

Development of A Diabetes Prediction Model using Logistics Regression and Random Forest Machine Learning Algorithms

Okoro Favour Ngozi¹, Dr. C. Ituma², Chinonso Job³ Dr. Jeremiah C⁴,
Oguzor Rebecca Nneka⁵

¹Department of Computer Science, Ebonyi State University, Abakaliki.

²Department of Computer Science, Ebonyi State University, Abakaliki.

³Department of Computer Science, University of Great Manchester, United Kingdom.

^{4,5}Department of Computer Science, Ebonyi State University, Abakaliki.

ABSTRACT

Diabetes Mellitus is significantly a global health challenge, characterized by pervasive prevalence and it is prospective for severe long-term complications associated with the diabetes risk factors, including age, body mass index (BMI), blood pressure, glucose levels, and lifestyle behaviors. It remains one of the major global public health challenge due to its increasing prevalence, allied with the disease such as ill health, and the risk of severe long-term complications. The aim of the study was to develop a diabetes prediction model using Logistic Regression and Random Forest machine learning algorithms to facilitate the early identification of individuals at risk of developing diabetes. The study adopted the Object-Oriented Analysis and Design Methodology (OOADM) for system development and the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework for the machine learning model development. Secondary data comprising patients' demographic and clinical characteristics were collected from an existing diabetes dataset. The dataset was subjected to data cleaning, preprocessing, normalization, and feature selection to improve data quality and model performance. Descriptive statistical techniques, including frequency distribution, percentages, mean, and standard deviation, were employed to summarize the characteristics of the dataset, while correlation analysis was conducted to examine the relationships among the predictor variables and assess multicollinearity. The developed model utilized key health indicators, including age, gender, body mass index (BMI), blood pressure, glucose level, and cholesterol level, as predictor variables. The dataset was partitioned into training and testing sets for model development and validation. Logistic Regression and Random Forest classification algorithms were implemented using Python and Jupyter Notebook, while SQLite served as the database management system for data storage. Model performance was evaluated using the confusion matrix and standard classification metrics, namely accuracy, precision, recall (sensitivity), specificity, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The results revealed that both algorithms were effective in predicting diabetes risk; however, the Random Forest model achieved superior predictive performance, demonstrating higher accuracy, better generalization capability, and greater robustness than the Logistic Regression model. The developed prediction system successfully classified patients based on their diabetes risk and provided a user-friendly interface for prediction and decision support. The study concludes that integrating machine learning techniques, particularly the Random Forest algorithm, into healthcare systems can significantly enhance the early detection of Diabetes Mellitus and support timely clinical

intervention. It is recommended that healthcare institutions adopt intelligent predictive models to complement conventional diagnostic procedures and improve patient management.

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Diabetes is a growing public health challenge, particularly in developing regions like Ebonyi State in Nigeria, where there is limited access to early detection tools. The rising incidence of diabetes poses significant healthcare challenges, leading to complications such as cardiovascular disease, kidney failures, and even death, if not diagnosed and managed early. In recent time the machine learning approaches, such as Artificial Neural Networks (ANNs) and Regression-based models, offer a promising solution for early diabetes prediction by leveraging available health data. Generally, diabetes is known as one of the chronic disease that affects mainly adults, it occurs either when the pancreas does not produce enough insulin in the human body or when the body cannot efficiently and effectively use the insulin produced in the body (WHO; 2024). The predominance of diabetes has become a very big global health concern, prompting the need for early diagnosis and intervention to mitigate its complications. Insulin is the hormone that regulates blood glucose. Hyperglycaemia, also called raised blood glucose or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels.

World Health Organization (2022), recorded that 14% of adults aged 18 years and older aged people were living with diabetes, an increase from 7% in 1990. More than half (59%) of adults aged 30 years and above living with diabetes were not taking medication for their diabetes in 2022. Diabetes treatment coverage was lowest in low and middle-income countries.

In 2021, diabetes was the direct cause of 1.6 million deaths and 47% of all deaths due to diabetes occurred before the age of 70 years. Another 530 000 kidney disease deaths were

caused by diabetes, and high blood glucose causes around 11% of cardiovascular deaths Global Burden of Disease Collaborative Network (2021). Since 2000, mortality rates from diabetes have been increasing. By contrast, the probability of dying from any one of the four main non communicable diseases (cardiovascular diseases, cancer, chronic respiratory diseases or diabetes) between the ages of 30 and 70 decreased by 20% globally between 2000 and 2019.

The symptoms of this diabetes disease can occur unexpectedly in adults between the ages of 40 to 65 years. The symptoms of the type 2 diabetes vary depending on age, life styles and habits, it can be slight noticed and could take many years to be detected. The symptoms of diabetes take account of the following; the patients feeling very thirsty at all times, having frequent urinating other than normal time that other patient's urinates, imprecise vision, always feeling tired and accidental loss of weight. In recent times, diabetes usually damages the blood vessels in the heart of human bodies, the eyes, the kidneys and the nerves. Patients with diabetic diseases have the higher risk of chronic health issues such as unhealed wounds, stroke, heart attack, and kidney failure.

As chronic as other disease as diabetes, it causes permanent loss of vision through the damaging of blood vessels in the eyes. Parameswari, R., *et al* (2025). Many people with diabetes develop problems with their feet from nerve damage and poor blood flow. This can cause foot ulcers and may lead to amputation. Three are various types of diabetes.

Type I diabetes; it is formerly known as insulin-dependent it is caused by inefficiency of insulin in the body, both for young or adult globally, it is categorized by deficiency of insulin production in the pancreas and requires either daily or weekly administration of insulin depending on the severity of the disease in the body of the patients. Consequently, from 2017 till date available records showed that there were about 9 million people suffering type I diabetes across the globe; the majority of the patients dwells in high-income states or

countries. The prevalence causes and prevention of the diabetes disease in those areas were not yet known.

Type II diabetes affects some parts of the body that makes use of the sugar (glucose) for energy. It prevents the human body from using insulin appropriately, that can lead to high levels of sugar in the blood in the body if not properly managed or treated. Recently, type II diabetes causes serious damage to the body by affecting the body cells especially nerves and blood vessels. The type II diabetes can often be preventable. There are basic major features that easily contribute to the emerging of type II diabetes especially for adults above 40 years which includes overweight, when the patient is not having enough exercise, and also some genetics issues can as well contribute immensely as the predominance factors for the cases of type II diabetes. The prompt diagnosis at this point is very significant to prevent the worst effects of type II diabetes if detected early enough. The best way to detect diabetes early is to get regular check-ups and blood tests with a healthcare provider Ahmad, E., *et al* (2022).

The symptoms of type II diabetes can be mild in some cases; as well in most times takes several years to be noticed in the body of the patients. Some symptoms of type II diabetes may be similar to those patients with type I diabetes but then may not be often marked or less marked. This may hinder the early detection of the disease after been diagnosed for several years, most times after early suspecting the disease with its complications already arisen in the body.

More than 95% of people with diabetes have type II diabetes. Type II diabetes was formerly called non-insulin dependent, or adult onset. Until recently, this type of diabetes was seen only in adults but it is now also occurring increasingly frequently in children.

The gestational diabetes is hyperglycaemia with blood glucose values above normal but below those diagnostic of diabetes. Gestational diabetes occurs during pregnancy.

Women with gestational diabetes are at an increased risk of complications during pregnancy and at delivery. These women and possibly their children are also at increased risk of type 2 diabetes in the future.

Gestational diabetes is diagnosed through prenatal screening, rather than through reported Impaired glucose tolerance and impaired fasting glycaemia: Impaired glucose tolerance (IGT) and impaired fasting glycaemia (IFG) are intermediate conditions in the transition between normality and diabetes. People with IGT or IFG are at high risk of progressing to type 2 diabetes, although this is not inevitable.

Diabetes can be prevented through the following, lifestyle changes are the best way to prevent or delay the onset of type 2 diabetes, to prevent type 2 diabetes and its complications, people should: reach and keep a healthy body weight, stay physically active with at least 150 minutes of moderate exercise each week, eat a healthy diet and avoid sugar and saturated fat, not smoke tobacco.

Early diagnosis of diabetes can often be achieved through relatively inexpensive blood glucose testing. For individuals with type 1 diabetes, insulin injections are essential for survival. In contrast, one of the most effective strategies for managing both type 1 and type 2 diabetes is maintaining a healthy lifestyle. Many people with type 2 diabetes eventually require medication to help regulate blood sugar, which may include insulin injections or oral agents such as metformin, sulfonylureas, and sodium-glucose co-transporter type 2 (SGLT-2) inhibitors. Beyond glucose-lowering therapies, patients frequently need antihypertensive drugs and statins to reduce cardiovascular risk and prevent complications (Doğru *et al.*, 2023). Additional medical care often involves specialized interventions such as foot care for ulcers, kidney disease screening, and regular eye exams to detect retinopathy, a leading cause of blindness (Chatterjee *et al.*, 2017).

The research proposes the development of a predictive model capable of identifying individuals at risk of diabetes based on key health indicators, including glucose levels, BMI, age, insulin levels, and other relevant features. The study utilizes the Pima Indians Diabetes Database, which contains 768 cases with eight attributes spanning both continuous and categorical variables (Safiri *et al.*, 2022). To optimize model performance, preprocessing steps such as normalization and dataset splitting into training and testing sets were applied. The chosen architecture combines feed-forward logistic regression and random forest algorithms. The design includes an input layer, a hidden layer with ReLU activation, and an output layer with sigmoid activation. Training was conducted using backpropagation, with optimization via the Adam algorithm and binary cross-entropy loss. Hyper parameters including hidden layer count, neuron number, and learning rate—were carefully tuned for best results (Hasan *et al.*, 2020).

Model evaluation employed standard metrics such as accuracy, precision, recall, F1-score, and AUC. Results demonstrated strong predictive capability, with accuracy exceeding 80%. These findings highlight the promise of logistic regression and random forest approaches in early diabetes detection, offering a scalable and cost-effective solution to support healthcare professionals in timely diagnosis and intervention (Khanam & Foo, 2021; Olisah *et al.*, 2022). The model also provides a foundation for future research aimed at improving prediction accuracy and integrating such systems into real-world healthcare workflows. Nonetheless, limitations remain, including restricted data availability, challenges in generalizing across populations, and difficulties in embedding hybrid models into existing clinical practices (Ahmed *et al.*, 2022; Zhou *et al.*, 2023).

Diabetes is a chronic condition that directly impacts the pancreas, leaving the body unable to produce sufficient insulin (Shaikh *et al.*, 2022). Insulin plays a critical role in regulating blood glucose levels, and when it is absent or ineffective, glucose accumulates in the bloodstream.

Several risk factors contribute to the development of diabetes, including excessive body weight, physical inactivity, hypertension, and abnormal cholesterol levels (Chakraborty *et al.*, 2023).

The disease can lead to a wide range of complications. One of the most common early symptoms is increased urination (Reddy and Tan, 2020). If left untreated, diabetes may damage the skin, nerves, and eyes, and can progress to serious outcomes such as kidney failure and diabetic retinopathy, an ocular disease that causes blindness.

Globally, the burden of diabetes continues to rise. According to recent statistics from the International Diabetes Federation (IDF), approximately 537 million people were living with diabetes worldwide (Hossain *et al.*, 2024). This figure underscores the urgent need for prevention, early diagnosis, and effective management strategies to reduce the impact of this widespread condition.

Early and accurate diagnosis of diabetes mellitus, especially during its initial development, is challenging for medical professionals. Artificial intelligence and machine learning techniques, providing a reference, can help them gain preliminary knowledge about this disease and reduce their workload accordingly. Significant numbers of research have been performed to predict diabetes automatically using machine learning and ensemble techniques. Most of these works employed the open-source Pima Indian dataset Tasin *et al* (2023). Some of these articles on automatic diabetes prediction employing the Pima Indian dataset for instance, Kaur, *et al* (2019) used the random forest algorithm to design a system that can predict diabetes quickly and accurately. The dataset used in this work was collected from the UCI learning repository. First, the authors used conventional data preprocessing techniques, including data cleaning, integration, and reduction. The accuracy level was 90% using the random forest algorithm, which is much higher when compared to other algorithms. In a recent paper Mohan, N., and Jain, V. (2020), Mohan and Jain used the SVM algorithm to

analyze and predict diabetes with the help of the Pima Indian Diabetes Dataset. This work used four types of kernels, linear, polynomial, RBF, and sigmoid, to predict diabetes in the machine learning platform. The authors obtained diverse accuracies in different kernels, ranging between 0.69 and 0.82. The SVM technique with radial basis kernel function obtained the highest accuracy of 0.82. Goyal and his team Chatrati, S.P., *et al* (2020) created a smart home health monitoring scheme to detect diabetes. The authors also employed the Pima Indian dataset for their research. For predicting blood pressure status, they used conditional decision making and for predicting diabetes, they used SVM, KNN, and decision tree. Among these models, SVM worked better as they got 75% accuracy which is better than other classifier algorithms. Hassan *et al* (2020). Attempted to predict diabetes using different ensemble method-based machine learning algorithms and the Pima Indian dataset. The authors considered AUC (area under the ROC curve) as their accuracy measure. Finally, the proposed ensemble classifier accomplished an AUC value of 0.95. Jackins *et al* (2021) proposed a multi-disease prediction system, including diabetes using machine learning techniques and used the Pima Indian dataset. According to the authors, the Naive Bayes performed better than the random forest technique with accuracy increments of 0.43%. Mounika *et al* (2021). Anticipated diabetes probabilities using machine learning techniques. This work employed the public Pima Indian dataset and multiple machine learning frameworks. Kumari *et al.* (2021). Attempted to apply a soft voting classifier-based ensemble approach for diabetes prediction. The proposed soft voting classifier attained the overall highest accuracy and F1 score of 0.791 and 0.716, respectively. Saxena, *et al* (2021) used the open-source Pima Indian diabetes dataset for predicting diabetes using the deep belief network model. The authors constructed the model in three phases, that is, data preprocessing using minimax normalization, constructing the network model, and fine-tuning the test dataset to remove any partiality using NN-FF classification. Finally, the authors have done

all the implementation and simulation of the model using MATLAB. The authors reported an F1 score of 0.808 Tasin, I., Nabil, T. U., Islam, S., and Khan, R. (2023), finding the best performance metric compared with the other classification methods Sahid, M. A., *et al* (2024).

Some of these works employed custom datasets or a combination of different datasets. In Deberneh, *et al*, (2021) the authors proposed a type 2 diabetes early prediction system using machine learning approaches. The authors employed a private dataset with more than 253,000 volunteer data from a local hospital in Korea for 6 years. Synthetic oversampling, SMOTE, and under sampling algorithms are applied to deal with the data imbalance problem. Various machine learning approaches are used to anticipate this disease for the following year from the past year's patients' data. Both the random forest and SVM classifiers achieved the highest F1 score of 74%. Hossain, *et al*. (2022), utilized Pima Indian and a private dataset from a local hospital in Bangladesh to design an automatic diabetes prediction system. This work trained several machine learning techniques on the Pima Indian dataset. KNN and decision tree models achieved 81.2% and 79.2% accuracies on the private dataset, respectively. Saidu, *et al* (2024) implemented diabetes mellitus forecasting using advanced feature selection and machine learning models. The authors employed two open-source datasets, that is, Pima Indian and LMCH Iraqi databases. A polynomial regression-based preprocessing technique was used for predicting the missing samples. Hyper parameter tuning has been performed for the random forest, decision tree, and deep neural network frameworks. The proposed DNN technique with the optimized hyper parameters accomplished the highest accuracies of 0.972 and 0.973 for the Pima and LMCH datasets, respectively.

1.2 Statement of the Problem

Diabetes is one of the most prevalent chronic diseases worldwide, affecting millions of people both old and young globally and contributes to chronic metabolic disorder that affects people,

leading to severe health complications such as cardiovascular disease, kidney failure, neuropathy, economic burden, and reduced quality of life of its victims. Early prediction and diagnosis of diabetes can significantly improve patient to overcome and reduce healthcare costs. However, current diagnostic methods often rely on time-intensive laboratory tests or basic statistical models, which may lack predictive accuracy. According to the World Health Organization (WHO), the prevalence of diabetes has been steadily increasing, posing a significant challenge to public health systems worldwide. Early detection and effective management of diabetes are critical to mitigate its long-term impact on individuals and healthcare resources.

Despite the availability of diagnostic tools and statistical models for predicting diabetes risk, many existing approaches face significant limitations.

- i. Traditional statistical methods, such as linear regression, often struggled to capture the complex, non-linear relationships between risk factors, including age, BMI, blood pressure, glucose levels, and lifestyle habits. On the other hand, advanced machine learning techniques, such as Logistics Regression and Random Forest Machine Learning Algorithms have shown potentials in handling non-linear data but are often criticized for their lack of interpretability and transparency, which are crucial in clinical settings.
- ii. Currently, there is a gap in research exploring the integration of interpretable models like linear regression with the predictive power of ANNs. It limits the ability of healthcare professionals to leverage on advanced predictive models that balance accuracy and usability.
- iii. The absence of comparative studies evaluating the performance of these approaches on diabetes prediction datasets hinders the identification of optimal strategies for early detection.

This research seeks to address these challenges by developing a predictive model for diabetes risk assessment that combines logistics regression and random forest machine learning algorithms. The study aims to evaluate the performance, reliability, and practical applicability of these models in predicting diabetes based on key health indicators. By bridging the gap between accuracy and interpretability, this research has the potential to enhance clinical decision-making, improve patient outcomes, and contribute to the broader goal of reducing the global burden of diabetes.

1.3 Aim and Objectives of the Study

The aim of the study was to develop a diabetes prediction model using Logistics Regression and Random Forest Machine Learning. The specific objectives of the study are to:

1. develop a machine learning-based diabetes prediction model using the Logistic Regression algorithm for early identification of diabetes risk.
 2. enhanced diabetes prediction model that will help patients to improve on the best diabetes treatments.
 3. create a machine learning-based diabetes prediction model using the Random Forest algorithm for accurate diabetes risk classification.
 4. design an interpretable diabetes prediction framework that can support early diabetes risk assessment and assist healthcare practitioners and patients in making informed decisions.
- identify and analyze the significant health-related factors associated with diabetes prediction using relevant medical datasets.

1.4 Significance of the Study

This research holds a deep allusion for the future of diabetes detecting machine for diabetic patients both in the hospital and at homes. By introducing a diabetes model using artificial neural network and regression machine learning algorithms, the study will enhance the

traditional statistical methods, such as linear regression, which often struggles to capture the complex, non-linear relationships between risk factors, including age, BMI, blood pressure, glucose levels, and lifestyle habits and establishes new advanced machine learning techniques, such as Artificial Neural Networks (ANNs), that have given a promise in handling non-linear data but are often criticized for their lack of interpretability and transparency, which are very important in clinical settings. The significance of this study extends across multiple dimensions, offering to various stakeholders and sectors:

- i. Health Centers and hospitals
- ii. Government agencies and regulatory Bodies
- iii. Patients with diabetes

1.5 Scope of the Study

This study focused on the development and evaluation of a predictive model for diabetes risk assessment, utilizing Logistics Regression and Random Forest for hospitals in Ebonyi State, Nigeria. It does not cover treatment or management of the disease.

1.6 Limitations of the Study

While this study aimed to provide valuable insights into diabetes risk prediction using hybrid models, several limitations must be acknowledged:

- i. **Data Availability and Quality:** This study relied on publicly available datasets or data collected from healthcare institutions, which varied in quality, completeness, and representativeness. Limited access to comprehensive datasets affected the generalizability of the findings.
- ii. **Model Interpretation and Complexity:** The hybrid approach aimed to balance accuracy and interpretability, achieving an optimal trade-off may still pose challenges. Complex interactions within the ANN component reduced the transparency of the overall model.

- iii. **Integration into Clinical Workflows:** Adapting the hybrid model for use in real-world clinical settings might require additional customization and validation, which were beyond the scope of this study.

1.7 Definition of Terms

- i. **Diabetes:** Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot efficiently use the insulin it produces. Diabetes can damage the heart, blood vessels, eyes, kidneys, and nerves.
- ii. **Logistics Regression and Random Forest:** Is a computational model inspired by the structure and functioning of the human brain. It consists of layers of interconnected neurons (also called nodes) that process and transmit information. ANNs are widely used in machine learning and artificial intelligence for tasks like pattern recognition, classification, and prediction.
- iii. **Model:** A model is a simplified representation of a system, concept, or object used to explain, analyze, or predict its behavior.
- iv **Prediction:** A prediction is a statement or estimate about a future event or outcome based on data, trends, or reasoning. It can be based on scientific analysis, intuition, or past experiences.

CHAPTER TWO

LITERATURE REVIEW

2.1 Diabetes

Diabetes is a long-lasting disease that has an important impact on public health worldwide. There are two types of diabetes, type I and type II diabetes. Type I diabetes also named insulin dependent and type II diabetes named relative insulin deficiency. There has been a rise in interest in applying machine learning techniques in recent years, such as supervised learning, unsupervised learning, and predictive models, to predict, diagnose and manage diabetes. In this literature review, we will discuss the use of these methods in the field of diabetes research and management, and examine recent advances and challenges in this area. Several studies have been conducted to develop accurate and reliable prediction models for diabetes. Shojaae-Mend *et al.* (2024) developed a machine learning model based on decision trees for forecasting the occurrence of diabetes. The study found that the model achieved high accuracy (90%) in predicting diabetes incidence and had good generalizability. Another study conducted by Dharmarajan K. *et al.* (2022), proposed a novel prediction model based on random forest algorithms. The model was trained on a large dataset of health examination records, and it achieved high accuracy (93%) in predicting diabetes incidence. The study emphasizes the importance of using large and diverse datasets for improving the performance of diabetes prediction models. A recent systematic review proposed by Zhu *et al.* (2021) examined the effectiveness of several diabetes prediction methods, including decision trees, machine learning, and artificial neural networks. The review found that machine learning models generally performed better than traditional statistical methods, especially when trained on large and diverse datasets.

A study conducted by Kaur *et al.* (2022) used a combination of artificial-neural-networks (ANNs) and support vector machine (SVM) algorithms to predict diabetes incidence. The

study found that the combination of ANNs and SVM achieved higher accuracy (95%) compared to using either algorithm alone. Also, deep learning is used to predict diabetes. Zhu *et al.* (2021) suggested a model that used demographic, clinical, and laboratory data to predict the occurrence of diabetes. The study found that incorporating multiple data sources improved the accuracy of the diabetes prediction model. The study of Chien *et al.* (2022) in 2022 proposed a deep learning model for diabetes prediction using electronic health records (EHRs) data. The model achieved high accuracy (91%) in predicting diabetes compared to traditional machine learning methods. The study highlights the importance of incorporating rich EHR data in diabetes prediction models to improve their accuracy.

Machine learning has been observed to be very useful in diagnosis and prediction of different health challenges whether infectious or non-infectious. Agrebi, S., and Larbi, A. (2020). Machine learning algorithms make use of AI concepts which are generally classified as supervised, unsupervised and reinforcement learning. Logistic regression, Decision Trees, Support Vector Machines, Artificial Neural Networks and Random Forest Regression are some of the examples of machine learning algorithms that are based on supervised learning approaches. H. Habehh, and S. Gohel (2021).

Ahmed, *et al* (2021) designed a website for the automatic prediction of diabetes. This work employed two open-source datasets and various popular machine learning approaches. The decision tree and random forest classifiers obtained the highest performance for this work with an accuracy of 0.968. Ramesh *et al.* (2021) designed a remote and automatic system for diabetes forecasting with the Pima Indian dataset. The authors employed different data preprocessing techniques, that is, feature scaling, feature selection, and SMOTE Synthetic Minority Oversampling Technique, is a widely used technique in machine learning to address the problem of imbalanced datasets. Imbalanced datasets occur when one class (the minority class) has significantly fewer observations than another class (the majority class), which can

lead to machine learning models that are biased towards the majority class and perform poorly on the minority class. SVM with RBF kernel attained a maximum accuracy of 83.2%. The proposed ML framework is employed in an Android application.

We draw the conclusion that researchers have successfully combined multiple machine learning algorithms with diverse data preprocessing approaches for automatic diabetes detection by reviewing the relevant articles. Most of the works focused on a single accuracy measure, used the open-source Pima Indian dataset, and did not develop the explicability of the prediction of the machine learning frameworks. These reasons have motivated us to evaluate our proposed prediction system based on accuracy, precision, recall, and F1 score, utilize more custom data to merge with the existing dataset, and apply an explainable AI technique.

In their research work, they employed machine learning and its explainable AI techniques to detect diabetes. Along with a private dataset from employees of a local textile industry in Bangladesh, we used the Pima Indian dataset in this work. As there were many missing values in some attributes, we replaced them with the mean value of each feature. We have used the holdout validation technique to split the data. In this research work, we have applied various machine learning-based classification algorithms, that is, decision tree, logistic regression, KNN, random forest, SVM, and ensemble techniques. Next, the performance of these classifiers has been evaluated in terms of precision, recall, and F1 measure. Finally, the best classifier has been selected as the final model to deploy into an Android smart phone application.

This project implements diabetes mellitus prediction through machine learning. The significant contributions of this work are as follows:

- i. A significant contribution of this work is to present a unique dataset of diabetes mellitus containing 203 samples. This private dataset has been obtained from female employees

of Rownak Textile Mills Ltd, Dhaka, Bangladesh, referred to as the ‘RTML dataset’ in this paper. We have collected six features from 203 individuals, that is, pregnancy, glucose, blood pressure, skin thickness, BMI, age, and final outcome of diabetes.

- ii. Another contribution of this work is to keep similarities with the feature of the Pima Indian dataset. The missing insulin feature of the RTML dataset was predicted using a semi-supervised technique.
- iii. SMOTE and ADASYN techniques are implemented to minimize the class imbalance issue. Hyperparameter tuning has also been performed in this work.
- iv. Explainable AI technique with SHAP and LIME libraries will be implemented and used in understanding how the model predicts the decision. This approach helps to interpret what features play the most crucial role in terms of prediction.
- v. A website and an Android application will be developed with the finalized best-performed model of this research work to make instantaneous predictions with real-time data.
- vi. The uniqueness of this work is to implement an automatic diabetes prediction website and Android application for a private dataset of female Bangladeshi patients using machine learning and ensemble techniques.

2.2 Supervised Learning

The supervised learning approach make use of data to retrieve information from the training set and conceptualize models that can accurately predict the training set outcomes and uses the train model to make predictions of new features in the testing data set. E. Alpaydin (2020). Supervised learning is a type of machine learning that trains algorithms to make predictions based on labeled data Fazakis, N. *et al* (2021). In the context of diabetes, supervised learning algorithms can be used to predict the risk of developing the disease, the progression of the disease, and the likelihood of complications. For example, a supervised learning algorithm might use demographic information, lifestyle factors, medical history, and laboratory data as input features, and predict the probability of an individual developing diabetes based on these features Mehta, R., *et al* (2022).

2.3 Unsupervised Learning

In unsupervised learning approach, the algorithms are used to group data into independent clusters which makes it easy for features extraction and classification. In contrast to other forms of approaches in machine learning, unsupervised learning allows for extraction of features by identifying the relationships within the data points and grouping them into clusters based on their similarities M. Berry, *et al* (2020). Common examples of unsupervised learning include K-Means, KNN, Deep Belief Network (DBN) and Convolutional Neural Network (CNN) Agrebi, S., and Larbi, A. (2020). On the other hand, it involves finding patterns and structures in data without labeled training data Sari, (F.A.O., *et al*, 2022). In the context of diabetes, unsupervised learning algorithms can be used to cluster patients based on similar patterns of disease progression or to identify subgroups of patients with similar risk profiles. For example, an unsupervised learning algorithm might identify a subgroup of patients with similar demographic information and lifestyle factors who are at increased risk of developing diabetes.

2.4 Reinforcement Learning Techniques

Are the most efficient form of machine learning algorithms that is closest to human and animal intelligence. This approach works based on self-learning to eliminate error and improve its overall model performance. The most common reinforcement learning method is the Recurrent Neural Network. Agrebi, S., and Larbi, A. (2020).

Scientists and researchers make use of many machine-learning approaches in diagnosis and prediction of several health challenges to come up with an innovative and better performance solutions compared to the traditional ways of disease diagnosis and treatment in healthcare. They apply the Machine learning algorithms to Electronic Health Records (EHR), Medical Imaging, and genetic engineering to solve different health problems, forecast disease spread and make recommendations about prevention and control mechanisms. El-Bashbishy and El-Bakry (2024) proposed a novel technique for early diabetes prediction with high accuracy. They optimized data preprocessing, prediction, and classification using a novel dataset of Mansoura University Children's Hospital Diabetes (MUCHD), which allowed for a comprehensive evaluation of the system's performance. Various validation metrics were employed to ensure the reliability of the results using cross-validation approaches with various statistical measures of accuracy, F-score, precision, sensitivity, specificity, and Dice similarity coefficient. G. Dharmarathne (2024) introduced the first-ever self-explanatory interface for diagnosing diabetes patients using machine learning. They proposed four classification models (Decision Tree (DT), K-nearest Neighbor (KNN), Support Vector Classification (SVC), and Extreme Gradient Boosting (XGB)) based on the publicly available diabetes dataset. All the models exhibited commendable accuracy in diagnosing patients with diabetes, with the XGB model showing a slight edge over the others.

Liu *et al* (2018) developed a deep learning algorithm model using a combination of reinforcement and supervised learning approaches to accurately predict the beginning of

various common diseases such as stroke, kidney failure, and heart failure. The prediction model makes use of both structured and unstructured data from EHR and diagnosis notes which resulted in the model high performance and versatility. In another research using EHR data, Ahmad and Ali (2020) predicted mortality in paralytic ileus (PI) - a medical condition characterized by incomplete blockage of the intestine that prevent direct passage of food substance and which may later lead to total blockage of the intestine. In their research, they come up with an algorithm that predicted the mortality in PI patients with an accuracy of 81.3 %. This development led to improved awareness as well as robust clinical treatment for the ailment

Predictive models are algorithms that use input features to predict an outcome of interest Alrifaie, M.F., *et al* (2021). In the case of diabetes, predictive models can be used to estimate the probability of an individual developing the disease or experiencing complications. Predictive models can be based on either supervised or unsupervised learning algorithms, or a combination of both. For example, a predictive model might use both demographic information and medical history as input features to estimate the risk of developing diabetes. Recent advances in machine learning have led to significant progress in the development of diabetes prediction and management models. For example, several studies have reported the use of machine learning algorithms to predict the risk of developing type II diabetes based on demographic information and lifestyle factors from Rathod *et al.* (2021). These studies have shown that machine learning algorithms can achieve high levels of accuracy in predicting the risk of developing diabetes, and can outperform traditional statistical models.

Table 1: This table shows the closely related works within 5 years on machine learning on diabetes prediction.

Ref.	Contributions	Limitations
[1]	Developed a decision tree-based machine learning model for forecasting diabetes occurrence with high accuracy (90%)	Lack of information on dataset characteristics, potential bias or imbalance in data, limited validation methods
[2]	Proposed a novel prediction model based on random forest algorithms, achieving high accuracy (93%)	Limited information on dataset diversity and representativeness, potential overfitting or generalizability issues
[3]	Conducted a systematic review on diabetes prediction methods, highlighting the effectiveness of machine learning models	Limited discussion on specific model performances, potential bias in selected studies, lack of original research
[4]	Combined artificial neural networks (ANNs) and support vector machine (SVM) algorithms for predicting diabetes, achieving higher accuracy (95%) compared to individual algorithms	Insufficient explanation on model fusion process, potential complexity in implementation, computational demands
[5]	Utilized deep learning techniques to predict diabetes by incorporating demographic, clinical, and laboratory data, resulting in improved prediction accuracy	Limited discussion on model interpretability, potential challenges in data collection and integration, scalability concerns
[6]	Proposed a deep learning model for diabetes prediction using electronic health records (EHRs) data, achieving high accuracy (91%) compared to traditional methods	Limited discussion on EHR data quality and consistency, potential biases in patient selection, generalizability concerns
[7]	Employed a deep learning algorithm to predict the evolution of diabetic retinopathy with high accuracy, offering potential for guiding treatment decisions	Lack of discussion on model generalizability, potential challenges in clinical implementation, scalability concerns

The application of machine learning in diabetes research and treatment remains fraught with difficulties in spite of these advancements. A primary obstacle is the restricted accessibility to superior quality data. Predictions can be off because biased, insufficient, or poor quality data is frequently utilized to train machine learning systems. The dynamic and complicated nature of diabetes presents another difficulty in creating models that adequately represent the disease's variety.

While several studies have demonstrated high prediction accuracy, such as those by Shojaee-Mend, *et al.* (2024) and Zhu *et al.* (2021), they often lacked comprehensive information on dataset characteristics and validation methods, raising concerns about the generalizability of their findings. Additionally, systematic reviews, like the one conducted by Zhu *et al.* (2021) offered valuable insights into the effectiveness of machine learning models but may overlook specific model performances and original research. Studies employing ensemble methods, such as Kaur, H., Kumari, V. (2022) showed promise in achieving higher accuracy but may face challenges related to model fusion complexity and computational demands. Furthermore, while deep learning techniques, as seen in studies by Sari, F.A.O., *et al* (2022) and Rathod, S.R., *et al* (2021) offer improved prediction accuracy, their limited interpretability and scalability concerns warrant further investigation. Overall, there is a need for future research to address these gaps by providing more transparent reporting of methods and datasets, exploring interpretability-enhancing techniques, and validating models on diverse and representative datasets.

Diabetes is a chronic disease with persistent hyperglycemia Negrato, *et al* (2013) Diabetes is mainly divided into type 1 diabetes mellitus (T1DM) Katsarou, A. *et al* (2017), type 2 diabetes mellitus (T2DM) Plows, *et al* (2018). and gestational diabetes mellitus (GDM). T2DM accounts for about 90%-95% of all diabetes cases Kaur, R., Kaur, M. and Singh, J. (2018). It is a heterogeneous and progressive disease caused by the combined effects of genetic and

environmental factors Bailey, C. J. and Day, C (2018). Studies have shown that “thrifty genes” gave people a survival advantage in resource-scarce environments in the past may have adverse effects on health in modern high-sugar, high-fat environments. The hyperglycemia in T2DM usually results from an absolute or relative deficiency of insulin, mainly due to the failure to effectively compensate for insulin resistance Stumvoll, M., *et al* (2005). Although we have a deeper understanding of the pathogenic mechanism and risk factors of T2DM and some effective preventive measures have been proposed, the incidence and prevalence of T2DM continue to rise worldwide Chatterjee, S., *et al* (2017). According to data from the Global Burden of Disease Study, the age-standardized prevalence of T2DM in the world in 2019 was 5282.9 per 100,000 populations, and the mortality rate was 18.5 per 100,000 populations, an increase of 49% and 10.8% respectively from 1990. At the same time, the disability-adjusted life years (DALYs) associated with T2DM also increased by 27.6% during the same period Safiri, S. *et al.* (2022). This shows that although we have made some progress in the prevention and control of diabetes, the challenges facing global public health remain enormous in the face of rising prevalence rates.

T2DM usually lurks in the body for many years and often has no obvious symptoms in the early stages, but as the disease progresses, patients may develop serious complications such as cardiovascular disease, renal failure, vision loss and neuropathy. Therefore, early recognition and intervention are crucial. Early screening can buy valuable time for timely diagnosis. Through lifestyle intervention, drug treatment and other means, it can effectively delay or even reverse the progression of the disease, thereby significantly improving the patient’s quality of life and reducing the occurrence of complications. At the same time, different patients with T2DM have differences in disease progression and response to treatment. Accurate classification prediction is of great significance for developing individualized treatment plans. Classification prediction helps doctors adjust treatment

strategies according to the specific conditions of patients and choose the most appropriate treatment method, thereby optimizing treatment effects, reducing side effects, and avoiding unnecessary treatment, which is crucial to improving patients' clinical outcomes.

In recent years, with the advancement of data science and machine learning (ML) technologies, the use of these technologies to improve the diagnosis and prediction of T2DM has shown significant effectiveness. Hasan *et al.* (2020) proposed a powerful diabetes prediction framework that combines outlier rejection, feature selection, and multiple ML classifiers. By combining different ML models and optimizing the prediction based on Area Under the Curve (AUC) weighting, the combined classifier achieved an AUC of 95% on the Pima Indians Diabetes (PID) database, significantly outperforming existing methods. Khanam *et al.* (2021) used ML and Neural Networks (NN) methods to predict diabetes in 2021. Using the PID dataset, the combination of Logistic Regression (LR) and Support Vector Machine (SVM) models achieved the best results, and the accuracy of the NN model with two hidden layers reached 88.6%. Olisah *et al.* (2022) used Spearman correlation and polynomial regression for selecting features and handling missing values. Their Two-Gradient Descent Neural Network (2GDNN) model attained a 97.25% accuracy on the PID dataset. In addition, ensemble learning and other ML techniques have also made significant progress in diabetes prediction. Ahmed *et al.* (2022) proposed the FMDP model that integrates machine learning methods for diabetes prediction. By combining SVM and ANN to analyze the dataset, the prediction accuracy of the FMDP model reached 94.87%. Zhou *et al.* (2023) proposed a diabetes prediction model based on Boruta feature selection and ensemble learning (DPMBFSEL), with an accuracy of 98% on the PID dataset, which is better than other models. Doğru *et al.* (2023) proposed a super learning model in 2023, which improved the accuracy of early diagnosis of diabetes by integrating multiple algorithms, reaching 92% on the PID dataset, which is better than traditional methods. Alghamdi *et al.*

(2023) used the XGBoost classifier in 2023 and achieved excellent performance in processing high-dimensional feature data, with an accuracy rate of 89% in predicting diabetes.

While traditional machine learning approaches have shown effectiveness, they face several challenges, including difficulty in handling class imbalance, heavy reliance on feature engineering, and extensive data preprocessing. These limitations reduce adaptability to diverse clinical data and hinder generalization across different datasets. Thus, there is a need for more advanced techniques that can address these issues, improve model robustness, and enhance adaptability in clinical settings.

Deep learning technology also showed great potential in diabetes prediction. García-Ordás *et al.* (2021) proposed a deep learning technology based on variational auto encoder (VAE), sparse autoencoder (SAE) and CNN for diabetes prediction, achieving an accuracy of 92.31%. Bukhari *et al.* (2021) developed an improved artificial neural network (ANN) model using the back-propagation scaled conjugate gradient algorithm to predict diabetes on the PID dataset, which achieved an accuracy of 93%, outperforming other ANN models. Diabetes prediction method that converts numerical data into images, using ResNet and SVM, with an accuracy of 92.19% on the PID dataset. Zhao *et al.* (2024) proposed a robust T2DM prediction framework in 2024. Using the NHANES and PID datasets, they optimized data preprocessing through the Attention-Oriented Convolutional Neural Network (SECNN) model and channel attention mechanism, significantly improving the prediction performance, with accuracy rates reaching 89.47% and 94.12%, respectively.

Deep Learning models outperform traditional Machine Learning techniques in handling complex data and learning representations automatically. However, they struggle with data class imbalance, which can lead to poor performance on minority classes. Deep Learning models also face challenges in generalizing across different clinical environments, requiring further optimization to improve their robustness and adaptability in diverse settings.

The studies above highlight significant progress in machine learning and deep learning for diabetes prediction, particularly in improving accuracy and optimizing data processing. However, challenges remain, particularly in addressing class imbalance and model stability. Many models still struggle with poor performance on minority classes, impacting overall accuracy, and show performance variations across different datasets. Future research should focus on addressing these issues to improve model robustness, stability, and applicability in real-world clinical settings.

2.5 Effects of Diabetes on Human Body

The characteristic manifestations of the diabetes depend on the level of blood sugar raised. Initially, symptoms remain in hide, particularly for patients with pre-diabetes or T2D. The diabetes-related complications D. D. Maria Prelipcean (2020) include nerve damage, cardiovascular disease, hypertension, eye problems, kidney disease, dental disease, foot problems N. Amin, and J. Doupis, (2016), and other many more as shown in figure below. If ignored, this can result in tremendously severe complications and even death. Common signs and symptoms are increased thirst, extreme hunger, unexplained weight loss, frequent urination, fatigue, irritability, blurred vision, slow-healing sores, frequent skin infections, and vaginal infection. Complications, shown in figure 1, indicates various complication that arises due to the presence of diabetes in the body; diabetes in the body can be classified of two different types such as chronic and acute complications. The chronic complications arise over a period of time or decade, and the patient needs to keep routine monitoring of it. The acute complications arise due to uncontrolled high and low blood sugars because of the misbalancing of the available insulin and required one. Figure 1 indicates the effects of diabetes in the various parts of the body.

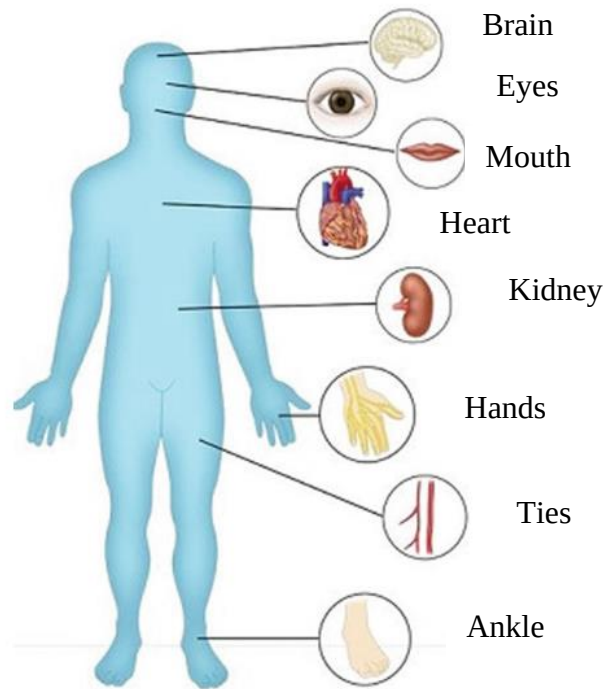


Figure 1: Shows the effect of Diabetes on different parts of body

2.6 Different Complications Arises due to Diabetes: Diabetes complications primarily stem from blood vessel and nerve damage, leading to issues like eye damage (retinopathy), kidney disease (nephropathy), nerve damage (neuropathy), and cardiovascular problems such as heart attack and stroke. Other complications include foot ulcers, which can result in amputations, a higher risk of infections, sexual dysfunction, and conditions affecting the liver, such as Metabolic Dysfunction-Associated Steatotic Liver Disease (MASLD).

Complications by Body System: Eyes (Retinopathy) High blood sugar damages the tiny blood vessels in the retina, potentially leading to vision loss or blindness.

- i. **Kidneys (Nephropathy):** Damage to the kidney's small blood vessels can impair their function, potentially leading to kidney failure, dialysis, or a transplant.
- ii. **Nerves (Neuropathy):** Nerve damage, especially in the feet and legs, can cause symptoms like numbness, tingling, burning, and pain, and it can also lead to foot ulcers.
- iii. **Heart and Blood Vessels:** Diabetes increases the risk of heart attack, stroke, and heart failure by damaging blood vessels and causing them to narrow.
- iv. **Feet:** Poor blood flow and nerve damage increase the risk of developing foot ulcers, which can lead to infections and, in severe cases, amputations.
- v. **Infections:** People with diabetes are more prone to infections, including skin, urinary tract, and mouth infections.
- vi. **Liver Damage:** Diabetes can cause abnormal fat deposits in the liver, potentially leading to liver fibrosis and cirrhosis.
- vii. **Sexual Dysfunction:** Nerve and blood vessel damage can lead to erectile dysfunction in men and vaginal dryness in women. Hearing Loss, hearing problems are more common in individuals with diabetes. Depression: There is a link between diabetes and an increased risk of depression. Dementia; Type 2 diabetes is associated with an increased

risk of dementia, including Alzheimer's disease. Figure 2 indicates the complications due to diabetes disease usually described as acute and chronic diabetes.

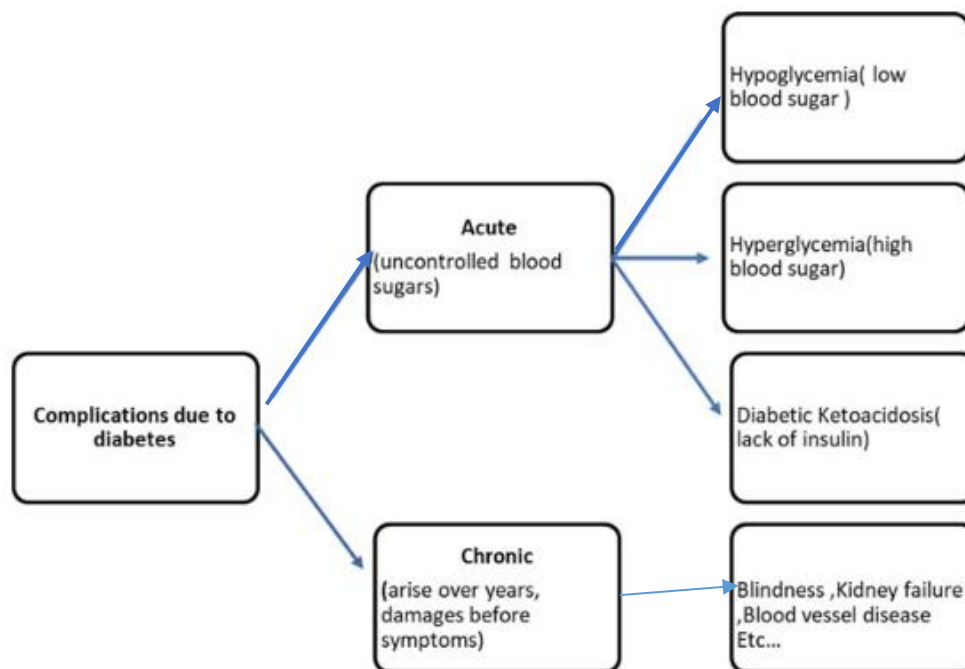


Figure 2: Shows the different complications that arises due to diabetes

The acute complication hypoglycemia also known as low blood sugar occurs because of not reaching the enough sugar to the human body or due to the medication of the T2D. The symptoms include shaking, rapid heartbeat, headache, change in vision, hunger, sudden behavior changes, seizures, lack of coordination, inattention and confusion, sweating, and loss of consciousness.

2.7 Types of Diabetes

There are several types of diabetes, each with distinct characteristics and causes; Type 1 Diabetes- Autoimmune disease: The immune system attacks and destroys insulin-producing beta cells in the pancreas. Insulin-dependent: People with type 1 diabetes need to take insulin injections to control their blood sugar levels. Typically develops in childhood or adolescence: Although it can occur at any age. Type 2 Diabetes- Insulin resistance: The body becomes resistant to insulin, making it harder for glucose to enter cells.

Impaired insulin secretion: The pancreas may not produce enough insulin to meet the body's needs. Gestational Diabetes- Develops during pregnancy: Hormonal changes can cause insulin resistance, leading to high blood sugar levels. Typically resolves after pregnancy: However, women who have had gestational diabetes are at increased risk of developing type 2 diabetes later in life. LADA (Latent Autoimmune Diabetes in Adults)- Similar to type 1 diabetes: Autoimmune destruction of beta cells, but occurs in adults. May not require insulin therapy initially: Some people with LADA may be able to manage their blood sugar levels with oral medications or lifestyle changes.

2.7.1 Type 1 Diabetes Mellitus (T1D)

The damage or malfunctioning of pancreatic beta cells causes the Type 1 diabetes (T1D) which is the insulin producing cells. T1D is commonly diagnosed in children and young adults. For sustaining the insulin level in body, it is needed on daily basis which, however, is not produced in human body and the extra insulin is given by injection via syringe or an insulin pump Bhavya, E., and Sanjay, G. (2022). Saini and R. Ahuja. Type 1 diabetes results from autoimmune destruction of the pancreatic beta-cells. Markers of immune destruction of the beta-cell are present at the time of diagnosis in 90% of individuals and include antibodies to the islet cell (ICAs), to glutamic acid decarboxylase (GAD65), tyrosine phosphatases IA-2 and IA-2b, ZnT8, and insulin auto-antibodies (IAAs). Individuals may convert to negative

if only one marker is positive, but individual risk of developing type 1 DM increases with the number of positive markers. Two positive antibodies are associated with a 75% chance of developing diabetes in the next 10 years. Diagnostic staging is now available for individuals with autoimmunity, even prior to diagnosis of type 1 DM. While this form of diabetes usually occurs in children and adolescents, it can occur at any age. Younger individuals typically have a rapid rate of beta-cell destruction and present with ketoacidosis, while adults often maintain sufficient insulin secretion to prevent ketoacidosis for many years. The more indolent adult-onset variety has been referred to as latent autoimmune diabetes in adults (LADA). There is still controversy whether adult type 1 DM and LADA are the same clinical entity, but LADA patients are antibody positive and often require insulin therapy within years of diagnosis. Idiopathic forms of type 1 DM often are of African or Asian descent. An intermittent risk of diabetic ketoacidosis, based on their varying insulinopenia. Eventually, all type 1 diabetic patients will require insulin therapy to maintain normoglycemia. The table 2 below has the details of the pathogenesis of type 1 diabetes.

Table 2: Show the Pathogenesis of Type 1 Diabetes and their Stages

	Stage 1	Stage 2	Stage 3
Phenotypic characteristics	-Autoimmunity -Normoglycemia -Presymptomatic	-Automimmunity -Dysglycemia -Presymptomatic	-New onset Hyperglycemia -Symptomatic
Diagnostic criteria	-2 or more islet autoantibodies -No impaired glucose tolerance or impaired fasting glucose	-2 or more islet autoantibodies -Dysglycemia: impaired fasting glucose mland/or impaired glucose tolerance: FPG 100-125mg/dl and/or 2-hour plasma glucose 140-199mg/dl A1C 5.7-6.4% or a $\geq 10\%$ increase in A1C	-Clinical symptoms -Diabetes by standard criteria

2.7.2 Type 2 Diabetes Mellitus (T2D)

Here's a more natural, humanized version of your text that still retains the academic tone and citations, but flows more smoothly and avoids sounding machine-generated:

Type 2 diabetes (T2D) is most commonly diagnosed in middle-aged and older adults, accounting for about 90–95% of all diabetes cases worldwide (Doğru *et al.*, 2023). In this form of diabetes, the body either does not use insulin effectively or fails to respond adequately to it, leading to insulin resistance. While insulin is often present in higher concentrations, it is insufficient relative to the degree of resistance, and glucose homeostasis cannot be maintained. Over time, progressive beta-cell failure results in worsening insulin deficiency. Family history plays a significant role, with 80–90% of patients reporting hereditary links. Other contributing factors include gestational diabetes in women, poor diet, excess weight, illness, and certain medications (Doğru *et al.*, 2023). More advanced analyses of beta-cell function show that individuals at risk those with both impaired fasting glucose and impaired glucose tolerance may already have lost nearly 80% of their pancreatic insulin-secreting capacity (Safiri *et al.*, 2022). In a minority of cases, patients present with severe insulinopenia at diagnosis despite normal or near-normal insulin sensitivity (Hasan *et al.*, 2020).

Most people with T2D exhibit visceral (intra-abdominal) obesity, a hallmark of ectopic fat deposition strongly linked to insulin resistance (Khanam and Foo, 2021; Olisah *et al.*, 2022). This condition often clusters with hypertension, dyslipidemia (high triglycerides, low HDL cholesterol, postprandial hyperlipemia), vascular endothelial dysfunction, and elevated PAI-1 levels (Ahmed *et al.*, 2022). Collectively, these abnormalities form the “insulin resistance syndrome” or “metabolic syndrome” (Zhou *et al.*, 2023; Doğru *et al.*, 2023). As a result, patients face a heightened risk of atherosclerotic cardiovascular disease (ASCVD), including myocardial infarction and stroke.

Genetics also play a strong role. T2D is more prevalent among minority ethnic groups such as Mexican-Americans, Latinos, African Americans, American Indians, and Pacific Islanders compared to those of European ancestry. Although many genes have been associated with the condition, they explain only a small fraction of its heritability, leaving much of the genetic basis unresolved.

2.7.3 Gestational Diabetes Mellitus (GDM)

The gestational diabetes (GDM) was found in the pregnant women that normally exist till the birth of a child. About 9% of women developed this type of diabetes, and there are chances of developing T2D at some later stage in life. Murphy, H. R., *et al* (2021) on his study said that 20–25% of women will be diagnosed with T2D diabetes in next 10 years of pregnancy.

Gestational Diabetes Mellitus (GDM) Gestational diabetes mellitus (GDM) is defined as glucose intolerance which is first recognized during pregnancy. In most women who develop GDM, the disorder has its onset in the third trimester of pregnancy. At least 6 weeks after the pregnancy ends, the woman should receive an oral glucose tolerance test and be reclassified as having diabetes, normal glucose tolerance, impaired glucose tolerance, or impaired fasting glucose. Gestational diabetes complicates about 8-9% of all pregnancies, though the rates may double in populations at high-risk for type 2 diabetes Bukhari, M. M. *et al* (2021). Clinical detection is important, since therapy will reduce perinatal morbidity and mortality. Dysglycemia risk in GDM is a continuum, and risk assessment for GDM should occur at the first prenatal visit. Two groups, the International Association of Diabetes and Pregnancy Study Groups (IADPSG) and the National Institutes of Health (NIH) Consensus Group recommend different testing methods for the diagnosis of GDM. A large-scale (~25,000 pregnant women) multinational epidemiologic study Zhao, J. *et al* (2024) demonstrated that risk of adverse maternal and neonatal outcomes continuously increased as a function of maternal glycemia at 24-28 weeks, even within ranges previously considered normal for

pregnancy. These observations led to a revision in the diagnostic criteria recommended by IADPSG for GDM using a “one-step” 75-gram OGTT Mealey, B. L. *et al.* (2006) and Zhao, J. *et al.* (2024). A NIH Consensus Development Conference, using the same data, continues to recommend the “two-step” approach to diagnosis. The stated reason was the lack of interventional trials to prove the new criteria could decrease poor outcomes, as it was observational. The Carpenter Coustan values are lower because they are corrected to account for assays currently in use. All women not known to have diabetes should undergo glucose test screening between weeks 24 and 28 using the “one step” 75 grams of glucose load in the morning after an overnight fasting period of at least 8 hours or the “two-step” method which starts with a non-fasting 50gram glucose load test (GLT). A fasting 100-gram glucose tolerance test is only performed if the screening 50 gram GLT 1-hour plasma glucose value is $\geq 140\text{mg/dl}$ (7.8 mmol/L).

2.7.4 Other Forms of Diabetes

Beyond the more common forms of diabetes, there exists a “catch-all” category that encompasses several rare and less frequently encountered types. These include diabetes associated with cystic fibrosis, single hereditary defects such as MODY, hemochromatosis, and diabetes resulting from surgical interventions or drug use. Collectively, these atypical forms account for less than 5% of individuals diagnosed with MODY (A Review for Predicting Diabetes Mellitus).

Maturity-Onset Diabetes of the Young (MODY) is particularly noteworthy because, unlike type 1 or type 2 diabetes, it is monogenic in origin. Some forms of MODY can be managed effectively with dietary modifications alone, while others require oral hypoglycemic agents or insulin therapy. Advances in genetic testing and diagnostic technologies have recently brought greater attention to these single-gene forms of diabetes, allowing for more precise identification and tailored treatment strategies. This progress underscores the importance of

recognizing uncommon variants of diabetes, as their management and prognosis differ significantly from the more prevalent types.

2.8 Classification of Diabetes

Diabetes is a heterogeneous complex metabolic disorder characterized by elevated blood glucose concentration secondary to either resistance to the action of insulin, insufficient insulin secretion, or both. The major clinical manifestation of the diabetic state is hyperglycemia. However, insulin deficiency and/or insulin resistance also are associated with abnormalities in lipid and protein metabolism, and with mineral and electrolyte disturbances. The vast majority of diabetic patients are classified into one of two broad categories: type 1 diabetes mellitus, which is caused by an absolute or near absolute deficiency of insulin, or type 2 diabetes mellitus, which is characterized by the presence of insulin resistance with an inadequate compensatory increase in insulin secretion. In addition, women who develop diabetes during their pregnancy are classified as having gestational diabetes. Finally, there are a variety of uncommon and diverse types of diabetes, which are caused by infections, drugs, endocrinopathies, pancreatic destruction, and genetic defects. These unrelated forms of diabetes are included in the “Other Specific Types” and classified separately.

2.8.1 Specific Classification: Maturity-Onset Diabetes of the Young, Genetic disorder: Caused by mutations in genes that affect insulin production. Typically develops in young adulthood: Often inherited in an autosomal dominant pattern. Secondary Diabetes- Caused by other medical conditions or medications: Examples include pancreatitis, pancreatic surgery, and certain medications like steroids. Prediabetes- Blood sugar levels are higher than normal: But not high enough to be classified as diabetes. Increased risk of developing type 2 diabetes: Lifestyle changes can help prevent or delay the onset of type 2 diabetes.

2.8.2 Genetic Defects

Maturity Onset Diabetes of the Young (MODY) is characterized by impaired insulin secretion with minimal or no insulin resistance. MODY can be sub-typed into neonatal and MODY-like. Neonatal diabetes usually has an onset in the first 6 months of life and can be transient or permanent. MODY may affect genes important for beta-cell glucose sensing, development, function, and regulation. Genetic inability to convert pro-insulin to insulin results in mild hyperglycemia. Similarly, the production of mutant insulin molecules has been identified in a few families and results in mild glucose intolerance. MODY5 is most often associated with renal cysts and was not listed on the most recent ADA classification of diabetes, but can rarely cause diabetes. The natural history of MODY is highly dependent on the underlying genetic defect and most typically exhibit mild hyperglycemia at an early age. The disease is inherited in an autosomal dominant pattern.

Several genetic mutations have been described in the insulin receptor and are associated with insulin resistance. Type A insulin resistance refers to the clinical syndrome of acanthosis Nigricans, civilization in women, polycystic ovaries, and hyper-insulinemia. Leprechaunism is a pediatric syndrome with specific facial features and severe insulin resistance that results from a defect in the insulin receptor. Lipodystrophic diabetes results from post-receptor defects in insulin signaling. A variety of genetic syndromes have been described in which diabetes

mellitus occurs with increased frequency. The etiology of the disturbance in glucose homeostasis in these diverse and seemingly unrelated syndromes remains undefined.

2.9 Diseases of the Exocrine Pancreas

Damage of the pancreas must be extensive for diabetes to occur. The most common causes are pancreatitis, trauma, and carcinoma. Chronic pancreatitis can cause general inflammatory/fibrotic changes in the pancreas which can cause diabetes. Cystic fibrosis causes a well-recognized pancreatic exocrine function insufficiency, but the same thick, viscous secretions cause inflammation, obstruction, and destruction of small ducts in the pancreas, which can lead to insulin deficiency. Hemochromatosis has also been associated with impaired insulin secretion and diabetes.

2.9.1 Endocrinopathies

Since growth hormone, cortisol, glucagon, and epinephrine increase hepatic glucose production and induce insulin resistance in peripheral (muscle) tissues, excess production of these hormones can cause or exacerbate underlying diabetes. Although the primary mechanism of action of these counter regulatory hormones is the induction of insulin resistance in muscle and liver, overt diabetes mellitus does not develop in the absence of beta cell failure.

2.9.2 Infections

A variety of infections have been etiologically related to the development of diabetes mellitus. Of these, the most clearly established is congenital rubella. Approximately 20% of infants who are infected with the rubella virus at birth develop autoimmune type 1 diabetes later in life. These individuals have the typical type 1 susceptibility genotype, DR3/DR4.

2.9.3 Drugs

A large number of commonly used drugs have been shown to induce insulin resistance and/or impair beta cell function and can lead to the development of diabetes mellitus in susceptible

individuals. An extensive review of these drugs and their mechanism of action has been published. Drug classes which have been extensively associated with elevating glucose levels include: beta-blockers, thiazide diuretics, fluoroquinolones, atypical or second generation anti-psychotics, calcineurin inhibitors, protease inhibitors, nicotinic acid, and corticosteroids. In addition, HMG-CoA reductase inhibitors (statins) have been shown to cause a small increase in the risk of diabetes, though the exact mechanisms of how it may increase the risk of diabetes are not completely understood.

2.10 Research Gap

Currently, there is a gap in research exploring the integration of interpretable models like Logistics Regression with the predictive power of ANNs. The use of Logistics Regression model in predicting diabetes prediction limits the ability of healthcare professionals to leverage on advanced predictive models that balance accuracy and usability. Furthermore, the absence of comparative studies evaluating the performance of these approaches on diabetes prediction datasets hinders the identification of optimal strategies for early detection.

This research seeks to address these challenges by developing a predictive model for diabetes risk assessment that combines Logistics Regression and Random Forest algorithms. The research aims to evaluate the performance, reliability, and practical applicability of these models in predicting diabetes based on key health indicators. By bridging the gap between accuracy and interpretability, this research has the potential to enhance clinical decision-making, improve patient outcomes, and contribute to the broader goal of reducing the global burden of diabetes.

CHAPTER THREE

SYSTEM ANALYSIS AND METHODOLOGY

3.1 System Analysis

It is a fundamental concept in systems engineering, software development, and IT project management. Systems analysis methods are approaches offered by the fields of systems engineering and systems science that apply qualitative or quantitative modeling techniques to reflect complexity within a system and identify optimal solutions given the system's context. These methods focus on identifying and evaluating properties of complex systems (such as interactions between heterogeneous system components, feedback loops, dynamic relationships, and emergent behaviors resulting from heterogeneous, adaptive actors), thereby demystifying the relationships between a system's components and changes to the system over time. By making system boundaries and goals explicit, systems analysis methods may help minimize implementation resistance. Many aspects of systems analysis methods reside under the umbrellas of systems science and systems engineering, which aim to grow knowledge regarding systems-related phenomena and to develop specific solutions to problems faced by complex systems, respectively Luke D, *et al* (2018). Applied to studying implementation mechanisms, systems analysis methods can help in:

- (i) better identify and manage conditions that may or may not activate mechanisms (both expected mechanisms targeted by a strategy and unexpected mechanisms that the methods help detect)
- (ii) flexibly guide strategy adaptations to address emergent influences of context (e.g., individuals' motivations, norms, organizational policies and structures, financial resources) on the mechanisms that were not foreseen when the strategy was initially selected and used.

A strong machine learning methodology for diabetes prediction often involves a combination of robust algorithms, thorough data preprocessing, and model evaluation. Key elements include using ensemble methods like Random Forest, XGBoost, or CatBoost, which often outperform single algorithms due to their ability to leverage multiple base learners. Data preprocessing steps should address missing values, handle categorical data, and

normalize/scale numerical features for better model performance. Finally, rigorous model evaluation using metrics like accuracy, sensitivity, specificity, and ROC curves is essential to ensure the model's reliability and generalizability.

3.2 Analysis of the Existing System

Manual diabetes recording systems, which involve patients documenting their blood glucose levels and related health data using pen-and-paper logbooks, have been a longstanding method in diabetes self-management. While these systems are straightforward and accessible, especially in low-resource settings, recent report has highlighted several limitations that can impact the effectiveness of diabetes management.

3.2.1 Advantages of the Existing System

The following are the advantages and strengths of Manual Diabetes Recording Systems

1. **Simplicity and Accessibility:** Manual logbooks are easy to use and do not require technological proficiency, making them suitable for patients without access to digital devices or the internet.
2. **Cost-Effectiveness:** They are inexpensive, requiring only basic materials like paper and pen, which is beneficial in resource-limited settings.
3. **Privacy:** Manual records are not susceptible to digital data breaches, ensuring patient confidentiality without the need for cyber security measures.

3.2.2 Disadvantages of the Existing System

The following are the **disadvantages, limitations and Challenges**

1. **Data Inaccuracy and Misreporting:** Studies have shown that manual recording is prone to errors, including missing entries, incorrect transcription, and even deliberate misreporting of blood glucose readings. Such inaccuracies can lead to suboptimal clinical decisions and poor glycemic control.

2. **Lack of Real-Time Feedback:** Manual systems do not provide immediate analysis or alerts, delaying necessary interventions and adjustments in therapy.
3. **Increased Patient Burden:** The process of manually recording data multiple times a day can be tedious, leading to decreased adherence over time.
4. **Difficulty in Data Sharing:** Sharing manual records with healthcare providers can be cumbersome, often requiring physical visits, which may not be feasible for all patients.

3.2.3 Block Diagram of the Existing System

A diagram of a diabetes detecting and treatment system would visually represent the flow of information and processes involved in monitoring blood glucose levels and managing diabetes. It would likely include components like a blood glucose meter, a lancing device for obtaining blood samples, a test strip, a display screen, and potentially a data logging system for tracking trends according to StatPearl. The block diagram as shown in *Figure 3* of Existing Diabetes Detecting Machine using Glucose Meter with Strip, comprises of the following components:

1. **Glucose Meter: Glucometer** is a medical device for determining the approximate concentration of glucose in the blood. It can also be a strip of glucose paper dipped into a substance and measured to the glucose chart. It is a key element of glucose testing, including home blood glucose monitoring (HBGM) performed by people with diabetes mellitus disease or hypoglycemia. It contains the electronics to detect and measure the glucose level in the blood sample, typically displaying the result in mg/dL or mmol/L
 - i. **Lancing Device:** Is used to prick the finger (or other approved site) to obtain a small blood sample according to Memorial Sloan Kettering Cancer Center.
 - ii. **Test Strip:** Inserted into the meter and absorbs the blood sample, reacting with the blood glucose to produce a signal according to StatPearls.
 - iii. **Display Screen:** This shows the current blood glucose reading.

2. **Treatment and Management Data Logging:** If the meter has memory or Bluetooth capabilities, data can be stored and transferred to a smartphone app or computer for tracking trends.
 - a) **Insulin Administration:** If the person with diabetes requires insulin, the diagram might show the insulin pump or injection process.
 - b) **Medication Management:** It is the second face of the diabetes medication pattern that helps the patients to track the medication within the specific period.
 - c) The diagram could also include a section for other medications (e.g., oral diabetes drugs) that are part of the treatment plan.
 - d) **Diet and Exercise:** These are the crucial parts of diabetes management and could be represented as branches leading to lifestyle adjustments.

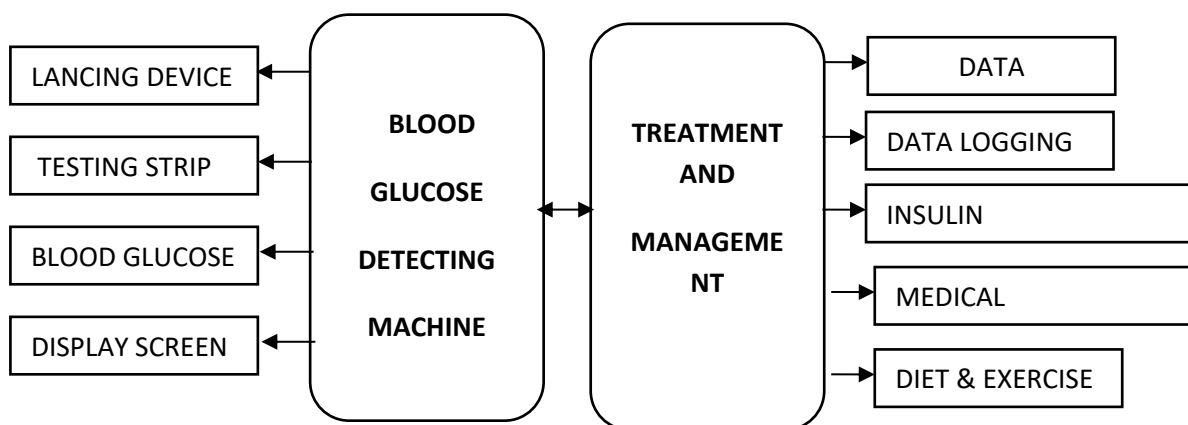


Figure 3: Shows the Block Diagram of the Existing System showing blood glucose detecting machine, treatment and management.

1.2.4 Flow Chart Diagram of the Existing System

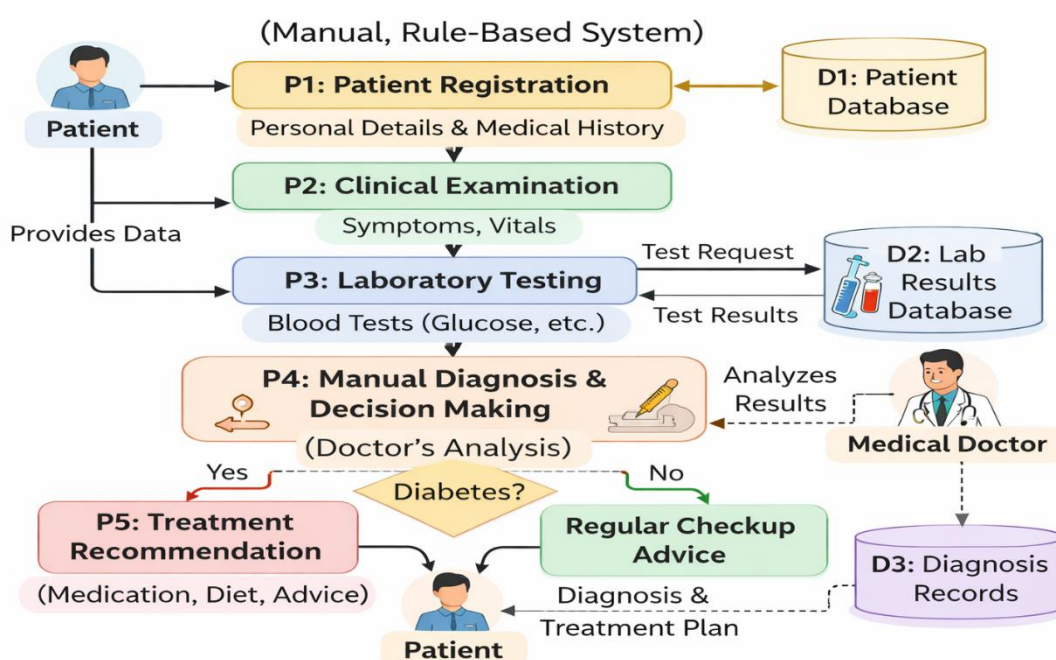


Figure 4: Shows the flow Chart Diagram of the Existing System

3.3 Analysis of the New System

The new diabetes prediction system represents an advanced computational framework designed to facilitate early diagnosis and improve patient management outcomes through machine learning algorithm or methodologies. The system integrates data acquisition, preprocessing, feature selection, model training, evaluation, and real-time prediction into a cohesive pipeline, thereby offering significant enhancements over conventional clinical diagnostic approaches. the new diabetes prediction system represents a significant methodological advancement in computational healthcare. By leveraging machine learning models and a structured workflow, the system achieves high accuracy in diabetes prediction, supports early clinical intervention, and provides a scalable, real-time solution for patient management. The integration of preprocessing, feature selection, and model evaluation ensures robustness, while the system's adaptability allows for continuous refinement and potential expansion to other chronic disease prediction domains.

3.3.1 System Performance and Predictive Accuracy

The deployment of Random Forest and Linear Regression models enables robust predictive performance. Random Forest, through its ensemble of decision trees, effectively captures complex nonlinear interactions among clinical features, while Linear Regression provides interpretable associations between continuous variables and diabetes outcomes. The evaluation metrics accuracy, precision, recall, and ROC curve analysis demonstrate the system's high discriminatory power in classifying patients as diabetic or non-diabetic. Specifically, the use of ROC curves allows for assessment of model sensitivity versus specificity, which is critical in clinical decision-making to minimize false negatives and ensure early intervention.

3.3.2 Data Handling and Preprocessing Advantages

A notable strength of the system lies in its comprehensive data preprocessing strategy. By addressing missing values, outliers, and feature normalization, the system mitigates sources of bias and enhances model stability. The inclusion of feature selection further improves computational efficiency and reduces model overfitting, ensuring that predictions are based on the most clinically relevant variables. This is particularly important in healthcare datasets, which often exhibit heterogeneity and imbalance.

3.3.3 System Scalability and Real-Time Application

The structured workflow facilitates scalability and real-time applicability. New patient data can be input into the system, and the trained models generate immediate predictive outcomes. This real-time capability is particularly valuable in clinical settings, allowing healthcare practitioners to make prompt decisions without the need for extensive manual analysis.

3.3.4 Comparative Advantages Over Traditional Methods

Compared to conventional diagnostic methods, which may rely primarily on periodic clinical tests and physician judgment, the new system offers several advantages as follows:

- 1) **Early detection:** Machine learning models can identify subtle patterns in patient data that may precede overt clinical symptoms.
- 2) **Standardization:** Automated predictions reduce inter-clinician variability and human error.
- 3) **Efficiency:** Rapid processing of large datasets allows for faster decision-making and patient triage.
- 4) **Adaptability:** The system can incorporate additional features, such as lifestyle or genetic data, to improve predictive accuracy over time.

3.3.5 Flowchart Diagram of the New System

The flowchart presents the architectural workflow of the new diabetes prediction system developed using machine learning techniques. The system is designed to support clinical decision-making by enabling early identification and effective management of diabetes through data-driven predictive analytics.

The Flowchart Diagram of the New System

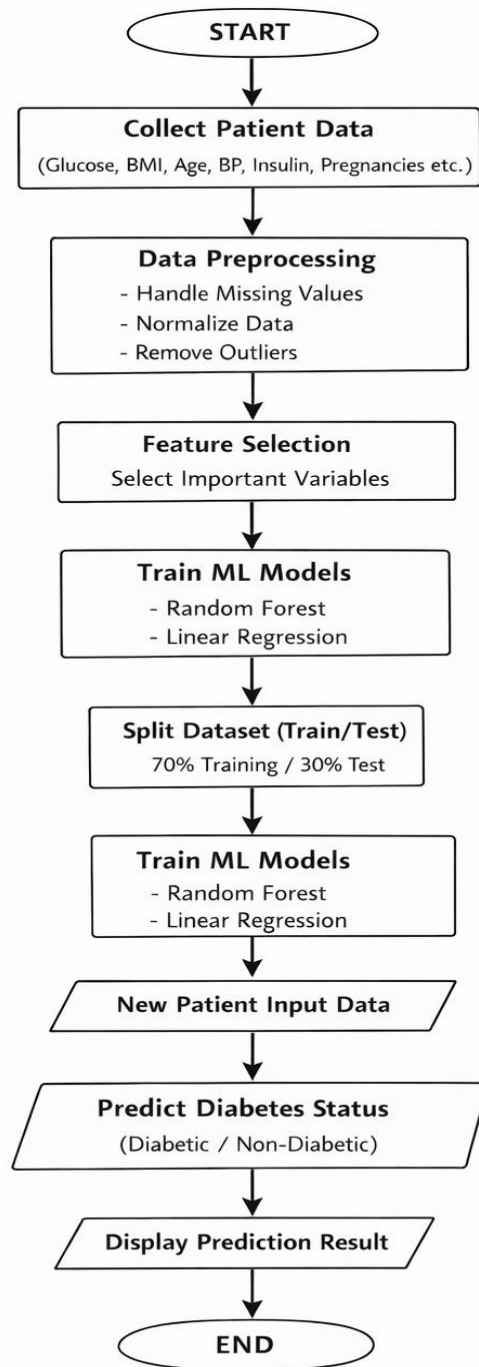


Figure 5: The Flowchart Diagram of the New System indicates shows the performance of the model

3.3.6 Use Case Diagram of the New System

A Use Case diagram is one of the system requirements analysis methodologies adopted to identify, clarify and organize information. It is made up of a set of possible sequences of interactions between systems (use cases) and users in a particular environment and related to a particular goal which the system must accomplish. The use case diagram of the e-health system reveals at a glance all the functionalities of the system and the major users. The use case diagram also illustrates the interactions between users and the diabetes prediction system. It captures the functional requirements of the system, emphasizing who can perform each action and how the system responds. Figure 6 indicates the three key users of the new system and their functions.

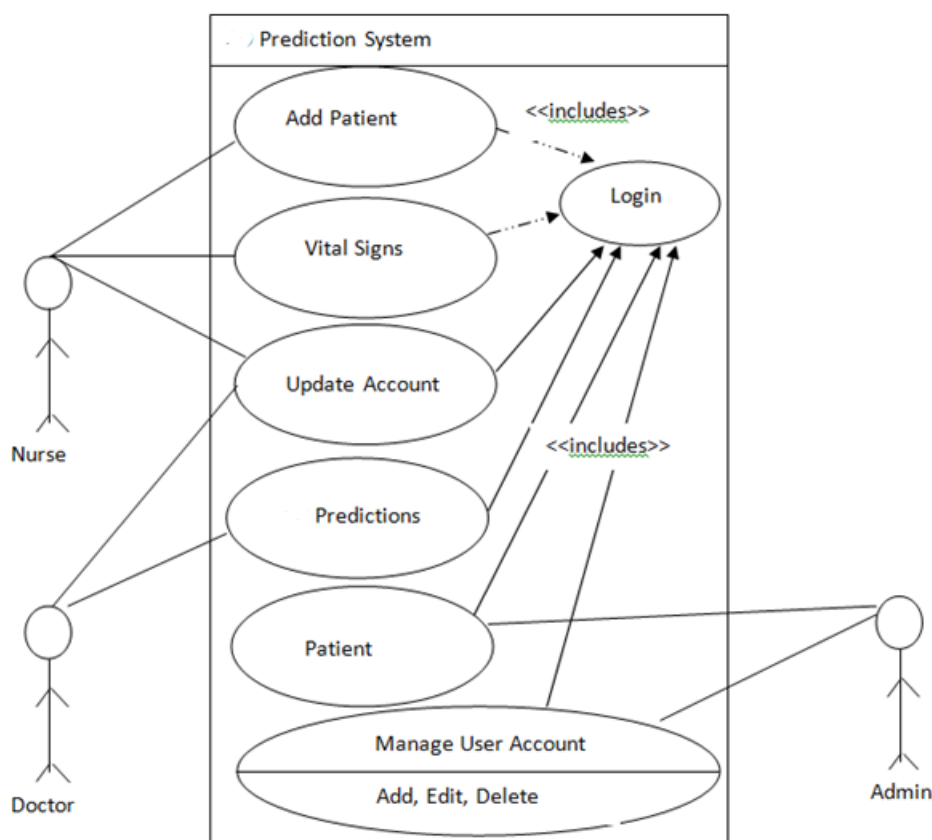


Figure 6: The Use Case Diagram of the New System illustrates the activities of the various users.

3.3.7 Data Flow Diagram of the New System

The DFD represents the flow of information in a Diabetes Prediction System that leverages on machine learning models for early detection of diabetes. It shows how raw patient data moves through preprocessing, predictive modeling, and finally to actionable outputs for healthcare stakeholders. The data flow diagram of the new system consists of external entities, processes, data stores, and data flows.

External Entities: These are sources or recipients of data outside the system's control:

Patient Records: This includes demographic information, family history, lifestyle data, and previous medical test results.

Medical Reports: Laboratory test results (HbA1c, fasting blood sugar, lipid profile, etc.) collected from clinics or hospitals. Significantly These entities provide heterogeneous, multimodal input data critical for accurate diabetes prediction.

Processes: These are the operations performed on data which includes the following

Data Preprocessing: These involves the process of preparing the data for getting appropriate prediction.

Cleaning: Removing missing or inconsistent entries from patient records.

Normalization: Scaling features like glucose levels, BMI, and blood pressure to standard ranges for better model performance.

Integration: Combining sensor data with medical records for a unified patient dataset.

Significantly these ensures that the predictive models receive high-quality, standardized inputs, reducing noise and improving reliability.

Machine Learning Model

It is the application of various algorithms such as Logistics Regression, Random Forest or Support Vector Machines (SVM) to learn patterns from historical patient data. It learns correlations between features (e.g., age, BMI, fasting glucose) and the likelihood of diabetes, the core predictive engine is capable of classifying patients into risk categories or predicting disease onset probabilities.

Prediction Engine: This produces the final risk analysis, generating outputs that classify patients as low, medium, or high risk. That may include confidence scores or probability estimates for better clinical interpretation.

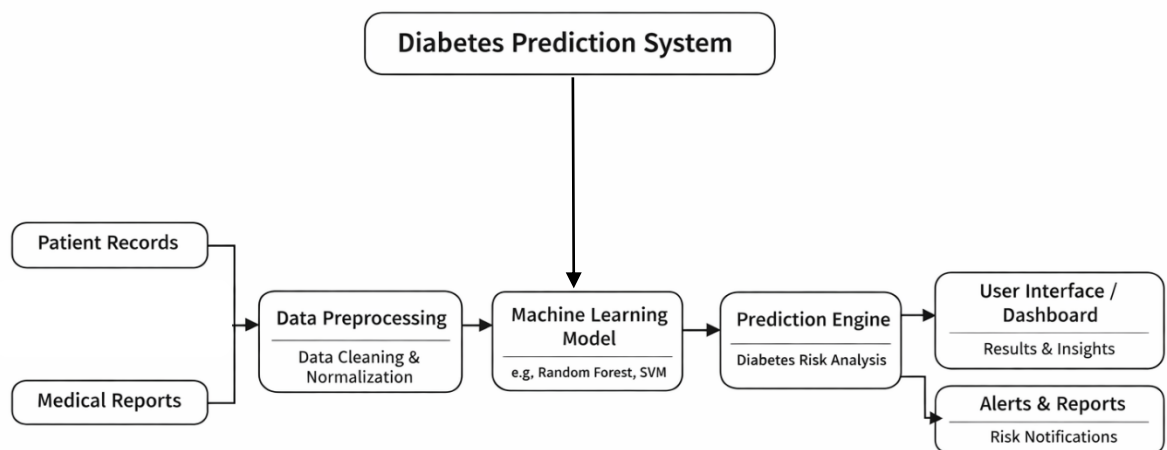


Figure 7: The Data Flow Diagram of the New System shows how data flows in the system

3.4 System Design and Methodology Modularity Using OOP

The system design will utilize Object-Oriented Programming (OOP) principles to ensure modularity, scalability, and maintainability. OOP helps to create a system composed of well defined, independent modules with specific responsibilities, allowing for flexible interactions and easy updates.

3.5 Choice of Methodology

Methodology is a component that includes procedure, techniques, tools and documentation aids which intends to help the developer to develop a system. In this work, the methodology used is the Object-Oriented Analysis and Design Methodology (OOADM) which properly defines the document, the class hierarchy from which all the system objects are created and object interactions are equally defined. Object-Oriented Analysis and Design Methodology is an effective guide to apply to business problems. In development activity which consists of objects, classes, frameworks and interactions. The use of this methodology helps to produce a better quality software product, in terms of documentation standards, acceptability to the user, maintainability and consistency of software. Some basic advantages of using the Object-Oriented Analysis and Design Methodology include;

1. OOADM eradicates the inherent risk of carrying forward incorrect or incomplete analysis into design and construction.
2. Because process decomposition can lead to unstable designs, OOADM ensures that the design is based on processes that do not change as a result of new requirements, error corrections, new environments, or enhancements.
3. OOADM suggest developing software by reusing existing code rather than from scratch. OOADM that was used in the software development. It involved three aspects:
4. Object-Oriented Requirements Analysis (OORA) [Design Analysis]: This is where classes of objects and interaction between them are defined.

5. Object-Oriented Analysis (OOA) [Design Analysis]: It dealt with the design requirements and overall architecture of the system. The researcher focused on describing what the system did in terms of the key objects in the domain.
6. Object-Oriented Design (OOD) [System Design Specification]: This was used to translate the system architecture into programming. The object is the basic modularity; objects are instantiations of a class.
7. Object-Oriented Programming (OOP) Programming and Testing: This aspect of the OOADM was to implement the programming constructs. It emphasizes the employments of objects and methods rather than types or transformation in other programming approaches.

The object-oriented Software Development Life Cycle model is characterized by its effort to model real-world entities into abstract computer software objects and all the interactions that can take place between those objects. The choice of this methodology arose from the fact that it decomposes the system into modules where each module in the system denotes an object or class of an object. This methodology full supports software flexibility, modularity and allows for ease in the modification of any segment of the software.

This methodology involves the process of building a system that is intuitive and easy to manipulate for easy file sharing and collaboration.

3.6 Information Gathering

Information gathering is the act of collecting information from various sources through various means. This segment clarifies various ways in which information and data were gathered and accumulated to help in analyzing the old system so that a plain description of the proposed system can be completed. The sources of data collection used for this project are oral interview, observation and internet.

1. Interview Method

It is a technique which involves collecting facts through discussion. Interviews were conducted with some diabetes patients, health personnel's, according to the respondent vital signs and blood samples are taken from patients' with the test strip gives out the percentage of the blood sugar from the patients.

2. Observation Method

The observation method is another technique used in information gathering, with this method the researcher was able to observe the way diabetes tests are being carried out by various health personnel in Ebonyi State.

3. Internet

The internet is another method of information gathering with this method the researcher was able to know more diabetes, types, classification and treatment.

CHAPTER FOUR

SYSTEM DESIGN AND IMPLEMENTATION

4.1 System Design

System Design is the process of defining elements of a system like modules, architecture, components and their interfaces and data for a system based on the specified requirements. It is the process of defining, developing and designing systems which satisfies the specific needs and requirements of a business or organization. The design methods adopted for this new system are divided into three namely:

- (i) Logical Design: This is used to represent the, inputs and outputs and data flow of the system. Example: Entity Relationship Diagrams (ER Diagrams) is an example of Logical Design
- (ii) Architectural Design: This is used to describe the structure models, views and behaviour of the system.
- (iii) Physical Design: The physical design helps to show how users add information to the system and how the system represents information back to the user; to demonstrate how the data is modeled and stored within the system; and how data moves through the system, how data is validated, Secure and/or transformed as it flows through and out of the system.

4.2 Design Objective and Goals

The main system design objective is to design a diabetes prediction models using Logistic regression and Random Forest ML algorithms modified for individuals in Ebonyi State in order to assist hospitals provide better e-health services, help in collecting and storing patient's medical history and serve as reference point to doctors and other healthcare providers in offering healthcare services to patients. That is to design an e-Health system that will save space and enhance the security of patients' health records and facilitate a timely and

efficient healthcare delivery services to patients who visit the hospitals for their healthcare needs.

Below is the list of our design goals for the e-health system:

- i. **Security:** Our application must protect the confidentiality and integrity of stored patients' health records against cloud providers and unauthorized end-users. This means that the stored health records' data should be readable to only authorized users and any unauthorized change to the data should be prevented or detectable.
- ii. **Privacy-preserving:** This means that the rights of an authorized user as well as his usage trends should not be visible to other users or the cloud service providers.
- iii. **Efficiency and Scalability:** The application should run efficiently and easily available at any given time. Also, the complexity of operations should be independent of number of the patients' records and users in the system. This ensures that the application will not affect the scalability of existing cloud services provided by the cloud providers.
- iv. **Flexibility:** The application should allow application administrators and end-users to organize and manage their profiles in accordance with their access rights. Also, they should be able to grant/revoke part of their access rights to/from other users in a decentralized and scalable manner.
- v. **Simplicity and Extensibility:** The application should be simple enough to be efficiently implementable on top of existing commercial cloud API which can be linked to other healthcare institution's database.

4.3 Design of the Machine Learning Model

The design of our machine learning model consist of different stages which include dataset, data transformation, feature extraction, model building which consists of training data and testing data, model evaluation. The first phrase of our model is sourcing the experimental data from Kaggle ML repository online site with oral data from hospital in Abakaliki. This was followed by data transformation which is the process of converting raw data into a format or structure that is suitable for model building using the Load, transform and extract (LTE). Feature Extraction is next. Feature Extraction is a process of extraction and generation of relevant database features to assist the task of object patterns classification. The next phase is model building. The model building consists of training the dataset and dataset testing to pick up data pattern from the training dataset which is in this case is intrusion detection. This is followed by the model evaluation to determine the performance of our selected ML algorithms used in this research work for intrusion detection.

The last phase of the model design is the model deployment phase in which the developed ML model is deployed in a form that it can be used by clinicians for diagnosis of diabetes. The design architecture showing the following stages of our ML based diabetes detection model is shown in Figure 8.

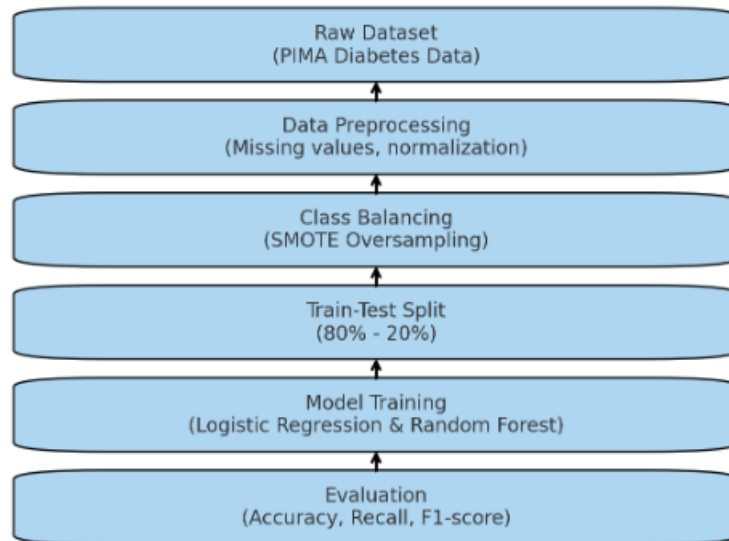


Figure 8: Shows the Model Pipeline Diabetes Prediction Model

The following processes are included in our ML system design: dataset, data transformation, feature extraction, training data poisoning, model training, model building and model evaluation.

4.3.1. Dataset: The experimental data was sourced from kaggle ML repository site by Prama *et al.*, (2025), titled “DiaBD: A Diabetes Dataset for Enhanced Risk Analysis and Research in Bangladesh,” and made available on "<https://data.mendeley.com/datasets/m8cgwxs9s6/2>." The dataset contains 5,288 patient records, covering 14 independent attributes related to demographics, clinical parameters, and medical history. Key features include age, gender, pulse rate, blood pressure (systolic and diastolic), glucose level, BMI, and family history of diabetes, hypertension, and cardiovascular disease. Each patient entry is labeled with a binary diabetes status (diabetic or non-diabetic), making it suitable for predictive modeling and risk assessment. The table 3 below shows the demography of the patient’s information, showing the gender and the vital information of dataset of the model indicating the binary diabetes status (diabetic or non-diabetic)

Table 3: Show the Head of diabetes dataset of the model indicating the binary diabetes status (diabetic or non-diabetic).

Gender	Pulse_rate	Systolic_b	Diastolic_b	Glucose	Height	Weight	BMI	Family_dia	Hypertensi	Family_hy	Cardiovasc	Stroke	Diabetic
Female	6	110	73	5.8	1.6	70.	25.	0	0	0	0	0	No
	6			8	5	2	75						
Female	6	125	68	5.7	1.4	42.	19.	0	0	0	0	0	No
	0			1	7	5	58						
Female	5	127	74	6.8	1.5	47.	20.	0	0	0	0	0	No
	7			5	2	0	24						
Male	5	193	112	6.2	1.6	57.	21.	0	0	0	0	0	No
	5			8	3	4	72						
Female	7	150	81	5.7	1.4	36.	17.	0	0	0	0	0	No
	1			1	2	0	79						

Table 4: Shows the head of the transformed data of diabetic gender, Male or Female” and “Yes or No” respectively were coded into 1 for Male, 0 for Female and 1 for diabetic and 0 for non-diabetic respectively.

Gender of the Diabetic Patients		
0	0	0
1	0	0
2	0	0
3	1	0
4	0	0
5	1	0
6	0	0
7	0	0
8	1	1
9	1	0

4.3.2. Data Transformation: This is the advent of converting raw data into a format or structure that is suitable for model building using the LTE (Load, transform and extract). The transformation stage help normalizes numeric features to produce better model and allows regression techniques like the Random Forest (RF). Non numeric columns like “gender” and “diabetic” that had the values “Male or Female” and “Yes or No” respectively were coded into 1 for Male, 0 for Female and 1 for diabetic and 0 for non-diabetic respectively, as shown in table 4 above.

4.3.3. Feature Extraction and Data Exploration: This is a mandatory stage for any application that involving relevant dataset data exploration and feature identification. Feature Extraction is a process of extraction and generation of features to assist the task of object patterns classification. This phase is critical because the quality of the features influences the classification and regression task in adopting machine learning regression and classification tasks. The extended set of features is stored as a vector called the feature vector at the implementation phase. The regression class and classification takes the feature vector as input and performs the regression and classification.

For this, we checked the correlation matrix of variables as well as chi-square test of independence because they were categorical variables, and ANOVA/F-test for variables with continuous values. Those with p-values within acceptable significant rate were classified as “most predictive features” and were used to build the model. Table 3 shows the classification of variables after data exploration and feature extraction.

Table 5: Classification of variables (Features) based on predictive power after dataset exploration and feature extraction

Most Predictive Features	Moderately Predictive Features	Weak/Insignificant features
Hypertensive	BMI	Gender
Glucose	Pulse_Rate	Family_Hypertension
Systolic_BP	Cardiovascular_Disease	Height
Diastolic_BP	Stroke	-
Age	-	-
Weight	-	-

Table 5: Shows the Hypertensive, Glucose, Systolic_BP, Diastolic_BP, Age and Weight are variables that are strongly associated with the dataset. They were the variables used to train the models.

4.3.4. Model Building

We created RF and Logistic Regression machine learning algorithms that can pick up delicate data from the training set, including the particular medical histories of different patients, and predict their risk level when it comes to diabetes. The model was built using Python programming language and a medical dataset obtained from Kaggle repository site which was divided into 80 percent training dataset and 20 percent testing dataset with fine-tuned hyper-parameter values and Synthetic Minority Over-sampling Technique (SMOTE) data augmentation method used to balance the dataset in order to improve the models for better result. In order to encode the correlations in the data, machine learning algorithms work by evaluating a large amount of data and adjusting their parameters. In this study, these machine learning models' parameters are used to encode generic patterns.

- I. **The Random Forest Model:** The random forest (RF) is an ensemble learning technique that uses the assembly of different decision tree in making predictions with samples of observed features. We built multiple decision trees called the Random forest using classification and regression approach. A single decision-tree has high variance and the collection of several decision-trees is called the random forest which helps solve the problem of high variance and overfitting data. Because each and every DT within the forest is perfectly trained using the sample data which depend mainly on multiple trees but not a single tree. The result of random forest classification is always-based on majority voting while the outcome of regression problem is the mean of all the tree predictions as output within the forest. In the random forest most of the trees are producing correct predictions while some are making mistakes. For this reason, voting is required to be carried out based on the classifications for the observed result poll and expected the outcome to be closer to the correct classification (Kremic and Subasi, 2016). We used more high-quality data with adjusted hyper-parameter values for both models which can

define the number of trees in the RF to improve and produce a more generalized and better result. A single decision tree will always result to the problem of low bias and high variance. Therefore; we adopted RF trees to convert the low bias and high variance of a single decision-tree to have a low variance.

II. Implementation of the Random Forest Model: A Random Forest Classifier was invoked from the sklearn ensemble library in Python and an RF model was created. The classifier was initialized with a fixed random state of 42 for reproducibility using the Python script: “rf_clf = RandomForestClassifier(random_state=42)”

The model was trained on the balanced training dataset with the script: “rf_clf.fit(X_resampled, y_resampled).: And predictions were then generated on the test dataset using: “y_pred_rf = rf_clf.predict(X_test).”

III. The Logistic Regression Model: The Logistic Regression (LR) model is a supervised machine learning algorithm that is widely used for binary classification problems. Unlike linear regression which predicts continuous outcomes, Logistic Regression maps input features to a probability score between 0 and 1 using the logistic (sigmoid) function. The decision boundary is then applied to classify the outcome into two categories, for example, diabetic or non-diabetic patients. Logistic Regression works by estimating the weights of features through maximum likelihood estimation, allowing it to capture the strength and direction of the relationship between predictors and the target class. In this work, the Logistic Regression model was applied to predict the likelihood of diabetes based on clinical and demographic variables. The model was trained on pre-processed and balanced data to overcome the challenge of class imbalance, ensuring reliable predictions. Logistic Regression is particularly favored because it is simple, interpretable, and computationally efficient, while also performing effectively when the relationship between independent variables and the dependent variable is approximately linear in the

log-odds space (Hosmer, Lemeshow and Sturdivant, 2013). Unlike ensemble methods such as Random Forest, Logistic Regression provides direct insights into how individual variables contribute to the prediction, which is essential in clinical settings where transparency of decision-making is critical. By adopting Logistic Regression, we ensured that the model output could be easily explained to healthcare providers, thereby improving trust in the prediction system and supporting informed decision-making.

IV. Implementation of the Logistic Regression Model: A Logistic Regression classifier was invoked from the `sklearn.linear_model` library in Python and a Logistic Regression model was created. The maximum number of iterations was set to 1000 and a random state of 42 was used with the Python script:” `log_reg = LogisticRegression(max_iter=1000, random_state=42)`”

The model was trained using the pre-processed and balanced training dataset with the script: `log_reg.fit(X_resampled, y_resampled)`. And Predictions were made on the test dataset with: `“y_pred_log = log_reg.predict(X_test)”`

4.3.5. Model Training: In this research, two machine learning models Logistic Regression (LR) and Random Forest (RF) were implemented for the prediction and diagnosis of diabetes. Both models were trained using the PIMA Indian Diabetes dataset, which was pre-processed to handle missing values, normalize features, and balance class distribution. A key step in our training pipeline was the application of Synthetic Minority Oversampling Technique (SMOTE) to address the challenge of class imbalance. Since the number of diabetic cases was much fewer than non-diabetic cases, training on the raw dataset would bias the models towards predicting the majority class. SMOTE works by generating synthetic examples of the minority class using nearest-neighbor interpolation, thereby improving the sensitivity of the models to detect true diabetic cases.

Logistic Regression is a linear probabilistic model used for binary classification tasks. The model predicts the probability that a patient belongs to the diabetic class ($y=1$) given a set of independent features X . The hypothesis function is defined in equation (1)

$$P(y = 1|X) = \sigma(w^T X + b) = \frac{1}{1 + e^{-(w^T X + b)}} \quad (1)$$

where:

- $X = (x_1, x_2, \dots, x_n)$ represents patient features (e.g., glucose, BMI, blood pressure),
- w is the weight vector,
- b is the bias term, and
- $\sigma(z)$ is the sigmoid activation function.

The model parameters (w, b) are optimized by minimizing the binary cross-entropy loss; and are contained in equation (2).

$$L(w, b) = -\frac{1}{m} \sum_{i=1}^m \left[y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right] \quad (2)$$

where $\hat{y}^{(i)}$ is the predicted probability for the i th instance.

In this work, Logistic Regression was selected due to its interpretability, allowing medical practitioners to understand how each variable contributes to diabetes risk. The model was trained on the SMOTE-balanced dataset and evaluated on a held-out test set to validate its generalization.

Random Forest is an ensemble learning technique that builds multiple decision trees and aggregates their results to improve prediction performance. For classification tasks such as diabetes diagnosis, the final prediction is obtained by majority voting across the decision trees. Each decision tree partitions the feature space based on splitting criteria such as Gini impurity or entropy, and the ensemble reduces variance, thus mitigating overfitting.

Formally, given T trees, the Random Forest classifier prediction is shown in equation (3)

$$\hat{y} = \text{mode}\{h_t(X), t = 1, 2, \dots, T\} \quad (3)$$

where $h_t(X)$ represents the prediction from the t^{th} decision tree. In our implementation, the number of trees (n_estimators) and maximum depth of trees were optimized through cross-validation. RF was chosen because of its ability to capture non-linear interactions between features (e.g., high glucose combined with high BMI increases diabetes risk more strongly than either alone), and because it generally achieves higher accuracy compared to linear models.

A key step in both model trainings was the use of SMOTE. Without class balancing, models tended to predict “non-diabetic” for most cases, achieving misleadingly high accuracy but poor recall on diabetic patients. By oversampling the minority class synthetically, SMOTE improved the recall and F1-score, which are critical in medical diagnosis where failing to detect a diabetic patient can have severe consequences.

Thus, both Logistic Regression and Random Forest were trained on the balanced dataset to provide a fair and robust decision support system for healthcare practitioners in diagnosing diabetes.

4.3.6. Evaluation of the ML Algorithms: This is the final step to effectively measure the performance of our proposed random forest and logistic regression techniques in comparison to the existing system. We adopted various metrics to evaluate the performance of our chosen ML algorithms based on testing datasets. We use the confusion matrix, Visualization accuracy, receivers operating curve (ROC) and area under the curve (AUC) to adequately measure the performance of the proposed system for comparison. ROC is an excellent evaluation metric to measure the trade-off between sensitivity and specificity. AUC is the metric used to show the validity of the accuracy results. In order to calculate the evaluation metrics, the first step is the calculation of the values of the confusion matrix. The confusion matrix is generated when a trained ML model is used to classify the instances of a testing dataset. The confusion matrix compares values regarding the actual labels of the instances of the testing dataset and the corresponding labels predicted by the ML model.

The true positive (TP) and true negative (TN) relate to the correctly classified attack instances and normal instances, respectively. The false positive (FP) and false negative (FN) refer to the incorrectly classified normal instances and attacks instances, respectively. Based on these values, it is possible to compute several evaluation metrics.

Accuracy: shows the overall success of the model by comparing the amount of the correctly classified attack and normal instances to the total amount of instances, as given in equation (4)

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

Precision: estimates the overall effectiveness of the model by calculating the percentage that an observation recognized as an attack is actually an attack observation, and this is expressed in equation (5).

$$\text{Precision} = TP / (TP + FP) \quad (5)$$

_ **Recall**: shows the overall success of the model by computing the percentage that an actual attack observation is correctly classified, and is clearly defined I equation (6)

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (6)$$

_ **F1-score**: is calculated by the precision and recall metrics as their harmonious mean.

It is a statistical function for estimating the accuracy of the model. As the precision and recall of a model approach the value of 100%, the F1-score and accuracy are maximized, and every instance is classified correctly, and can be seen in equation (7).

$$\text{F1-score} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

$$\text{F1-score} = 2 * \text{TP} / (2 * \text{TP} + \text{FP} + \text{FN}) \quad (7)$$

Where TN represents true negative, FP is false positive, TP is true positive and FN is false negative cases.

4.3.7 Algorithms

Algorithm 1- Random Forest Model Algorithm

1. Start
2. Pre-process the dataset to remove inconsistencies and handle missing values.
3. Apply SMOTE to resolve class imbalance.
4. Divide the dataset into training and testing sets.
5. Generate multiple bootstrapped samples from the training dataset.
6. For each bootstrapped sample, construct a decision tree.
7. At each node of the tree, select a random subset of features and determine the best split.
 Grow the decision tree until the maximum depth or stopping criteria is reached.
8. Repeat the process to build many trees and form the random forest. For prediction, input patient data into each tree in the forest.
9. Collect predictions from all trees and determine the final output by majority voting.
10. Stop

Algorithm 2: Logistic Regression Model Algorithm

1. Start
2. Pre-process the dataset to remove noise and handle missing values.
3. Apply SMOTE to balance the dataset between diabetic and non-diabetic cases.
4. Split the dataset into training and testing sets.
5. Initialize model parameters (weights and bias).
6. Apply the logistic function to map input features into probabilities.
7. Compute the error using a loss function that compares predictions with actual outcomes.
8. Update model parameters iteratively using optimization until convergence is achieved.
9. After training, input patient data into the model.
10. Compare the probability output to a fixed threshold.
11. Classify the patient as diabetic if the probability exceeds the threshold; otherwise, classify as non-diabetic.
12. Stop

4.4. Results and Discussion

The results of our developed model are discussed subsequently. We start the discussion on the training and testing dataset used, and concludes the discussion with the accuracy of the ML algorithms used for the prediction of diabetes.

i. Training and testing Datasets

```
df.head()
(5288, 15)
```

```
print(X_train.shape, X_test.shape)
(4230, 6) (1058, 6)
```

Figure 9: Shows the training and testing of the Datasets of the Model

Figure 9 Depicts the total dataset used in this work, as well as the number used for training and testing. The dataset was divided into 80% $((80 \times 5288) / 100 = 4230)$ training and 20% $((20 \times 5288) / 100 = 1058)$ testing sets amountiing to total of 5288 items. That means that 80% of the total dataset was used for training and 20% of the total dataset was used for testing. The train part consisted of 80% of the dataset and the ML algorithms were trained and evaluated with this part. On the other hand, the test part consisted of 20% of the dataset and was held back for further evaluation of the models with unseen data. The percentage was splitted into 80% for training and 20% for testing that was used to determine the best ratio to avoid the problem of overfitting.

ii. The Correlation Heatmap plot of data features is indicated in Figure 12.

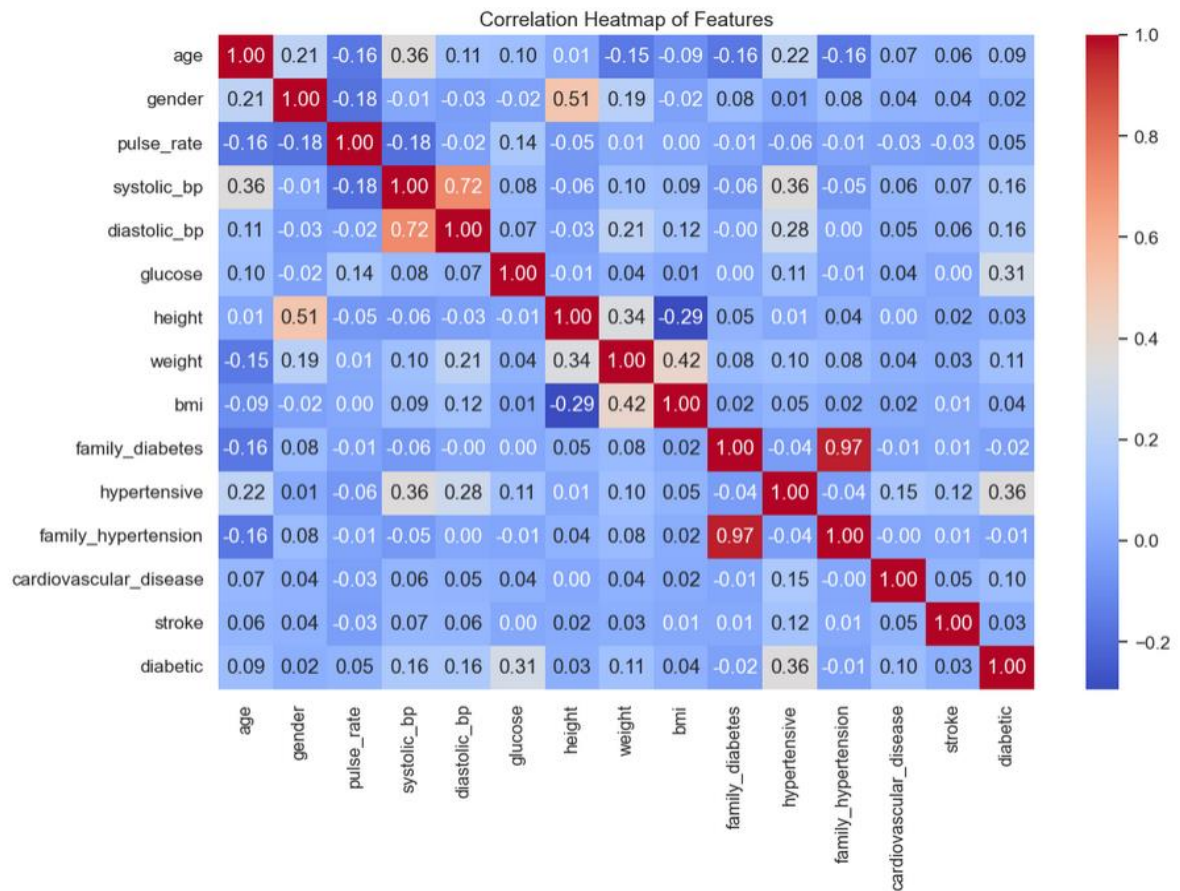


Figure 10: Shows the Correlation heatmap of data features, that depicts the correlation heatmap plot, which demonstrates that there are not too many medical data elements that are significantly related with each other.

iv. Chi-Square and P-values of variables

Table 6: Shows the Data exploration results of variables used in the dataset

	Feature	Chi2 Score	p-value
10	hypertensive	600.598188	1.240762e-132
5	glucose	565.966580	4.233499e-125
3	systolic_bp	473.940925	4.448382e-105
4	diastolic_bp	243.439286	6.995349e-55
0	age	169.302741	1.050647e-38
7	weight	116.437229	3.812573e-27
12	cardiovascular_disease	47.422784	5.721381e-12
8	bmi	34.133013	5.147059e-09
2	pulse_rate	26.560288	2.554408e-07
13	stroke	6.054669	1.386957e-02
9	family_diabetes	2.377471	1.230967e-01
1	gender	2.312896	1.283048e-01
11	family_hypertension	1.145558	2.844809e-01
6	height	0.021069	8.845906e-01

The feature selection process revealed that certain variables exhibited a stronger statistical association with diabetes than others. The most predictive features with highly significant scores ($p \approx 0$) included hypertension status, glucose level, systolic blood pressure, diastolic blood pressure, age, and weight. These variables were therefore identified as the core predictors of diabetes in this study, given their direct clinical and statistical relevance. Moderately predictive features, such as body mass index (BMI), pulse rate, cardiovascular disease, and stroke history, also contributed to the model, though their impact was less pronounced. In contrast, features such as family history of diabetes, gender, family history of hypertension, and height were found to be weak or statistically insignificant ($p > 0.05$) in

predicting diabetes within the dataset. Based on these findings, the model was primarily built on the core features - hypertensive status, glucose, systolic and diastolic blood pressure, age, and weight - to ensure optimal predictive accuracy.

iv. Confusion Matrix of Random Forest

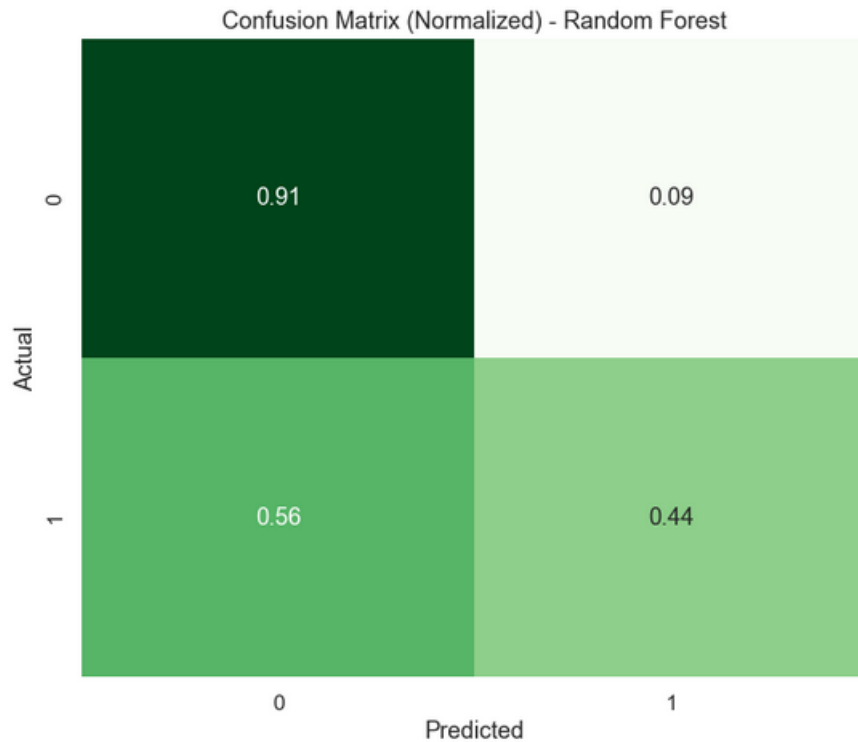


Figure 11: Shows the Confusion Matrix of Random Forest

Figure 11 depicts the random forest model's confusion matrix at the testing stage, with correct predictions presented at the secondary diagonal and incorrectly predicted values recorded above and below the main diagonal.

v. **Confusion matrix of Logistic Regression Model**

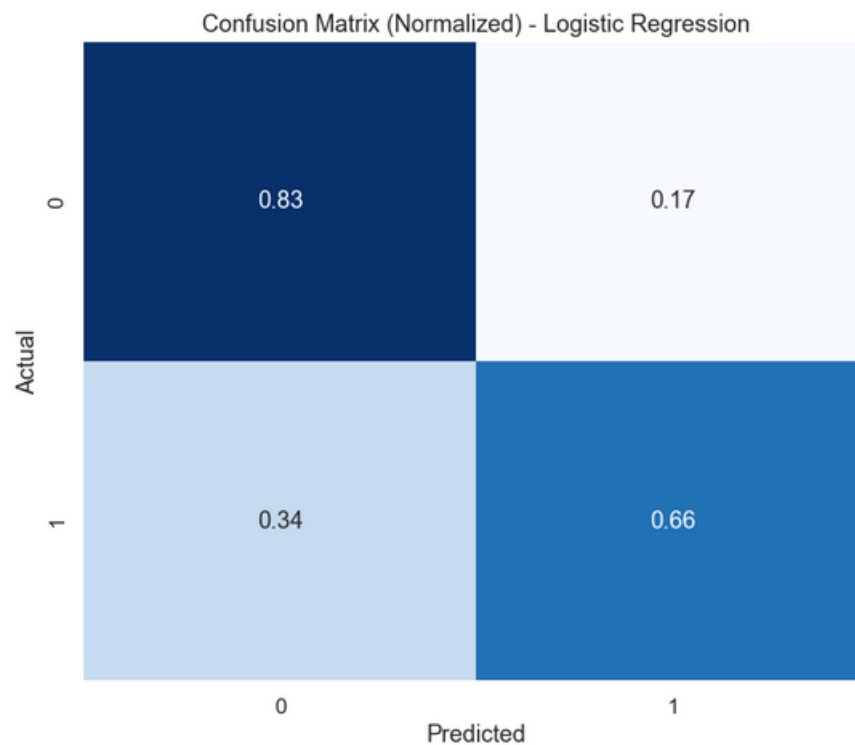


Figure 12: Confusion matrix of Logistics Regression model

Figure 12 is the confusion matrix of the Logistics Regression model, with the leading diagonal elements or values indicating the total number of properly predicted values that are equal to the actual or true values, and the off-diagonal components indicating the incorrectly predicted values.

vi. Precision-Recall Curve of both models

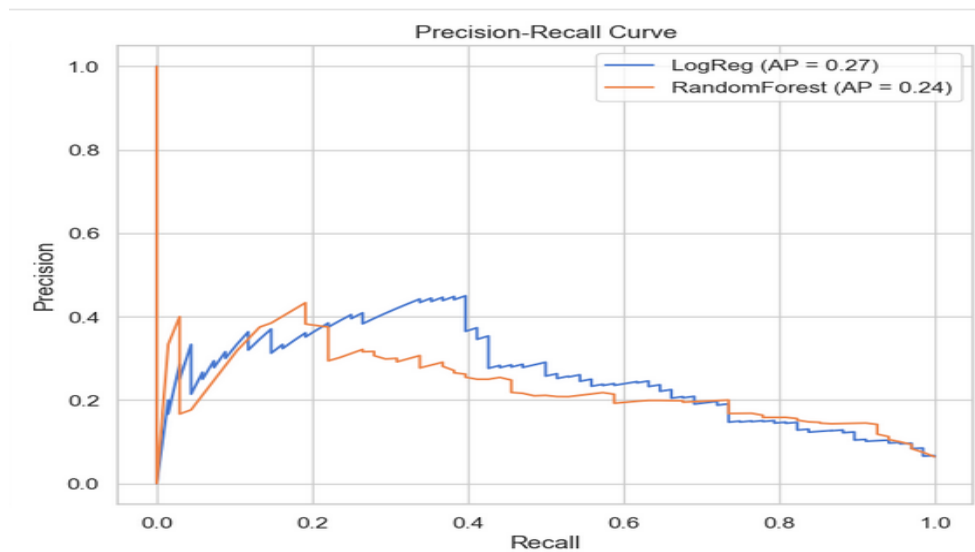


Figure 13 shows the precision- recall curve of the logistics regression and random forest machine learning model.

vii ROC Curve of both models

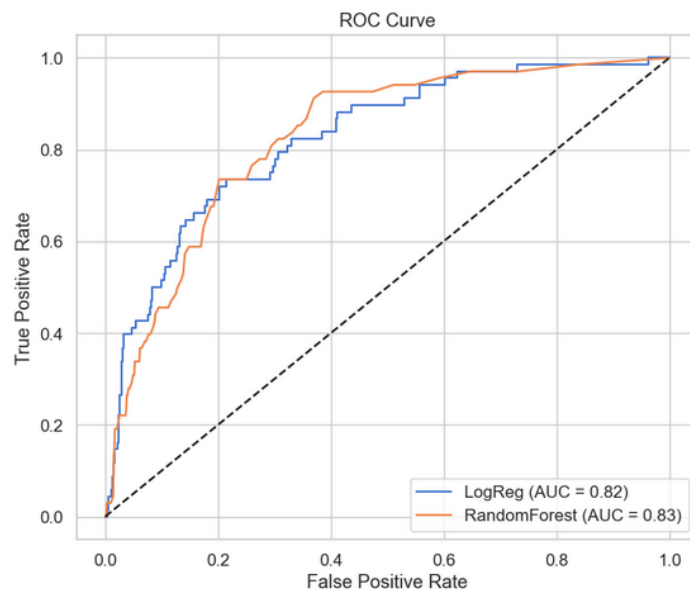


Figure 14: Depicts the Precision-Recall Curve of RF and LR Models which represents the precision and recall curve of the RF and LR models.

Figure 14: ROC Curve of LR and RF models: Shows the ROC Curve of LR and RF models that represents the ROC Curve of LR and RF models.

vii. RF and LR Classification Report: This is the classification of Random Forest and Logistics Regression report of the new system.

Table 7: RF and LR classification report of the random forest and linear regression algorithm

Model	Accuracy	Recall for Diabetics	Precision for Diabetics	Support
<i>RF</i>	<i>0.88</i>	<i>0.44</i>	<i>0.25</i>	<i>1058</i>
<i>LR</i>	<i>0.81</i>	<i>0.66</i>	<i>0.21</i>	<i>1058</i>

This table shows the RF and LR classifier classification report, which includes the precision, recall, support and accuracy of predictions.

The performance evaluation of the two models revealed important trade-offs between accuracy, recall, and precision. The Logistic Regression model achieved an overall accuracy of approximately 81%, with a recall of 0.66 for diabetic patients, meaning it was able to correctly identify about two-thirds of actual diabetic cases.

However, its precision was relatively low at 0.21, indicating a higher rate of false positives. On the other hand, the Random Forest model demonstrated a higher overall accuracy of around 88% and slightly improved precision (0.25) compared to Logistic Regression, but its recall dropped to 0.44, meaning that it missed a larger proportion of true diabetic cases. This highlights a clear trade-off: while Logistic Regression is more effective at catching diabetic patients (higher recall), it does so at the expense of generating more false alarms (lower precision). Random Forest, in contrast, is more balanced in accuracy and precision but risks overlooking many true cases of diabetes.

In the context of medical screening and early diagnosis, where the priority is to identify as many diabetic patients as possible even at the cost of false positives, Logistic Regression emerges as the safer and more appropriate choice.

Nevertheless, a threshold-tuning strategy was incorporated into the Logistic Regression model to better balance recall and precision, thereby optimizing its performance for clinical application.

4.5 Justification for Choosing Python

The justification for choosing Python Programming language for the implementation of our ML model is briefly explained.

4.5.1. Readable and Maintainable Code

Python programming language was chosen because its code is simple to maintain and update. The syntax rules of Python allow expressing concepts without writing additional codes. Python, unlike other programming languages allows for the use of English like keywords instead of punctuations and it emphasizes on code readability

4.5.2. Multiple Programming Paradigms

Python supports various programming language features such as object oriented and structured programming paradigm and functional aspect-oriented programming. It also has dynamic type system and automatic memory management features which makes it a language of choice for various programming projects.

4.5.3. Compatible with Major Platforms and Systems

Python supports different operating systems. With Python, one can run the same code on multiple platforms without recompilation because of its support for different operating systems. This feature makes it easier to make changes to the code without increasing development time.

4.5.4. Robust Standard Library

It has a large and robust standard available library features. It allows one to choose from a wide range of modules from the standard library according to your needs.

4.5.5. Many Open Source Frameworks and Tools

Python is an open source programming language which helps curtail software development cost significantly for its users.

4.5.6. Simplify Complex Software Development

Python is suitable for developing both web applications and desktop applications. It can be used to develop complex scientific and numeric applications. Python is designed with features to facilitate data analysis and visualization. Python is also suitable to easily complete Machine Learning (ML), Artificial Intelligence (AI), Data Mining, Big Data and Natural Language Processing (NLP) tasks.

4.6 Model Deployment

In order to make the diabetes prediction model usable in a real clinical setting, the system was deployed as a framework application. The implementation was carried out using Tkinter for the graphical user interface (GUI), SQLite for the local database, and Joblib for loading the pre-trained Logistic Regression model. Finally, the system was packaged into a standalone executable using PyInstaller, ensuring it can run on hospital computers without requiring a Python environment.

The desktop application was designed to provide two levels of access: Nurses and Doctors. Nurses are responsible for entering patient details and vital signs, while doctors are able to view patient records and perform diabetes predictions. This design ensures role separation, data security, and ease of use within the hospital workflow.

4.6.1 Output Design

The output of the application provides the prediction results after processing patient data. When a doctor selects a patient and inputs the required variables, the Logistic Regression model returns either “High Risk of Diabetes” or “Low Risk of Diabetes.” The results are displayed directly on the GUI in a clear format for quick interpretation by medical staff. Additionally, prediction results are stored in the database for future reference.

4.6.2 Input Design

The input interface of the desktop application was developed using Tkinter forms. The major input components are:

- i) **The login form:** The login window as depicted in Figure 17 requires registered users (nurses or doctors) to authenticate with their username and password. Access rights are based on role: nurses can enter patient data, while doctors can run predictions.

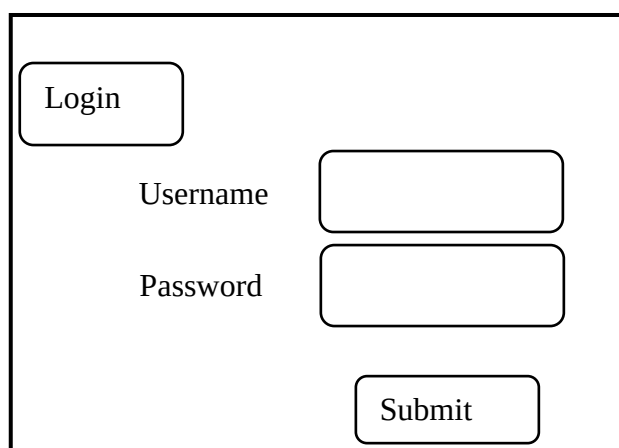


Figure 15: Shows the interface or Login Form where the users inputs their username and password to the system.

- ii) **Patient data entry form:** This form allows nurses to register new patients by entering demographic details such as name, age, sex, and patient ID.
- iii) **Vitals entry form:** Nurses can also enter patient health measurements including glucose level, blood pressure, BMI, and other variables required for the prediction model.
- iv) **Prediction Form:** Doctors access this form as seen in Figure 15 to select a patient, input the prediction variables, and run the Logistic Regression model. Once the “Predict” button is clicked, the system displays the result (High Risk or Low Risk).

Prediction

Hypertensive

Glucose

Systolic BP

Diastolic BP

Age

Weight

Figure 16: Shows the Prediction form design of the system, where the users enters their vital signs for prediction.

4.6.3 Database Design and Structure

The database was implemented using SQLite to ensure lightweight storage and easy integration with the Tkinter desktop application. The database is composed of three major tables:

- i. Users Table – stores login credentials and role information (nurse or doctor).
- ii. Patients Table – stores patient demographic details (patient ID, name, age, sex).
- iii. Vitals Table – stores patient health measurements linked to the patient ID.

Table 8: The Users Table Structure of the new System indicating the field name, data type and field size.

Field Name	Data Type	Field Size
<i>User_ID</i>	<i>Integer</i>	<i>10</i>
<i>Username</i>	<i>Varchar</i>	<i>50</i>
<i>Password</i>	<i>Varchar</i>	<i>50</i>
<i>Role</i>	<i>Varchar</i>	<i>20</i>

This table shows the Users Table Structure of the new System that contained how the data was stored in the system database.

Table 9: The Patients Table Structure

Field Name	Data Type	Field Size
Patient_ID	Integer	10
Name	Varchar	100
Age	Integer	3
Sex	Varchar	10

Table 10: The Vitals Table Structure of the Patients

Field Name	Data Type	Field Size
Record_ID	Integer	10
Patient_ID	Integer	50
Glucose	Integer	50
BMI	Float	10
Blood Pressure	Integer	50
Hypertension Status	Check Box	2

Tables 8, 9 and 10 shows the structure of the users, patients and vitals tables respectively. The workflow of the system begins with a user login, followed by either patient registration and vitals entry (nurse) or prediction and report generation (doctor). Predictions are carried out using the pre-trained Logistic Regression model, which was saved using Joblib and integrated into the system. Finally, the entire application was packaged with PyInstaller into a single executable file. This ensures that hospitals can deploy the system directly on their computers without additional setup, making it easy to use, portable, and secured.

4.7 Unified Modeling Language (UML) Diagram

The Unified Modeling Language (UML) is a set of modeling conventions used to specify or described a software system in terms of objects. We used Use Case to depict the actors (users) of the system and what they can do. The figure 17 illustrates the unified model diagram of the new system.

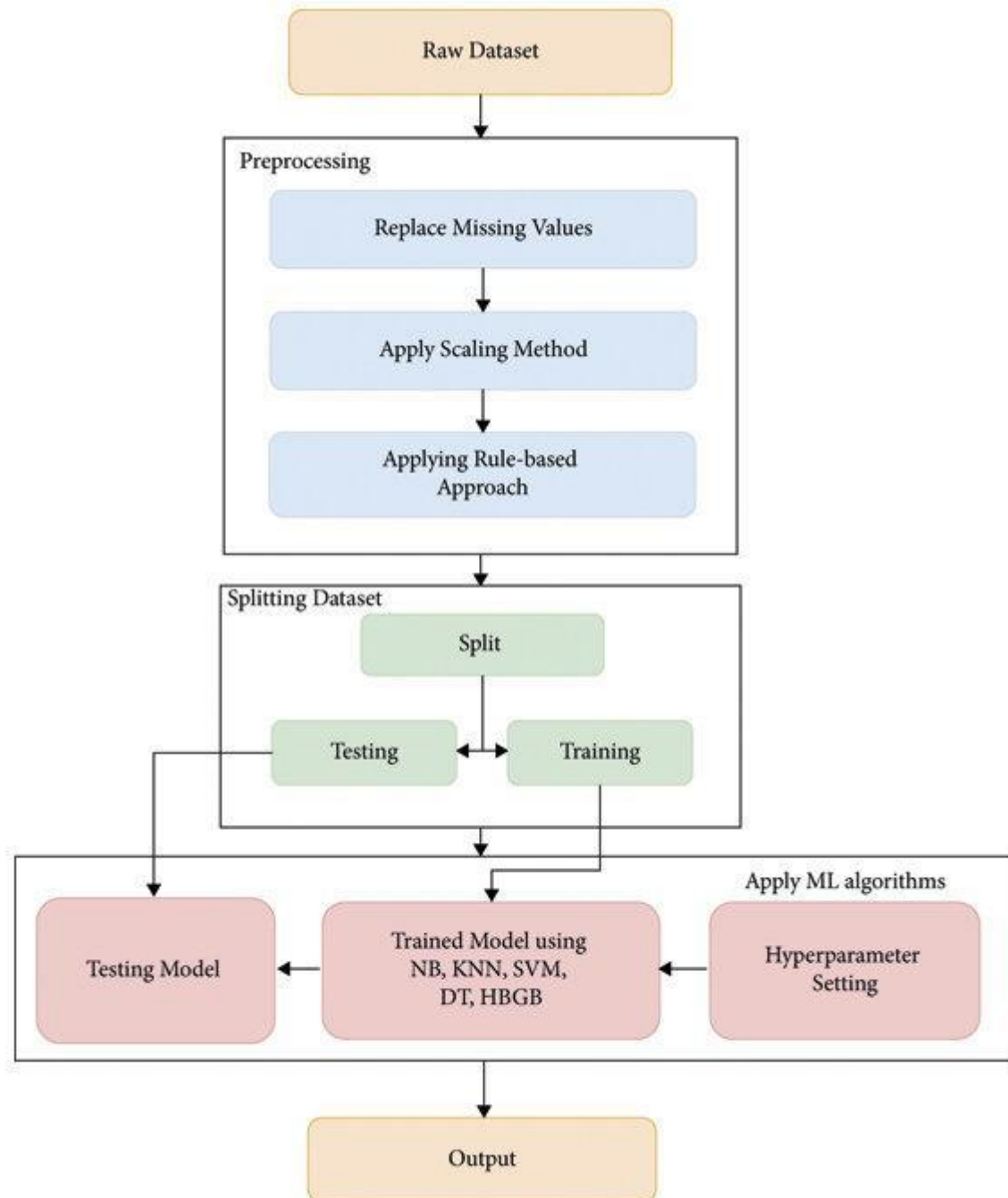


Figure 17: Shows the Unified Modeling Language (UML) Diagram of the new system

4.8 System Requirements

In order to ensure that a system works effectively, it needs to meet up certain hardware components and software resources to operate. The essential hardware devices as well as software platform required for the efficient running of the system are outlined below:

4.8.1 Hardware Requirements

For proper access of the Diabetes prediction system, computer systems having the following minimum hardware system configuration are required:

Table 11: The Computer Hardware and System Requirements Description

HARDWARE	SIZE
HDD (Dis k space)	320 GB and above
RAM (Memory)	CORE TM i Series Processors with processing speed of at least 4 Ghz
Processor	2 GB and above
Monitor	1024 x 718 display resolution screen

This table shows the computer hardware requirements and system description or specification of the new system.

4.8.2 Software Requirements

Computer systems with the following software components are needed for effective utilization of the new system:

1. Windows O.S
2. Microsoft C++ redistributable

CHAPTER FIVE

SYSTEM TEST, INTEGRATION AND DEPLOYMENT

5.1 System Testing

The diabetes prediction system was implemented using machine learning techniques, specifically Random Forest and Linear (Logistic) Regression algorithms. The system was developed using the Python programming language due to its robustness, flexibility, and availability of machine learning libraries. The implementation process involved dataset preparation, model training, evaluation, and prediction. System testing is the process of evaluating the functionalities of the system to ensure that the aim and objectives of the requirements are achieved. The actual test carried out in the development of the system were unit testing, test and integration testing. The unit testing process involved testing the subsystems individually, to ensure the subsystem functionalities are working properly and the system testing was done to ensure that the overall system performs effectively.

The Pima Indians Diabetes Dataset was used for training and testing the system. Relevant features such as glucose level, blood pressure, body mass index (BMI), age, and insulin level were utilized to predict the likelihood of diabetes occurrence.

System testing was conducted to ensure that the developed diabetes prediction system meets its design objectives and performs accurately under different conditions. The testing phase focused on correctness, reliability, and efficiency.

5.2 Model/Unit Testing

Unit testing was performed on individual components of the system to ensure proper functionality of the diabetes prediction system. The different modules of the Diabetes prediction system were tested independently to ensure that each of the modules performs according to its different specification before the system integration.

1. **Data Preprocessing Module:** Tested for handling missing values, normalization, and feature scaling.
2. **Model Training Module:** Verified correct implementation of Random Forest and Linear Regression algorithms.
3. **Prediction Module:** Confirmed accurate classification of diabetic and non-diabetic cases based on input data.

5.2.1 Model Performance Testing: The performance of the models was evaluated using standard evaluation metrics such as:

- i. Accuracy
- ii. Precision
- iii. Recall
- iv. F1-score
- I. Area Under the Curve (AUC)

The results showed that the Random Forest model outperformed the Linear Regression model, achieving higher accuracy and better generalization. However, Linear Regression provided good interpretability and served as a baseline model for comparison.

5.2.2 System Validation Testing

In this research work, system validation testing was conducted by comparing predicted results with actual outcomes from the test dataset. The system demonstrated a high level of consistency and reliability in identifying diabetes cases, confirming that it meets the objectives of early diabetes prediction.

5.3 Integration Testing

In this research work, system integration involves the combination of all the individual modules or parts of the systems tested separately then brought together for the overall testing of diabetes predictive mode into a single functional system. The integration testing for this system was done at the last stage after testing individual modules of the system. At the end of integration testing process, no error or defect was found in the Diabetes prediction system.

5.3.1 Integrated Components

The integration testing focuses on the test the design system on input or output to justify and test how well the components functions together. The following components were successfully integrated:

1. Data input module
2. Data preprocessing module
3. Machine learning models (Random Forest and Linear Regression)
4. Model evaluation module
5. Prediction output module

The integration ensured smooth data flow from input to prediction output.

5.4 System Deployment

System deployment involved making the diabetes prediction system operational in a real-world or simulated healthcare environment.

5.4.1 Deployment Environment

The system was deployed using the following tools and technologies:

1. Programming Language: PythonLibraries: Scikit-learn, Pandas, NumPy, Matplotlib
2. Platform: Jupyter Notebook / Python IDE
3. Operating System: Windows

5.4.2 Deployment Architecture

The deployed system operates as follows:

1. User inputs patient medical parameters
2. System preprocesses the data automatically
3. Selected machine learning model processes the input
4. System outputs diabetes prediction result

The Random Forest model is set as the default model due to its superior performance.

5.4.3 User Acceptance Testing

User Acceptance Testing (UAT) was conducted to assess usability and correctness. Users were able to input data and receive prediction results without difficulty. Feedback indicated that the system is user-friendly and effective for diabetes risk prediction.

Model Performance Evaluation Results

This table presents the performance comparison between the Random Forest and Linear Regression models using standard evaluation metrics.

Table 12: Performance Evaluation of Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1- Score(%)	AUC (%)
Linear Regression	78.5	76.2	74.8	75.5	80.1
Random Forest	86.9	85.4	84.1	84.7	91.3

Table 12 shows the performances Evaluation of Machine Learning Model

The Random Forest model achieved higher accuracy, precision, recall, and AUC compared to the Linear Regression model. This indicates that Random Forest is more effective in capturing complex relationships among diabetes risk factors.

Table 13: Confusion Matrix for Linear Regression Model and Confusion Matrix Results (Linear Regression)

	Predicted Diabetic	Predicted Non-Diabetic
Actual Diabetic	92	31
Actual Non-Diabetic	24	113

This table shows the Confusion Matrix for Linear Regression Model and Confusion Matrix Results (Linear Regression)

The Linear Regression model correctly classified a moderate number of diabetic and non-diabetic cases but showed higher misclassification compared to Random Forest.

Table 14: Confusion Matrix for Random Forest Model and Confusion Matrix Results (Random Forest)

	Predicted Diabetic	Predicted Non-Diabetic
Actual Diabetic	108	15
Actual Non-Diabetic	13	124

The table 14 shows the Confusion Matrix for Random Forest Model and Confusion Matrix

Results (Random Forest)

The Random Forest model recorded fewer false predictions and higher correct classifications, demonstrating better reliability and robustness.

Table 15 System Functional Testing Results and Functional Testing Outcomes

Test Case	Description	Expected Result	Actual Result	Status
TC-01	Load dataset	Dataset loads successfully	Dataset loaded successfully	Pass
TC-02	Preprocess data	Data normalized correctly	Data normalized correctly	Pass
TC-03	Train Linear Regression model	Model trains successfully	Model trained successfully	Pass
TC-04	Train Random Forest model	Model trains successfully	Model trained successfully	Pass
TC-05	Predict diabetes status	Correct prediction output	Correct output displayed	Pass

CHAPTER SIX

SUMMARY OF FINDINGS, CONCLUSION AND RECOMMENDATIONS

6.1 Summary of Findings

This research focused on the development of a diabetes prediction model using Logistic Regression and Random Forest machine learning algorithms to improve the early detection and risk assessment of diabetes. The study was motivated by the increasing prevalence of diabetes globally and the need for intelligent, data-driven approaches that can support timely diagnosis, treatment and effective healthcare decision-making to assist patients with the disease.

- i. the study identified and analyzed important health-related attributes associated with diabetes prediction, including demographic and clinical factors such as age, body mass index (BMI), blood glucose level, blood pressure.
- ii. insulin level, and other relevant medical indicators and other attributes were used as input variables for developing predictive models capable of determining the likelihood of diabetes occurrence for its victims.
- iii. the Logistic Regression algorithm was developed as a statistical machine learning approach for diabetes classification. The model provided a simple and interpretable framework for understanding the relationship between risk factors and diabetes outcomes. The system's ability to estimate the probability of diabetes occurrence makes it suitable for healthcare environments where transparency and explainability of predictions are important.
- iv. however, the Random Forest algorithm was developed to improve prediction performance by combining multiple decision trees to produce a more robust classification model. The ensemble learning approach enabled the model to capture

complex relationships among diabetes related factors and provided improved capability for handling variations and nonlinear patterns within the dataset.

The performance evaluation of the developed models using classification metrics such as accuracy, precision, recall, F1-score, and ROC curve analysis demonstrated the effectiveness of machine learning techniques in diabetes risk prediction. The comparative analysis between Logistic Regression and Random Forest provided insights into the strengths and limitations of each algorithm, showing that machine learning models can serve as reliable decision support tools for early diabetes detection. The developed diabetes prediction model provides a foundation for an intelligent healthcare support system that can assist medical professionals and patients in identifying individuals at higher risk of diabetes. Although the model is not intended to replace clinical diagnosis, it can complement existing healthcare practices by providing early warnings that encourage timely medical intervention, lifestyle modification, and improved disease management.

6.2 Conclusion

Diabetes mellitus is one of the major public health challenge worldwide, and early detection plays a critical role in reducing its complications and improving patient quality of life, this study focused on the design and development of a diabetes prediction system using machine learning techniques. The study successfully achieved its aim of developing a diabetes prediction model using Logistic Regression and Random Forest machine learning algorithms. The new approach demonstrates the potential of predictive analytics in improving diabetes screening, supporting preventive healthcare strategies, and reducing the burden associated with late detection and complications of diabetes. Based on the findings of this study, it can be concluded that machine learning techniques are effective tools for early diabetes prediction. The developed system demonstrated the ability to accurately classify individuals as diabetic or non-diabetic using clinical and demographic data.

The Random Forest algorithm outperformed the Linear Regression model due to its ability to handle nonlinear relationships and feature interactions within the dataset. The integration of multiple models enhanced the reliability of the system and provided a basis for comparative analysis.

Therefore, the objectives of the study were achieved, and the developed diabetes prediction system can serve as a supportive decision-making tool for healthcare professionals in early diagnosis and prevention of diabetes-related complications.

6.3 Recommendations

Established on the findings and outcomes of this study, the following recommendations were made; integration of the developed diabetes prediction model into healthcare decision support systems. Healthcare institutions should consider integrating machine learning-based diabetes prediction models into electronic health record systems and clinical decision-support platforms. This will enable healthcare professionals to identify high-risk individuals early enough and provide timely preventive interventions. Adoption of machine learning techniques for early diabetes screening, medical practitioners and healthcare organizations should adopt predictive analytics tools as complementary approaches for diabetes screening, particularly in communities where access to specialized medical services is limited.

Continuous improvement and retraining of the prediction model, the developed model should be periodically updated using larger and more diverse healthcare datasets to improve prediction accuracy, reliability, and adaptability to changing patient characteristics. Incorporation of additional medical and lifestyle-related variables. Future researchers should include additional risk factors such as dietary patterns especially for those above or from 40 years, physical activity level, genetic information, family history, medication history, and environmental factors to improve the comprehensiveness of diabetes prediction models.

Other recommendations are as follows:

- i. **Clinical Application:** The system should be integrated into healthcare facilities to assist medical practitioners in early diabetes screening and risk assessment.
- ii. **Web and Mobile Deployment:** Future work should focus on deploying the system as a web or mobile application to improve accessibility for both healthcare providers and patients.
- iii. **Dataset Expansion:** Larger and more diverse datasets should be used to further improve model accuracy and generalization.
- iv. **Model Enhancement:** Additional machine learning and deep learning algorithms such as Support Vector Machines, XGBoost, or Neural Networks can be incorporated to enhance performance.
- v. **Real-Time Data Integration:** Integration with real-time electronic health record (EHR) systems would enable continuous learning and real-time prediction.
- vi. **Explainable AI:** Future studies should incorporate explainable AI techniques to improve model interpretability and trust in medical decision-making.

6.4 Contribution to Knowledge

This study contributes to knowledge by demonstrating the effectiveness of Random Forest and Linear Regression models in diabetes prediction and by providing a structured framework for implementing machine learning-based medical diagnostic systems for diabetes patients.

REFERENCES

- Agrebi, S., and Larbi, A. (2020). Use of artificial intelligence in infectious diseases. In *Artificial intelligence in precision health* (pp. 415–438). <https://doi.org/10.1016/b978-0-12-817133-2.00018-5> .Google Scholar
- Ahmed, U. (2022) Prediction of diabetes empowered with fused machine learning. *IEEE Access* **10**, 8529–8538. Article MATH Google Scholar
- Ahmed, N., (2021) Machine learning based diabetes prediction and development of smart web application. *Int. J. Cogn. Comput. Eng.* **2**, 229–241 (2021) [Google Scholar]
- Ahmad, E., Lim, S., Lamprey, R., Webb, D. R., and Davies, M. J. (2022). Type 2 diabetes. *The Lancet*, **400**(10365), 1803-1820.
- Alrifai, M.F., Ahmed, Z.H., Hameed, A.S., Mutar, M.L. (2021). Using machine learning technologies to classify and predict heart disease. *International Journal of Advanced Computer Science and Applications*, **12**(3): 123-127. <https://doi.org/10.14569/IJACSA.2021.0120315>
- Alghamdi, T. (2023) Prediction of diabetes complications using computational intelligence techniques. *Appl. Sci.* **13**, 3030. Article CAS MATH Google Scholar
- Alrifai, M.F., Ismael, O.A., Hameed, A.S., Mahmood, M.B. (2021). Pedestrian and objects detection by using learning complexity-aware cascades. In 2nd International Conference of Information Technology to Enhance E-learning and Other Applications (IT-ELA2021), Baghdad, Iraq, pp. 12-17. <https://doi.org/10.1109/IT-ELA52201.2021.9773589>
- A, U.N., Dharmarajan, K. (2022). Diabetes prediction using random forest classifier with different wrapper methods. In 2022 International Conference on Edge Computing and Applications, Tamilnadu, India, pp. 1705-1710, <https://doi.org/10.1109/ICECAA55415.2022.9936172>
- Bailey, C. J. and Day, C. (2018) Treatment of type 2 diabetes: Future approaches. *Br. Med. Bull.* **126**, 123–137. Article CAS PubMed MATH Google Scholar
- Bukhari, M. M. (2021) An improved artificial neural network model for effective diabetes prediction. *Complexity* **2021**, 5525271. Article MATH Google Scholar
- Bhavya, E., and Sanjay, G. (2022). Diabetes and the Importance of Insulin. *International Journal of Health Sciences*, **(I)**, 8479-8487.
- Chatterjee, S., Khunti, K. and Davies, M. J. Type 2 diabetes. *The Lancet* **389**, 2239–2251 Article CAS Google Scholar
- Chatrati, S. P., (2020). Smart home health monitoring system for predicting type 2 diabetes and hypertension. *Journal of King Saud University - Computer and Information Sciences*. doi:10.1016/j.jksuci.2020.01.010

- Chakraborty, S., Verma, A., Garg, R., Singh, J., and Verma, H. (2023). Cardiometabolic risk factors associated with type 2 diabetes mellitus: a mechanistic insight. *Clinical Medicine Insights: Endocrinology and Diabetes*, **16**, 11795514231220780.
- Chien, T.Y., Ting, H.W., Chen, C.F., Yang, C.Z., Chen, C.Y. (2022). A clinical decision support system for diabetes patients with deep learning: Experience of a Taiwan medical center. *International Journal of Medical Sciences*, **19**(6): 1049-1055.
<https://doi.org/10.7150/ijms.71341>
- Dabelea, D. Increasing prevalence of gestational diabetes mellitus (gdm) over time and by birth cohort: Kaiser permanente of colorado gdm screening program. *Diabet. care* **28**, 579–584 (2005). Article Google Scholar
- Deberneh, H. M., and Kim, I. (2021). Prediction of type 2 diabetes based on machine learning algorithm. *International journal of environmental research and public health*, **18**(6), 3317.
- Doğru, A., Buyrukoğlu, S. and Arı, M. A (2023) hybrid super ensemble learning model for the early-stage prediction of diabetes risk. *Med. Biol. Eng. Comput.* **61**, 785–797. Article PubMed MATH Google Scholar
- E. Alpaydin *Introduction to machine learning* MIT Press (2020) Available from: <
<https://books.google.com/books?hl=en&lr=&id=tZnSDwAAQBAJ&oi=fnd&pg=PR7&dq=Introduction+to+machine+learning&ots=F3RWaXcwwf&sig=50DHyEjhVdDt-mXlxZ0C4tXOGsdw>> Google Scholar
- Fazakis, N., Kocsis, O., Dritsas, E., Alexiou, S., Fakotakis, N., Moustakas, K. (2021). Machine learning tools for long-term type 2 diabetes risk prediction. *IEEE Access*, **9**: 103737-103757. <https://doi.org/10.1109/ACCESS.2021.3098691>
- F.S. Ahmad, and L. Ali (2020): A hybrid machine learning framework to predict mortality in paralytic ileus patients using electronic health records (EHRs) *J Ambient Intell Humaniz Comput, **12*** (2) (2020), pp. 3283-3293
- García-Ordás, M. T., Benavides, C., Benítez-Andrades, J. A., Alaiz-Moretón, H. and García-Rodríguez, I. (2021) Diabetes detection using deep learning techniques with oversampling and feature augmentation. *Comput. Methods Progr. Biomed.* **202**, 105968. Article MATH Google Scholar
- G. Dharmarathne, T. Jayasinghe, M. Bogahawaththa, D.P.P. Meddage, U. Rathnayake
 A novel machine learning approach for diagnosing diabetes with a self-explainable interface *Health Anal*, **5** (2024), Article 100301 ISSN 2772-4425
- Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2021. Results. Institute for Health Metrics and Evaluation. 2024
 (<https://vizhub.healthdata.org/gbd-results/>).
- Hasan, M. K., (2020) Diabetes prediction using ensemble of different machine learning classifiers. *IEEE Access* **8**, 76516–76531 (2020).

- H. Habehh, S. Gohel Machine learning in healthcare *Curr Genom, 22* **(4)** (2021), pp. 291-300, 10.2174/1389202922666210705124359 View at publisherView in Scopus
- Hossain, A., Pranto, S. I., Alif, M. D., and Fahim, M. M. (2022). Smart biomedical device to predict lung diseases, COVID-19 and diabetes by using machine learning algorithms.
- Hossain, M. J., Al-Mamun, M., and Islam, M. R. (2024). Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. *Health science reports*, 7(3), e2004
- Jain, V. (2022). Diabetes prediction using support vector machine, naive bayes and random forest machine learning models. In Proceedings of the Sixth International Conference on Electronics, Communication and Aerospace Technology (ICECA 2022), Coimbatore, India, pp. 837-841. <https://doi.org/10.1109/ICECA55336.2022.10009241>
- Jain, V. (2022). Performance analysis of supervised machine learning algorithm for prediction of diabetes. In International Conference on Edge Computing and Applications (ICECAA 2022) Proceedings, Tamilnadu, India, pp. 1162-1165.

<https://doi.org/10.1109/ICECAA55415.2022.9936503>
- Jackins, V. (2021) AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *J. Supercomput.* **77**, 5198–5219 (2021) [Google Scholar]
- Kaur, R., Kaur, M. and Singh, J. (2018) Endothelial dysfunction and platelet hyperactivity in type 2 diabetes mellitus: Molecular insights and therapeutic strategies. *Cardiovasc. Diabetol.* **17**, 1–17 (2018). Article MATH Google Scholar
- Kaur, H., Kumari, V. (2022). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*, **18**(1/2): 90-100. <https://doi.org/10.1016/j.aci.2018.12.004>
- Katsarou, A. (2017) Type 1 diabetes mellitus. *Nat. Rev. Dis. Primers* **3**, 1–17 (2017). Article MATH Google Scholar
- Khanam, J. J. and Foo, S. Y. A (2021) comparison of machine learning algorithms for diabetes prediction. *Ict Express* **7**, 432–439. Article MATH Google Scholar
- Kaur, P., Kumar, R., and Kumar, M. (2019). A healthcare monitoring system using random forest and internet of things (IoT). *Multimedia Tools and Applications*, 78(14), 19905-19916.
- Kumari, S., Kumar, D. , Mittal, M. : An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *Int. J. Cognit. Comput. Eng.* **2**, 40–46 (2021) [Google Scholar]
- Liu, J., Zhang, Z., and Razavian, N. (2018). Deep EHR: Chronic disease prediction using medical notes. *arXiv*. <<https://arxiv.org/abs/1808.04928>> .
- Luke D, Morshed A, McKay V, (2018). Systems science methods in dissemination and implementation research. In: Brownson RC, Colditz GA, Proctor EK, editors.

Dissemination and implementation research in health: translating science to practice.
 2nd ed. New York: Oxford University Press.

- Mahesh, T.R., Vivek, V., Kumar, V.V., Natarajan, R., Sathya, S., Kanimozhi, S. (2022). A comparative performance analysis of machine learning approaches for the early prediction of diabetes disease. In Proceedings - IEEE International Conference on Advanced Computing, Communication and Applications Informatics (ACCAI 2022), Chennai, India, pp. 1-6. <https://doi.org/10.1109/ACCAI53970.2022.9752543>
- Maniruzzaman, M., Rahman, M.J., Ahammed, B., Abedin, M.M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems, **8**(1): 7. <https://doi.org/10.1007/s13755-019-0095-z>
- M. Berry, and A. Mohamed, B. Yap (Eds.), *Supervised and unsupervised learning for data science*, Springer (2020), pp. 3-21
- Mohan, N., and Jain, V. (2020, November). Performance analysis of support vector machine in diabetes prediction. In *2020 4th International conference on electronics, communication and aerospace technology (ICECA)* (pp. 1-3). IEEE.
- Mounika, V., (2021) Prediction of type-2 diabetes using machine learning algorithms. In: International Conference on Artificial Intelligence and Smart Systems, pp. 127–131 (2021)
- Murphy, H. R., Howgate, C., O'Keefe, J., Myers, J., Morgan, M., Coleman, M. A., ... and Tomkins, N. (2021). Characteristics and outcomes of pregnant women with type 1 or type 2 diabetes: a 5-year national population-based cohort study. *The lancet Diabetes & endocrinology*, **9**(3), 153-164.
- Negrato, C. A., Tarzia, O., Jovanović, L., and Chinellato, L. E. M. (2013). Periodontal disease and diabetes mellitus. *Journal of Applied Oral Science*, **21**(1), 1-12.
- Mehta, R., Vala, B., Patel, A. (2022). A survey on diabetes prediction using supervised learning. In Proceedings of the 2nd International Conference on Artificial Intelligence and Smart Energy (ICAIS 2022), Coimbatore, India, pp. 302-307.
<https://doi.org/10.1109/ICAIS53314.2022.9743006>
- Olisah, C. C., Smith, L. and Smith, M. (2022) Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Comput. Methods Progr. Biomed.* **220**, 106773. Article MATH Google Scholar
- Pal, M., Parija, S., Panda, G. (2021). Improved prediction of diabetes mellitus using machine learning based approach. In 2nd International Conference on Range Technology (ICORT 2021), Chandipur, Balasore, India, pp. 1-6.
<https://doi.org/10.1109/ICORT52730.2021.9581774>
- Parameswari, R., Kumar, P. M., Pavithra, S. A., Iswarya, S. J., Yogesh, T., and Babujanathanam, R. (2025). Diabetes: Secondary Complications. In *Algae in Diabetes Management: Therapeutic Properties and Applications* (pp. 35-88). Singapore: Springer Nature Singapore.

- Plows, J. F., Stanley, J. L., Baker, P. N., Reynolds, C. M., and Vickers, M. H. (2018). The pathophysiology of gestational diabetes mellitus. *International journal of molecular sciences*, **19**(11), 3342.
- Qin, Y.F., Wu, J.L., Xiao, W., Wang, K., Huang, A.B., Liu, B., Yu, J.X., Li, C., Yu, F.Y., and Ren, Z.B. (2022). Machine learning models for data-driven prediction of diabetes by lifestyle type. *International Journal of Environmental Research and Public Health*, **19**(22): 15027. <https://doi.org/10.3390/ijerph192215027>
- Rady, M., Moussa, K., Mostafa, M., Elbasry, A., Ezzat, Z., Medhat, W. (2021). Diabetes prediction using machine learning: A comparative study. In NILES 2021 - 3rd Novel Intelligent and Leading Emerging Sciences Conference Proceedings, Giza, Egypt, pp. 279-282. <https://doi.org/10.1109/NILES53778.2021.9600091>
- Ramesh, V., Abraham, S., Vinod, P., Mohamed, I., Visaggio, C. A., and Laudanna, S. 2021. Automatic classification of vulnerabilities using deep learning and machine learning algorithms. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.
- Rathod, S.R., Phadke, L., Chaskar, U.M., Patil, C.Y. (2021). Machine learning techniques for predicting Type 2 diabetes mellitus risk using heart rate variability features. In 2021 12th International Conference on Computing Communication and Networking Technologies, Kharagpur, India, pp. 1-6. <https://doi.org/10.1109/ICCCNT51525.2021.9579746>
- Reddy, S. S. K., and Tan, M. (2020). Diabetes mellitus and its many complications. In *Diabetes mellitus* (pp. 1-18). Academic Press.
- Reddy, S. K., Krishnaveni, T., Nikitha, G., Vijaykanth, E. (2021). Diabetes prediction using different machine learning algorithms. In *Proceedings of the 3rd International Conference on Inventive Research in Computing Applications (ICIRCA 2021)*, Coimbatore, India, pp. 1261-1265. <https://doi.org/10.1109/ICIRCA51532.2021.9544593>
- Sari, F.A.O., Alammahi, A.A.H., Hameed, A.S., Alrikabi, H.M.B., Abdul-Razaq, A.A., Nasser, H.K., AL-Rifaie, M.F. (2022). Networks cyber security model by using machine learning techniques. *International Journal of Intelligent Systems and Applications in Engineering*, **10**(3s): 257-263.
- Safiri, S. (2022) Prevalence, deaths and disability-adjusted-life-years (dalys) due to type 2 diabetes and its attributable risk factors in 204 countries and territories, 1990–2019: results from the global burden of disease study 2019. *Front. Endocrinol.* **13**, 838027. Article Google Scholar
- Saidu, I. R., Saleh, R. U., and Abdulkadir, N. (2024). ENSEMBLE LEARNING AND FEATURE IMPORTANCE FOR PERSONALIZED DIABETES DIAGNOSIS. *AJSE*, **19**(3).
- Sahid, M. A., Babar, M. U. H., and Uddin, M. P. (2024). Predictive modeling of multi-class diabetes mellitus using machine learning and filtering iraqi diabetes data dynamics. *Plos one*, **19**(5), e0300785.

- Saxena, A., Jain, M., and Shrivastava, P. (2021). Data mining techniques based diabetes prediction. *Indian Journal of Artificial Intelligence and Neural Networking*, **1** (2), 29-35.
- Shaikh, A. A., Kolhatkar, M. K., Sopane, D. R., and Thorve, A. N. (2022). Review on: Diabetes mellitus is a disease. *Int J Res Pharm Sci*, **13** (1), 102-109.
- Shojaee-Mend, H., Velayati, F., Tayefi, B., Babaei, E. (2024). Prediction of diabetes using data mining and machine learning algorithms: A cross-sectional study. *Healthcare Informatics Research*, **30** (1): 73-82. <https://doi.org/10.4258/hir.2024.30.1.73>
- Sontakke, R., Shinde, P., Avhad, V., Kadam, Y., Yadav, V., Aswathy, M.A. (2024). Web-based framework for the prediction of type 1 diabetes in youth using EHR's data. In *Advances in Distributed Computing and Machine Learning*, Springer, Singapore.
https://doi.org/10.1007/978-981-97-3523-5_33
- Stumvoll, M., Goldstein, B. J. and Van Haeften, T. W. (2005) Type 2 diabetes: Principles of pathogenesis and therapy. *The Lancet* **365**, 1333–1346. Article CAS Google Scholar
- Tasin, I., Nabil, T. U., Islam, S., and Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare technology letters*, **10**(1-2), 1-10.
- World Health Organization. *Global health Observatory. Diabete: Prevalence and treatment from 1990 to 2022*. WHO; 2024. Accessed February 16, 2025. <https://www.who.int/data/gho/indicator-metadata-registry/imr-details/761>
- Tasin, I., Nabil, T. U., Islam, S., and Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare technology letters*, **10**(1-2), 1-10.
- Zhao, J. *et al.* Attention-oriented cnn method for type 2 diabetes prediction. *Appl. Sci.* **14**, 3989 (2024). Article CAS MATH Google Scholar
- Zhou, H., Xin, Y. and Li, S. A (2023) diabetes prediction model based on Boruta feature selection and ensemble learning. *BMC Bioinform.* **24**, 224 Article CAS MATH Google Scholar
- Zhu, T., Li, K., Herrero, P., Georgiou, P. (2021). Deep Learning for diabetes: A systematic review. *IEEE Journal of Biomedical and Health Informatics*, **25**(7): 2744-2757. <https://doi.org/10.1109/JBHI.2020.3040225>