or socio-economic barriers.

Another important contribution from this research into the area of personalized medicine is related to patient-empowered healthcare. Machine learning models can look at a variety of factors: genetics, lifestyle, environment—almost all data points—to provide personalized risk assessment and give customized risk-reduction proposals. Such a touch of

personalization may motivate him or her to take steps toward a healthy lifestyle with precautionary measures that reduce their chances of developing diabetes or other related diseases.

Further, predictive modeling can be used to ensure maximum efficiency for clinical trials and other research studies aimed at treating diabetes. By spotting the people at high risk, it is the people who matter that are being recruited into the study. Therefore, it will ensure that research is done on relevant populations. It can enable the fast development of new treatments and interventions for the good of patients and medical knowledge.

The inclusion of machine learning in diabetes detection creates room for continuous monitoring, hence putting early warning systems into place. Wearable devices and health monitoring applications can collect real-time data and run machine learning algorithms to detect the first signs of diabetes or their complications. This will allow systems to warn patients and health providers of possible problems for early action, hence reducing the risk of serious events in health.

In other words, predictive modeling with the purpose of early detection of diabetes using machine learning brings the potential to revolutionize health care as a whole through early interventions that will reduce health complications, thus cutting down health costs and improving patient outcomes. It enhances accessibility to diagnostic services, promotes personalized health care, and accelerates medical research. These technologies, as they evolve further and become a part of healthcare systems, can go on to unlock key improvements for people at risk of developing diabetes, ensuring.

5.2 Impact on Environment

The application of predictive modeling toward early diabetes detection using machine learning is a great health technology that poses some very notable environmental implications. In the first instance, the large computational resources needed for the training and deployment of machine learning models, especially deep-learning algorithms, present a major first interaction with the environment. These computational demands often translate into an increased energy signature in data centers where these models are processed and maintained. This consists of the adoption of

energy-efficient hardware, cooling system optimization, and powering data centers with renewable energy sources to bring down their overall carbon footprint. Another important issue related to the environment is electronic waste emanating from the obsolescent computing hardware used in these different technologies. Reaching sustainable e-waste management through recycling programs and responsible disposal is important to reduce associated environmental impacts. Moreover, predictive models can massively support more sustainable healthcare practices by facilitating the early identification of patients at risk from a wide range of diseases, from diabetes to many others. It enables personalized treatment plans and preventive measures to be put in place, which might reduce the need for resource-intensive medical interventions and lower general healthcare-related emissions.

Moreover, the diffusion of telemedicine, facilitated by predictive modeling, eliminates the burden of traveling by patients and, as such, reduces the carbon dioxide emissions from travel. Optimizing resource allocation in health care institutions by true risk assessments, predictive models increase efficiency and decrease medical supply chain wastage. Such alliances among actors in the healthcare sector, technology developers, and environmental initiatives can further multiply benefits by advocating works on sustainability, promoting renewable energy works, and urging health care toward environmentally responsible practices.

Therefore, while predictive modeling for early diabetes detection is primarily attempted to improve health outcomes, the environmental impact that crops up therein presents the inclusion of sustainable practices in healthcare technology. Ensuring the attention is given to energy efficiency and e-waste management, besides developing sustainable models of healthcare delivery, would allow stakeholders to maximize the positive impacts of health and the environment these technologies can bring about in society.

5.3 Ethical Aspects

The Predictive modeling for early detection of diabetes using machine learning is subject to ethical considerations since it deals with very sensitive and personal information about health conditions that may cause serious impacts on people and society. Out of all the ethical concerns associated with such data, one on the very front would be data privacy and security. Machine learning models

demand large data sets, sometimes with personal health information. This shall be supplemented by robust data anonymization, encryption of data during transmission and storage, and compliance with strict regulations on the protection of personal data such as GDPR or HIPAA to protect patients' privacy.

It runs hand in glove with transparency and accountability while developing and deploying the model. What predictive models do, their limitations, and the potential for biases are to be given due attention by health-care providers and researchers, along with what the model's predictions mean for taking care of the patient. Algorithms should be put under very strict trials for the reduction of bias, especially regarding demographic groups or miss-diagnosis.

An important ethical consideration while making use of predictive modeling has to do with its ability to facilitate fair access and distribution of healthcare resources. Most importantly, these technologies need to be available for underserved populations, and predictive models must not perpetuate existing disparities in health care. Therefore, fairness in model development, deployment, and delivery of health care should be emphasized so that just outcomes for all—irrespective of their socioeconomic background or geographical location—are promoted.

It also includes continuous monitoring of the impact that these predictive models could have on clinical decision-making. Physicians should maintain independence in the interpretation of model predictions and exercise discernment in their incorporation into patient care plans, which should include administering individual patient preferences and the ethical standard of care. Of course, the most important part is: that predictive modeling in health care must result in continuous ethical dialogue and multidisciplinary cooperation to be able to sail through the ethical landscape. Stakeholder dialogue between patients, health providers, ethicists, and policymakers ensures that such technologies are developed and rolled out in ways that pursue the good life, promote welfare, and protect privacy while building human well-being

.

5.4 Sustainability Plan

The predictive modeling sustainability plan for the early detection of diabetes aims at having long-term viability, impact, and continuous improvement. Most important in the plan is scalable cloud-

based infrastructure, whereby the computing resources will be allocated on a dynamic basis. This will result in optimum cost efficiency and minimize environmental impact by doing away with high-power computing hardware to be on all the time.

Regular updates will include new and diversified datasets so that models retain their accuracy and applicability. This kind of continuous improvement will ensure the learning process of the predictive model with emerging trends and variations in patient data and its shaping of the robustness and generalizability of the model. Adherence to strong ethical standards, in the same light, will be followed to ensure the privacy and security of patients over their data, gaining the trust of the users and stakeholders. Such a project will, therefore, require involvement with a wide range of stakeholders, from healthcare providers and researchers to policymakers. We envision the development of an ecosystem culturally infusing knowledge sharing and collaboration in seamlessly bringing this predictive model into clinical practice. This will allow for broader-store development of standardized protocols and guidelines on how one works using this model in various healthcare settings. Implementation and user documentation will be developed, training schedules, and other materials that shall support the predictive model. These resources will facilitate the continuous empowerment of future researchers and practitioners by availing appropriate knowledge and skills to them in order to be in a good position to improve on our work. We will enhance capacity building and promote continuous education, which ensures that the relevance and impact created by the project on medical and technological evolving landscapes remains stable.

A feedback mechanism will also be put in place to sustain the plan. This will ensure that there is continuous monitoring and checking of the model's performance. There will be a regular collection of user feedback and performance metrics to iteratively drive improvements and adapt based on the needs of different groups of users. Applications in early diagnosis and management of diabetes will contribute greatly to improving patient outcomes and healthcare efficiency.

CHAPTER 6
Conclusion and Future Work

6.1 Summary of the Study

The present study was focused on exploiting machine learning techniques for the development of prediction models that could enable the early detection of diabetes. Further, the study was instigated by the critical requirement of identifying diabetes at an early stage, thereby allowing for timely interventions to improve patient outcomes. In this study, different machine learning algorithms—Logistic Regression, Decision Tree, Random Forest, SVM, KNN, ANN, XGBoost, and Naive Bayes—were implemented to build various prediction models and compared in terms of accuracy, precision, recall, and F1-score.

It was initiated by literature research and a review of methodologies for diabetes predictions and machine learning in healthcare applications. In this paper, multiple features related to the diagnosis of diabetes were used in a diversified dataset. The dataset was preprocessed, including normalization and handling missing value entries, to ensure high-quality input before training the models.

Each of these machine learning models was trained and evaluated with proper care upon splitting the data into test and train sets. Computation of performance metrics was done to evaluate the efficiency of every algorithm. Among all the machine learning algorithms, Random Forest and XGBoost showed higher accuracy since they had a great capacity for handling complex patterns and interactions within details. The result proved that these models can offer a potentially robust tool for the diagnosis of diabetes at an early stage in clinical settings.

It also used advanced computational resources and other tools for the implementation of neural networks, such as TensorFlow and Keras, and Jupyter Notebook and Google Colab for running the codes and training models. One found an assertion for the importance of data augmentation and transfer learning methods in order to improve model performance.

6.1 Conclusions

The research was oriented to make use of machine learning techniques in developing predictive models with the intention of early detection for diabetes. Determined by the critical necessity of identifying diabetes at an early stage, timely interventions could subsequently be carried out that could improve patient outcomes. In this regard, several machine learning algorithms—Logistic Regression, Decision Tree, Random Forest, SVM, KNN, ANN, XGBoost, Naive Bayes—have been implemented to create and compare models with respect to their accuracy, precision, recall, and F1-score.

Literature research and methodologies on diabetes prediction and machine learning applications in healthcare were places to start. Experimental Methods In this work, different experimental methods are used with a diverse dataset of various features relevant for the diagnosis of diabetes. The data preprocessing phase consisted of the handling of missing values and normalization for the preparation of high-quality input for model training.

All machine learning models were trained and evaluated with respect to a split of data for training and testing. Further, performance measures were computed in order to estimate the algorithms' performance. Random Forest and XGBoost models turned out to be most accurate, thus showing that they were more efficient with the complex patterns and interactions hidden within the data. According to the results, these models could turn out to be quite effective tools in clinical practice for the early detection of diabetes.

6.2 Implication for Further Study

This study in predictive modeling for the early detection of diabetes using machine learning has laid a good foundation, although it holds a lot of avenues for future research. The optimization and improvement of the used models are among the important aspects. While in this study, XGBoost and Random Forest performed well, further tuning of their hyper-parameters and advanced feature engineering, coupled with the integration of ensemble methods, can result in better performance. Also, increasing the dataset of patients with respect to demographics, previous medical conditions, and other factors related to different lifestyles would enhance the generalizability and strength of the models. Data augmentation techniques would further improve class imbalance and training

results; this can be realized through the generation of synthetic data. Further investigation into feature importance may involve the detection of some important variables that predominantly affect predictive accuracy, and this will enable the construction of more accurate and specific models. It could also be developed with new features from continuous glucose monitoring, additional genetic information, and more detailed dietary habits. Finally, models able to analyze in real time and longitudinally would set up the continual monitoring with early intervention needed to improve patient outcomes and move the field of diabetes prediction and management forward.

REFERENCES

References

[1] Wikipedia contributors, "Diabetes," *Wikipedia, The Free Encyclopedia*, 29-Jun-2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Diabetes&oldid=1231626058.

[2] "Doctor penguin," *Doctorpenguin.com*. [Online]. Available: https://doctorpenguin.com/newsletters. [Accessed: 30-Jun-2024].

[3] "GetStat," *GetStat*. [Online]. Available: https://getstat.site/?i. [Accessed: 30-Jun-2024].

[4] "Open journal systems," *Sersc.org*. [Online]. Available: http://sersc.org/journals/. [Accessed: 30-Jun-2024].

[5] E. Ceh-Varela, L. Maes, and S. R. Shakya, "Machine Learning Analysis of Factors Contributing to Diabetes Development," *Cloud Computing and Data Science*, pp. 157–182, 2024.

[6] S. M. Ganie and M. B. Malik, "An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators," *Healthcare Analytics*, vol. 2, no. 100092, p. 100092, 2022.

[7] L. Zhang, Y. Wang, M. Niu, C. Wang, and Z. Wang, "Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study," *Sci. Rep.*, vol. 10, no. 1, p. 4406, 2020.

[8] H. El Massari, Z. Sabouri, S. Mhammedi, and N. Gherabi, "Diabetes prediction using machine learning algorithms and ontology," *Journal of ICT Standardization*, vol. 10, no. 2, pp. 319–337, 2022.

[9] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *J. Big Data*, vol. 6, no. 1, 2019.

[10] U. Ahmed *et al.*, "Prediction of diabetes empowered with fused machine learning," *IEEE Access*, vol. 10, pp. 8529–8538, 2022.

[11] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in *Computer Vision and Machine Intelligence in Medical Image Analysis*, Singapore: Springer Singapore, 2020, pp. 113–125.

[12] A. K. Yadav, D. Sagar, and N. Rani, "A study on non-invasive diabetes causing variables and their covariance relationship in diabetes prediction using machine learning algorithms," in *Lecture Notes in Networks and Systems*, Singapore: Springer Nature Singapore, 2024, pp. 365–375.

[13] P. Rani, R. Lamba, R. K. Sachdeva, P. Bathla, and A. N. Aledaily, "Diabetes prediction using machine learning classification algorithms," in *2023 International Conference on Smart Computing and Application (ICSCA)*, 2023.

[14] A. M. Posonia, S. Vigneshwari, and D. J. Rani, "Machine Learning based Diabetes Prediction using Decision Tree J48," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, 2020.

[15] M. Woźniak, J. Szczotka, A. Sikora, and A. Zielonka, "Fuzzy logic type-2 intelligent moisture control system," *Expert Syst. Appl.*, vol. 238, no. 121581, p. 121581, 2024.

[16] N. Thomas, *A practical guide to diabetes mellitus*, 8th ed. New Delhi, India: Jaypee Brothers Medical, 2018.

[17] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 40–46, 2021.
[18] G. Scheiner, *Think like a pancreas (third edition): A practical guide to managing diabetes with insulin*. Cambridge, MA: Da Capo Lifelong, 2020.
[19] S. Thériault, "Early prediction of gestational diabetes: a practical model combining clinical and biochemical markers," *Clinical Chemistry and Laboratory Medicine (CCLM)*, vol. 54, pp. 509–518, 2016.
[20] M. Hassan, "Diabetes prediction in healthcare at early stage using machine learning approach," in *12th International conference on computing communication and networking technologies (ICCCNT)*, IEEE, 2021.
[21] L. Chaves and G. Marques, "Data mining techniques for early diagnosis of diabetes: a comparative study," *Applied Sciences*, vol. 11, no. 5, 2021.
[22] R. K. Bernstein, *Dr. Bernstein's diabetes solution: The complete guide to achieving normal blood sugars*. Little Brown and Company, 2016.

Appendix A

Course Outcomes, Complex Engineering Problems (EP), and Complex Engineering Activities (EA) Addressing

Title:

PREDICTIVE MODELING FOR EARLY DETECTION OF DIABETES USING MACHINE LEARNING TECHNIQUES

Student ID: [213-15-4413]

CO Description for FYDP

CO	CO Descriptions	PO
Phase -I		
CO1	Integrate recently gained and previously acquired knowledge to identify a **p** problem for the Final Year Design Project (FYDP)	**PO1**
CO2	Analyze different aspects of the goals in designing a solution for this FYDP	**PO2**
CO3	Explore diverse problem domains through a literature review, delineate the issues, and establish this goals for the FYDP	**PO4**
CO4	Perform economic evaluation and cost estimation and employ suitable project management procedures throughout the development life cycle of the FYDP	**PO11**
Phase -II		
CO5	Design and develop technical solutions and system components or processes that meet specified requirements, ensuring compliance with public health and safety standards, as well as considering cultural, socioeconomic, and environmental factors in this FYDP	**PO3**
CO6	Choose and apply appropriate methodologies, resources, and contemporary engineering and IT technologies to address complex engineering processes, encompassing prediction and modeling, while adhering to relevant constraints in this FYDP	**PO5**
CO7	Analyze societal, health, safety, legal, and cultural considerations, along with associated responsibilities, in the context of professional engineering practice and the resolution of this problem, employing logical reasoning guided by contextual understanding.	**PO6**
CO8	Comprehend and evaluate the enduring sustainability and impact of professional engineering endeavors in addressing intricate engineering challenges within social and environmental frameworks.	**PO7**
CO9	Implement ethical principles and adhere to professional standards and norms in this FYDP	**PO8**
CO10	Capable of operating proficiently both individually and as a team member or leader across diverse teams and interdisciplinary settings in this FYDP.	**PO9**

CO11	Proficiently communicate with the engineering community and broader society regarding complex engineering endeavors, including the ability to comprehend and generate comprehensive reports and design documentation, as well as provide and receive clear instructions throughout this FYDP.	**PO10**
CO12	Acknowledge the importance of self-directed and life-long learning within the evolving landscape of technology, and possess the readiness and capability to engage in lifelong learning endeavors.	**PO12**

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP), and Attainment of Complex Engineering Activities (EA)

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP):

SN	EP Definition	Attainment	CO	Justification (with Knowledge Profile)	References
1.	**EP1: Depth of Knowledge required**	Yes	CO1, CO2, CO3, CO5, CO6, CO7 and CO8	This project demonstrates **fundamental engineering (K3)** principles by employing machine learning algorithms such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Logistic Regression, Decision Tree, Random Forest, and XGBoost for predictive modeling tasks. The project further showcases **specialist knowledge (K4)** by applying advanced techniques in machine learning, including model training, and the evaluation of model performance metrics, to accurately predict the early onset of diabetes based on a diverse set of features.	**Page no:** [17-21] **Section:** [3.2] **Page no:** [1] **Section:** [1.1]
				The project applies **engineering practice & design (K5)** by the figure of the process of experiments. The project addresses **engineering practice & technology (K6)** by employing Various machine learning model	**Page no:** [17-21] **Section:** [3.2] **Page no:** [25-28]

					Section: [3.4]
				This project aligns with **K8 (Research Literature)** by synthesizing insights from recent studies to advance the early prediction of diabetes using machine learning. It demonstrates a comprehensive understanding of current methodologies in predictive modeling and leverages state-of-the-art techniques to enhance the accuracy and reliability of diabetes prediction..	**Page no:** [10-14] **Section:** [2.2, 2.3]
2.	**EP2: Range of Conflicting Requirements**	Yes	CO2, and CO7	This project addresses **EP-2** by recognizing the hurdles in the early prediction of diabetes, including the limitations of traditional methods and the complexities of integrating various machine learning algorithms. Through comparative analysis, it confronts challenges in identifying key predictive features and model performance, offering insights for refining predictive methodologies and improving the accuracy and efficiency of diabetes detection.	**Page no:** [15] **Section:** [2.4, 2.5]
3.	**EP3: Depth of analysis required**	Yes	CO2, and CO6	This project addresses **EP-3** by meticulously comparing experimental outcomes, highlighting Transfer Learning as the chosen significant solution for enhancing early prediction of diabetes amidst multiple potential machine learning approaches..	**Page no:** [31-45] **Section:** [4.2]

4.	**EP4: Familiarity of Issues**	Yes	CO8	This project's interdisciplinary approach extends beyond traditional boundaries, integrating machine learning into medical diagnostics for early detection of diabetes. By applying advanced computational techniques, it aims to advance healthcare practices, spend specifically addressing **EP-4** by enhancing disease prediction and management. This research endeavors to contribute significantly to public health initiatives through improved diagnostic accuracy and proactive healthcare interventions.	**Page no:** [10-15] **Section:** [2.2, 2.4]
5.	**EP5: Extends of application codes**	No	CO5	N/A	N/A
6.	**EP6: Extends of stakeholders involved and conflicting requirements**	No	CO8	N/A	N/A
7.	**EP7: Interdependence**	Yes	CO5	This project adopts a comprehensive approach that integrates diverse components spanning data collection, statistical analysis, and methodological frameworks. By addressing these aspects synergistically, it aims to provide a holistic solution to intricate challenges in medical diagnostics, emphasizing **EP-7.** This approach ensures that the research not only tackles high-level problems effectively but also contributes to advancing the field of healthcare diagnostics through rigorous and integrated methodologies.	**Page no:** [17-28] **Section:** [3.2, 3.3, 3.4]

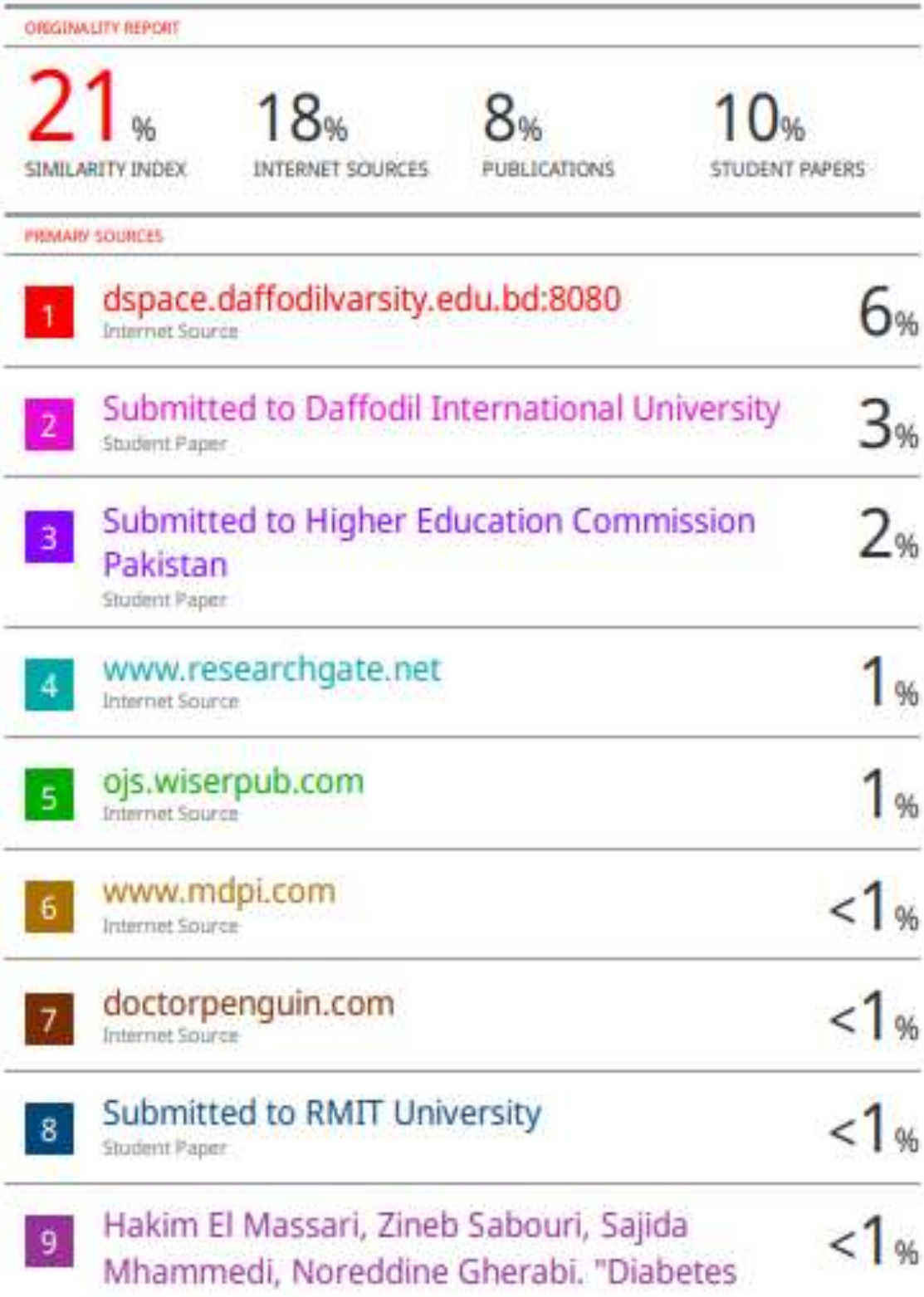
Addressing CO11 with Complex Engineering Activities (EA) [Some or all of the following]:

SN	EA Definition	Attainment	CO	Justification	References
1.	**EA1: Range of resources**	Yes	CO11	The objective of the study is to make an early diagnosis of diabetes using machine learning, wherein the researchers had 520 instances from Sylhet Diabetes Medical Hospital. The most important resources entail high-performance computing, software, and a dataset. Since accurate labeling was directly proportional to the quality of training, metrics used for the evaluation process were accuracy, precision, recall, F1-score, ROC curve, and AUC. Hyperparameter tuning was done by cross-validation. Ethical considerations include patient privacy and data security..	**Page no:** [29] **Section:** [3.5]
2.	**EA2: Level of interaction**	No		N/A	N/A
3.	**EA3: Innovation**	No		N/A	N/A

5.	**EA-5: Familiarity**	Yes		This project advances current research by exploring a pioneering method for early prediction of diabetes through machine learning techniques. It includes an exploration of foundational terminologies and an extensive comparative analysis, unveiling fresh perspectives in the domain.	**Page no:** [9-14] **Section:** [2.1, 2.3]

Addressing CO (4, 9, 10, and 12):

SN	COs	Attainment	Justification	References
1	CO4	Yes	This project addresses **CO4** by integrating effective project management and financial oversight, ensuring meticulous planning, resource allocation, and budget estimation for optimal resource utilization throughout the research lifecycle.	**Page no:** [5-6] **Section:** [1.6]
2	CO9	Yes	The project demonstrates adherence to ethical principles by prioritizing patient privacy, obtaining informed consent, and transparently documenting the research process, ensuring responsible knowledge dissemination and societal well-being through the ethical application of advanced healthcare technologies that comply with **CO9**.	**Page no:** [48-49] **Section:** [5.3]
3	CO10	No	N/A	N/A
4	CO12	Yes	The project's dedication to continuous learning **(CO12)** and adaptation within the dynamic technological landscape is reflected in its comprehensive data collection, rigorous statistical analysis, meticulous methodology development, and thorough experimental results and analysis, showcasing a commitment to staying updated and refining techniques to address modern challenges.	**Page no:** [17-29, 31-45] **Section:** [3.2, 3.3, 3.4, 3.5, 4.2]



ORIGINALITY REPORT

21%	18%	8%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	6%
2	Submitted to Daffodil International University Student Paper	3%
3	Submitted to Higher Education Commission Pakistan Student Paper	2%
4	www.researchgate.net Internet Source	1%
5	ojs.wiserpub.com Internet Source	1%
6	www.mdpi.com Internet Source	<1%
7	doctorpenguin.com Internet Source	<1%
8	Submitted to RMIT University Student Paper	<1%
9	Hakim El Massari, Zineb Sabouri, Sajida Mhammedi, Noreddine Gherabi. "Diabetes	<1%