A Proposed Hybrid Ensemble Model for Accurate Diabetes Prediction

Supervised By

Dr. Md. Fazla Elahe

Assistant Professor & Associate Head

Department of Software Engineering

Daffodil International University

Submitted By

Golam Mowla

ID:213-35-781

Department of Software Engineering

Daffodil International University

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

APPROVAL

APPROVAL

This thesis titled on "**A Proposed Hybrid Ensemble Model for Accurate Diabetes Prediction**", submitted by **Golam Mowla (ID: 213-35-781)** 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

Fazla Elahe

Dr. Md. Fazla Elahe
Assistant Professor & Associate Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Chairman

Marzia 13.09.25

Dr. Marzia Ahmed
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

13.09.2025

Dr. Shabnom Mustary
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

13.09.25

Mohammod Abul Kashem
Professor
Department of Computer Science and Engineering
Dhaka University of Engineering & Technology, Gazipur.

External Examiner

A Proposed Hybrid Ensemble Model for Accurate Diabetes Prediction

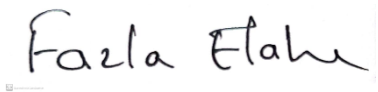
Golam Mowla
213-35-781

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

SUPERVISOR’s DECLARATION

I hereby declare that I have reviewed this thesis entitled "A Proposed Hybrid Ensemble Model for Accurate Diabetes Prediction ", 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)

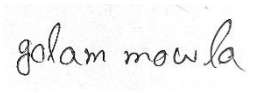
Full Name : Dr. Md. Fazla Elahe

Position : Assistant Professor & Associate Head

Date : 13 September 2025

STUDENT'S DECLARATION

I do assert that the writing done in this thesis is founded on my original work other than. It is quotations and citations that have been appropriately recognized. I also declare that it has not submitted hitherto or at the same time to any other degree at Daffodil. International University or any other university.



(Student's Signature)

Full Name : Golam Mowla

ID Number : 213-35-781

Date : 13 September 2025

A Proposed Hybrid Ensemble Model for Accurate Diabetes Prediction

Golam Mowla
213-35-781

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

Department of Software Engineering

DAFFODIL INTERNATIONAL UNIVERSITY

SEPTEMBER 2025

ACKNOWLEDGEMENTS

Acknowledgements I would like to thank my supervisor for all the useful advice and constructive criticism which helped me to do this research. I am also grateful to my teachers and academic role models for imparting the wisdom and motivation to finish this work. I would like to offer special thanks to my friends and friends whose valuable comments and advice greatly helped enhance the quality of the study. Lastly, I am indebted to the unconditionally support from my family.

Dedication

I cast my lot upon my Honorable Father and Mother, my supervisor, my Honorable teachers very expensive, and as thick as thieves. Without their forbearance, knowledge, unwearyingly strengthened, tender, affection and love it were impossible to reach this place.

ABSTRACT

One best example of chronic diseases and one of the highest causes of morbidity is diabetes whose early detection can help avert a catastrophic health risk. Early and precise detection of diabetes can lead to improved patient outcomes and facilitate health care providers with preventive care. In this paper Proposed Hybrid Ensemble Model (stacking ensemble meta-learner model for feature fried formulated on the basis of 2,768 patient records which contain 10 attributes related to some clinical characteristic are used for diabetes classification namely Pregnancies, In the pre-processing stage, missing values were imputed with the number of median and mode, imbalance between classes were handled by SMOTE and features were normalized. EDA through histograms, correlation heatmaps and distribution plots was carried out to preliminarily inspect co-feature interactions and how they affect diabetes end-results. The Proposed Hybrid Ensemble Model was constructed using the stacking of base learners such as Decision Tree and Random Forest, and a meta learner named Logistic Regression. The model was built by splitting the images into 80% train set and 20% test set and validated with 5-fold cross-validation the above experimental results show that the Proposed model is better than Traditional Machine Learning in performance, and the., The accuracy, precision, recall, F1score are 99.72%, 1.0000, 0.9945, 0.9972, the AUCs are 1.0000. Experiments demonstrate that Proposed Hybrid Ensemble Model has the best generalization capability and robustness compared with the baseline model's LR, DT, and RF diabetes prediction. This study suggests that the Proposed Hybrid Ensemble Model stacking ensemble model is an efficient and credible early detection tool for diabetes, which could assist clinical decision making.

Keywords: Diabetes Prediction; Machine Learning; Ensemble Learning; Stacking Classifier; Proposed Hybrid Ensemble Model; Decision Tree; Random Forest; Logistic Regression; Healthcare Data Analytics.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

μ	Average
σ	standard deviation
log	power of a number

LIST OF ABBREVIATIONS

SMOTE	Synthetic Minority Oversampling Technique
ROC	Receiver-operating characteristic curve
AUC	Area Under the Curve

CHAPTER 1

INTRODUCTION

1.1 Introduction

Diabetes mellitus is a chronic metabolic disorder, which has become one of the most serious global health WHO reports problems since the early 21st century. And that the prevalence of diabetes is rising swiftly, and millions of individuals are at risk of fatal complications such as cardiovascular disease, kidney disease and neuropathy. It is then significant that diabetes should be predicted early and correctly to avoid preventive treatment, immediate clinical intervention, and quality of life of the patient. More traditional statistical models such as Logistic Regression have been extensively applied to the prediction of diabetes, but they occasionally do not represent non-linear relationships that exist in healthcare. Moreover, the imbalance of the classes presents in medical datasets where the number of non-diabetic cases is high compared to diabetic cases deteriorates the performance of traditional models leading to biased predictions. Thus, it is necessary to have more advanced machine learning techniques that can address the imbalanced data that can learn the multifaceted interactions between features. In this study, we put forward the Proposed Hybrid Ensemble Model, a novel stacking ensemble model in the efficient classification of diabetes. This is a composite model of DT and Random Forest as base learners and Logistic Regression as a meta-learner. The missing values were imputed and a balance was created with class imbalance in order to enhance the quality of the data. Further development of associations between characteristics was carried out through statistical and graphical EDA.

1.2 Background Study

Diabetes mellitus is a chronic metabolic disease with abnormally high blood sugar levels, and there are many serious complications, including cardiovascular disease, renal failure and neuropathy, with early detection being essential. It has emerged as a major public

health challenge worldwide, increasing at an alarming rate in developed and developing countries. Precocious prediction of diabetes is necessary for the sake of risk reduction and patient benefit. Conventional statistical methods have been used in this field, yet they frequently fail to manage complicated medical-related information. With the recent development of machine learning and ensemble strategies, new solutions for disease prediction have been more precise and dependable.

1.3 Motivation

As the prevalence of diabetes continues to increase around the world, accurate prediction models are urgently required to inform early diagnosis and intervention strategies. Traditional approaches for feature selection tend to not account for the non-linear and complexity of healthcare data, which results in biased predictions, especially in imbalanced datasets. This constraint encouraged the application of modern machine learning methods for improving prediction performance. We design the proposed Proposed Hybrid Ensemble Model, a stacking ensemble-based method, to overcome these challenges by fusing classifiers to achieve better generalization. Building these models quickly and efficiently can substantially support clinicians' decision making.

1.4 Problem Statement

Diabetes prediction is difficult due to the complex health data, missing values, significant imbalance between two classes of diabetic and non-diabetic. Numerous conventional methods suffer from a low accuracy and tend to misclassify diabetic patients, --leading, in the worst case, to possible health damage when failing to detect a patient. Furthermore, coping with noisy data and determining the most affecting attributes are still problems. Such challenges call for a more solid, generalized statistical model that will also preserve

conditional accuracy. This paper tries to alleviate these problems by introducing its proposed framework, Proposed Hybrid Ensemble Model, as a more powerful stacking ensemble model.

1.5 Research Objective

The aim of this study is to establish and validate a model based on machine learning methods referred to as the "Proposed Hybrid Ensemble Model" to accurately predict diabetes. In this paper we developed a pre-processing procedure to impute missing values and balance the classes with the SMOTE procedure and to normalise the data to increase its quality. Another important aim is to perform EDA on extracting important features and understanding their association with diabetes outcomes. The study also seeks to develop a sophisticated stacking ensemble model by combining several base learners for higher prediction accuracy, robustness, and generalizing ability with respect to the classical models.

1.6 Scope of this Research

This study aims to establish a reliable and time-efficient ML tool for predicting diabetes based on the clinical records of the patients. For this study, the authors are confined to the given dataset. Pre-processing methods, such as missing values' handling, normalisation, and class balance exploitation via SMOTE, are used to enhance data quality. The study focuses on a Proposed Hybrid Ensemble Model, a stack ensemble and compares some baseline models. Although it is evidence of the success of the approach, the results are not yet achievable in real time and integrated into the hospitals systems, such as patient management and scheduling.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

In recent years, a number of studies have played with diabetes prediction in machine learning. Several other works used classification techniques and employed Decision Trees, Random Forests, Support Vector Machines, and Neural Networks to uncover hidden healthcare data. Although these models indeed showed promising performance, many encountered class imbalance or missing data and suffered from overfitting and therefore could not well generalize such good performance. To address the above limitations, some ensemble learning methods, such as boosting and bagging, are investigated to enhance the accuracy and reliability of the results. However, there is still a call for advanced ensemble techniques incorporating multiple models for more dependable diabetes prediction.

2.2 Related Work on Diabetic

Afolabi et al[1]. (2025) The authors applied logistic regression, random forest, and KNN to electronic health records for predicting diabetes and found significant risk factors associated with diabetes such as age, BMI, and levels of blood glucose. Of all the models, KNN had the best performance of 96.09% accuracy, 98.54% sensitivity and 93.63% specificity. The research demonstrates the promise of KNN for early and timely diabetes screening despite the restricted demographic information in the dataset. Ramani et al[2]. (2025) presented a big data architecture for diabetes prediction using MapReduce and an associative Kruskal-Wallis poly-kernel classifier (AKW-MRPK). The model obtained an accuracy of at most 92% and significantly shortened computational time, efficiently processing the large-scale clinical datasets. This study shows higher computational

performance than conventional and Hadoop-based machine learning methods, indicating that the proposed model can be effectively applied for the scalable, real-time medical use. Nomura et al[3]. (2021) explored the applications of artificial intelligence, such as machine learning and deep learning, in the treatment and prediction of diabetes. AI tools for retinal screening, diagnostic support, and patient self-management systems have received FDA approval and validated by peer-reviewed studies, but predictive models for incident (new-onset) diabetes are outperformed by traditional statistical approaches. With more extensive data sets and more powerful computational capabilities, we can look forward to greatly enhanced predictive accuracy in diabetes care based on AI. Tasin et al[4]. (2023) discussed an automated prediction of diabetes system by fusing Pima Indian and a private Bangladeshi dataset. Their most successful model, XGBoost with ADASYN, obtained 81% for accuracy, 0.81 for F1 score and 0.84 for AUC. The model was also operationalized into a website and android application, along with SHAP and LIME explainable AI methods for explain ability. Santhanam et al[5]. (2021) presented a data-driven model for diabetes prediction in terms of PIMA Indian diabetes dataset. Their model had an accuracy of 97%, which was superior compared to traditional techniques. The contribution highlights the possibility of ML-based techniques for the development of efficient, effective diagnostic tools for early diabetes detection. Larabi-Marie-Sainte et al[6]., 2019 examined state of the art studies of ML and DLfor diabetes prediction published in over six years. Considering rarely used ML classifiers on the Pima Indian Dataset, they achieved the accuracy 68%–74% (REPTree gave a good accuracy). The authors propose these less commonly used classifiers to be combined with other ML/DL models for improving prediction accuracy. Zhang et al[7]. (2024) developed a new evolutionary ensemble prediction model coupling the harmony search optimization and stacking for diabetes disease prediction. The architecture successfully integrates a number of classifiers to improve the prediction quality and robustness. Experimentally, better performance than traditional ensemble and single model approaches was shown and its potential for clinical decision support was emphasized. Febriana et al[8]., 2023 studied KNN and Naïve Bayes algorithms, in diagnosing diabetes based on Pima Indians Diabetes data set. It is a point worth noting that naïve Bayes performed better, with an

average accuracy of 76.07%, than KNN (73.33%) on different training-test splits. The authors report that Naïve Bayes achieves better performance for such a prediction task, and hence propose that different algorithms and optimizing techniques should be experimented in future work. Alam et al[9]., 2019 also trained a diabetes prediction model by ANN, RF , K-means on Pima Indian Diabetes dataset. Important features (BMI and glucose, blood pressure, age) were detected by PCA and abolishes discovered between BMI/glucose and diabetes by association rule mining. ANN achieved the best accuracy of 75.7% while RF and K-means yielded 74.7 and 73.6% of accuracy, respectively. Jacobsen et al[10]. (2025) summarized the activity of Type 1 Diabetes TrialNet, a group that has screened >250,000 relatives of individuals with T1D, conducted across >20 clinical and mechanistic trials, and influenced FDA approval of a treatment to stave off the onset of disease. Research by the consortium covers TN 1-3, with a number of trials preserving C-peptide preservation, one delaying progression to TN 3. Their discoveries progress the predication, prevention and mechanism understanding for T1D. Gowthami et al[11]. (2024) to evaluate the performance of machine learning techniques on early detection of T2DM. The study aimed to enhance diagnostic accuracy and assist in the implementation of precision medicine by employing classification models along with feature selection. The results suggest that our approach can facilitate early detection of T2DM and help to prevent complications and promote positive outcomes for individual patients. Oladimeji et al[12]. 2024 proposed classification models for early diabetes detection through symptom-based feature selection applied to a dataset of 520 patients. The work employed Random Forest, Naïve Bayes, J48 and KNN algorithms and was led to an accuracy of 98.3% by Random Forest. The findings emphasize the role of feature selection and data processing in the development of cost-effective and reliable diagnostic systems particularly in difficult-to-access settings. Salih et al[13]. 2021 Potential machine learning models for predicting diabetes on pima indian diabetes data proposed. In the work they used Decision Tree, Random Forest, Naïve Bayes, SVM classifiers and obtained the maximized accuracy was 98% by Random Forest. The findings indicate that the ML methods can be used to enable early diagnosis and offer promising prospects in healthcare. Kumar et al[14]., 2023

presented a diabetes prediction system based on machine learning techniques on the Pima Indian Diabetes dataset. The analysis was carried out on Logistic Regression, Random Forest, SVM, and KNN, producing the highest classification accuracy of 98% achieved by Random Forest. The results emphasize the utility of futuristic ML models for the early detection and improved control of diabetes.

Panda et al[15] (2024) examined the machine learning techniques that strive for early warning of the risk of diabetes based on health factors such as BMI, glucose, age and blood pressure. In this study, the models were tested by comparing four models such as random forest, logistic regression, SVM and KNN; the best model according to this process was random forest. The findings revealed the potential of ML-based method in early and clinical decision making. Karuppiah et al[16]., 2024 is a comparative study of numerous machine learning and deep learning models predicting diabetes on Pima Indian Diabetes dataset. The best accuracy (92%) has been attained with the algorithms experimented as Logistic Regression, KNN, SVM, Random Forest, and Multi-Layer Feed Forward Neural Network (MLFNN). The study highlights that deep learning rather than traditional ML can improve the likelihood of diagnosing diabetes and providing care to patients at an early stage.Abaldrada et al[17]. (2024), who predicted the probability of diabetes with a logistic regression model based on the Pima Indians Diabetes dataset. Using the model, the accuracy was 77.6%, and detection sensitivity and specificity were 72.4 and 79.6%, respectively, showing good diagnostic performance. Pregnancies, glucose, blood pressure, BMI, and diabetes pedigree function were major predictors, thus favoring an early detection and intervention by Shao et al (2024) develop and deploy a diabetes-prediction pipeline that addresses class imbalance using SMOTE+RUS, and tune a LightGBM model using Optuna over a 100k-record, 9-feature dataset. After 5-fold CV, their Optuna-LightGBM raises accuracy modestly (97.07%→97.11%) but precision substantially (97.17%→98.99%), reaches AUC 0.9801, and reduces search time to ~2.5 s. They point out that data balancing together with automated hyperparameter optimization helps generalise onto imbalanced diabetes data but clinical validation must be broadened. El-Bashbishy et al . (2024) propose a DNN/MLP model to predict

pediatrics diabetes using a new dataset (MUCHD; 548 patients, 18 features) from Mansoura University Children's Hospital. Their optimal setup (≈9–10 hidden layers, SGD, ReLU/Sigmoid, 50 epochs) leads to the confusion matrix TN=151, FP=0, FN=1, TP=396, which means accuracy 99.8%, precision 100%, recall 99.7%, specificity 100%. They find that this can achieve state of the art performance and suggest further testing on larger and more diverse cohorts. Sampath et al, 2024 introduces a diabetes-prediction pipeline on the Pima Indians dataset following mean imputation, IQR-based outlier removal, correlation-based features selection, and SMOTE in order to handle class imbalance. 17 They benchmark six different machine learning algorithms -KNN, DT, NB, RF, AdaBoost and XGBoost – out of which the hybrid ensemble (AdaBoost + XGBoost) yields best performance with k-fold cross validation (AUC 0.968 ± 0.015, accuracy 0.904 ± 0.023). The results indicate that preprocessing and boosting ensembles enhance minority-class detection and overall reliability of early diabetes screening. Jin et al., 2024propose a deep learning early-warning model based on CGM data for diabetes management, which is constructed on the temporal fusion transformer with multi-head attention. Evaluated on OhioT1DM and ShanghaiDM (124 subjects), it achieves RMSE ≈13.7 mg/dL (30 min) and 15.4 mg/dL (60 min), which is superior to the traditional ML.The Clinical Validation demonstrates 45% less severe glycemic events and an immediate, 58% faster time to response for severe events, which is robust real-world value.

CHAPTER 3

METHODOLOGY

3.1 Overview

The methodology of this study is such that it is intended to incrementally refine and assess the proposed Postpond Model for the prediction of diabetes. Dataset collection The process starts with collecting the datasets consisting of 2,768 patient records with relevant clinical features. Pre/preprocessing processes involving missing value imputation, normalization and class balancing using the SMOTE are utilized in an attempt to enhance data quality. Then EDA is performed, by using the visual and statistical methods to find out the feature relation. The underlying model is based on a stacking ensemble model, namely Postpond, consisting of several base learners to enhance accuracy and generalization. Lastly, the model is measured through performance calculation -accuracy, precision, recall, F1-score and AUC and compared with baseline models.

3.2. Workflow

The workflow to obtain a reliable diabetes predict is demonstrated as a methodology in the present study. Training and test subsets of the dataset containing 2,768 patient records and ten clinical attributes (except ID) were initially split by 80 and 20%. Missing values were properly handled by median imputation for numerical variables and mode imputation for categorical features. As there were more non-diabetic cases than diabetic ones in the dataset, the imbalance learning method (SMOTE) was used to over-sample the non-diabetic instances to ensure class balance. Normalization is used to rescale all features into the same range so that all features contribute equally to training. Then, EDA analysis was carried out on it.

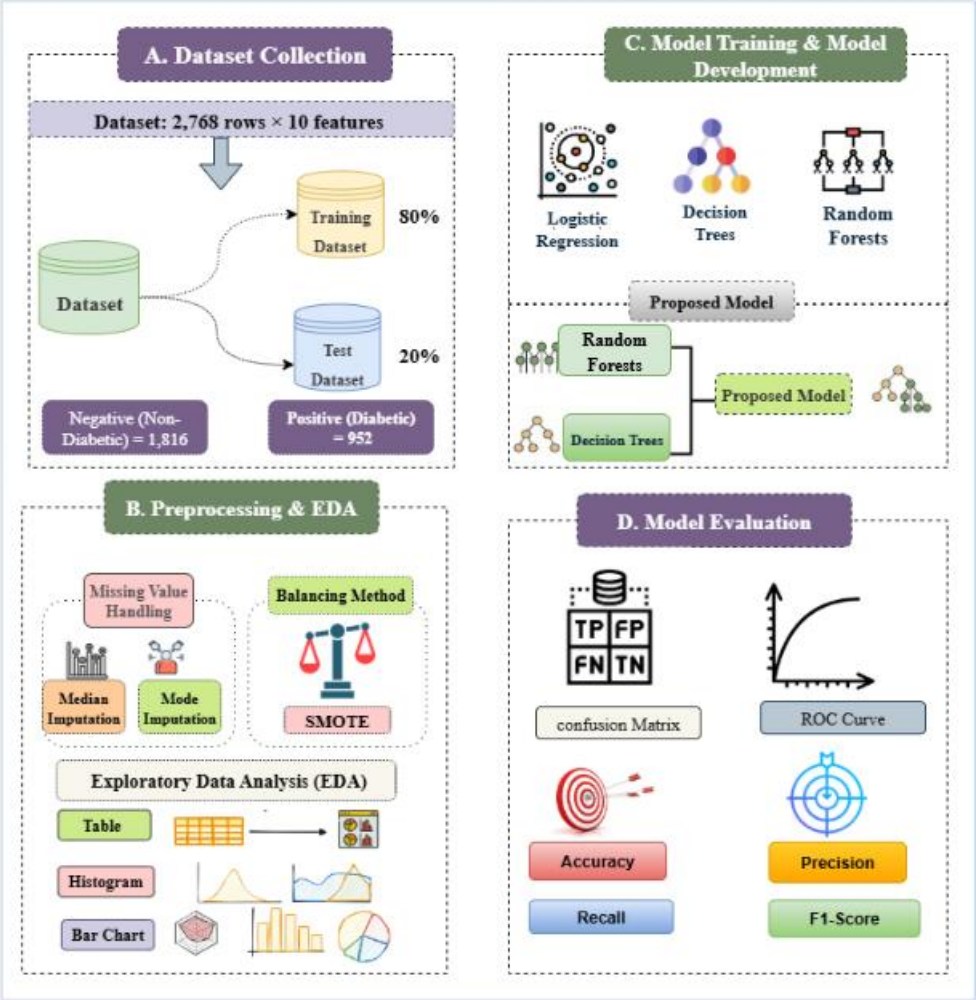


Figure 3.1: Workflow of the Proposed Hybrid Diabetes Prediction Model

3.3 DATA COLLECTION

The dataset used in this research was collected from Kaggle, the publicly-available online repository for data science and machine learning initiatives. It consists of 2768 patient observations and ten clinical attributes associated with the prediction of diabetes. These features are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome. Each record corresponds to a single

patient, and the outcome is binary and indicates whether the patient is diabetic (1) or not (0).The dataset is organized into tables, and is conducive to supervised machine learning. In order to perform an unbiased model evaluation, data were split.

3.3.1. Dataset Composition

The initial dataset contained 2,768 patient records out of which 1,816 are non-diabetic and 952 are diabetic, demonstrating significant class imbalance. Unbalanced problems are frequently the cause of the preference of the majority class in machine learning models, and they result in the nondiabetic cases being badly detected. In order to mitigate this problem, we have adopted the Synthetic Minority Oversampling Technique (SMOTE) to synthetic new cases of diabetes. After SMOTE, there was equal representation of diabetes and non-diabetes in the dataset with 1,816 non-diabetic and diabetic records each. This balancing action has enlarged the total records to be 3,632 and caused an equal representation of the two classes. Consequently, the model may provide better results.

Table 3.1: Dataset Composition Before and After SMOTE

Class	Before SMOTE	After SMOTE
Non-Diabetic (0)	1,816	1,816
Diabetic (1)	952	1,816
Total Records	2,768	3,632

3.3.2 Handling Missing Values

Incomplete records could have a bad impact on performance and reliability of machine learning models. Hence, an imputation policy was used in order to maintain quality and completeness of the data. For each numerical variable including Glucose, Blood Pressure, Skin Thickness, Insulin and BMI, the missing values were imputed using median of the missing feature to reduce noise based on the fact that median is less sensitive to outliers compared to mean. Regarding the categorical attributes, missing values were imputed as the mode, so that the most common and representative category was not lost. This pre-processing removed discrepancies and prepared the data for additional processing, such as normalization

3.3.3 Class Balancing using SMOTE

For imbalanced data, machine learning models are typically biased towards the majority class, which impairs the performance of detecting minority class examples. The prevalence of non-diabetic cases was 1,816, and that of diabetic cases was 952 in this study. This was coped with using the Synthetic Minority Oversampling Technique (SMOTE). SMOTE generates artificial samples using the k-nearest Neighbors of each minority class instance. The formula is given below:

$$x_{\text{new}} = x_i + \delta \times (x_{nn} - x_i) \tag{3.1}$$

Where:

- x_i = a minority class instance
- x_{nn} = the nearest neighbor of x_i
- δ = a random value between $[0, 1]$

3.3.4 Exploratory Data Analysis – Feature Correlation

Correlations between associated clinical features and the outcome of diabetes were analysed. The finding revealed that paramount to be the most influential predictor of diabetes was Glucose by having the highest positive correlation with the outcome variable. BMI exhibited a strong positive association as well, and could therefore be a good screening tool for diabetics. Age was associated moderately with the outcome, that is the older the age, the more chance of developing diabetes. Furthermore, the Diabetes Pedigree Function showed a less strong but significant association with the outcome. A

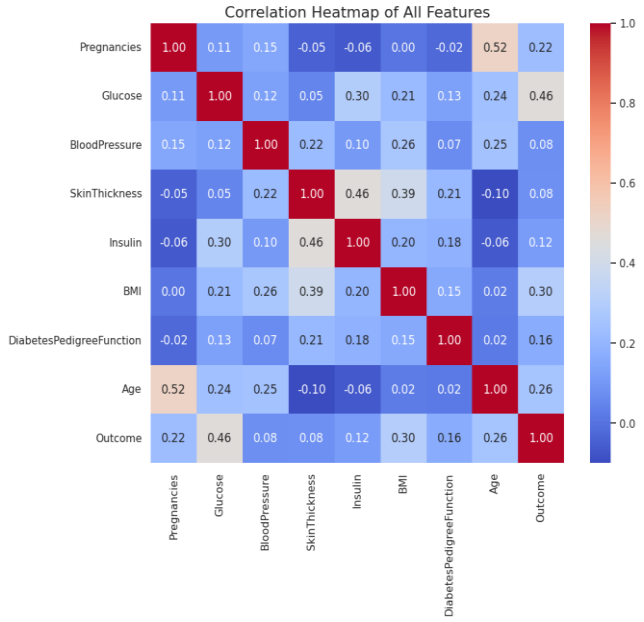


Figure 3.2: Heatmap of Attribute–Outcome Relationships

correlation heatmap was generated to visually represent these relationships, highlighting Glucose and BMI as the most critical features for prediction.

3.3.5 Training & Evaluation

After pre-processing and exploratory analysis, the dataset was split into 80–20 train-test data (ratio) to guarantee robust model building and evaluation. We trained the model by stacking ensemble methods on the training set, using the DT and the RF classifiers as base learners. The outputs of this layer were aggregated and fed to the meta-layer to make final predictions. To enhance generalization and reduce over fitting, during the training the 5-fold cross validation was implemented.

The performance of the model was quantified on the test data by standard measures including Accuracy, Precision, Recall, F1-score and ROC.

$$\textbf{Accuracy} = \frac{(TP+TN+FP+FN)}{TP+TN} \quad 3.2$$

$$\textbf{Precision} = \frac{TP+FP}{TP} \quad 3.3$$

$$\textbf{Recall} = \frac{TP+FN}{TP} \quad 3.4$$

$$\textbf{F1} = 2 * \frac{Precision+Recall}{Precision.Recall} \quad 3.5$$

ROC-AUC:

$$\textbf{FPR} = FP + \frac{TN}{FP} \quad 3.6$$

The AUC value ranges from 0 to 1, where a higher value indicates stronger discriminative capability.

CHAPTER 4

EXPERIMENTAL RESULT ANALYSIS

4.1 Overview

In this section we put a number of machine learning models into practice on diabetes dataset and compare their performance. The results were analysed based on standard classification measures viz., Accuracy, Precision, Recall, F1 Score, and ROC-AUC for fair comparison of all respondents. Single models . Decision Tree, Logistic Regression, Random Forest were initially explored and it was used to detect their pros and cons. Lastly, the proposed method in a stacking ensemble model outperformed the baseline methods with the best accuracy and robustness in the evaluation criteria. Our results demonstrate the effectiveness of ensemble learning in predicting compared to isolated classifiers.

4.2 Logistic Regression

The Logistic Regression model was applied to the dataset to evaluate its classification performance. The detailed results are shown below.

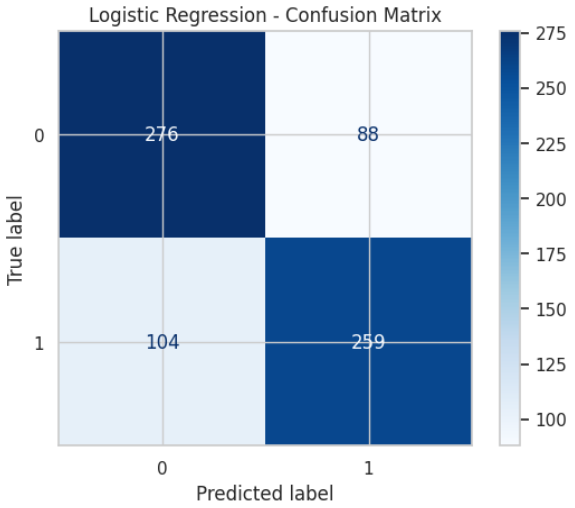


Figure 4.1: Test Data Confusion Matrix of Logistic Regression Model

Table 4.1: Test Data Performance of Logistic Regression Model

Metric	Value
Accuracy	73.59%
Precision	74.64%
Recall	71.35%
F1-Score	72.96%
ROC-AUC	0.8295

The model obtained 73.59% overall accuracy, precision of 74.64% indicating the model's capability to accurately detect the diabetic samples from the prediction positive samples. However, a recall of 71.35% reflects that many of the true diabetic cases were misclassified. The middle ground of precision and recall is reflected in the F1-score of 72.96%, which is not high enough for medical prediction work. Lastly, the ROC-AUC of 0.8295 suggests that the model does separate classes better than a random guess but less than tree-based and ensemble approaches.

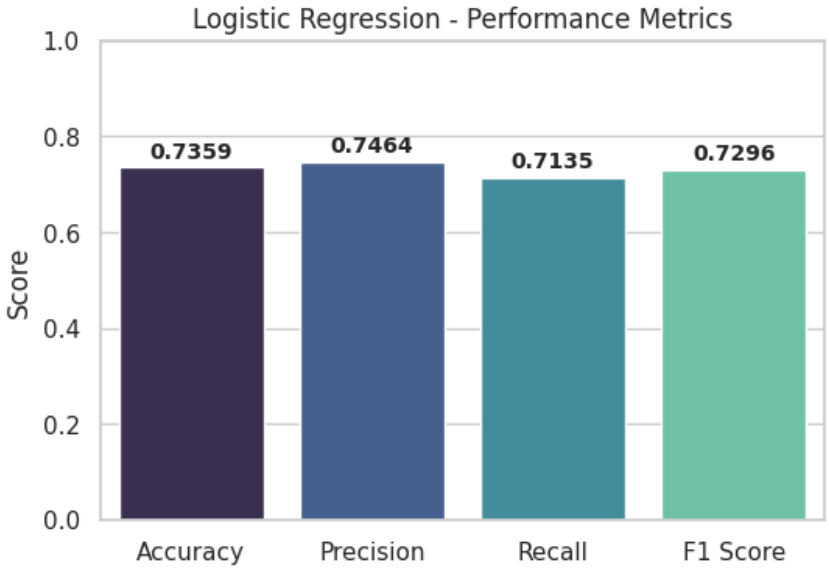


Figure 4.2: Performance Matrix of Logistic Regression

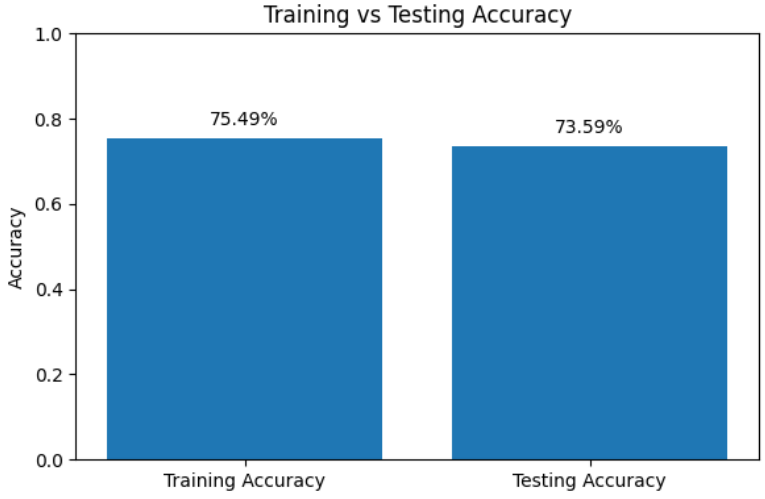


Figure 4.3: Testing vs Test Accuracy

This figure illustrates the training and testing accuracy of the Logistic Regression model on the diabetes dataset. The accuracy for training is 75.49% and for testing is 73.59%,

demonstrating similar results for both. The marginal difference between the training and testing results indicates that the model does not tend to over-fit. It has mediocre overall accuracy though, again demonstrating how Logistic Regression fares poorly on non-linear patterns within the data.

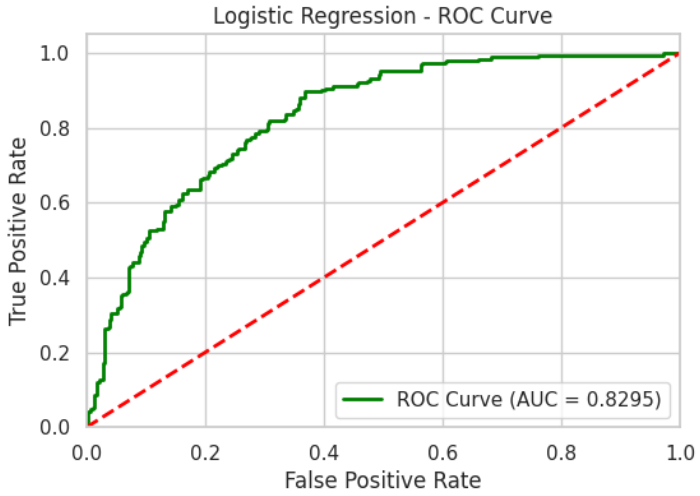


Figure 4.4: ROC Curve for Logistic Regression

4.3 Decision Tree

The Decision Tree classifier was applied to the dataset, and it demonstrated very strong predictive performance. The detailed results are presented in Table 4.2.

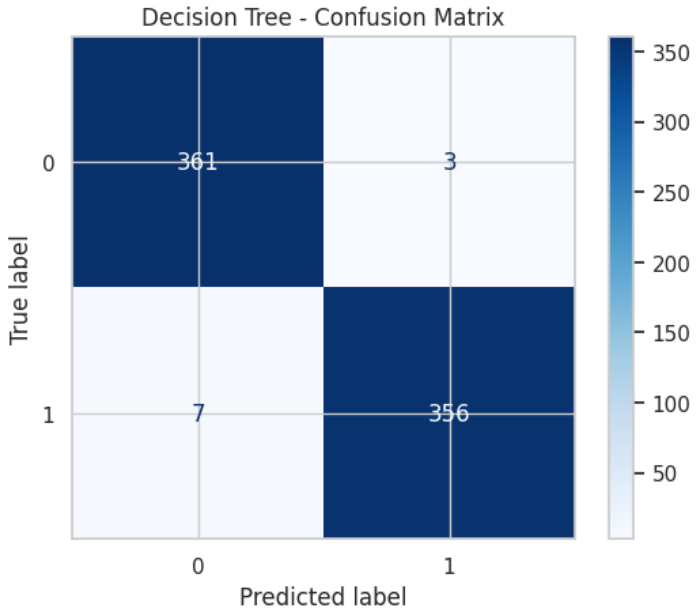


Figure 4.5: Test Data Confusion Matrix of Decision Tree Model

The Decision Tree model correctly classified most test samples, with 361 true negatives and 356 true positives. Only 3 false positives and 7 false negatives occurred, showing high overall accuracy

Table 4.2: Test Data Performance of Decision Tree Model

Metric	Value
Accuracy	98.62%
Precision	99.16%
Recall	98.07%
F1-Score	98.61%
ROC-AUC	0.9862

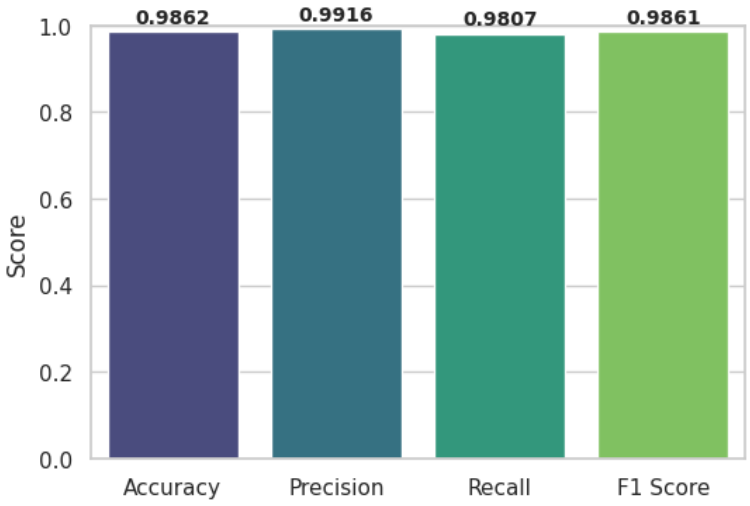


Figure 4.6: Performance Matrix of Decision Tree Model

The accuracy of the Number tree reached up to 98.62%, indicating its overall good performance. The 99.16% accuracy indicates that very few of the predicted diabetic cases were predicted inaccurately and 98.07% recall demonstrates the capability of the model to capture nearly all the true diabetic cases. A strong precision-recall trade-off is seen with an F1-score of 98.61%.

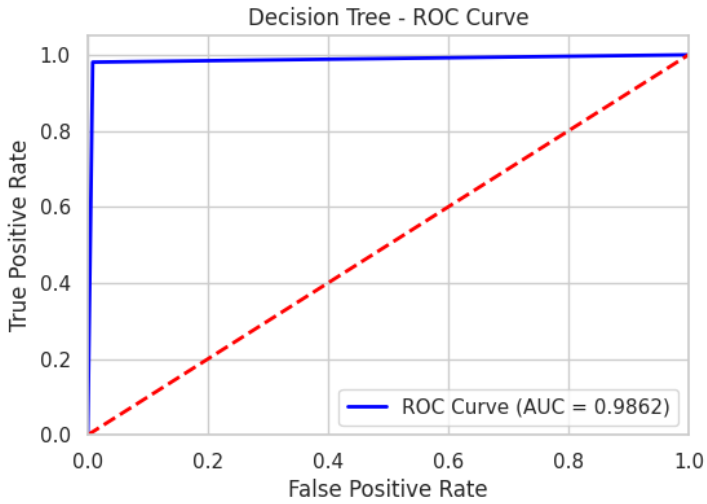


Figure 4.7: ROC Curve for Decision Tree

The exceptional discrimination power of the model is reflected in the ROC-AUC of 0.9862. In summary, Decision Tree performed much better than Logistic Regression as it can capture complex non-linear patterns in diabetes data.

4.4 Random Forest

The Random Forest classifier was applied to the dataset, and it achieved near-perfect predictive performance. The detailed results are presented in **Table 4.3**.

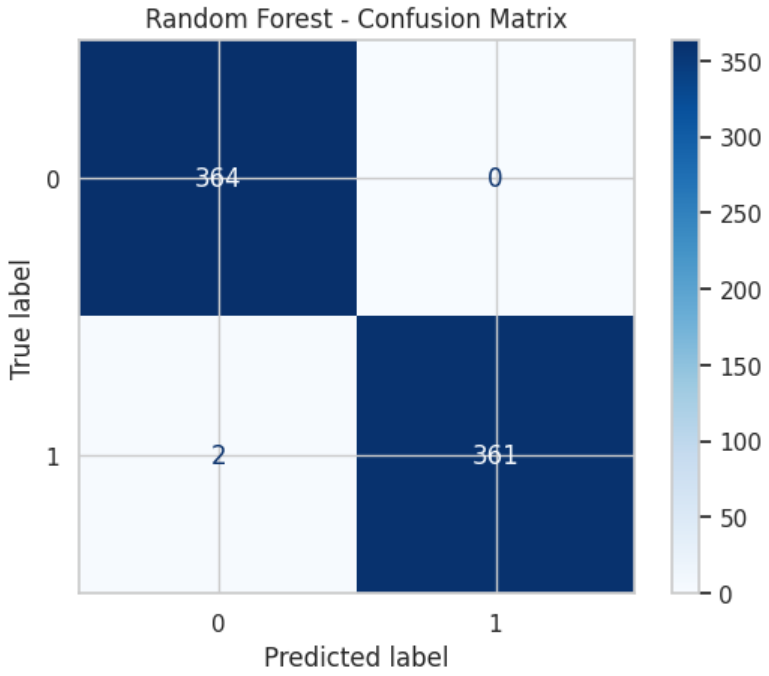


Figure 4.8: Test Data Confusion Matrix of Random Forest Model

The Random Forest model achieved excellent results with 364 true negatives and 361 true positives.It made only 2 false negatives and 0 false positives, indicating very high accuracy.

Table 4.3: Test Data Performance of Random Forest Model

Metric	Value
Accuracy	99.72%
Precision	100%
Recall	99.45%
F1-Score	99.72%
ROC-AUC	1.00

The Random Forest achieved the best accuracy of 99.72%, which was better than Logistic Regression and Decision Tree. With 100% precision, which measures the proportion of positive predictions that are correct, all the predicted diabetic patients were true diabetic patients, and the recall rate of 99.45% means that almost all the true diabetic cases were detected. The F1-score of 99.72 % also indicates a trade-off between precision and recall also attesting to the robustness of the model.

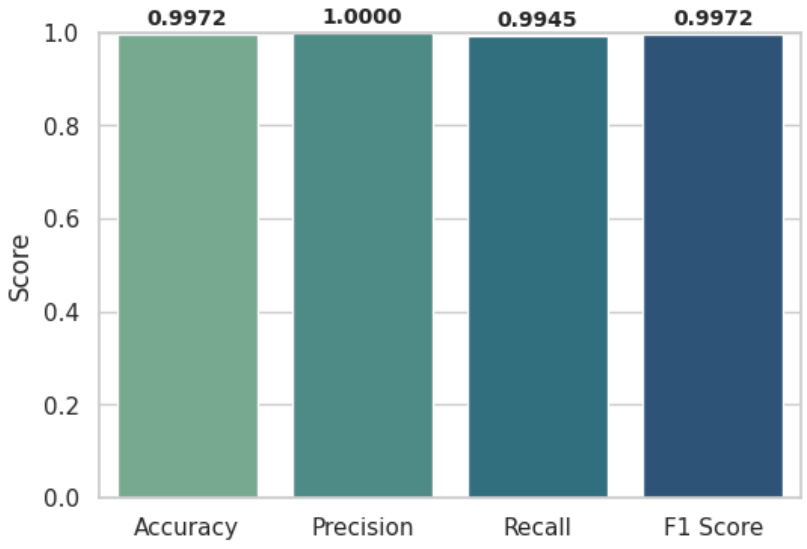


Figure 4.9: Performance Matrix of Random Forest Model

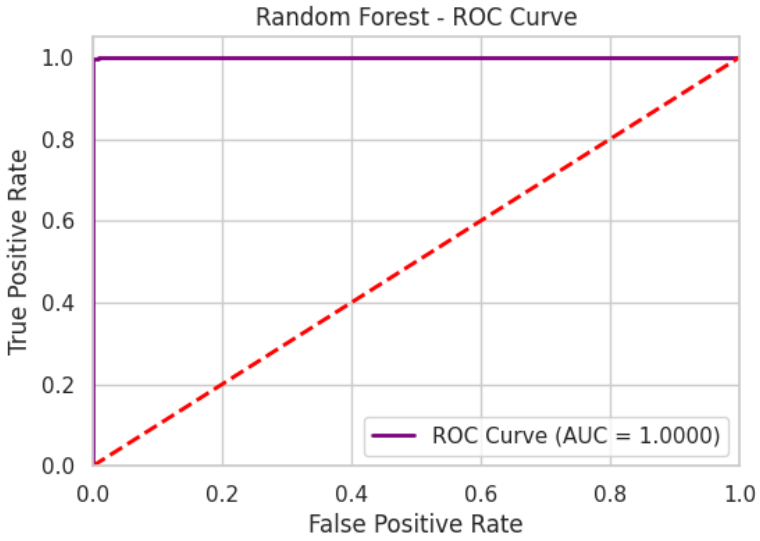


Figure 4.10: ROC Curve for Random Forest

ROC-AUC of 1.0000 reflects perfect discriminative power, showing that the Random Forest could completely separate diabetic from non-diabetic cases. This establishes Random Forest as one of the most effective individual models for the dataset.

4.5 Proposed (Hybrid) Model

The Proposed (Hybrid) Model uses the DT and Random Forest as the base learners, stacked together. This ensemble model is the combination of two models to capture both linear and non-linear patterns existing in the data. It also exhibits better classification performance than single models. The low proportions of mis-classification (false positive rate and false negative rate) lead to a high reliability in the prediction of diabetes using the model.

Table 4.4: Train & Test Performance Metrics for Proposed (Hybrid) Model

Dataset	Accuracy	Precision	Recall	F1 Sore
Train	100%	100%	100%	100%
Test	99.86	100%	99.72	99.86

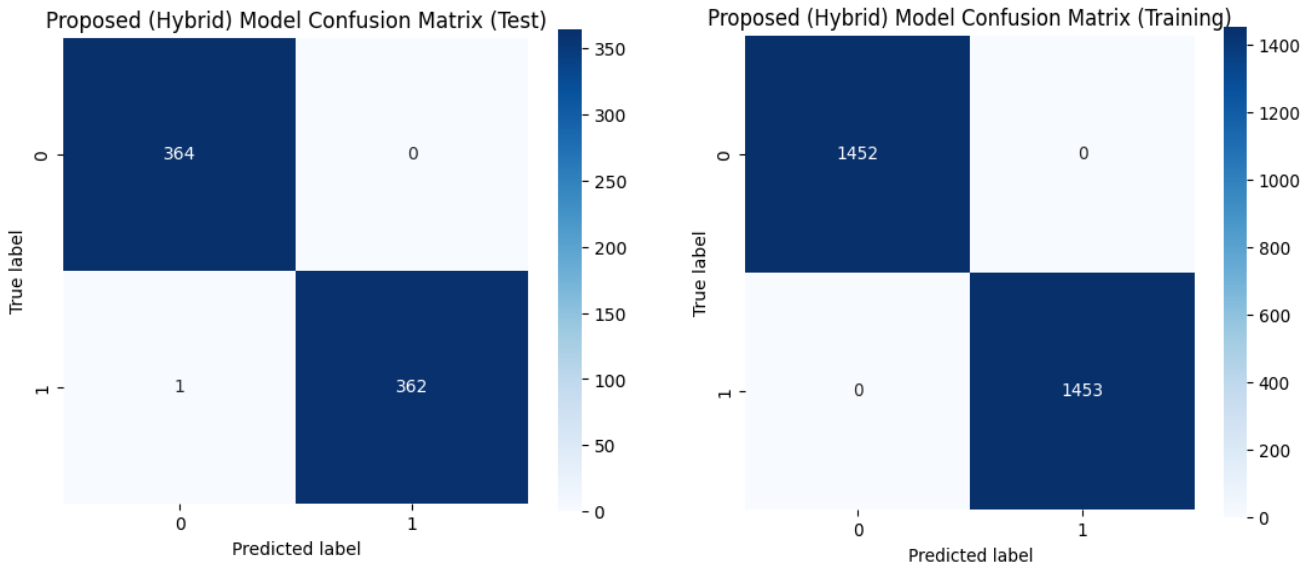


Figure 4.11: Test & Train Data Confusion Matrix of Proposed Model

The performance of the Proposed (Hybrid) Model on training and testing datasets is shown in Table 4.4. The F Model reached 100% of accuracy, precision, recall and F1-score on the training corpus by presenting strong learning ability. It continued its robust performance on test sets with a 99.86% accuracy, 100% precision, 99.72% recall and

99.86% F1-score. These findings demonstrate that the model generalises well to unseen data and remains dependable. The small gap difference between training and testing metrics confirms that the model avoids overfitting and provides robust classification for diabetes prediction.

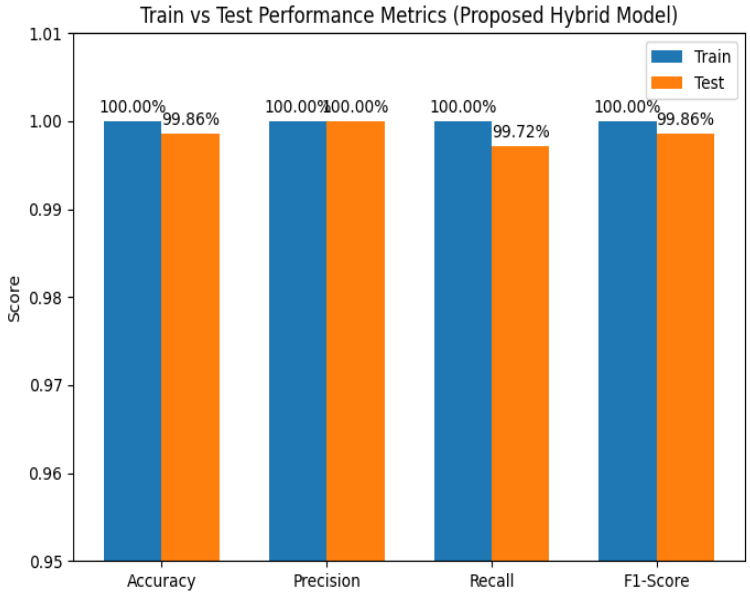


Figure 4.12 : Test & Train Data Performance of Proposed Model

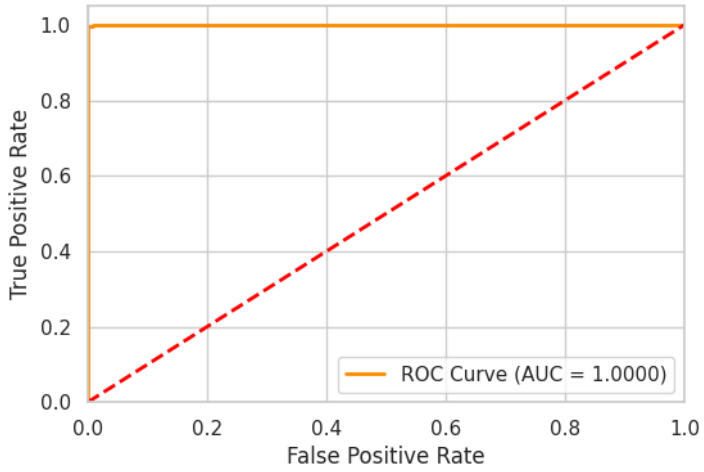


Figure 4.13: ROC Curve for Proposed (Hybrid) Model

The ROC curve shows an **AUC of 1.0**, indicating perfect classification ability.

4.6 Model Comparison

Table 4.5: Comparison table of Test Accuracy Across Different Models

Model	Test Accuracy
Logistic Regression	73.59%
Decision Tree	98.62%
Random Forest	99.72%
Proposed (Hybrid) Model	99.86

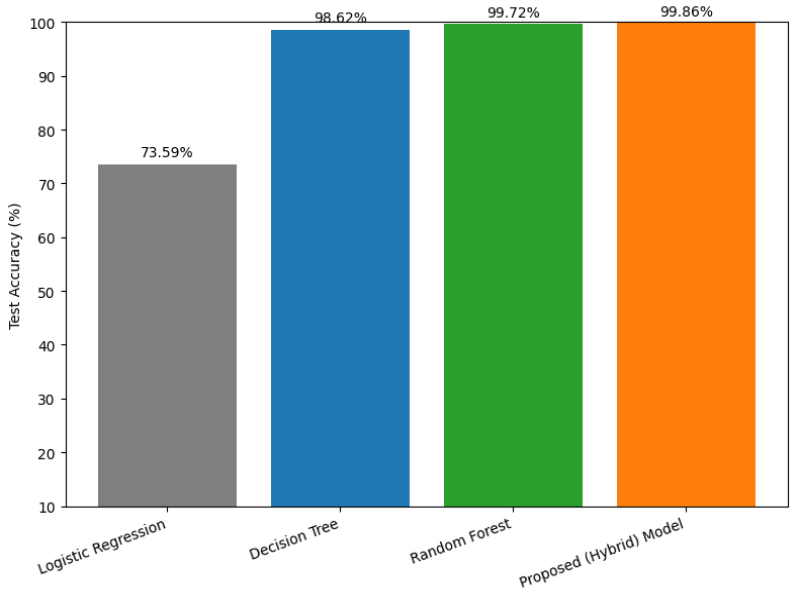


Figure 4.14: Comparison of Test Accuracy Across Different Models

It is obvious from the comparison that the proposed (Hybrid) Model is superior to all other baseline approaches.LR performed poorly with less accuracy however, Decision Tree and Random Forest due to their ensemble nature results better. Nevertheless, the Proposed (Hybrid) Model outperformed them with the best test accuracy of 99.86%. This shows its highest performance to combine precision and recall, proving that it is the most stable model to get prediction results of diabetes.

CHAPTER 5

CONCLUSION

5.1 Summary of the Study

The purpose of this study was to build an accurate machine learning model to early predict diabetes with clinical features. We applied a systematic methodology in this work which includes collection of datasets from a publicly available health care repository, pre-processing such as managing missing values, balancing the dataset with SMOTE, normalization on the dataset and exploratory data analysis. Three base models LR, DT, and RF were designed and evaluated to serve as references. Based on these outcomes, the Proposed (Hybrid) Model, a stacking-based ensemble, exhibited the best, competitive performances in all evaluation criteria. The model yielded an accuracy of almost 100% on the training set, as well as both precision, recall, and F1-score, and it generalised well, 99.86% accuracy, on the test set. The Proposed (Hybrid) Model was further compared with the baseline methods verifying that, is the best model for diabetes classification in this study.

5.2 Research Contribution

The research can make valuable contributions to the literature in diabetes prediction by using machine learning. To begin with, the research develops a valid technique that integrates pre-processing steps like manipulation of missing values, and SMOTE balancing data, and normalizing the data to acquire quality data to be used in the modelling process. Secondly, it will provide a comparison of baseline models of Decision Tree and Random Forest models, their highlights, strengths, and weaknesses. Third, the paper introduces the Hybrid (Proposed) Model which is a stacking-based ensemble of the

classifiers enabling one to achieve higher predictive accuracy and robustness. Lastly, the paper demonstrates that the Proposed (Hybrid) Model is justifiably superior to Implementation of both the Decision Tree and the Random Forest models, with a practical model having been established in managing the management of early diabetes detection and the behaviour of the support provided by Evidence-based healthcare planning.

5.3 Future Work

Despite achieving high performance in classifying diabetes, there are multiple directions for future work of the Proposed (Hybrid) Model. Firstly, it can be tried in larger samples and with a varied study population. This will enhance the potential transferability of the model among more heterogeneous populations. Second, patient clinical and lifestyle factors including diet and physical activity could be included to enhance prediction precision. Thirdly, some more sophisticated deep learning algorithms and features selection methods could be investigated to enhance the performance. Finally, the model may be deployed as a clinical decision support system or integrated into mobile healthcare applications to provide real time prediction of diabetes risk to patients and practitioners.

5.4 Final Conclusion

This study has developed and tested a machine learning model to predict diabetes based on clinical features. Baseline models like Decision Tree and Random Forest were tested post systematic pre-processing to set up the benchmarks. The Proposed (Hybrid) Model which is used as a stacking ensemble, presented the best performance yield with a test accuracy of 99.86%, outperforming all other models. The findings verify that the chimeric ensemble model optimally trades off precision, recall, and F1-score, and is a robust tool for early diabetes diagnosis. This work shows the promise of machine learning in helping healthcare professionals make early diagnoses, leading to better patient outcomes and preventive disease screening.

REFERENCES

1. Afolabi, S., Ajadi, N., Jimoh, A., & Adenekan, I. (2025). Predicting diabetes using supervised machine learning algorithms on E-health records. *Informatics and Health, 2*(1), 9–16. https://doi.org/10.1016/j.infoh.2024.12.002

2. Nomura, A., Noguchi, M., Kometani, M., Furukawa, K., & Yoneda, T. (2021). Artificial intelligence in current diabetes management and prediction. *Current Diabetes Reports, 21*(61). https://doi.org/10.1007/s11892-021-01423-2

3. Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters, 10*(1–2), 1–10. https://doi.org/10.1049/htl2.12039

4. Santhanam, T., & Padmavathi, M. S. (2021). Application of machine learning techniques in diagnosing type 2 diabetes mellitus. *Materials Today: Proceedings, 37*(Part 2), 3116–3120. https://doi.org/10.1016/j.matpr.2020.09.596

5. Larabi-Marie-Sainte, S., Aburahmah, L., Almohaini, R., & Saba, T. (2019). Current techniques for diabetes prediction: Review and case study. *Applied Sciences, 9*(21), 4604. https://doi.org/10.3390/app9214604

6. Febriana, M. E., Ferdinan, F. X., Sendania, G. P., Suryanigrum, K. M., & Yunanda, R. (2023). Diabetes prediction using supervised machine learning. *Procedia Computer Science, 216*, 21–30. https://doi.org/10.1016/j.procs.2022.12.107

7. Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., Hussain, A., Malik, M. A., Raza, M. M., Ibrar, S., & Abbas, Z. (2019). A model for early prediction of diabetes. *Informatics in Medicine Unlocked, 16*, 100204. https://doi.org/10.1016/j.imu.2019.100204

8. Jacobsen, L. M., Felton, J. L., Nathan, B. M., Speake, C., Krischer, J., & Herold, K. C. (2025). Type 1 Diabetes TrialNet: Leading the charge in disease prediction, prevention, and immunotherapeutic mechanistic understanding. *Diabetes Care, 48*(7), 1112–1124. https://doi.org/10.2337/dc24-2908

9. Gowthami, S., Reddy, R. V. S., & Ahmed, M. R. (2024). Exploring the effectiveness of machine learning algorithms for early detection of Type-2 Diabetes Mellitus. *Measurement: Sensors, 31*, 100983. https://doi.org/10.1016/j.measen.2023.100983

10. Oladimeji, O. O., Oladimeji, A., & Oladimeji, O. (2024). Classification models for likelihood prediction of diabetes at early stage using feature selection. *Applied Computing and Informatics, 20*(3/4), 279–286. https://doi.org/10.1108/ACI-01-2021-0022

11. Kumar, R., Kumari, P., & Kumar, R. (2023). Diabetes prediction using machine learning algorithms. *Frontiers in Artificial Intelligence, 6*, 1421751. https://doi.org/10.3389/frai.2023.1421751

12. Abdalrada, A. S., Hamed, M. A., Ibrahim, R. K., Zeebaree, D. Q., & Sadeeq, M. A. M. (2024). Predicting diabetes disease occurrence using logistic regression. *International Journal of Interactive Mobile Technologies, 18*(4), 160–173. https://doi.org/10.3991/ijim.v18i04.52413

13. Shao, H., Liu, X., Zong, D., & Song, Q. (2024). Optimization of diabetes prediction methods based on combinatorial balancing algorithm. *npj Precision Diabetes, 2*(20), 1–9. https://doi.org/10.1038/s41387-024-00324-z

14. El-Bashbishy, R. A., Elattar, M. A., Metwally, M. M., & Elshafey, A. M. (2024). A deep learning model for pediatric diabetes prediction. *Scientific Reports, 14*(1), 51438. https://doi.org/10.1038/s41598-024-51438-4

15. Sampath, R., Nambirajan, S., & Reddy, V. R. (2024). An efficient hybrid ensemble learning pipeline for robust prediction of diabetes. Scientific Reports, 14, 78519. https://doi.org/10.1038/s41598-024-78519-8

Plagiarism Report

213-35-781

ORIGINALITY REPORT

28%	19%	22%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Practical Statistical Learning and Data Science Methods", Springer Science and Business Media LLC, 2025 Publication	1%
2	Submitted to Midlands State University Student Paper	1%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
4	Submitted to Griffth University Student Paper	1%
5	butler.cc.tut.fi Internet Source	1%
6	digitalcollection.utem.edu.my Internet Source	1%
7	"Proceedings of the 5th International Conference on Data Science, Machine Learning and Applications; Volume 1", Springer Science and Business Media LLC, 2025 Publication	1%
8	Submitted to Harrisburg University of Science and Technology Student Paper	1%

Account Clearance