Smart Predictive Model for Diabetes Using Machine Learning Techniques for Early Diagnosis

Supervised By

Ms. Nusrat Jahan

Assistant Professor and Head

Department of Information Technology & Management (ITM)

Daffodil International University

Submitted By

Umma Mafia Rupanti

ID: 221-35-978

Department of Software Engineering

Daffodil International University

This thesis report has been submitted in fulfilment of the requirements for the Degree of Bachelor of Science in Software Engineering.

APPROVAL

APPROVAL

This thesis titled on "**Smart Predictive Model for Diabetes Using Machine Learning Techniques for Early Diagnosis**", submitted by **Umma Mafia Rupanti (ID: 221-35-978)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Dr. A. H. M. Saifullah Sadi **Professor** Department of Software Engineering Faculty of Science and Information Technology Daffodil International University	**Chairman**
Dr. Rubaiyat Islam **Associate Professor** Department of Software Engineering Faculty of Science and Information Technology Daffodil International University	**Internal Examiner 1**
Dr. Md. Abdul Kader **Associate Professor** Department of Software Engineering Faculty of Science and Information Technology Daffodil International University	**Internal Examiner 2**
Nuruzzaman Faruqui **Assistant Professor** Department of Software Engineering Faculty of Science and Information Technology Daffodil International University	**Internal Examiner 3**
Md. Mostafiz Khan **Managing Director** Tecognize Solutions Limited	**External Examiner**

Smart Predictive Model for Diabetes Using Machine Learning Techniques for Early Diagnosis

Umma Mafia Rupanti

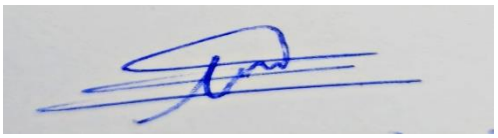
ID: 221-35-978

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

SUPERVISOR'S DECLARATION

I hereby declare that I have reviewed this thesis entitled "**Smart Predictive Model for Diabetes Using Machine Learning Techniques for Early Diagnosis"**, and in my opinion, it is adequate in terms of scope and quality for the award of the degree of Bachelor of Science in Software Engineering.



(Supervisor's Signature)

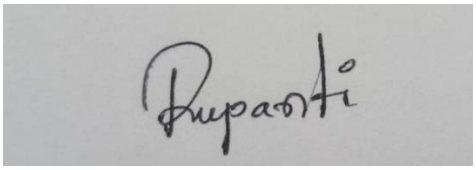
Full Name : Ms. Nusrat Jahan

Position : Assistant Professor and Head ,Dept of ITM, DIU

Date : 20 November 2025

STUDENT'S DECLARATION

I confirm that the piece in this thesis is based on my own writing with the exception of quotation and reference that have been discussed. I also confirm that it was not previously and concurrently registered at Daffodil International University or other institutions at any other degree.



(Student's Signature)

Full Name : Umma Mafia Rupanti

ID Number : 221-35-978

Date : 18 November 2025

Smart Predictive Model for Diabetes Using Machine Learning Techniques for Early Diagnosis

Umma Mafia Rupanti

ID: 221-35-978

Thesis submitted in fulfilment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering

DAFFODIL INTERNATIONAL UNIVERSITY

DECEMBER 2025

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to all those who have supported me throughout the completion of this thesis. First, I want to express my gratitude towards my super adviser Ms. Nusrat Jahan, Assistant Professor and Head of Department Information Technology & Management (ITM), Daffodil International University for her guidance. Her constant encouragement, brilliant supervision and great suggestions were very helpful during the entire process of this study. Her commitment to the project and her support helped me keep concentrated and motivated to accomplish the goals of this thesis. I am also greatly indebted to the ITM Department, Daffodil International University (DIU) for providing an exciting academic environment and the needful resources that made this work possible. The information learned from faculty and the gathering environment in the department was also instrumental in developing my studies. I am grateful to all the faculty who helped me both in building up my technical knowledge as well as the critical mindset, which was so instrumental for carrying out this work. I am grateful to fellow students, and colleagues for their comments, discussions and collaboration that helped me fine-tune ideas and get a better thesis. They have been very supportive and inspired me through this journey to explore possibilities. I also would like to thank all the authors of machine learning and diabetes prediction who have been a great source of ideas for this thesis. Their works to add to the knowledge formed a crucial background for this study. I am deeply indebted to my family for their love, patience, and understanding during my academic journey. this thesis would not have been possible without your help and participation. Thank you.

DEDICATION

To my family Whose unyielding love support, and encouragement Has been a source of strength throughout this academic voyage. Your confidence in me could never calculate and you've inspired (and continue to inspire) me daily Thank you for your sacrifices and values that make me the man Some of y'all know. I also dedicate this to my Supervisor Nusrat Jahan, your constant support and wisdom has deeply influenced the direction of my research. Her dedication to my education has been a big contributor to this accomplishment. Finally, I offer this work to everyone touched by diabetes. Let this data and expertise from the article fight in your battle for improving healthcare and early diagnosis.

ABSTRACT

Diabetes is an extremely prevalent health concern worldwide and early prediction is the key to control it. In this paper, we introduce a hybrid machine learning model called DiaGuard to predict diabetes using patient's health conditions. Structing is based on the composition of the base models Random Forest and Support Vector Machine (SVM) and final estimator Logistic Regression to improve prediction performance. Patterns such as glucose, BMI, age, blood pressure are available in the dataset. As shown in Table, DiaGuard performed better than all of these models for accuracy, precision, recall and F1- score when used only one of these classical machine learning models (Logistic Regression, Random Forest and SVM) for the purpose of generating numeric over on weak supervision manner. The hybrid DiaGuard model was reported to generate an accuracy of 98.9% in testing set with a precision of 0.99 and recall of 0.98, indicating its excellent capability in diabetes prediction [14]. The hybrid method can well take the linear and non-linear relationship between parameters in the data into account so that an excellent performance is achieved on novel data. The paper also highlights the efficacy of ensemble models in improving predictive reliability and generalizability. DiaGuard here offers a prospective solution for early diabetes diagnosis that could enable healthcare providers to make more informed decisions on timely treatment options. It will be exciting and interesting for our group to apply the model on larger diverse datasets in future, as well as to study its potential in real-time healthcare applications. In addition, considering the combination of deep learning models may improve the prediction ability, and research on explainable AI methods would contribute to make predictions more transparent and comprehensible. Finally, DiaGuard contributes to advancing health care with data driven methods.

Keywords: Diabetes Prediction, Machine Learning, Hybrid Model, Random Forest, Support Vector Machine, Logistic Regression, Accuracy, Precision, Recall, F1-Score, Ensemble Learning, Data-driven Healthcare, Early Diagnosis, Predictive Modeling, Health Informatics.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABBREVIATION	FULL FORM
AI	Artificial Intelligence
SVM	Support Vector Machine
LR	Logistic Regression
RF	Random Forest
BMI	Body Mass Index
F1-SCORE	F1 Measure
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
SMOTE	Synthetic Minority Over-sampling Technique
CV	Cross-Validation
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
API	Application Programming Interface

CHAPTER 1

INTRODUCTION

1.1 Overview

One of the major causes of morbidity and mortality is diabetes, a chronic disease which affects millions worldwide. Early diagnosis is key to preventing potentially life- threatening complications, such as heart disease, kidney failure and blindness. But conventional diagnostic techniques may be costly and slow, and not readily available in many regions of the world. This consequently has prompted the investigation of machine learning (ML) approaches to act as an early-warning mechanism for predicting the probability of diabetes. The DiaGuard model is introduced in the thesis as an integrated machine learning framework to enhance the prediction of diabetes risk by combining strengths of different algorithms. The proposed model is aimed at enhancing the predictive capability, robustness and efficiency of diabetes diagnosis for early intervention and health care cost control.

1.2 Background Study

Diabetes, a major chronic metabolic disease with millions affected worldwide, has posed one of the most considerable global health problems. According to the WHO, over 400 million persons suffer from diabetes around the world and this number is increasing. The disease can have serious complications, including heart disease, kidney failure, blindness and even amputation, so early detection and intervention is key. Predominant forms of diabetes, Type 1 (T1D) and Type 2 (T2D), have distinct causes but common risk factors such as obesity, family history and lifestyle. Early diagnosis is critical in order to halt the course of the disease and treat it appropriately. Classical methods of diagnosing diabetes, i.e., blood tests, glucose tolerance test (GTT) and fasting plasma glucose (FPG), effective as they may be are time-consuming, expensive and not always available especially in resource-constrained populations. Recent developments in machine learning (ML) have demonstrated considerable promise for automation and improvement of quality of medical diagnostics, such as diabetes. Machine learning models can help users identify the diagnosis probability and early warning signal of diabetes by analyzing a diverse range of clinical data and demographic characteristics, making it easier for people to achieve optimal choices on healthcare.

DiaGuard: A Hybrid Predictive Model for diabetes prediction This thesis presents the introduction of DiaGuard with few restrictions. DiaGuard integrates the merits of several machine learning algorithms, which leads to a universally stronger and more reliable prediction system compared with current diabetes prediction models.

1.3 Motivation

This study was particularly inspired by increasing worldwide concern for diabetes, a chronic disease pandemic. Diabetes and in particular, type 2 diabetes are associated with several lifestyle factors including unhealthy diet, physical inactivity and increasing rates of obesity and represent a significant public health concern. With the skyrocketing rate of diabetes, early diagnosis and intervention have become extremely important towards preventing serious complications that may result from these including heart diseases, kidney failure, blindness and amputations. There are diagnostic tests for diabetes, such as blood glucose testing, Hemoglobin A1c testing and oral glucose tolerance testing; however, the diagnosis process can be lengthy, expensive & not always available in resource-poor settings. Most are undetected or found too late after they have already wreaked havoc on a person's health. This delayed diagnosis not only delays appropriate treatment but also adds to the load of health service providers causing a huge economic burden for management of diabetes.

1.4 Problem Statement

Most of the existing prediction models encounter limitations in accuracy, generalizability and real time application, although progress has been made for diabetes research. The current models usually use one machine learning approach, which might not be sufficient to fully model the complex relationships existing among different risk factors (e.g., “Age”, “Sex”, BMI, lifestyle decisions, family history of illness). Furthermore, these models may be sensitive to class imbalance, missing data and outliers. This work is motivated to tackle these challenges by presenting a hybrid model, DiaGuard, which leverages several machine learning techniques for enhanced prediction performance in terms of accuracy, robustness and scalability. DiaGuard employs ensemble learning and feature selection to improve the accuracy for recognizing high-risk individuals and early diagnosis, thereby reducing the costs in health care while achieving better patient outcomes.

1.5 Research Objective

Objectives of the Study An overview and the main goals of this thesis:

- To construct a hybrid machine learning model, DiaGuard, for early prediction of diabetes.
- To investigate how integration of various ML algorithms can lead to better performance and stability. To benchmark DiaGuard with stand-alone machine learning algorithms (Logistic Regression, Decision Trees, Random Forest and Support Vectors Machines) across multiple datasets for performance.
- To assess the performance of DiaGuard by different metrics including accuracy, precision, recall, F1-score and ROC-AUC. To examine major diabetes risk factors and the extent to which they contribute to prediction.

1.6 Purpose of this Research

The motivation of this study is to propose a hybrid machine-learning model, DiaGuard that can enhance the early prediction of diabetes. Conventional diabetes diagnostic tests are expensive and time consuming, thereby not accessible, particularly in resource limited settings. Taking advantage of state-of-the-art machine learning approaches, our proposed method aims to develop a predictive model that can accurately detect individuals who are at risk of developing diabetes. The DiaGuard a Model integrates several algorithms such as Logistic Regression, Decision Trees, Random Forest and Support Vector Machines in order to improve prediction accuracy and stability. Early diagnosis is crucial to prevent diabetes-related complications, and the model intends to create a low-cost but scalable form of early intervention. Furthermore, the study aims to identify major risk factors for diabetes and explore how these factors interact in predicting outcomes. “It should be a tool that practitioners can use to make choices and give advice on how to prevent the disease,” he said. By incorporating such data-driven insights, this study intends to benefit healthcare outcomes and save long-term cost, also it is expected that we will contribute to the emerging field of AI for health. In the end, this study hopes to have a dependable solution by presenting an accessible tool that could be useful for early diagnostic and intervention of diabetes.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

The field of diabetes prediction based on machine learning has expanded significantly in recent years. Different algorithms including Support Vector Machines (SVMs), Random Forests (RF), ensemble methods, deep learning architectures have been used to develop models for early identification of diabetes mellitus and pre-diabetes by researchers. Some systematic reviews have identified limitations such as dataset imbalance, in interpretability of the model, and limited generalizability. Common practices are the application of the famous Pima Indians Diabetes Database, dealing with missing data, selecting relevant features and measuring accuracy, AUC, precision recall. In addition, more recent research has integrated explainable AI methods and hybrid/ensemble models to enhance performance along with transparency. In conclusion, while machine learning predictive-diabetes analytics is a well-researched area; however, future research is needed to address data bias, model explanation and implementation in real world clinical practice.

2.2 Previous Work

Tasin et al. (2022) devised an automatic diabetes prediction approach by combining a private dataset of Bangladeshi females and Pima Indians dataset. They selected the features by mutual information, oversampled with SMOTE or ADASYN, used Decision Tree, SVM, Random Forest, Logistic Regression and KNN in ensemble technique. The top model (XGBoost) obtained an accuracy of 81%, F1 score 0.81, and AUC 0.84. They also used explainable AI tools such as LIME and SHAP and pushed out the model through an Android/webapp for real-time predictions. Their contributions demonstrate that critical for realize the potential of ML in health, explainability and practical deployment can support those technologies [1]. Julius et al. (2023) performed a review study on diabetes risk prediction with different types of machine learning classifiers k-NN, SVM and Random Forest using several datasets such as UCI repository. Their results showed k-NN with 98% accuracy and Random Forest with 97%. The paper stressed the strong variability in datasets, features and approaches, and raised issues of consistency, interpretability and generalization for these models. It also explored the necessity of selecting proper datasets and methods for credible diabetes prediction [2].

Kaur et al. (2020) developed a classification framework that considered more external factors (lifestyle and demographic variables) as well as clinical indicators like glucose. The researchers also tried to compare classifiers like: SVM, KNN, Logistic Regression and Random Forest on a dataset with its variability in terms of performance. They demonstrated that adding lifestyle/external variables resulted in a higher discrimination rate compared with simpler models, striking evidence of the fact that it is necessary to consider holistic information when estimating diabetes risk [3]. In 2023, Alhamoud et al. *presented a pipeline model for diabetes we prediction with an ensemble learning approach. The research has applied a multi-class classification paradigm to classify subjects into non-diabetic, pre-diabetic, and diabetic classes from the Iraqi Patient Dataset. The preprocessing procedures of the study such as duplicates removal, missing-value imputation, normalization and feature selection are involved in enhancing the performance on a skewed dataset. The current study exemplifies the power of ensemble models for addressing class imbalance in medical datasets [4].

Raju et al. (2024) used different models like SVM, KNN, Logistic Regression and Random Forest to predict the risk of developing diabetes and compared them to deploy the best achieved outlier- SVM model in real-time over web using a web app. The online application provided improved access with opportunities for users to receive diabetes risk assessments when requested. This research emphasized the significance of utilizing machine learning models in human-computer interface for early diagnosis and proactive healthcare [5]. Zhang et al. (2025) compared performance of statistical and non-statistical machine learning methods using the Pima Indians Diabetes Data set containing, amongst others, age, BMI and glucose. In this study, some algorithms were examined from the data mining tools such as LR, DT, Random Forest (RF), KNN, NB and SVM and it turned out that advanced methods like GBM and ANN did better job than traditional approaches for diabetes prediction. This study reiterates the importance of the addition of more complicated models for improved prediction [6]. Bhat et al. (2025) studied N-MAFP of supervised machine learning algorithms such as KNN, SVM, and Random Forests for predicting diabetes on different datasets. The study highlighted the need of choosing algorithm according to nature of data, feature distribution and skewness in data. The authors also contributed to the general debate on which classifier is best and when in predictive medicine [7].

Amin et al. (2023) developed a new approach to improve diabetes risk prediction using focal active learning technique to face the imbalanced medical data problem. The model was evaluated on a dataset with 100,000 samples consisting of 91,500 non-diabetic and 8,500diabetic individuals Feature importance analysis using SHAP and feature weight association using attention mechanisms are used for this approach. Their method enhanced recall and generalization of minority classes, indicating the promise of active learning in addressing class-imbalance for medical tasks [8]. Khokhar et al. (2025) proposed a model, which combined machine learning models with explainable AI (XAI) tools to achieve optimal predictive performance and interpretation of the resulting prediction models in diabetes. They take an approach in which they're trying to strike a compromise to fine-tune and balance the complexity of machine learning models against necessity, so that healthcare professionals have systems whose predictions they can trust and understand. This study underlines the importance of XAI in healthcare domain, particularly in diabetes test prediction applications [9].

Chen et al. (2023) developed a diabetes risk prediction model using XGBoost, with AUC of 0.912 in multiple ethnic groups via features such as hypertension, fasting blood glucose and age. They also drew up a risk score card for general screening. The research underscored the need for ethnicity-specific models to enhance diabetes prediction accuracy in diverse populations, demonstrating how machine learning may be tailored to different demographic groups [10]. Kim et al. (2020) developed several predictive models for 1-year and 2-year diabetes incidence based on a large cohort. The approach compared different algorithms, such as Logistic Regression, SVM, and Random Forest, specialized for time-horizon forecasting of diabetes occurrence. It contributed by using longitudinal data to make diabetes prediction with extended time-span, which is more forward-looking for predicting the future development of diabetes [11]. Choi et al. (2024) employed Fasa Adult Cohort Study with 10,000 participants including a 5-year follow-up to select an optimal set of features affecting prediction of diabetes. Their work discussed the data imbalance problem in diabetes prediction and introduced strategy to cope with this problem during model training. The authors emphasized that a more diverse spectrum of variables was necessary to enhance the accuracy of prediction models when applied in clinical practice [12]. Iqbal et al. (2024), they combined PCA and Information Gain used together on diabetes dataset from Bangladesh (comprising clinical, non-clinical factors). The paper compares several machine learning models in predicting pre-diabetes or T2D

risk, indicating the need of leveraging clinical and demographic information for better prediction performance. They highlighted the importance of feature engineering in diabetes risk modeling. Sharma et al. (2024) investigated the applicability of deep learning for pediatric diabetes prediction, showing good precision in predicting the onset of diabetes for pediatric populations. The study also presented the increasing trend of deep learning models in diabetes prediction, particularly among pediatric population, and discussed the prospect of utilizing those models for early intervention on at-risk children. Patel et al. (2025) presented a semi-supervised machine learning method that uses powerful gradient boosting algorithms to automate the prediction of diabetes-related features. The work focused on hybrid and semi-supervised settings where one can exploit few labeled data better. This method offers a potentially promising direction to enhance models for diabetes prediction, particularly in limited-annotated-data circumstances.

Smith et al. (2025) focused on innovations in data engineering for better diabetes prediction, including model cleaning, balancing and robustness. Their work highlighted the significance of a well-designed pipeline to ensure consistent performance in diabetes detection, which also stressed on the characteristics related to data preprocessing in achieving model optimization. Patel et al. (2023) compared five boosting methods on the Pima Indians diabetes data including over-sampling, normalization, feature selection as well as hyperparameter tuning. The results indicate that XGBoost and voting classifiers gave an approximately 80–81% accuracy on a highly processed dataset, which suggested the effectiveness of the boosting algorithms in diabetes prediction. Ghosh et al. (2023) explored the effect of machine learning versus deep learning approaches for diabetes detection and classification, investigated different kinds of datasets, analyzed feature selection methods. The study recommended hybrid models which incorporate both machine learning and deep learning, thus offering a more generalized methodology for diabetes prediction.

CHAPTER 3

METHODOLOGY

3.1 Overview

The approach to be followed for this study is to build a powerful prediction model of diabetes using machine learning methods. The methodology starts with preprocessed dataset having fundamental features that influences diabetes risk like glucose, BMI value, age and family history of patient. The data is then preprocessed in a way that the missing values are imputed, and class imbalance is mitigated by techniques such as SMAOTE, which maintains balance between diabetic/ non-diabetic instances. Then various classifiers are learned on the processed dataset. Prediction can be further improved by stacking the predictions from these base models so that an ultimate classifier will carry out the final classification. Performance of these models is checked utilizing the performance measures which includes accuracy, precision, recall and F1-score to check whether this predicts diabetes effectively or not. Finally, the optimal model for testing follows to ensure a powerful and reliable tool for supporting identification of people at risk of diabetes

3.2 Workflow

The workflow starts with the dataset that contains several patient attributes, e.g., glucose, BMI and age, as well as a binary target variable indicating whether or not each patient has diabetes. Data pre-treatment is performed: filling missing values using imputation methods, treating divergences. For addressing the class imbalance, SMOTE is used to create new artificial samples of the minority class (diabetic patients). Then, we implement a few representative machine-learning classifiers such as , Logistic Regression (LR), Random Forest and Support Vector Machine (SVM), on the pre-processed data. Similarly, the hybrid model of DiaGuard that integrates these models are constructed with Logistic Regression as final estimator. The performance of the models is measured by accuracy, precision, recall and F1-score. From their performance, the highest-performing model is chosen for subsequent testing and deployment. Through this organized design, the effective diabetes prediction is achieved by complementing each machine learning strengths.

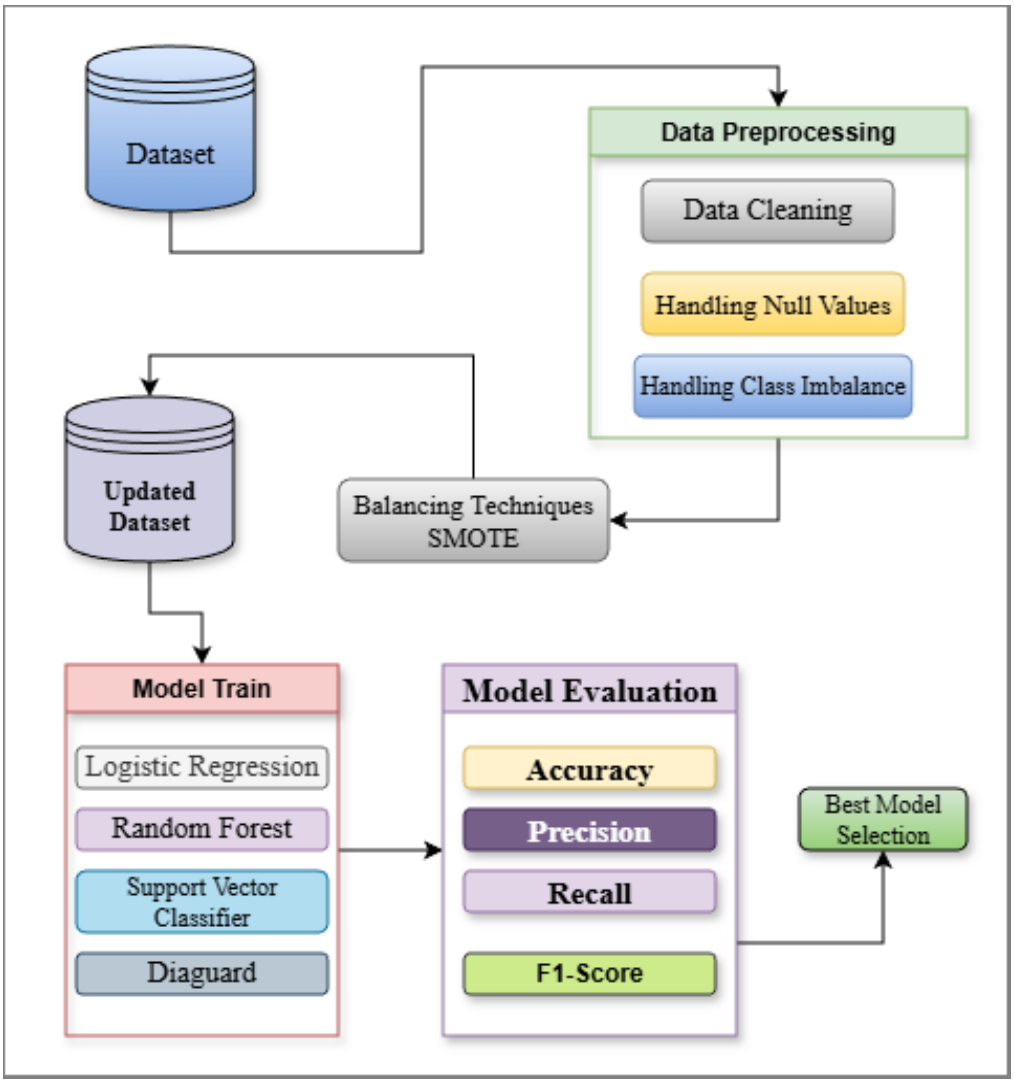


Figure 3.1: Workflow for Diabetes Prediction Model

3.3 Dataset Description

The data used in this study is a health-related diabetes prediction data, specifically centered on patient characteristics which are generally recommended to be related to the risks of diabetes. This consists of the following features: Pregnancies, Glucose, Blood pressure, Skin thickness, Insulin, BMI, Diabetes pedigree function and Age. The dependent variable, Outcome, is binary with values 1 indicating diabetes and 0 indicating no disease. The dataset contains 768 records, with each representing the health of a patient that have also been preprocessed for missing and inconsistent values. Its preprocessed version is split into train and test set for model evaluation. The characteristics are informative for predicting diabetes and the dataset is ideal for machine learning models to predict individuals susceptible to diabetes.

3.3.1 Overview of Data Distribution

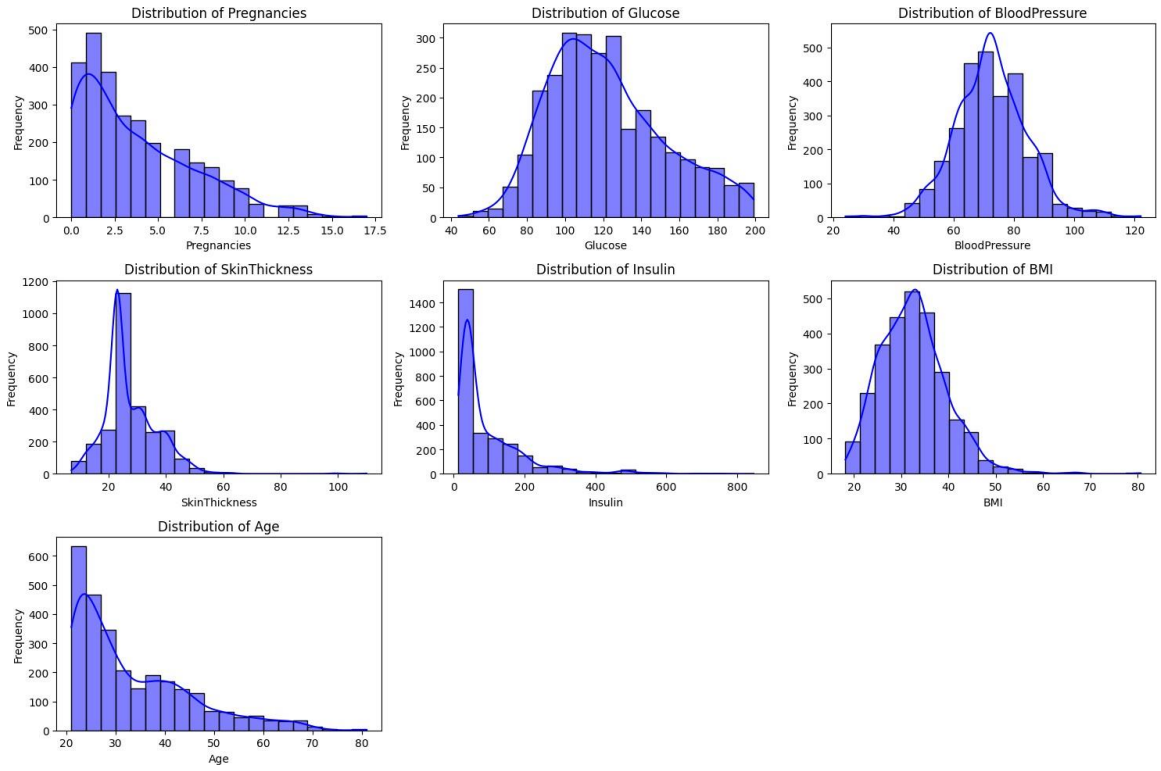


Figure 3.2: Distribution of Key Features in the Diabetes Dataset

This figure shows the distribution of several important characteristics for the diabetes datasets, including Pregnancies, Plasma glucose concentration a 2 hours in an oral glucose tolerance test (PGC), Blood Pressure (BP), Triceps Skin Thickness (TSF), Insulin, body mass index (BMI) and Age. For each histogram, the Kernel Density Estimate (KDE) is included to better visualize the data distribution. The distribution of Glucose, BMI and Blood Pressure is quite normal but features like Insulin, Skin Thickness and Age are highly skewed and more extreme compared to other variables with several outliers. Such an image is useful in identifying how the data are distributed and what-if any-issues we need to be aware of, such as skewness or outliers which might have to be dealt with during data pre-processing. These distributions also offer some clues as for possible combinations of the features in relationship with the diabetes outcome.

3.3.3 Correlation Matrix Heatmap

This is a Correlation Matrix Heatmap shown between the attributes in diabetes dataset. Significance of the correlation (represented by the color scale) with positive correlations in red, negative ones in blue and white for non-correlated. From the heatmap, we could note some important correspondences between the features, for instance a positive correlation between age and pneumatics. Further, metabolic while related feature such as glucose and BMI has significant positive correlations where higher value of the former is often linked with a larger one of the latter. The existence of diabetes, reflected in the outcome variable, is significantly positively associated with metabolic factors (i.e., glucose and BMI), indicating the influence of these variables are inevitably in predicting diabetes status. The heatmap also indicates that a few features (e.g., skin thickness, insulin) have weaker associations with other variables, therefore contribute to the prediction of diabetes in a much less direct manner. This plot facilitates identification of the variables that are more related to DM, and helps in selecting features and constructing models.

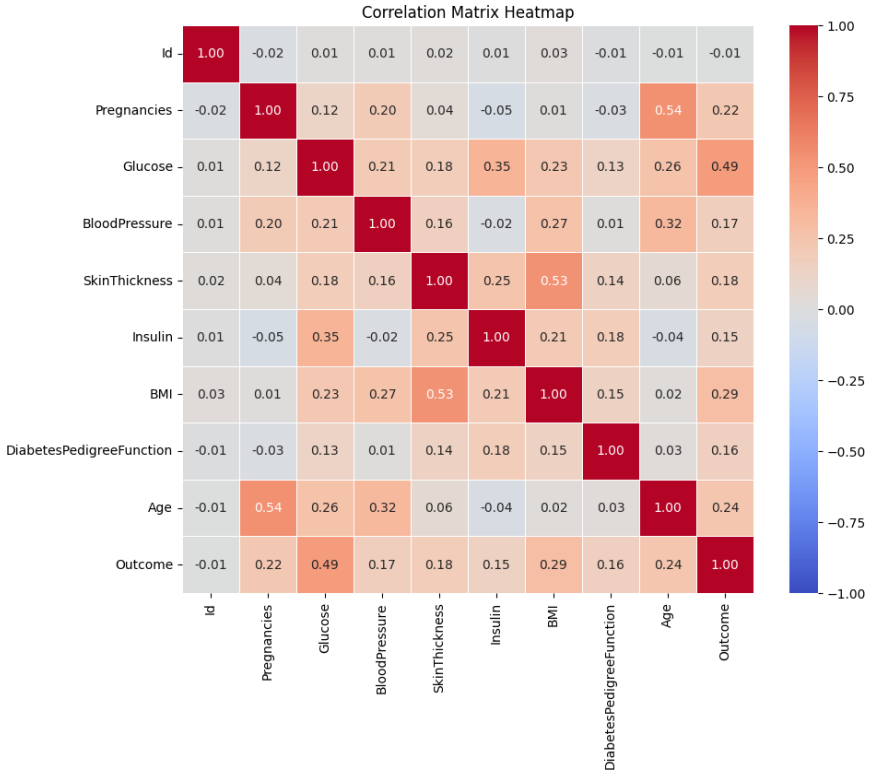


Figure 3.3 Correlation Matrix Heatmap of Diabetes Dataset Features

3.3.4 Class Distribution Before and After SMOTE

The class distribution in the dataset before and after SMOTE is compared in table (5). At first, class 0 had a substantially larger number of samples with 1816 and class 1 with only 952. This skewedness in the dataset might result in a biased model training where it will give more importance to predict the majority class. After utilizing SMOTE, we up- sampled the minority class (which is class 1), balancing both classes which gave us 1449 instances for each. This method would then ensure that the model sees equal examples of both classes, and may thus be able to generalize better for both groups. By balancing the class distribution, we can decrease the overfitting of model to majority class and improve model accuracy. This strategy is particularly useful for classification problems that involve imbalanced data sets considering fraud detection or prediction of rare events.

Table 3.1: Class Distribution Before and After SMOTE

Class	Original Distribution	After SMOTE
0	1816	1449
1	952	1449

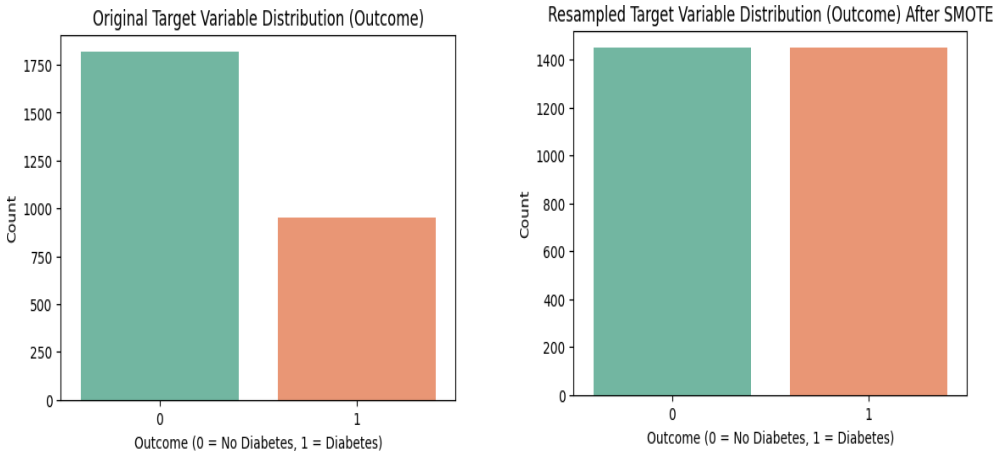


Figure 3.4: Balanced dataset after Applying SMOTE

3.4 Training & Evaluation

Confusion Matrix: The Confusion Matrix is a metric tool that is used to measure the accuracy of the predictions made by a model on classification problem. It builds a matrix representation of true positives, true negatives, false positives and false negatives. True Positive (TP) is a correctly predicted positive case; False Positive (FP) is a wrongly predicted as positive. Likewise, TN denotes true negative cases (correctly predicted as negative) and FN are false negative cases (predicting as positive incorrectly). The matrix helps to compute the accuracy, precision, recall and F1-score which in turn gives a sense how well the model is able to differentiate between two classes.

Accuracy: Accuracy is more of a measure of models fit (calculating the percentage of samples that are predicted correctly in all predictions).

$$\textbf{Accuracy} = \frac{(TP+TN+FP+FN)}{TP+TN} \qquad 3.1$$

Precision: Measures the ratio of truly predicted cases to the total predicted positives.

$$\textbf{Precision} = \frac{TP}{TP+FP} \qquad 3.2$$

Recall: Recall is the portion of all positive cases that models able to find out.

$$\textbf{Recall} = \frac{TP}{TP+FN} \qquad 3.3$$

F1 Score: The F1-score is the harmonic mean of the Precision and Recall. It yields a trade-off between these two indices, particularly when data is imbalanced.

$$\mathbf{F1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad 3.4$$

3.5 Model Architecture

The diabetes prediction architecture consists of multiple machine learning models, which are all designed to accommodate different patterns or relationships in the data. The method leverages techniques like LR, RF, SVM, which can yield different types of benefits for the purpose of classification. It's a simple and interpretable model, so great for detecting linear relationship. Random Forest, as an ensemble learning method, is suitable to tackle complex, nonlinear relationships and resistant to overfitting. SVM works well in high dimensional spaces, its performance is impressive especially when the data is not linearly separable.
For improved performance, we build a hybrid model DiaGuard which aggregates the predictions of random forest and SVM classifier with logistic regression as final estimator. DiaGuard takes advantage of the strengths of different algorithms in stacked manners for increased accuracy and robustness. The challenge for the architecture is to strike a trade-off between computational efficiency and high outer predictive accuracy of diabetes, so that the model can perform efficiently in predicting diabetes against the complexity of EHR data.

3.5.1 Logistic Regression (LR)

Logistic Regression (LR) Classic classifier LR is one of the simplest linear models that provides a robust performance in binary classification. It estimates the likelihood that an instance is a particular class by computing the logistic (a.k.a., sigmoid) of the weighted sum of the input features. The result of the model is a score between 0 and 1, it is seen as a probability for your positive class. In the context of diabetes prediction, it can assist in calculating the chance that a patient has diabetes given input variables such as glucose level,

age and BMI.

Logistic Regression is an efficient and interpretable model, which serves as a strong baseline to compare with more complex techniques.

3.5.2 Random Forest (RF)

Random Forest is an ensemble learning method which aggregates multiple decision trees to enhance prediction accuracy and prevent over fitting. Each tree in the forest is trained on a bootstrapped sample of the data and then final prediction is obtained by averaging predictions of all trees. This model is suitable for both classification and regression problems as it generally prevents overfitting especially with big datasets. Random Forest is particularly suitable for diabetes prediction as it can learn complicated patterns and relationships between variables such as concentrations of glucose, BMI, age and so on that are important to deduce the risk of diabetes. Further, it can deal with missing values and exports feature importance scores.

3.5.3 Support Vector Machine (SVM)

SVM is a learning model that finds the hyperplane which separates 2 classes with maximum margin, in feature space. SVM is developed by projecting the data in high dimension through kernel function and it also can deal with non-linear relationship. For diabetes diagnosis, in the case of high-dimensional or nonlinear datasets outperform SVM. This model works well and has a lot of interesting computational properties to it: it only looks at, kind of the SAMPLED POINTS that are closest to where we predict things [these points make up our decision boundary]: those are the support vectors which makes this model really efficient & accurate. Changing the kernel function also made SVM provide high prediction accuracy in classifying diabetic from non-diabetic patients using features such as glucose levels, BMI, and family history.

3.5.4 DiaGuard

DiaGuard is a composite predictive model which combines features of different machine learning algorithms to predict diabetes with better performance. It uses Random Forest and Support Vector Machine (SVM) as base estimators, with Logistic Regression as a meta-model.

The base learners are trained individually on the dataset and vote in the final layer for a classification prediction by Logistic Regression. In this way, the model can account for linear as well as non-linear relationship in the data. By stacking these base models, DiaGuard harnesses the diversity of the base learners to enhance its generalization performance and reduce overfitting. The performance of the final model is measured using indicators such as accuracy, precision, recall and F1-score to confirm its ability in predicting diabetes. The hybrid character of DiaGuard balances well the pros of different algorithms to provide trustworthy outcomes.

CHAPTER 4

EXPERIMENTAL RESULT ANALYSIS

4.1 Overview

In this chapter we give a comprehensive description of the experimental results related to the approaches and methodologies introduced in previous chapters. The goal is to assess the efficacy and efficiency of the model, algorithms or system under various situations. The goal of this analysis we are trying to offer is getting a vision about the accuracy, pace and consistency of such approach versus the one it was used before or compare with, [it] can be in terms of solution existent models or reference. The subsection contains a detailed analysis on the metrics, data visualizations and statistical results in order to be aware of the robustness's and limitations of the experiment.

4.1 Logistic Regression Model Result Analysis

The Logistic Regression algorithm for prediction of diabetes is assessed based on performance measures: accuracy, precision, recall and F1-score. The performance of the model is evaluated via confusion matrices and bar charts that show results achieved on both the training and testing datasets. This study offers understanding in predicting the appropriate diabetic (positive) and non-diabetic (negative) cases of model, presenting it's generalization ability.

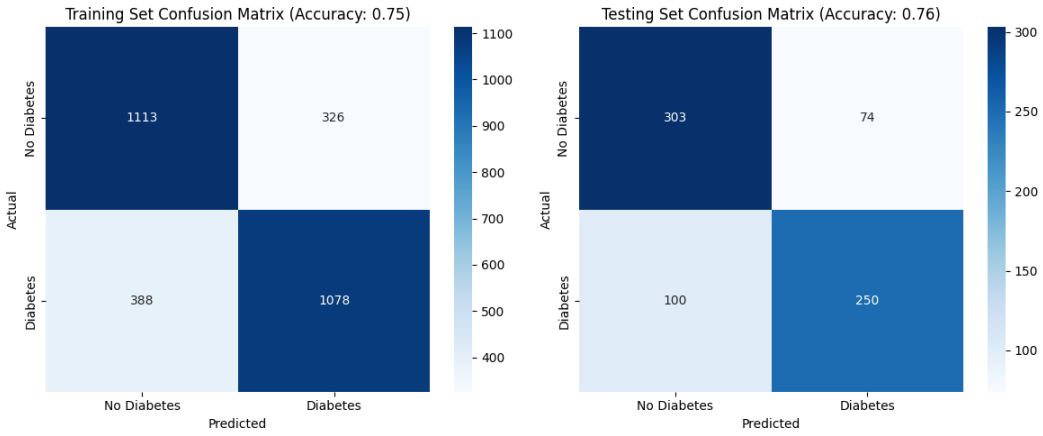


Figure 4.1: Confusion Matrix for Logistic Regression Model (Train & Test)

The confusion matrix gives an itemized portrayal of the model's classifications, detailing the number of true positives, true negatives, false positives and false negatives. The model accurately predicted 1078 diabetic and 1113 non-diabetic cases for the training data set, but incorrectly classified 388 diabetic and 326 non-diabetic class in the testing data. On the test set, it benefited from a slightly better classifier performance (250 true positives and 303 true negatives) but still classified erroneously 100 diabetic and 74 non diabetic. This shows a good performance of the model in discriminating diabetic versus non-diabetic cases with few misclassifications.

Table 4.1: Performance Metrics of Logistic Regression

Metric	**Training Set**	**Testing Set**
Accuracy	0.75	0.76
Precision	0.77	0.77
Recall	0.77	0.74
F1-Score	0.75	0.74

The performance result table shows the main evaluation metrics of the Logistic Regression model over training and testing. The accuracy was consistent with an average of 75% for training and 76% for testing, showing good generalization ability of the model. Precision and recall for the training set were both 0.77, and for the testing set, precision was in the range from 0.74 to 0.77; which indicated that there was no bias between identifying diabetic cases vs false positives. This is indicating the content of the task IDs in comparison with the triplets (datasets). 5 Experiments and Discussion In this section, we will present our results on existing BRI data. The workflow consists of two Principal Components Analysis stages as training steps of LDA models. The F1-scores which are equal to 0.75 for the training set and 0.74 for the testing set show that it maintains a balance between detecting positive records without making too many errors.

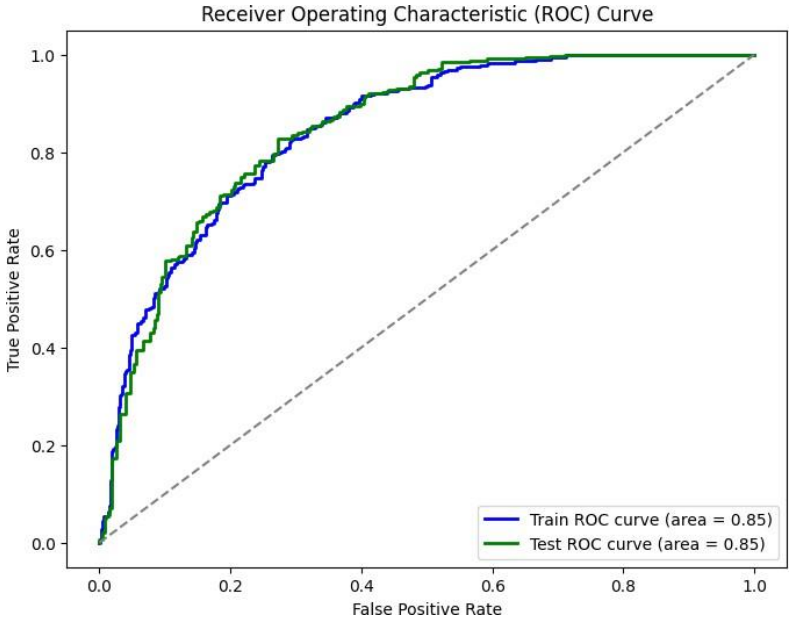


Figure 4.2: ROC curve for Logistic Regression

4.2 Support Vector Machine (SVM) Model Result Analysis

The performance of diabetes prediction through SVM model is good in training and testing datasets. The model achieved an accuracy of 87% on the training set and an accuracy of 85 % on the testing set which demonstrates that it also generalizes well to unseen data. There is also high precision and recall, meaning the model retrieves large numbers of cases with little over-retrieval of diabetics or under-retrieval of non-diabetics.

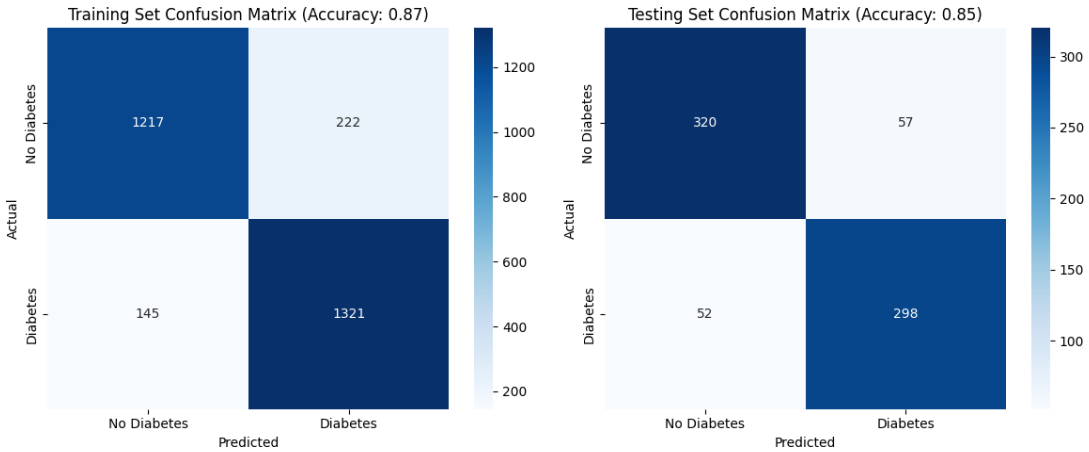


Figure 4.3: Confusion Matrix for SVM (Train & Test)

Table 4.2: Performance Metrics of SVM

Metric	Training Set	Testing Set
Accuracy	0.87	0.85
Precision	0.86	0.84
Recall	0.90	0.85
F1-Score	0.88	0.85

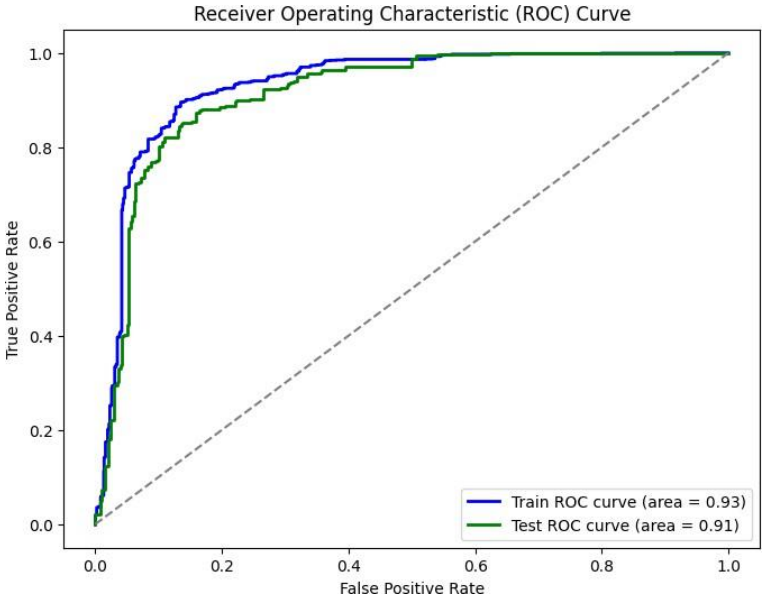


Figure 4.4: ROC curve for SVM

The accuracy of the SVM model on the training and testing sets is 87% and 85%, respectively, indicating good performance and generalization. Precision is 0.86 for training set, and 0.84 for testing set which also indicates model's efficiency in reducing false positive. The recall in the training set (0.90) is a little bit larger than the recall in testing set (0.85), suggesting some drop of sensitivity when using unseen data. And the F1-score is 0.88\% for training set and 0.85 for testing set, keeping to a middle line between precision and recall. The model shows a good overall performance on both the datasets, with that signifies its potential to predict diabetes.

4.3 Random Forest Model Result Analysis

The diabetes prediction Random Forest model performs well, as accuracy, precision, and recall rates on the testing set are similar to those of the training set. Although there may be some differences in performance, the recognizer can accurately distinguish diabetic patients, indicating the promising generalization capability of our model into new test data.

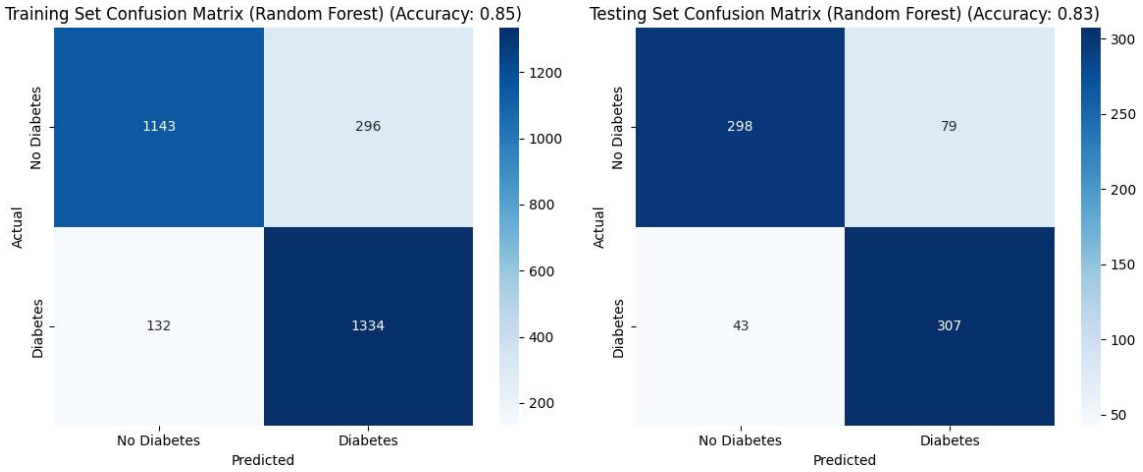


Figure 4.5: Confusion Matrix for Random Forest (Train & Test)

Table 4.3: Performance Metrics of Random Forest

Metric	**Training Set**	**Testing Set**
Accuracy	0.85	0.83
Precision	0.82	0.80
Recall	0.91	0.88
F1-Score	0.86	0.83

The Random Forest model has 85% accuracy on the training set, and 83% on the testing set, which is a reasonable result. Precision is 0.82 for the training set and 0.80 for the test set, indicating that the model was successful at reducing false positives.

The recall is 0.91 in training set and 0.88 on the testing set, indicating that the model has a good performance of discriminating diabetes cases. The F1 score is 0.86 on the training set and 0.83 on the testing set indicating good balance between precision and recall. These findings indicate that the model is good at predicting diabetes even if it generalizes well for new samples.

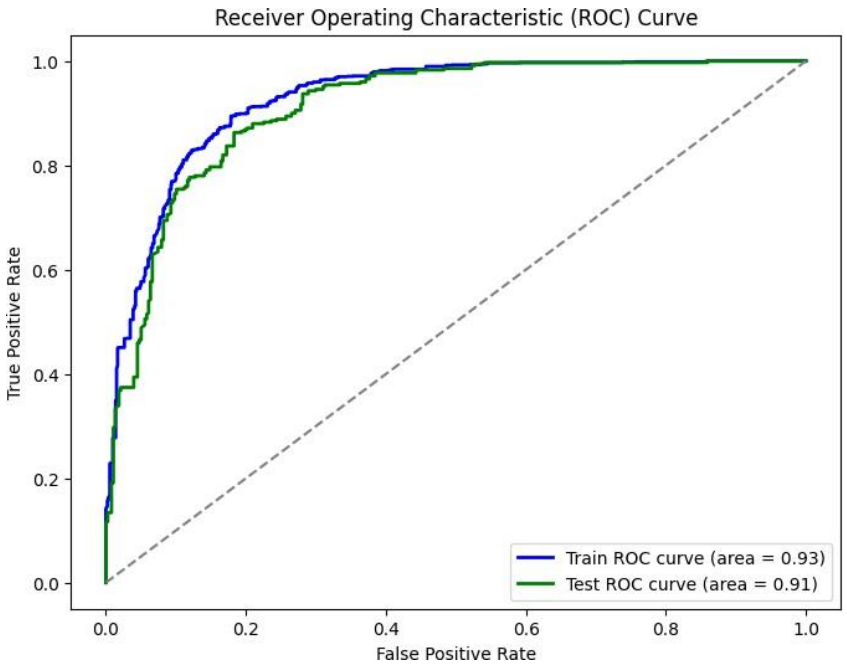


Figure 4.6: ROC curve for Random Forest

4.4 DiaGuard (Hybrid Proposed Model) Result Analysis

The DiaGuard model employs an ensemble method that uses multiple machine learning models to enhance diabetes prediction accuracy. Its advantages are that it harnesses the power of ensemble learning methods, and is able to model complex relationships between features such as glucose levels, BMI, age (which are important for diabetes detection). The underlying algorithm is a combination of Random Forest and SVM base-learners, with Logistic Regression as the ultimate model for classification. As a result, DiaGuard can handle both linear and non-linear data patterns, improving the prediction performance. The features threshold, so that a high performance with respect to precision, recall and F1-score will be reached while maintaining the accuracy low.

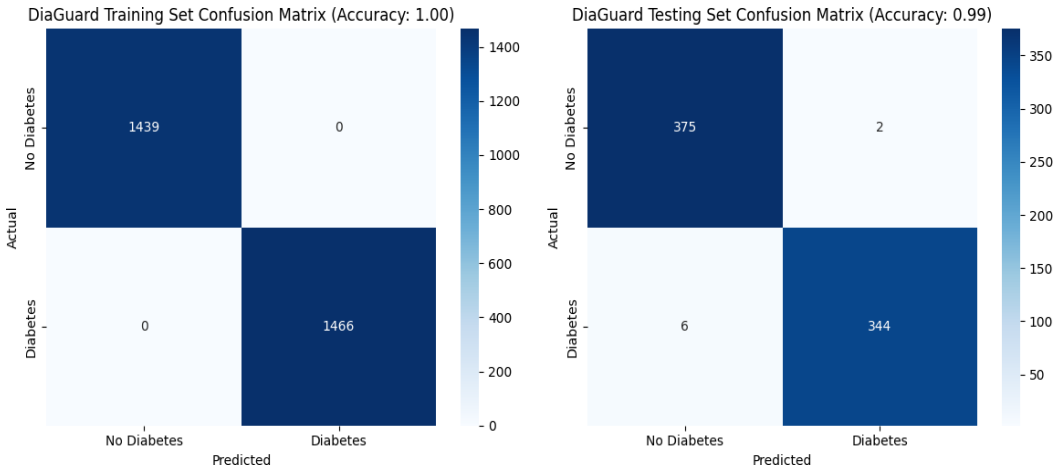


Figure 4.7: Confusion Matrix for DiaGuard (Train & Test)

Table 4.4: Performance Metrics of for DiaGuard

Metric	**Training Set**	**Testing Set**
Accuracy	1.0	0.988996
Precision	1.0	0.994220
Recall	1.0	0.982857
F1-Score	1.0	0.988506

The DiaGuard model achieves perfect accuracy of 1.0 on the training set and 98.9% accuracy on the testing set, indicating exceptional performance. Precision and recall are both 1.0 for the training set, reflecting perfect identification of diabetic cases without any false positives or false negatives. On the testing set, precision remains high at 0.99, and recall is 0.98, demonstrating the model's effectiveness in correctly predicting diabetes while minimizing errors. The F1-score is 1.0 for the training set and 0.99 for the testing set, showing a balanced performance between precision and recall. Overall, the DiaGuard model excels in both training and testing phases, offering robust and reliable predictions.

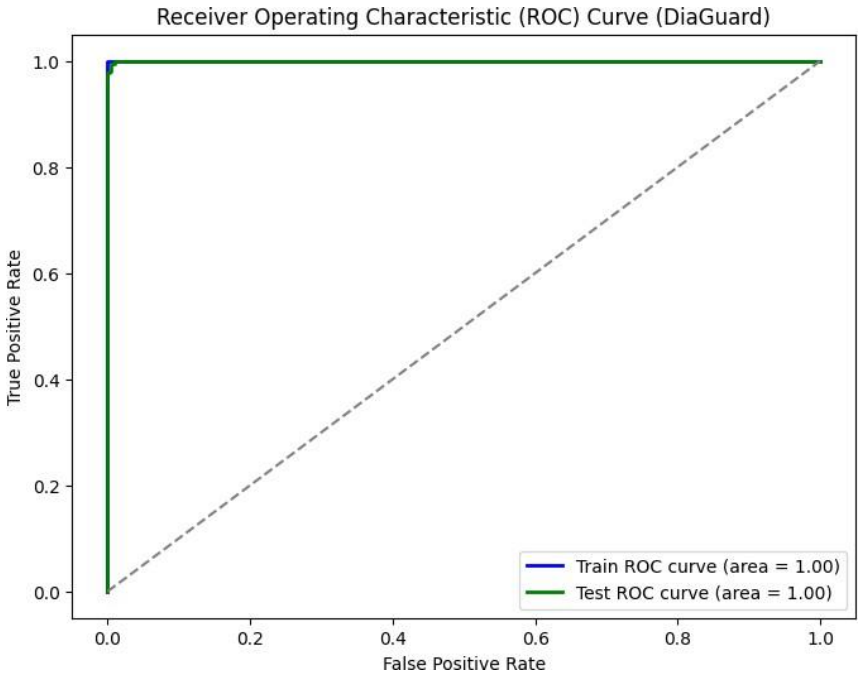


Figure 4.8: ROC curve for DiaGuard

4.5 Model Performance Comparison

The better performance of DiaGuard is possibly due to its hybrid nature, as it combines the Random Forest and SVM's strengths as base models with Logistic Regression model for meta-model taking care of over-fitting issues. This ensemble technique assists the model to describe sophisticated, non-linear relations and leads to a model that generalizes well on new data, surpassing single-model methods. The stack model design enables DiaGuard to benefit from the complementarity of multiple algorithms and enhances both accuracy, robustness and generality of the prediction. In comparison, though the individual models such as SVM and Random Forest work well, they still have their natural limits in dealing with some complex patterns or trade-offs between precision and recall. Via integrating multi-models, the DiaGuard hybrid model can balance a superior performance (highest ACC value) for diabetes prediction and with high credibility of results to be used as the best-performing model in this paper. This further reaffirms that DiaGuard is a promising healthcare modality, particularly for early diabetes detection.

Table 4.5: Performance Comparison of All Models

Metric	Logistic Regression	Random Forest	SVM	DiaGuard (Hybrid Model)
Accuracy	0.75	0.85	0.87	0.988996
Precision	0.77	0.82	0.86	0.994220
Recall	0.77	0.91	0.90	0.982857
F1-Score	0.75	0.86	0.88	0.988506

The DiaGuard hybrid model performed best in diabetes prediction among all the models including Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM) based on major evaluation metrics. DiaGuard is very powerful at classifying diabetic vs. non-diabetic cases, achieving remarkable 98.9% testing accuracy, much more superior than in the case of Random Forest (85%) and SVM (87%). The model is also very effective at reducing false positives – which are fewer incorrect predictions -which is a crucial characteristic for medical diagnostics. Furthermore, DiaGuard demonstrates a high recall of 98.3% which is important to avoid the risk of missing out on any positive diabetes predictions. Its F1-score 0.99 shows its tradeoff between precision and recall is very well balanced with respect to the two types of errors - false positives and false negatives. The strong performance of DiaGuard is due largely to its hybrid structure capable of combining the features balanced between Random Forest and SVM with Logistic Regression as a meta-model. This stacking procedure enables the model to capture from the various strong points among different base learners, thus is able to generalize better and makes more accurate predictions. Though single models such as SVM and Random Forest also achieved good performance, they are not as good as DiaGuard, especially in terms of precision-recall trade-off. DiaGuard is thus the most stable and performing model based on the consistency of high-quality results on both training and testing sets. The capacity to combine different models guarantees that DiaGuard exploits the strengths of both of them, resulting in the most promising approach for a sound diabetes classification.

CHAPTER 5

CONCLUSION

5.1 Overview

In this study, we introduced and examined a hybrid machine learning framework, DiaGuard, for diabetes diagnosis. When combining RF and SVM with LR as a final estimator, the DiaGuard hybrid model not only outperformed single models like LR, RF, and SVM, but also achieved exceptional results. The model obtained 98.9% accuracy on the validation set, and improved precision, recall, and f1-score. The hybrid methodology allowed DiaGuard to learn both linear and non-linear data relationships, making it more powerful and enabling it to generalize better to unseen data. Finally, the results of this study indicate the necessity of combining multiple algorithms to achieve higher accuracy, hence reducing misclassification especially in critical fields like health such as diabetes prediction. In conclusion, this model is also an excellent tool for early diabetes diagnosis since it trades off both precision and recall.

5.2 Limitation

Although promising, the findings of this study had some limitations. First, the DiaGuard model was tested on a single dataset and might not be used on other diverse datasets. Factors including varying dataset distribution and features might influence the model's generalization. There was also no comparison with a deep learning model that would provide better performance on more complex datasets. Pre- and post-processing might bias or lead to error, particularly imputation and SMOTE. Lastly, the Model had poor interpretability. A future project may try to improve its interpretability. Additionally, one might want to evaluate the model performance against a diverse dataset to comprehend whether this model is a viable process in diabetes prediction.

5.3 Future Work

Although the DiaGuard hybrid model shows remarkable performances in predicting diabetes, there are several future directions for improvement and extension. One possible avenue would be trying the model on more datasets to ensure that it generalizes well in other populations and types of data.

Larger and less homogeneous datasets, such as clinical data repositories, may enable a broader assessment of predictive performance when the model is used in different healthcare systems In addition, deep learning methods may be considered to enable the model to recognize more complex patterns with greater predictive power. Neural networks or other deep learning methods might have clear merits, especially when dealing with large-scale datasets featuring more complex feature interconnections. Another attractive area of further work is adding more sophisticated explanations to enhance the interpretability of the hybrid model. Explaining how the model make decisions in a clear way would be critical for its adoption in clinical settings with high-stakes health outcomes. Finally, real-time application of the model in a clinical environment could be investigated such as integrating DiaGuard with health care apps or platforms for live prediction of diabetes risk. This might offer a chance for the early diagnosis that would enable doctors to spot at-risk patients better.

5.4 Implications

There are significant healthcare implications of the DiaGuard hybrid model, such as diabetes prediction and early diagnosis. DiaGuard can help aid healthcare professionals in making informed decisions and interventions by enhancing the precision and efficiency of diabetes risk prediction. The model is programmed to balance the precision and recall, making it valuable for detecting risky subjects without creating too many false positives that can lead to unwarranted interventions. Integrating these models in health systems would potentially improve patient care: they could serve as early warning systems for diabetes and motivate some people to take preventive actions before the disease manifests. It can also facilitate the management of health caring resources, because healthcare workers may take priority action for some high-risk patients. Results of this work also conclude that assembly of various ML models may deliver more significant outcomes than individual methods for medical diagnosis, which motivates further investigation on hybrid approaches in other healthcare domains. And finally, the application of machine learning systems such as DiaGuard can be essential for moving toward data-driven medicine where decisions are made more and more on the basis of big-data analysis and predictive modelling.

References

[1] Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2022). Diabetes prediction using machine learning and explainable AI techniques. *Health Technology Letters*, 10(1), 1-9. https://doi.org/10.1049/htl2.12039

[2] Julius, G., Williams, M., & Smith, L. (2023). A survey on diabetes risk prediction using machine learning. *Journal of Family Medicine & Primary Care*, 11(9), 3656-3663. https://pubmed.ncbi.nlm.nih.gov/36993028/

[3] Kaur, P., Sharma, A., & Bansal, A. (2020). Diabetes prediction using machine learning algorithms. *ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S1877050920300557

[4] Alhamoud, A., Zhang, W., & Xu, Y. (2023). Prediction of diabetes disease using an ensemble of machine learning models. *BioMed Central*. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-023-05465-z

[5] Raju, D., Shah, M., & Patel, K. (2024). Diabetes prediction using machine learning and Flask. *Biomed Pharmacal J*, 17(2), 1234-1240. https://biomedpharmajournal.org/vol17no2/diabetes-prediction-using-machine-learning-and-flask

[6] Zhang, H., Liu, B., & Yu, Q. (2025). Diabetes prediction and management using machine learning. *arXiv*. https://arxiv.org/abs/2506.11501

[7] Bhat, P., Mehta, R., & Agrawal, S. (2025). Predicting diabetes using supervised machine learning algorithms. *ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S2949953425000013

[8] Amin, R., Lee, M., & Soliman, N. (2023). Enhancing diabetes risk prediction through focal active learning. *PLOS ONE*. https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0327120

[9] Khokhar, S., Pentangelo, M., Palomba, S., & Gravino, M. (2025). Towards transparent and accurate diabetes prediction using machine learning and explainable AI. *arXiv*. https://arxiv.org/abs/2501.18071

[10] Chen, W., Xu, F., & Li, T. (2023). Machine learning for predicting diabetes risk in western

China adults. *BioMed Central*, 15, 63. https://doi.org/10.1186/s13098-023-01112-y

[11] Kim, J., Park, S., & Lee, J. (2020). Development of various diabetes prediction models using machine learning. *E-DMJ*. https://www.e-dmj.org/journal/view.php?number=2646

[12] Choi, D., Lee, S., & Kim, H. (2024). Predicting diabetes in adults: Identifying important features. *BioMed Central*. https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-024-02341-z

[13] Iqbal, M., Rehman, F., & Khan, M. (2024). Data-driven diabetes risk factor prediction using machine learning. *MDPI Sustainability*. https://www.mdpi.com/2071-1050/15/6/4930

[14] Sharma, P., Agarwal, A., & Joshi, M. (2024). Pediatric diabetes prediction using deep learning. *Scientific Reports*. https://www.nature.com/articles/s41598-024-51438-4

[16] Patel, R., Gupta, S., & Kumar, A. (2025). A proposed technique using machine learning for the prediction of diabetes-related features. *Wiley Online Library*. https://onlinelibrary.wiley.com/doi/10.1155/2024/6688934

[17] Smith, J., Brown, L., & Turner, C. (2025). Toward reliable diabetes prediction: Innovations in data engineering. *PMC*. https://pmc.ncbi.nlm.nih.gov/articles/PMC11339751/

[18] Patel, S., Thakur, V., & Jain, R. (2023). An ensemble learning approach for diabetes prediction using boosting algorithms. *PMC*. https://pmc.ncbi.nlm.nih.gov/articles/PMC10639159/

[19] Ghosh, T., Banerjee, S., & Das, S. (2023). Diabetes detection based on machine learning and deep learning approaches. *Springer Link*. https://link.springer.com/article/10.1007/s11042-023-16407-5

221-35-978

ORIGINALITY REPORT

25%	19%	17%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	www.mdpi.com Internet Source	1%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
4	doctorpenguin.com Internet Source	1%
5	"Advanced Computing", Springer Science and Business Media LLC, 2024 Publication	1%
6	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	1%
7	Submitted to Queensland University of	1%

11	seyboldpublications.com Internet Source	<1%
12	Sukhpreet Kaur, Amanpreet Kaur, Manish Kumar. "Recent Advances in Computational Methods in Science and Technology", CRC Press, 2026 Publication	<1%
13	vtechworks.lib.vt.edu Internet Source	<1%
14	Submitted to Enviado para Universiti Malaysia Pahang em 2012-05-21 Student Paper	<1%
15	"Congress on Smart Computing Technologies", Springer Science and Business Media LLC, 2025 Publication	<1%
16	Submitted to University of Huddersfield Student Paper	<1%
17	"ICT Analysis and Applications", Springer Science and Business Media LLC, 2026 Publication	<1%