# Machine Learning-based System for an early and accurate Detection of Diabetes

# Abstract

Diabetes is a most common and severe chronic disease, which is also known as a second killer. Currently, there is no perfect cure or treatment for this disease. The patients can only depend on the medications for the auxiliary treatment. Most people are not serious about their physical health and opt out of the necessary treatment due to lack of money, which ultimately leads to the most severe complications. Therefore an accurate preduction and detection of this disease at the early stage is essential to prevent any severe impact on human health. This research study mainly aims to develop an efficient and robust diabetic detection system based on machine learning methodology. Further, an ensemble learning method and stacking method along with Boruta features is used to enhance the process of data classification. This stacking method uses a Logistic regression, K-nearest neighbor, and Naïve Bayes are the base model and a Support Vector Machine as the meta model. The experiments are conducted on the Kaggle PIMA Indian diabetes dataset. The performance of the proposed system is evaluated using key performance metrics, including precision, accuracy, recall, and precision. The proposed ensemble learning-based diabetic detection system attained an accuracy of 98% and provided excellent prediction results for diabetes.

Acknowledgment

I would like to express my heartfelt thanks to my supervisor [name] for this guidance and support throughout this research thesis. I don’t think I can complete this research without his assistance and support. I also want to acknowledge the support of the University to provide me with access to the desired resources to complete this research. I like to extend my gratitude to other staff members whose constrictive expertise and experience helped me to better complete this research. I also want to thank my family and friends for their love and support during this research thesis.

Dedication

I want to dedicate this research thesis to my family, especially my mother for her support and motivation to do better in life and become succeed.

Chapter 1: Introduction

# Introduction

This is the first chapter of this dissertation, which is based on a critical introduction of this whole project along with setting the research context, aim & objectives, research questions, and rationale. Currently, digital technologies are emerging in various fields at a great pace in different industries, such as healthcare, manufacturing, education, information technology, and so on. Among these healthcare industry is greatly impacted by this digital transformation by converting the manually performed tasks to automatically eliminate human involvement to avoid human errors and attain more accuracy (Talukder, 2023). The emergence of ML (Machine Learning) and AI (Artificial Intelligence) gained huge popularity in making the process of medical diagnosis more easier and accurate. The detection of diabetes in patients is one of the most significant use cases of this technology. Thus, this research project is based on the development of an ML-based system for the early detection of diabetes in patients to reduce the risk of this disease and prove timely precautions or cure. In this chapter, we will discuss the research background, aim & objectives, research questions, artifact description, research rationale, and thesis structure.

## Research background

Diabetes is considered one of the leading chronic diseases or health issues that cause death. Generally, diabetes is caused by to increased consumption of fast foods or processed foods that become a common part of our diet. Processed food items like bread, meat, rice, and cereal might cause high blood pressure and cholesterol that may result in diabetes (Liu et al., 2023). Further, all the artificial forms of sweetness and processed sugar may cause diabetic issues, which can significantly impact people. Thus, early recognition and intervention are very important to manage this disease and prevent possible complications.

The early detection of this disease poses several challenges, such as the symptoms of diabetes are probably absent in its early stage. The traditionally used methods for its recognition depend on the blood tests that may not be very frequent and accessible. It addresses an immediate need for an effective, accurate, non-invasive, and readily available method for early detection of diabetes.

Currently, machine learning has become one of the most popular and frequently used methods for medical diagnosis, especially in the detection of diabetes to provide the most promotion outcomes. Several research studies demonstrate good results for using machine learning in the diagnosis of various chronic diseases, such as diabetes. The use of various machine learning methods can significantly contribute to the early and accurate detection of diabetes. There are several machine learning methods, such as Support Vector Machines, Neural Networks, Random Forest, K-nearest neighbors, Logistic Regression, and Naïve Bayes algorithms that can efficiently analyze the data related to body mass index, hemoglobin, blood glucose level, and lifestyle factors (Abdulhadi & Al-Mousa, 2021). These methods help in recognizing the associated patterns with the early stages of diabetes and can substantially predict the risk of diabetes in the future to facilitate an early and timely intervention in prevention of the diabetes.

## Aim and Objectives

### Aim

The main aim of this research is to develop an ML-based system for early and accurate recognition of diabetes among patients by combining the obtained results from diverse machine learning methods.

### Objectives

The key objectives of this research project are as follows:

* To conduct an in-depth literature review to address the applications of machine learning in the process of medical diagnosis, especially diabetes detection.
* To execute, assess, and compare diverse machine learning methods in the accurate prediction of early-stage diabetes.
* To design a highly accessible and user-friendly system interface to cater to the needs of both patients and healthcare professionals.
* To ensure an adequate level of data security and privacy throughout this research project.
* To evaluate the proposed system using different performance metrics like precision, accuracy, recall, and F -score (Liu et al., 2023).
* To integrate the proposed system with existing clinical practices and healthcare infrastructure while addressing potential adoption barriers.

## Research questions

Research questions are a very important part of any research which provides a direction to facilitate the research to answer the questions while attaining the specified research aim and objectives.

* How can ML methods be used for developing an accurate and consistent system for the early recognition of diabetes?
* How can the accessibility and accuracy of conventional diagnostic methods be changed with the ML-based diagnostic system?
* How could this proposed system be integrated with the current healthcare systems to facilitate a better and early detection of diabetes?
* What are the probable challenges and limitations that might be faced by this proposed system in real-world healthcare settings?
* What are the main user data security and privacy concerns associated with this proposed ML-based diagnostic system?

## Problem statement

This research addresses the increased risk of the most dangerous chronic disease, namely diabetes which can have a severe impact on human health and even cause death if it is timely detected and treated. However, several methods are currently used in the diagnosis of diabetes but they depend on HbA1c measurements and blood tests that might cause a delay in the detection and might detect individuals who do not have any symptoms (Sharma & Sharma, 2020). Also, these methods lack accessibility due to restricted healthcare access, higher costs, and inconvenience. Any delay in the diagnosis of diabetes may increase the risk of most severe complications, such as cardiovascular diseases, kidney failure, or blindness. It may also lead to higher costs to manage the advanced stage of diabetes and cause complications. It can create significant social and economic burdens on both healthcare systems and individuals. Thus, the use of machine learning methods can enable healthcare professionals to better analyze the larger datasets from different sources for recognizing the associated patterns with the early stage of diabetes. It can detect and predict diabetes with higher accuracy while offering a more accessible, non-invasive, and personalized system for early recognition of diabetes.

## Artifact description

This artifact is based on the development of a machine learning-based detection system for the early and timely detection of diabetes to prevent or reduce the risk of diabetes (G. & K., 2019). This system can be available in either the mobile app or web application that could be used by individuals who are at risk of diabetes and healthcare professionals who are involved in the prevention and management of diabetes. This system can import data from various sources along with realizing the integration of various ML models for an accurate prediction of diabetes. This system can conduct a critical risk assessment while offering actionable insights in addition to a user-friendly interface, effective data security & privacy measures, and easy integration with the existing healthcare infrastructure (Abdulhadi & Al-Mousa, 2021). This system can provide several benefits, such as early recognition of diabetes, increased health awareness, enhanced healthcare system, and decreased healthcare costs.

## Research rationale and significance

Nowadays, with the increased consumption of processed food items and imbalanced diets, the risk of diabetes is increasing at a great pace which might even cause death. However, in the early stage, it might lack symptoms that lead to delayed diagnosis of this disease and increase the risks of severe complications. The frequently used methods like HbA1c measurements and blood tests are expensive, inconvenient, and might miss the cases in the early stages. It addresses a crucial need for a proactive and accessible strategy. Thus, this research proposed an ML-based diabetic detection system to better analyze the data from different sources and recognize vital patterns to increase the detection accuracy as compared to the traditionally used detection methods. Early detection of diabetes enables early interventions through prevention strategies, medications, and lifestyle changes to enhance an individual’s health and reduce probable complications (Lukmanto et al., 2019). It can also facilitate proactive diabetes management in alleviating the social and economic burden on healthcare professionals and individuals. Moreover, the comprehensive adoption of this system significantly empowers individuals to take control over their health, enhancing healthcare proficiency, and contributing toward a healthier population,

## Thesis structure

This research dissertation comprises six different chapters, introduction, literature review, research methodology, design, development & implementation, testing, result, & evaluation, and discussion & conclusion.

In the first chapter, I critically introduced this research project, including research background, aim & objective, research questions, problem statement, artifact description, research rationale & significance, and thesis structure. In the second chapter, I will critically review and discuss the existing or previously published research papers or journal articles to analyze other author’s works and gain a better understanding of this research (Raja & Pandian, 2020). Then in the third chapter, I will discuss the chosen research methodology for conducting this research thesis, including the research approach, data collection, and project management methods. In the fourth chapter, I will create an efficient architecture or design of the proposed system along with its implementation with the created codes for its implementations. In the fifth, chapter, I will provide the obtained results, testing methods to test system effectiveness and performance evaluation. In the sixth chapter, I will conclude this research by summarizing this research and key insights and providing recommendations to conduct future research.

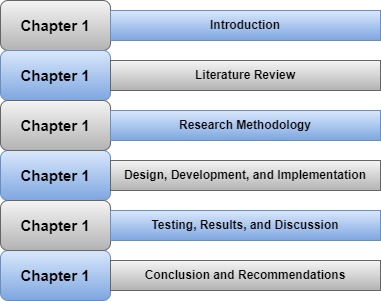


Figure : Thesis structure

# Conclusion

This research chapter provides a critical introduction to the research thesis that focuses on developing an efficient ML-based diabetic detection system for detecting diabetes in its early stages to timely diagnose and cure this disease for reducing the risk of severe complications and death. Here I provided background information, problem statement, aim & objectives, research questions, research rationale & significance, and thesis structure.

Chapter 2: Literature Review

# Introduction

This is the second chapter of this research dissertation, which is based on conducting a detailed literature review of the relevant and previously published research papers or journal articles. Here I will focus on comprehensively exploring the applications of machine learning in the detection of diabetes. This chapter will provide the details related to the most frequently used data types and machine learning algorithms in diabetes detection along with evaluating the performance of these algorithms following key performance metrics, such as precision, accuracy, recall, and F1-score. Here I will address the key research gaps and unmet research needs in the current landscape of this research. Further, the associated opportunities and challenges will also be addressed to attain key directions for future innovations and advancements. Moreover, it will also address the related challenges with feature engineering and data preprocessing because of imbalanced medical datasets, which need oversampling methods for balancing these data before data training.

# Literature Review

## Applications of ML Model for early diabetes recognition and accurate classification

According to Iparraguirre-Villanueva et al., (2023), diabetes refers to a chronic disease that occurs due to increased sugar levels in blood due to metabolic issues. This disease can severely damage many body organs, such as the eyes, heart, and blood vessels. Thus, early recognition of diabetes is very important for preventing its severe complications in patients. This research mainly intends to perceive and categorize type 2 diabetes in patients with the help of the ML method and choose the most effective and robust ML model for predicting diabetes. The author used 5 different ML models, such as decision tree, k-nearest neighbor, logistic regression, Naive Bayes, and support vector machine to predict the risk of diabetes. Further, a Pima Indian dataset hosted by Kaggle is used including about 768 patients, including different variables like concentration of blood glucose, family history of diabetes, age, no. of pregnancies, etc. The obtained research results demonstrated that the Naïve Bayes and K-nearest neighbor outperform other ML models. The k-NN model attained about 79.6% accuracy, whereas Naïve Bayes attained about 77.2% accuracy in the detection of diabetes (Iparraguirre-Villanueva et al., 2023). Finally, it can be concluded that the ML methods provide the most promising results in the detection of diabetes.

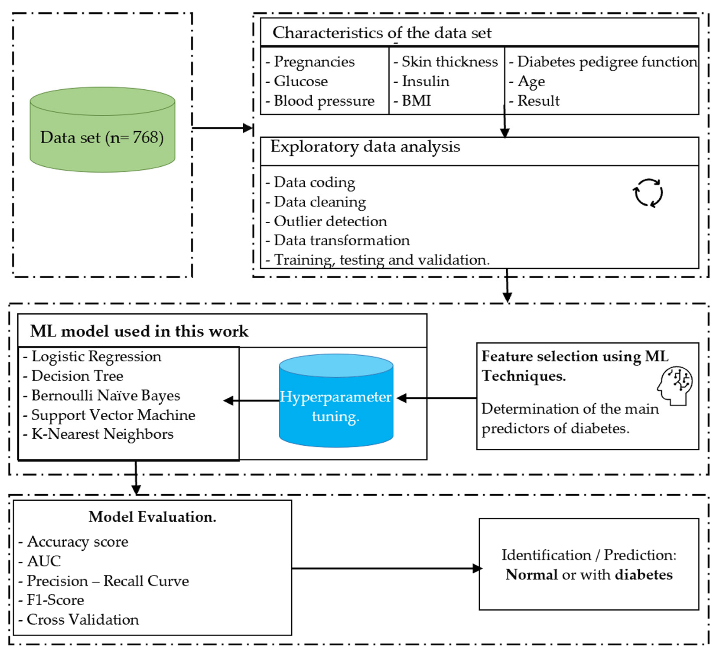


Figure : ML model's development process

## Diabetes mellitus detection following element level and ensemble modeling analysis

Chen et al., (2014) state that diabetes mellitus represents the most frequent and severe chronic disease which is mainly caused due to deficiency of insulin. This research explored and analyzed relations between diabetes and levels of different elements present in the samples of urine/hair in the diagnosis of diabetes. A dataset, including 8 element concentrations and 211 specimens is used. Three age groups are selected for analyzing the impact of age and it is observed that the most obvious difference is the impact of age on the levels of iron and zinc. With age, the concentration of iron declined in hair which is opposite to the trends in urine. Both SVM (support vector machine) and ensemble learning algorithms are used as the classifiers that demonstrate an average sensitivity, accuracy, and specificity of 100%, 99%, and 99% for hair samples and 89%, 97%, and 99% for urine samples. Finally, this research demonstrates by conducting several model evaluation metrics that the hair samples represent superiority as compared to the urine samples in the diagnosis and deterrence of diabetes (Chen et al., 2014). The hair samples provide more accurate and valuable information related to the research. diagnosis, prevention, and treatment of diabetes.

## Data mining methods in the prediction of diabetes

According to Perveen et al., (2016), diabetes is mainly associated with an abnormal increment in glucose levels in the blood, owing to excessive insulin production by the pancreas, or failure of cells while responding to produced insulin by the pancreas. Currently, the healthcare community is taking a huge interest in detecting and preventing diabetes. The existing studies proposed many kinds of clinical decision support systems that comprise different data mining algorithms for the detection and prediction of diseases. These previously used systems rely on a single classifier. Recently huge endeavors have been made to improve the accuracy of these systems with the help of ensemble classifiers. Thus, this research proposed an integrated method of bagging and AdaBoost methods using the J48 (c4.5) as the fundamental classifier and merging standalone data mining methods for classifying diabetic patients. This research tested the performance of these methods on the Canadian primary care surveillance dataset and the obtained results demonstrated that the proposed integrated method outperformed independent J48 decision tree and bagging methods (Perveen et al., 2016). Moreover, this research suggests using a support vector machine and Naïve Bayes algorithms in the ensemble learning framework for the diagnosis of other diseases, such as coronary heart disease, hypertension, etc.

## Type 2 diabetes prediction based on data mining methods

Wu et al., (2018) addressed that many families and individuals’ lives are largely impacted by diabetes. Most patients do not know about the quality of their health and the associated risk factors faced by them before the diagnosis. This research proposed a novel model to predict type diabetes with the help of data mining methods, including logistic regression and K-means algorithms. The use of the K-means algorithm addresses the seed value’s randomness by using a procedure for recording and sorting the values known as the sum of the squared errors into clusters in ascending order. The smaller values provide better results. Then the proposed model was evaluated based on the PIMA Indian diabetes dataset along with the two other datasets to attain excellent experimental results. The accuracy of diabetes prediction in the proposed model is about 3.04% higher as compared to other models. The research results also demonstrated the usefulness of the proposed model in real-world healthcare settings in the detection of diabetes (Wu et al., 2018).

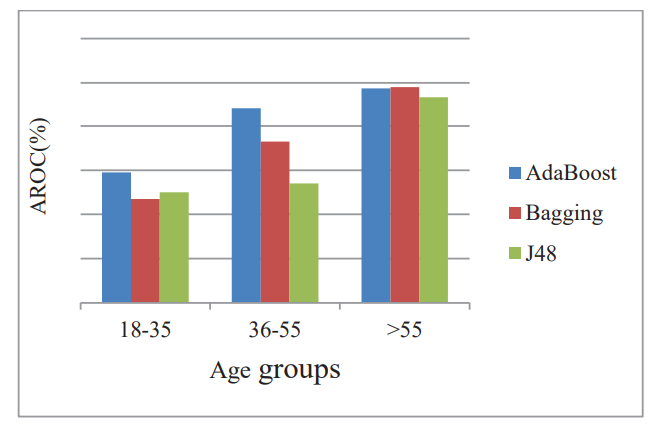


Figure : Comparison of J48 and ensembles against three diverse age groups

## Optimized logistic regression model for diabetes prediction

This research proposed a novel model based on data mining methods for an early prediction and diagnosis of diabetes using the Pima India Diabetic dataset. However, the K-means algorithm is modest and could be used for different data types but it is sensitive to the primary position of the cluster centers which helps in determining the ultimate cluster results for providing an efficient and sufficient dataset for the logistic regression mode and providing minimal data amount resulting from the original dataset’s wrong clustering that might limit the performance measures of the proposed logistic regression model (Zhu et al., 2019). The main goal of this research is to determine the potential ways to optimize the accuracy of logistic regression and K-means clustering algorithms. The proposed model is an integrated system of K-means and PCA (principal component analysis) methods for realizing robust clustered datasets. This proposed model comprises 3 components, namely data normalization, principal component analysis, and logistic regression & K-means algorithms. The obtained experimental results demonstrate that the use of PCA optimized the accuracy of both logistic regression and K-means classifiers. The accuracy of logistic regression is improved by 1.98% and K-means offers 25 accurately categorized data points.

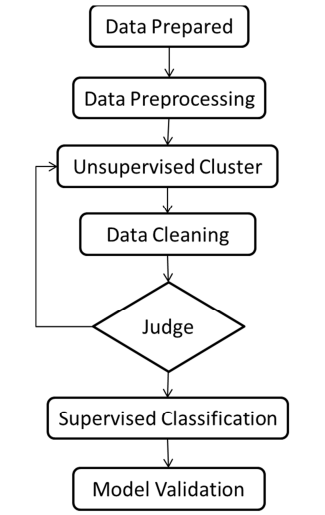


Figure : The chosen algorithm model

## Fuzzy SVM and feature selection-based model for diabetes classification

According to Lukmanto et al., (2019), the number of diabetes patients is frequently increasing with the increased consumption of processes and fast food products. This research focuses on developing a system for correctly detecting and classifying diabetes. It leverages fuzzy support vector machine and F-exponential feature selection methods for the recognition and categorization of diabetes. The features selection method is utilized for extracting the most vital and valuable features from the considered PID (Pima Indian Diabetes) dataset. Then this dataset was trained with the help of a support vector machine while generating effectual fuzzy rules. At last, the obtained output is categorized through the fuzzy inference technique. The proposed model attained about 89.02% accuracy for the PID dataset. Moreover, this proposed model offers an extremely optimized fuzzy rule count along with maintaining an appropriate level of accuracy.

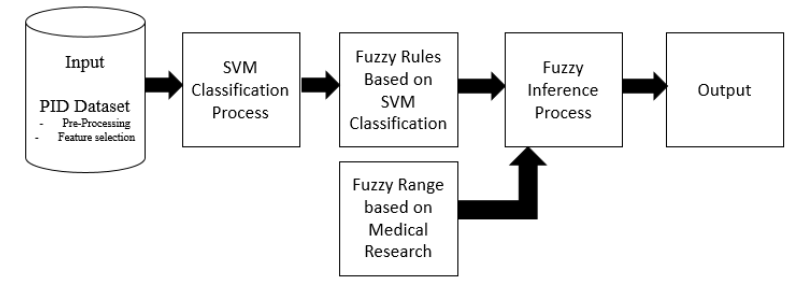


Figure : The proposed fuzzy SVM-based model

## Grey wolf optimization algorithm in diabetes detection

According to G. & K., (2019), diabetes increased due to increased unhealthy workplace culture, lifestyle, and absence of physical activities. However, diabetes is a chronic disease and incurable but early detection, monitoring, and diagnosis can prevent further harmful impacts. This research study proposes an efficient prediction model for diabetes, which is based on the GWO (grey wolf optimization) algorithm with fuzzy logic. This proposed model works on the concept of fuzzy rules and ant colony optimization, which do not provide an efficient level of accuracy because it only optimizes the local features and provides an accuracy of about 81% against the considered dataset. The selection of the most appropriate features is done by wrapper-based techniques for selecting optimal-level features and improving the obtained classification results. This proposed model can optimize the global features and offers an accuracy higher than the ant colony algorithm.

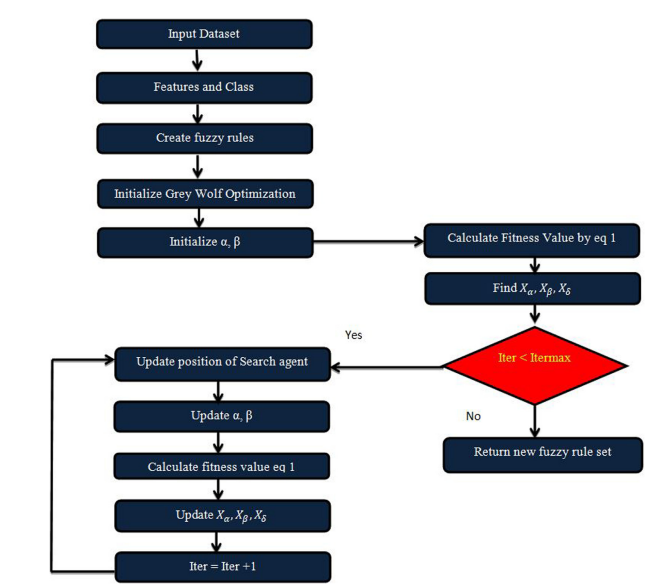


Figure : Flowchart of GWO algorithm

## ANN for diabetes prediction with the time domain features

The prediction of future levels of blood sugar can help patients avoid diabetes and its impact on their health. This research demonstrates the utilization of ML algorithms in the prediction of future blood glucose levels based on the existing blood glucose values. This research proposes an ANN (artificial neural network)-based model with time domain attributes for better predicting the glucose levels in the blood. First, the model features are chosen before 30 minutes of blood glucose measurements (Alfian et al., 2020). Then these features are combined with the time-domain attributes for giving extract inputs to the proposed ANN-based model. This research tested the proposed model on twelve patients and compared the obtained results with the other ML models such as Adaptive Boosting, support vector machine, decision tree, C4.5, K-nearest neighbor, and eXtremeGradient Boosting model. The results demonstrate that the proposed ANN-based prediction model outperformed all these ML models with an average RMSE (root mean square error) of 15.33, 10.65, 6.31, 2.82 mg/dL for the prediction horizons of 60, 45, 30, 15 min, respectively. This research addressed the combined time-domain attributes in the input data resulting in the optimized performance of the proposed prediction model.

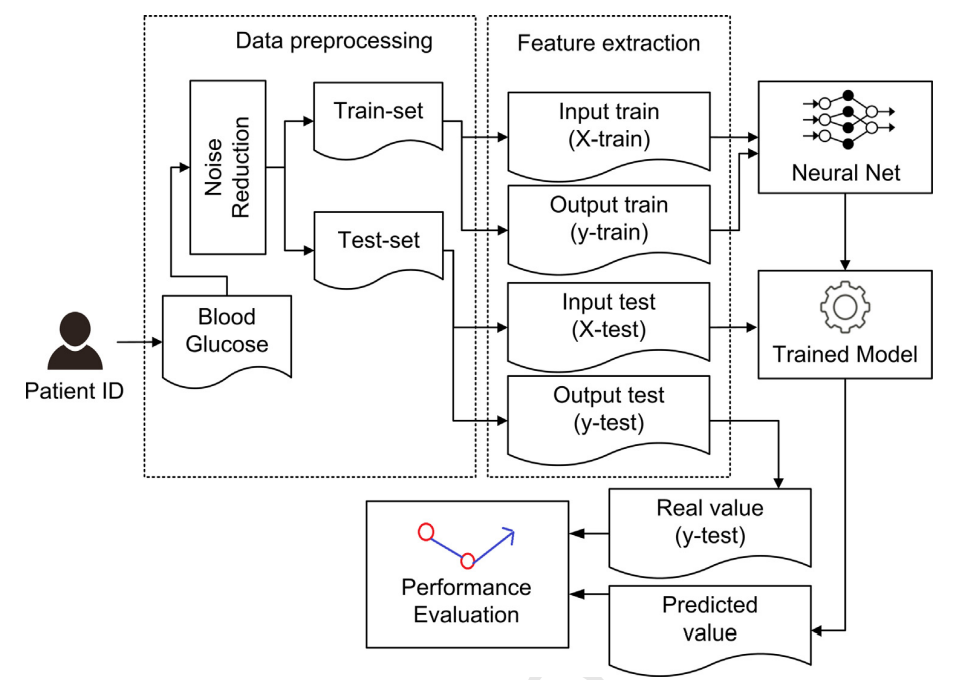


Figure : General schema of the proposed model

## PSO-FCM-based data mining model for diabetes prediction

According to Raja & Pandian, (2020), diabetes is typically contained a higher level of sugar levels. Recently, the patients of with diabetes have increased frequently across the world due to the increased consumption of processed foods, hectic schedules, and inactive lifestyles. Many research studies conducted in-depth research on various data mining methods for determining the accuracy of diabetes prediction. Data mining could be utilized with parametric modeling for the health dataset for synthesizing expertise in this particular field. This research study proposed a novel predictive model based on the data mining methods, namely FCM (fuzzy clustering method) and PSO (particle swarm optimization). The performance of this proposed model is evaluated using the PID dataset following the key performance metrics, including accuracy, sensitivity, and specificity. The attained research results demonstrated that the accuracy of this proposed model is improved by 8.26% than the other methods and it outperforms the other methods.

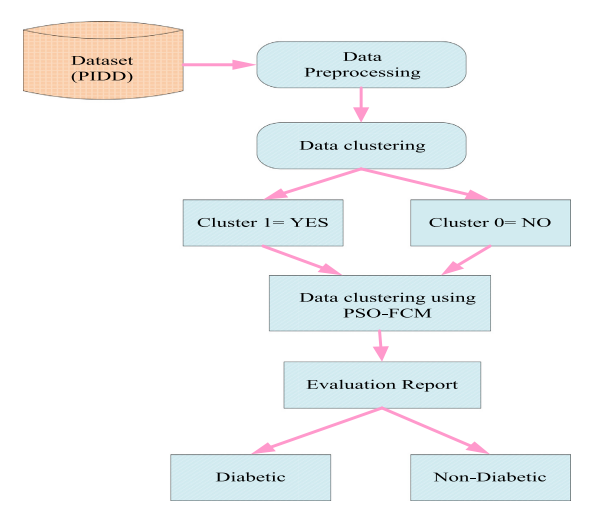


Figure : Structure of the proposed model

## A hybrid approach for diabetes detection using SVM and FF algorithms

Howsalya Devi et al., (2020) state that the human body needs energy to remain in an active state. The eaten food by use is converted into glucose and sugar to produce energy. Pancrease generates insulin, which consists of glucose in the body cells and further turns into energy for proper body functioning. The main purpose of this research is to diagnose diabetes using data mining methods. The use of data mining methods in the medical field is very important, valuable, and efficient for enhancing medical knowledge for making the most effective decisions. This research proposes an integrated system of the SMO (sequence minimum optimization) classifier and FF (farthest first) clustering algorithms. The clustering algorithm splits data into diverse cluster sets to decrease the dataset size and greatly shorten the required computation time. The obtained clustering time is utilized as an input for CVM for completing the classification task. This proposed integrated system attained excellent results on the PID dataset. The obtained experimental results demonstrate that the ensemble approach poses an accuracy of 99.4% in diabetes prediction (Howsalya Devi et al., 2020). These results also validate that the proposed hybrid approach of both data mining techniques can help healthcare professionals make better clinical decisions in the diagnosis and treatment of diabetic patients.

## An innovative hybrid approach for diabetes prediction with customized regulations

Rajagopal et al., (2022) argued that the initiation of data analytics and big data poses an immense influence on the areas of technology and medicine. This massive amount of medical data requires the use of frontier technologies, like deep learning, machine learning, and cloud computing to fully exploit big data and systematize the medical research’s computation processes. This research presents an innovative and customized hybrid model based on genetic algorithms and artificial neural networks for realizing an accurate prediction of diabetes. This model utilizes a spontaneous method to detect more evident patterns of the relationships among variables. First, a normalization method is used to preprocess the medical dataset. Then the proposed model recognizes the extent of each variable’s importance to impact the output and priority is given to the variables that are deemed most significant. Further, a regularization method is implemented to realize customized diabetes predictions. This proposed model deals with the missing numbers by considering a distinct identity from the numerical entities (Rajagopal et al., 2022). This proposed model and respective decision-making algorithm attained about 80% accuracy in the diabetes prediction for the PID dataset.

## Data-driven ML methods for the prediction of diabetes

Diabetes is a very common metabolic disease that is categorized by a higher level of blood sugar or disturbance in the metabolism of fats, carbohydrates, and proteins. The most critical disorder among all diabetes forms is hyperglycemia. The modern lifestyle and way of living have increased the cases of diabetes. Therefore, there is a critical need for an early and accurate detection & diagnosis of this disease. Machine learning methods have gained huge popularity among physicians and healthcare providers due to their immense potential to develop robust tools for the prediction, diagnosis, treatment, and management of different conditions  (Dritsas & Trigka, 2022). This research study utilizes the supervised ML method for creating effective risk prediction tools with higher proficiency. Further, a feature analysis is conducted to assess their significance and explore the diabetes association. The extracted features represent the most critical symptoms that develop slowly with this disease and can be utilized for training and testing different ML models. This research evaluated several ML models based on performance metrics, such as precision, accuracy, recall, F1-score, and AUC metrics under the ten-fold data splitting and cross-validation. The obtained results demonstrate that both K-NN and random forest outperformed other ML models.

## CNN and LSTM in the detection and description of diabetic retinopathy

Diabetes represents a very common and severe disease that is categorized based on the metabolic disorders caused by to deficiency of insulin. DR (diabetic retinopathy) refers to an eye disease from which diabetes patients may suffer and can cause blindness if it is not timely treated. Early detection of diabetes can help patients in getting effective treatment as per the severity of the disease (Amalia et al., 2021). Several researchers conducted many research studies to identify the most effective methods for the detection of diabetic retinopathy using deep learning and shallow learning methods, but these methods are not effective in detecting and preventing this disease. Thus, this research proposed a novel method by combining two deep learning approaches, namely LSTM (Long Short Term Memory) and CNN (Convolutional Neural Network). Here CNN helps in detecting the lesions on the images of the retinal fundus and LSTM helps to generate the descriptions for those lesions. The main target of this training process is to generate a model for better mapping the fundus images into a specific sentence. The research results based on the MESSIDOR dataset offer about 90% accuracy.

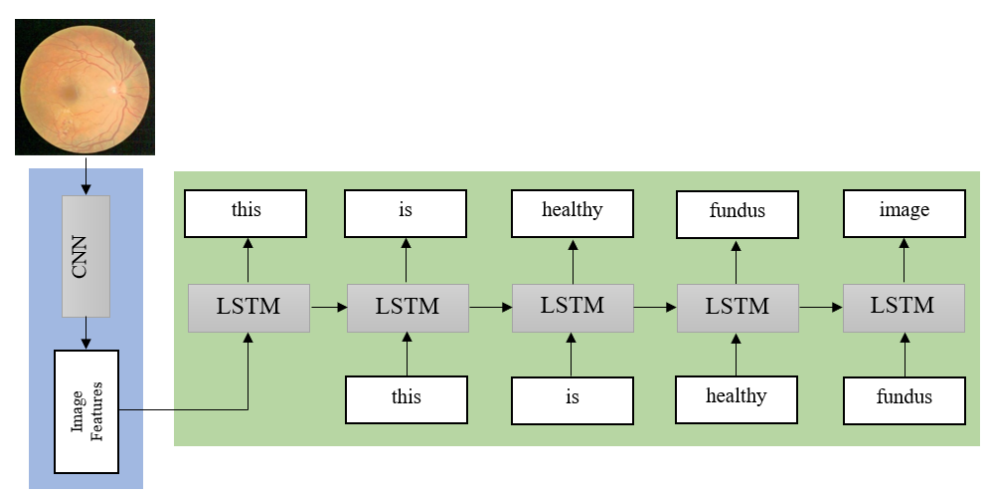


Figure : Diagram of the proposed model

## ML and DL methods for diagnosing diabetic retinopathy

Das et al., (2022) describe DR (diabetic retinopathy) as a health condition that is caused due to diabetes. Many macro and microvascular abnormalities and diminished glucose metabolism lead to this disease. Diabetic retinopathy represents the most frequent and severe complication of diabetes that leads to blindness due to the human retina’s disfigurement. According to the statistics, about 80% of patients suffer from DR who battling diabetes for more than 15 to 20 years. However, the manual diagnosis is feasible to overcome DR but to overcome and defend simultaneously needs a recolutionary method. Thus, these health conditions require initial detection and diagnosis to avert DR from leading to severe stages and prevent blindness. Several research studies proposed many ML methods for attaining this objective along with many feature extraction methods to extract vital features of DR for early recognition. The conventional ML methods demonstrate inadequate generalization during the whole feature extraction and categorization to deploy small datasets and consume huge training time that might cause incompetence for the prediction in the case of large datasets. Hence, a novel domain of machine learning, namely deep learning is proposed for better handling the small datasets using effective data processing methods (Das et al., 2022). They usually consist of large datasets for enhancing performance measures of image classification and feature extraction.

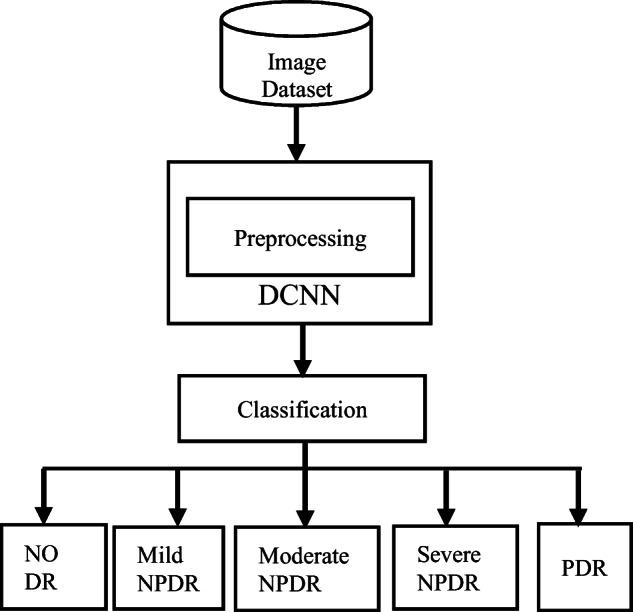


Figure : Working process of the proposed system

## Temporal DL framework for diabetic retinopathy prediction

The adoption of the HERs (electronic health records in hospitals ensures the availability of larger datasets that could be utilized for predicting medical complications. The patient trajectories are varied in real-world scenarios that create challenges in longitudinal data modeling. In the past few years, deep learning has emerged in medical applications its use remains limited in the IMTS (irregular medical time series). To address this issue, this research developed a generic DL framework to model IMTS for facilitating irregular rime representation and sequential neural network methods  (Rabhi et al., 2022). This research conducted a validation study on the 1207 patients for predicting the associated complications with retinopathy on the French database with the help of historic glycated hemoglobin measurements without data imputations or data aggregation. The proposed transformer-enabled model can provide the highest performance with 85.56% specificity, 83.33% sensitivity, 88.65% receiver operating characteristic curve, and 11.7% improvement in the architecture without any time information. This study is the first research that focuses on predicting the associated complications with diabetic retinopathy using longitudinal data gathered from patient visits and deep learning methods. Moreover, this research emphasized the importance of modeling time gaps between different medical records for improving the performance of prediction and addressing the utilities of a generic framework to conduct comparative research studies.

## An ensemble approach for diabetes prediction and classification using soft voting classifier

Kumari et al., (2021) state that diabetes represents a dreadful disease that is caused by to escalated level of blood glucose. Machine learning methods are very helpful in the early-stage recognition and prediction of diabetes. The main focus of this research is to predict diabetes with higher accuracy with the help of an ensemble ML algorithm. The PID dataset is used to conduct experiments for gathering the details of patients with and without diabetes. The present soft voting classification provides optimized binary classification and utilizes a group of 3 different ML algorithms, such as logistic regression, random forest, and Naïve Bayes for diabetes classification. Further, an empirical evaluation is presented for the proposed model using the state-of-art classifiers, including XGBoost, AdaBoost, random forest, support vector machine, logistic regression, Naïve Bayes, CatBoost, and GredientBoost considering the precision, accuracy, recall, and F1-score as the evaluation criteria with 73.4%, 79.04%, 71.45%, and 80,6%, correspondingly on the PID dataset. Moreover, the efficiency of this proposed system is analyzed and compared with the breast cancer dataset which provides 97.02% accuracy (Kumari et al., 2021).

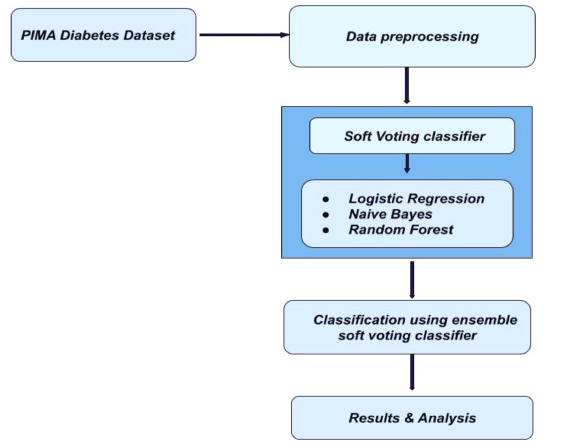


Figure : Flow diagrams of the proposed ensemble method using soft voting classifier

## An ensemble approach for diabetes prediction and classification using a soft voting classifier with the explicable AI

Kibria et al., (2022) argued that diabetes is the most severe chronic disease that spreads worldwide. In the past few years, several researchers endeavored to develop an efficient and reliable model for diabetes prediction using ML algorithms. These research studies pose fewer impacts on clinical practices as the current research studies focus on enhancing the ML model’s performance while avoiding explainability in clinical situations. Thus, healthcare professionals face difficulties in understanding these ML models and don’t trust them for clinical use. This research proposes an efficient, interpretable, and carefully constructed detection method with an explainable AI. A PID dataset is used consisting of 768 instances where 500 cases are non-diabetic and 268 are diabetic with different diabetic attributes. Here siz ML algorithms, namely support vector machine, random forest, logistic regression, AdaBoost, XGBoost, and artificial neural network are used with ensemble classifiers for diagnosing diabetes. Further SHAP (shapely additive explanations) are provided for each ML algorithm to help healthcare professionals better understand the used methods. The accuracy of the proposed ensemble model is about 90% with an F1-score of 89% for 5-fold cross-validation (Kibria et al., 2022). The proposed model can better enhance the clinical understanding of healthcare professionals about diabetes diagnosis and help them to take the best actions at the early stage of the disease.

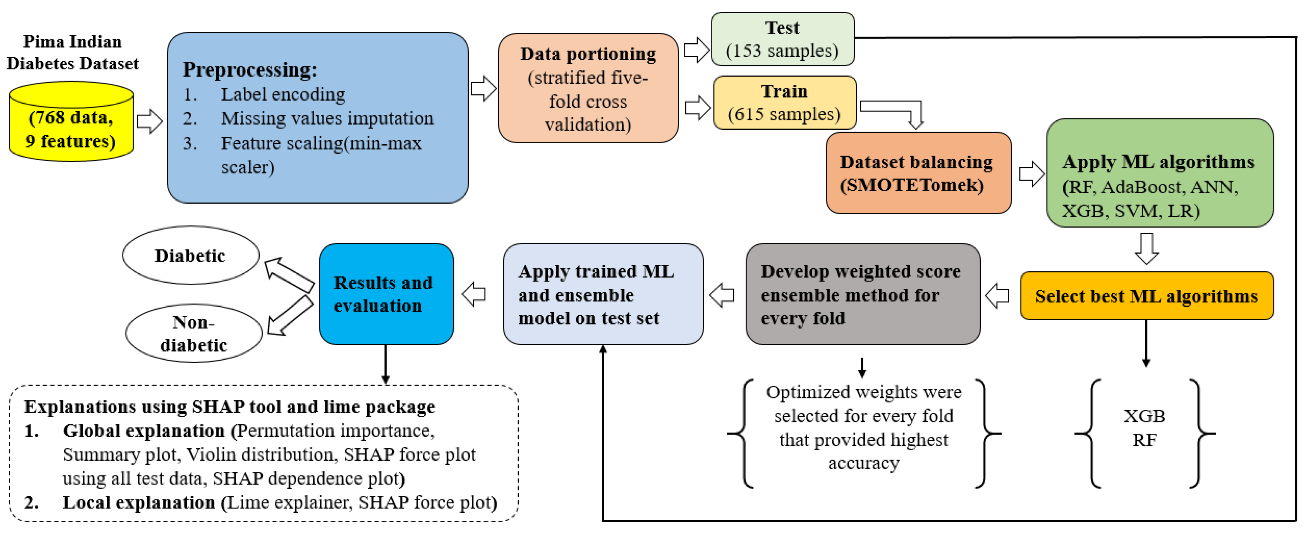


Figure :Workflow of the presented diabetic detection model

## An innovative stacking ensemble learning method for diabetes detection

According to Gollapalli et al., (2022), glucose represents a major source of energy for the human cells that are considered the most vital building blocks of life. The imbalance in glucose level represents a sign of diabetes, which is a very common chronic disease. It might lead to long-lasting complications, such as kidney failure, blindness, and heart disease. In the last three years, Saudi Arabia documented about a 10-fold increment in diabetic cases. Diabetes can be broadly categorized into three types, pre-diabetes, type 1 diabetes, and type 2 diabetes. The diagnosis of the correct type of this disease sometimes creates ambiguity for healthcare professionals. However, there are huge efforts have been made in predicting type 2 diabetes but there is a major lack of studies that emphasize to predict pre-diabetes and type 1 diabetes. Thus, this research study mainly focuses on distinguishing and predicting diabetes and controlling its progression. This research conducted about 4 different experiments using different ML algorithms, such as decision tree, support vector machine, random forest, bagging, stacking, and k-nearest Neighbor. Further, SMOTE (synthetic minority oversampling technique) is applied in balancing the chosen dataset (Gollapalli et al., 2022). The experimental results demonstrate the most promising results of the proposed novel stacking model, combining k-NN, bagging k-NN, and bagging DT achieving a prediction, recall, accuracy, and Cohen’s kappa score of 94.70%, 94.48%, 94.48%, and 0.9172.

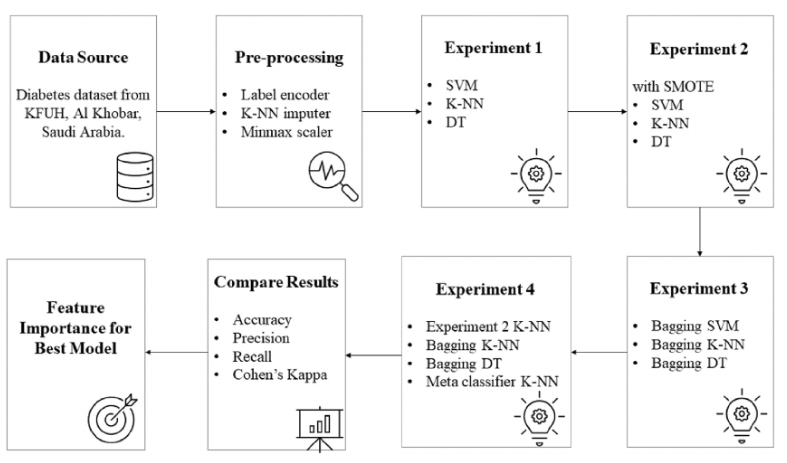


Figure : Research study's framework

## Secure federated learning in the prediction of multi-party diabetes

According to Su et al., (2023), the increased development of populating aging, social economy, and increased strength of risk factors can cause NCDs (chronic non-communicable diseases) that severely impact human health. This research addresses a major lack of the exchange of data between the municipalities and medical institutions in the anticipation and management of chronic diseases like diabetes. This research study aims to analyze the prediction of diabetes based on securely shared data between the medical investigation centers and hospitals, where the federated learning methods are utilized and associated security challenges are addressed. This research designed and proposed a customized and secured protocol for the tree and regression family algorithms considering the prediction problem of general disease for solving the issue of data leakage. Further, the author utilized a joint data modeling approach by combining logistic regression, neural network, XGBoost, and LightGBM among diverse organizations (Su et al., 2023). The experimental results conducted on the two different datasets, concerning precision, recall, and computational cost metrics, demonstrate that the federated learning models can better utilize patient data among different organizations for delivering an enhanced and reliable system for diabetes prediction.

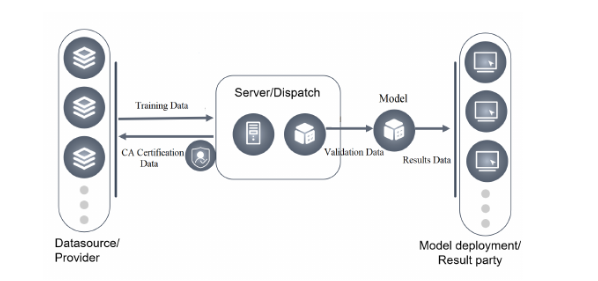
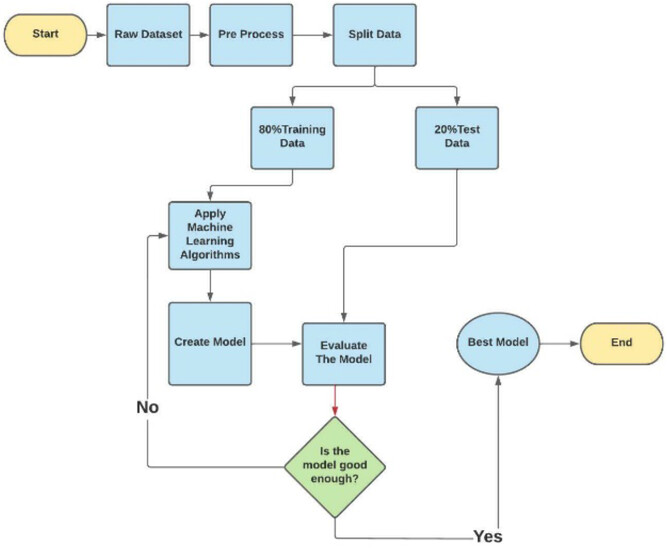


Figure : Federate learning platform's system layout

## KFPredict, an ensemble learning framework to predict diabetes following key feature fusion

According to Qi et al., (2023), diabetes is most likely to occur at the last stage of this disease, so it needs early recognition and treatment to avoid the caused complications. This research proposed an ensemble learning model, namely KFPredict for combining the multi-input models with the ML algorithms and other key features. First a multi-input neural network (KF\_NN) model is used for fusing the key features and using a decision-tree-enabled recursive features exclusion algorithm and correlational coefficient technique for screening the key input features and other secondary input features within the model. Further, KF\_NN is ensembled with three other ML algorithms, such as random forest, support vector machine, and K-nearest neighbor for realizing soft voting for forming predictive classifiers for more accurate diabetes prediction. This proposed model offers excellent prediction results on the testing set with a 93.5% accuracy, 0.85 sensitivity, and 0.98 specificity. This accuracy is about 18.18% higher than the KFPredict when used alone. Moreover, this research also compared KFPredict with existing methods and it outperformed the other methods with an improved accuracy of 14.93% (Qi et al., 2023).



# Conclusion

Here I presented a critical review and discussion of the existing and relevant research papers and journal articles to gain a better understanding of this research topic. Here I provided a critical review of 20 research papers to form a solid foundation of this research topic and analyze the other researcher’s work and the current research gaps along with the key directions to overwhelm those research gaps.

Chapter 3: Research Methodology

# Introduction

This is the third chapter of this thesis, which will provide an in-depth discussion of the selected research methodology and other methods or procedures that are employed in this research for gathering and analyzing research data with the rationale behind their selection. It can serve as the roadmap for the whole research by outlining the key steps for addressing the research objectives and questions. In this research part, I will discuss the chosen research design, research philosophy, and research approach along with the other techniques and tools used in this research. Moreover, here I will discuss the programming language and other libraries that are used to develop the necessary codes for the implementation of the proposed system.

# Research Methodology

Research methodology is an important part of any research, which assists how to better research while addressing the key research objective and questions. An efficient research methodology can offer a structured and systematic approach to conducting the research and serves as the potential guide for better gathering, analyzing, and interpreting research data validly and consistently. It helps in ensuring the validity and reliability of the collected research data along with realizing a clear roadmap while outlining the key steps that need to be followed throughout the research. It also helps the researcher to remain focused and organized to better address and interpret research problems, questions, and objectives.

## Mixed research methodology

In this research, I used **mixed research methodology**, which is a very powerful method to conduct this research by combining the vital research insights of both qualitative and quantitative research methods to offer an in-depth and nuanced understanding of this research along with associated problems and suitable solutions (Team, 2023). The use of mixed research methodology will offer a more holistic understanding of this research topic by providing subjective perspectives, experiences, and motivations along with the statistical analysis.

### Justification

The mixed research methodology is a very effective and suitable research methodology for this research as it can help to realize data triangulation, complementary insights, comprehensive exploration, improved validity, improved generalizability, and practical relevance. Also, this research methodology leverages the strengths of both quantitative and qualitative methods to attain a comprehensive and robust understanding of this research that leads to more expressive and impactful research results. The combination of both research methods can strengthen the validity of research findings by offering a detailed and nuanced understanding of this research topic. Moreover, the use of mixed research methodology can help to explore specific research contexts along with generalizing research findings for strengthening the validity and generalizability of the research results.

## Research design

The research design is very important in any research work as it offers a suitable blueprint to conduct this research study. In this research, I will use a **convergent or concurrent research design**. It will help me highlight the overall research methodology, structure, and approach that could be used for addressing the research objectives and questions (George, 2023). This research design is suitable for this research study because it enables an inclusive exploration of both objective measurements and subjective experiments. It also facilitates data triangulation from different methods and sources to enhance the reliability and validity of the obtained research findings.

### Justification

I used this research design because it can provide a clear structure or plan for this research by considering both qualitative and quantitative aspects of this research and putting equal emphasis on both data types. It also depicts the most critical steps that need to be followed throughout the use of data collection and analysis methods. It also helps keep the researcher focused and organized while addressing the research research problems. It can also help in ensuring data validity and reliability along with reducing errors and biases while ensuring that the obtained research findings correctly deliberate the research phenomenon.

## Research Philosophy

In this research, I used **pragmatism research philosophy** to highlight the usefulness and applicability of the obtained research findings. It helps in recognizing the importance of both qualitative and quantitative methods and seeks to combine them to offer an in-depth understanding of this research topic (Allemang et al., 2021). This research philosophy can be effectively used to develop an effective and practical system for the detection of diabetes. It also enables the utilization of different methodologies and data sources, including both qualitative and quantitative data for realizing an inclusive understanding of the research problem while considering the user’s needs and perspectives to ensure that the proposed system is adequate and relevant.

### Justification

This research philosophy is most suitable for this research as it enables to remain flexible in the adopted research approach and methods as per the selected research objectives and questions. This flexibility is especially significant in the mixed research methodology where several research approaches are utilized for gaining an immense understanding of this research topic. Also, it emphasizes the usefulness of research findings for generating robust knowledge to enhance the used practices. It also seeks a phenomenon from a more holistic perspective by considering the multiple dimensions and viewpoints. Moreover, it emphasizes the integration of theoretical and practical knowledge related to this research topic.

## Research approach

A **transformative research approach** is used in this research to develop the proposed ML-based diabetes detection system. This research approach emphasizes potential advancements in the knowledge related to this research along with addressing its impacts on healthcare professionals, individuals, and communities (Mertens, 2021). It helps to clearly describe the societal challenges relayed to diabetes on an individual’s health considering the social and economic impacts. It guides in building a user-friendly and user-centered system design for ensuring an in-depth exploration. It ensures that the proposed system is accessible, fair, and culturally suitable for different populations.

### Justification

The transformative research approach is an efficient approach that mainly focuses on creating positive changes and empowering communities or individuals. It identifies that this research has immense potential to address the associated challenges, and social inequalities, and promote social justice. It highlights the significance of the understanding of the context in which the research is conducted. Also, it values the active participation of communities or individuals in the research study to enhance the decision-making process. Moreover, this research approach will enable an analysis of the considered issue of diabetes at multi-stage levels, such as organizational, individual, societal, and interpersonal levels.

## Materials and Methods

### Dataset description

In this research, we used the Kaggle dataset, and the required data was gathered from (Akturk, 2020). After the collection of data, it is cleaned and pre-processed to recognize missing values and unbalanced classes. This dataset is taken from the National Institute of Diabetes & Digestive and Kidney Diseases. The main focus of this dataset is to predict diabetes based on diagnostic measurements to ensure whether or not the patient suffers from diabetes. Few constraints are placed to select the most relevant instances from this large dataset. Particularly, all of the involved patients are females of at least 21 years (Kibria et al., 2022). This dataset comprises about 768 instances and 8 attributes for the detection of diabetes in formulating binary classification problems. This dataset comprises the details about only female patients. I selected this dataset because it is one of the most common and publicly available datasets for predicting diabetes and several research studies used this dataset for developing this model.

### Data preprocessing

Data mining methods are used for data pre-processing. It plays an important role in evaluating model performance. The taken dataset contains several missing values and imbalanced data.

#### Imputation of missing values

The figure depicts all missing values presented in the dataset. This figure addresses that insulin comprises the highest missing values as compared to other features. Some of the features have 0 values that do not make any sense. These values are referred to as missing values in this dataset (Kibria et al., 2022). The features in which 0 is treated as the missing values are blood pressure, glucose, insulin, BMI, and skin thickness. For replacing the feature’s missing values, the median is taken to conform target value.

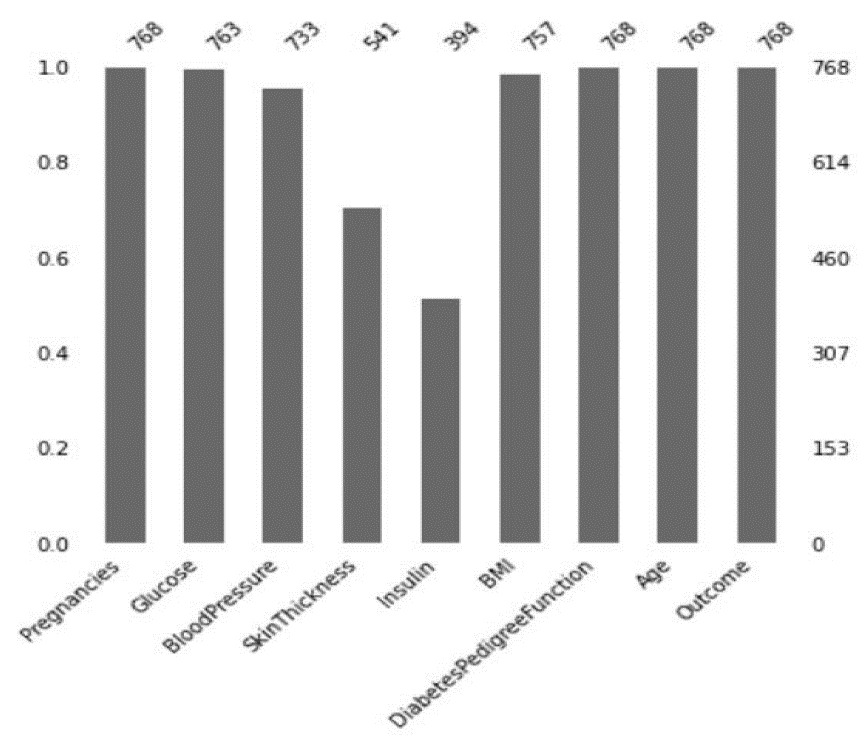


Figure : Feature's missing values before imputation

#### Data segmentation

The whole dataset is graded and 5-fold cross-validation is applied on this dataset.

#### Treating imbalanced dataset classes

In this research, a SMOTETomek method is used for balancing the dataset classes. This method comprises Tomek and SMOTE algorithms. Here first SMOTE is applied for creating a novel synthetic minority sample for obtaining a classes’ balanced distribution. Then Tomek is used for removing the closer sample with the class boundary to increase the separation between classes (Kibria et al., 2022).

Table : No. of classes before and after applying SMOTETomek on the training dataset

|  |  |  |
| --- | --- | --- |
|  | Before applying SMOTETomek | After applying SMOTETomek |
| Numbers is class 1 (diabetic) | 214 | 393 |
| Numbers in class 0 (non-diabetic) | 400 | 393 |

#### Feature scaling

I used a min-max normalization method on the selected dataset except for the tree algorithms like the random forest algorithm. the logistic regression and support vector machine algorithms required normalization. The min-max scaler can be illustrated as follows:

Where represents normalized values, whereas h represents the original value.

#### Boruta feature selection

The Boruta algorithm is a feature selection algorithm, which is based on the random forest classifier. The main aim of this algorithm is to select a set of most relevant and vital features to a dependent variable instead of any specific model https://www.techtarget.com/searchcio/definition/Agile-project-management. In this research (Zhou et al., 2023), I used the Boruta feature selection algorithm for choosing the most relevant features with the higher predictive relevance from the other eight associated features with the prediction of diabetes, such as age, insulin, BMI, glucose, and diabetes spectrum function.

### Methods

In this research, I used several machine learning methods, such as NB (Naïve Bayes), KNN (K-Nearest Neighbor), DT (Decision Tree), RF (Random Forest), and SVM (Support Vector Machine) (Jaloli & Cescon, 2022). These all research methods are effective and capable enough to better conduct this research based on the development of a system for early detection and prediction of diabetes.

#### Random Forest

This machine learning method represents a collaborative learning method for estimating the classification (class) or regression (number) according to the problem’s type by creating a larger no. of decision trees throughout the entire training phase for regression, classification, and other tasks. It also refers to the boosting method. Here a data sample is provided to each learner model for building any decision trees and marging them for obtaining a more stable and adequate prediction. In this method, the fundamental learner models are decision trees. It can dramatically reduce the errors by averaging the independent base learners.

#### Naïve Bayes

This is another ML method that can be used in this research and it can better handle larger datasets with multiple variables & features due to its simplicity and higher computational efficiency. This ability of Naïve Bayes makes it a suitable method for this type of medical research. This method is based on the probability theory for modeling relationships among different variables and calculating the probabilities (Bhattacharya & Roy, 2022). Also, it helps recognize and estimate the missing value from the dataset for providing more accurate diabetes predictions while minimizing the influence of missing values on the overall system performance. It realizes interpretable outcomes based on the conditional dependencies among the variables for gaining vital insights related to diabetes prediction.

#### Decision Tree

This ML algorithm represents an interpretable and clear structure for representing an effective decision-making process. The model structure similar to trees enables the researchers to easily understand and understand the rules and regulations that lead toward a specific classification. It can effectively handle the non-linear relationships among different variables. It poses lower computational costs throughout the training phase, which makes it appropriate to develop a robust system for handling massive no. of variable and patient records. However, it may not perform well with the imbalanced datasets.

#### Logistic Regression

Logistic regression represents a statistical modeling method, which is mainly used for predicting the probability of any activity based on the independent variables (Arora et al., 2022). This is an interpretable and simple method for easily understanding and explaining the obtained results. It can tackle both continuous and categorical variables. However, it may not correctly capture the non-linear relationships or interactions among the independent variables. In this research, this ML method can be used for predicting the likelihood of a person having diabetes considering various relevant risk factors.

#### K-Nearest Neighbor

K-NN represents a non-pragmatic algorithm i.e. it cannot make any assumptions against the considered data. It makes this method suitable for the medical field, where diabetes and its risk factors do not follow any specific distribution. Also, it is simple to use and can efficiently handle larger datasets. It can better capture the complicated relationships that exist in the medical dataset. Thus, it can be considered an efficient and informed selection for the research based on early recognition of diabetes.

#### Support Vector Machine

The support vector machine represents a linear model used for regression and classification problems. It can solve both linear and non-linear problems. It works similarly to linear regression. In this ML model, this algorithm categorizes data based on a hyperplane with the maximum peripheral distance. SVM is capable of categorizing and differentiating diverse classes related to diabetes with multiple complex factors (Ahmed et al., 2021). It addresses the most relevant and critical risk factors related to diabetes for early recognition of diabetes while assisting physicians in better diagnosing and treating this disease. Moreover, this ML method can better handle imbalanced data by comprising the class weights and using undersampling or oversampling methods.

### Project management approach

In this research, I used the **Agile project management approach** because it is an interactive and flexible approach to significantly enhance this research for developing a robust ML-based diabetic detection system for an accurate detection and classification of diabetes disease. It can help to better manage this project by dividing it into small tasks, known as sprints (Gillis et al., 2023). Using it, a personalized list of the vital features, functionalities, and tasks can be created. Further, its core principles can be adapted in the context of this research to realize fast progress and enhanced stakeholder management. Moreover, it can promote consistent collaboration and improvements throughout this research thesis to better manage this project.

### Programming language and libraries

In this research, we used **Python programming language**, which is a comprehensively used programming language largely used in the field of machine learning and data science. It is versatile enough and could be used for different purposes, such as scientific computing, web or app development, automation, and so on. Its vibrant community vigorously contributes to the development of necessary tools, libraries, and other resources. In this research, we used the following libraries.

#### Scikit-learn

It provides a wide range of supervised and unsupervised ML algorithms, including random forest, support vector machine, and decision tree. It offers required tools for model training data processing, feature selection, and performance evaluation. Further, it can simplify working with diverse algorithms and facilitate model sharing and comparison.

#### Matplotlib

Matplotli library represents a comprehensive library to create interactive, static, and interactive visualizations in the Python programming language (M, 2021). It makes hard things to be possible and easy things to be easier. It can create effective and publication-quality plots.

#### Numpy

It represents Python’s scientific computing library that can offer the most powerful tools to better manipulate and analyze numerical data. It can also be used for linear algebra, array-centric computations, and Fourier transform arbitrary number functions.

#### Pandas

Pandas represent a data analysis library that can offer strong data analysis tools and high-level data structures. It can be used for data cleaning, data wrangling, and data preparation (AlmaBetter, 2023). It is designed to make the process of data analysis and data manipulation easier and more intuitive.

#### Seaborn

Seaborn represents a data visualization library that offers higher-level plotting functions to create the most informative and attractive visualizations. It can be optimized to work with pandas’ data structures and integrated with the scikit-learn and NumPy libraries.

# Conclusion

In this chapter, I discussed the chosen research methodology to conduct this research along with other methods, tools, and techniques. Here I selected mixed research methodology along with the convergent research design, pragmatism research philosophy, and transformative research approach. Then I described the chosen Kaggle dataset with 768 instances and 8 attributes. Then I discussed the ML methods, such as NB, KNN, DT, RF, and SVM. Then I discussed the Python programming language along with the libraries, such as scikit-learn, matplotlib, NumPy, pandas, and seaborn libraries.

Chapter 4: Design, Development, and Implementation

# Introduction

This chapter is based on a detailed discussion of the design and implementation of the proposed diabetes detection and prediction system. Here I will outline the design of the proposed system, including the key components and behaviors. Then I will provide a discussion of the system implementation by providing evidence of the performed practical work. This chapter will demonstrate this research project’s practical aspects while demonstrating that the proposed system is designed, developed, and eventually implemented. It can serve as the bridge between the discussed theoretical aspects of this research and with practical outcomes of this research.

# Design, Development, and Implementation

## Design

The system design represents one of the most critical parts of this research because it guides to determining an effective strategy to carry out the entire research by realizing an effective research planning. The system design consists few key components that need to be converted into actions to create a low-level and high-level architecture of the proposed system. Here I will provide the low-level architecture of the proposed system, which comprises four key components, namely the user, the proposed diabetic detection system, machine learning models, and the database server (Aslan & Sabanci, 2023). The use of machine learning models can drastically improve the efficiency and accuracy of diabetes prediction and detection to timely cure it for prevent any severe impacts on patient health. The use of Scikit-learn, NumPy, Matplotlib, Pandas, and Seaborn libraries provide specific benchmarks for conveniently implementing this proposed system while making this system more compatible and user-friendly to enhance the user experience. Figure 16 represents a low-level architecture of the proposed ML-based diabetic detection system.

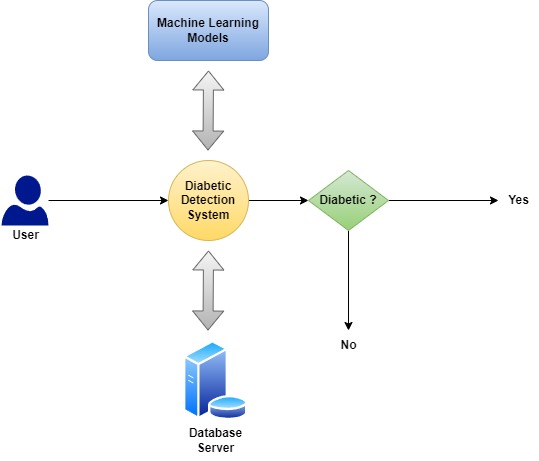


Figure : Low-level architecture of the proposed system

## Development and Implementation

In this research, I developed a diabetic detection system using machine learning methods and ensemble learning methods to improve the accuracy of the prediction of diabetes using this proposed system.

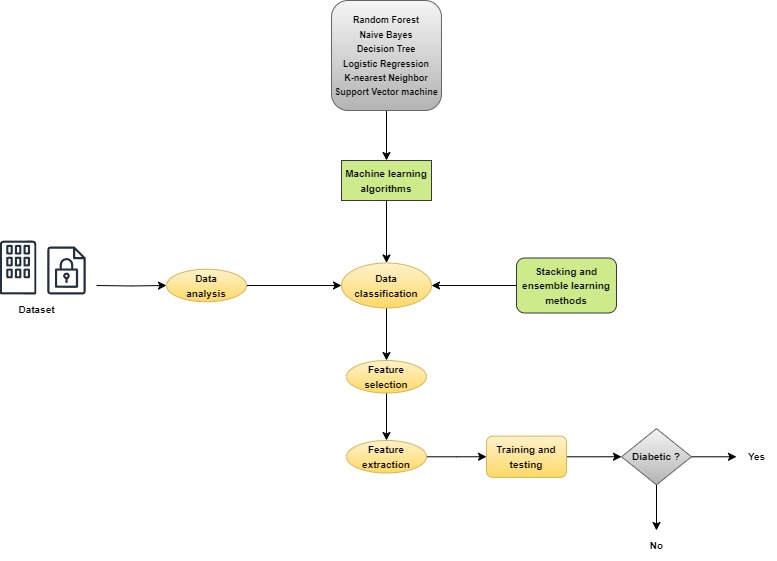


Figure : High-level architecture of the proposed diabetic detection system

The above-illustrated diagram represents a high-level architecture of the proposed system. According to this architecture, a dataset is selected, which comprised health-related data of patients (Butt et al., 2021). This dataset has 768 instances and 9 attributes. Then this dataset is analyzed using effective data analysis methods and data classification is performed using machine learning algorithms, such as Random Forest, Naïve Bayes, Decision Tree, Logistic Regression, K-nearest Neighbor, and Support Vector Machine algorithms. Further stacking and ensemble learning methods are applied to enhance the data classification process. Then vital features from the classified health data are extracted and selected. Then these extracted features are tested and trained to ensure whether the user is a diabetic patient or not. If the patient is diabetic, then appropriate medication, diet, and physical activities are suggested accordingly.

### Evidence of performed practical work

The below-illustrated figure depicts the details of the imported libraries for developing and implementing the proposed system. Here we imported pandas, numpuy, matplotlib, and seaborn libraries.

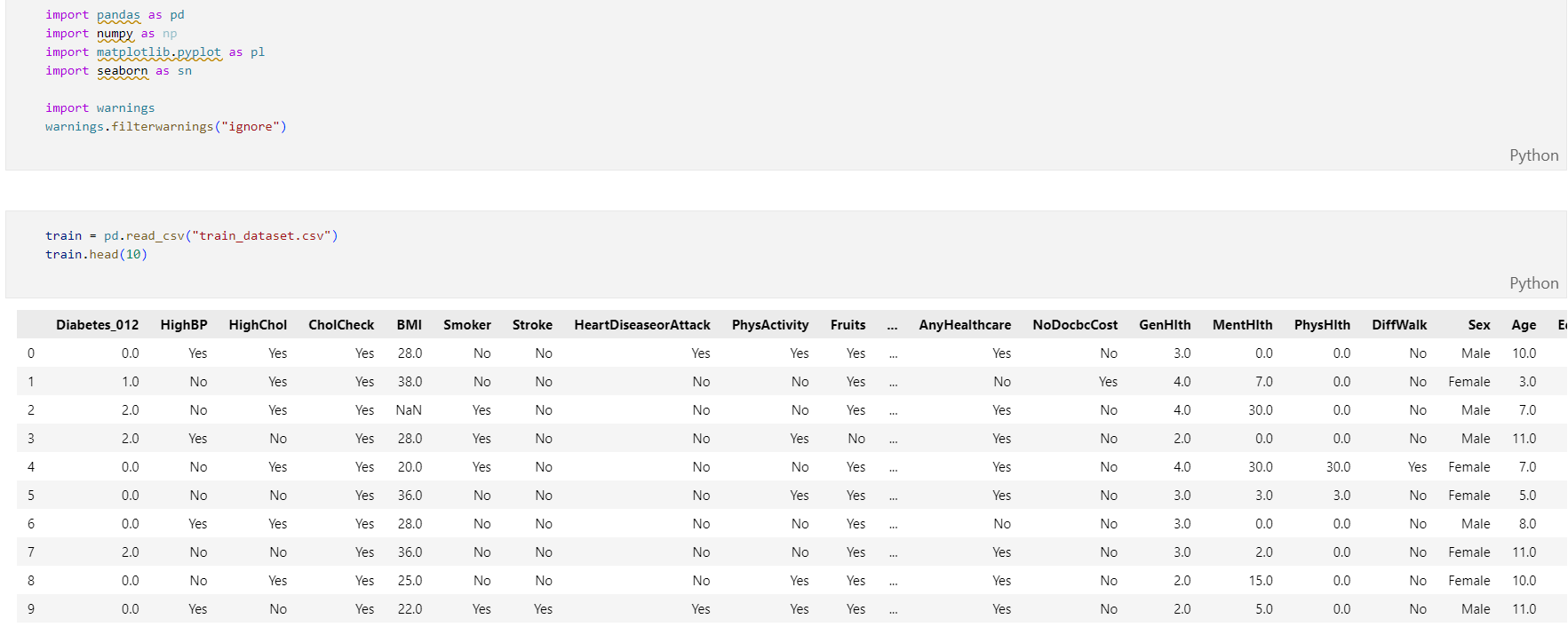


Figure : Importing necessary libraries

The below-illustrated figure depicts the details of the attributes of the selected dataset, including their name and data type.

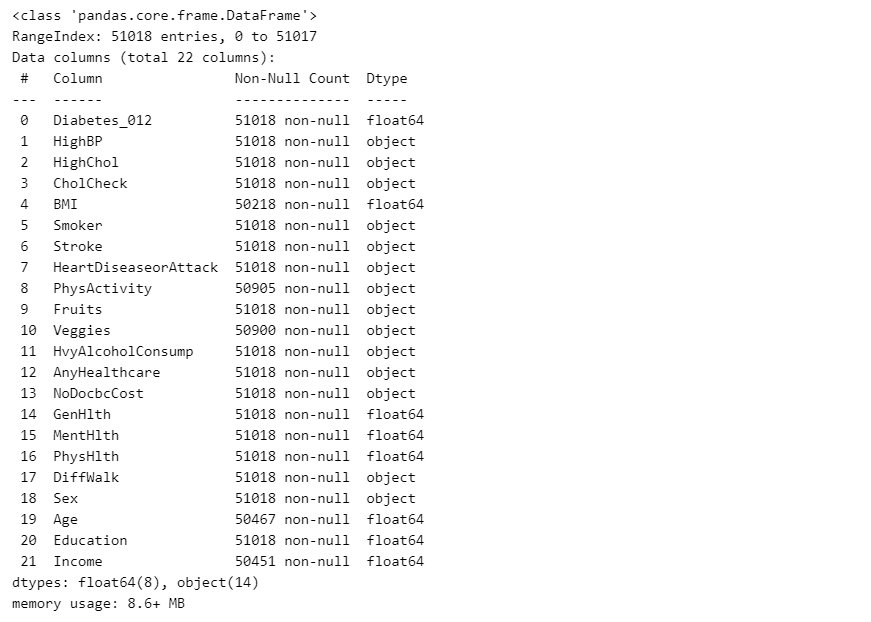


Figure : Attributes of the dataset

This diagram represents details of the Numerical data frame, including general health, mental health, physical health, age, education, & income, and Object data frame, such as high BP, High cholesterol, Smoking, Stroke, Heart disease, and physical activities (Ahmed et al., 2021).

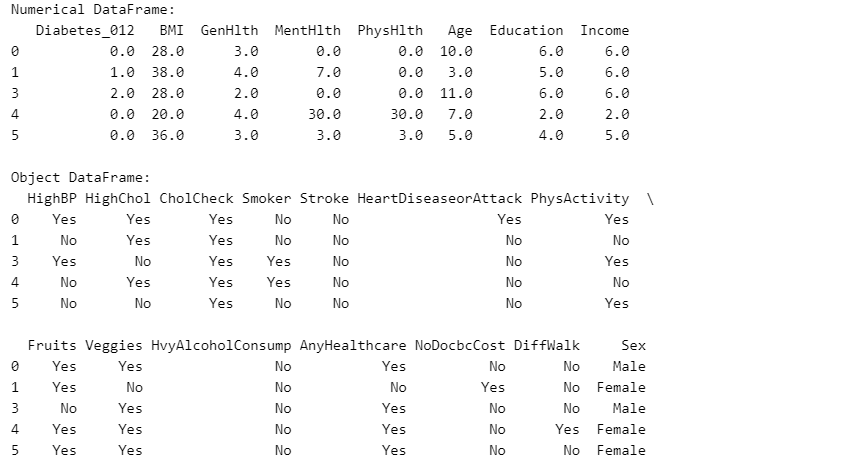


Figure : Numerical and object data frame

The below-illustrated figure represents the importing of the standard scale from the preprocessing of the sklearn library. Here numerical data frame is presented in columns with their respective values.



Figure : Importing standard scaler

A train\_df function is used for training all the attributes of the numerical data frame with their values.

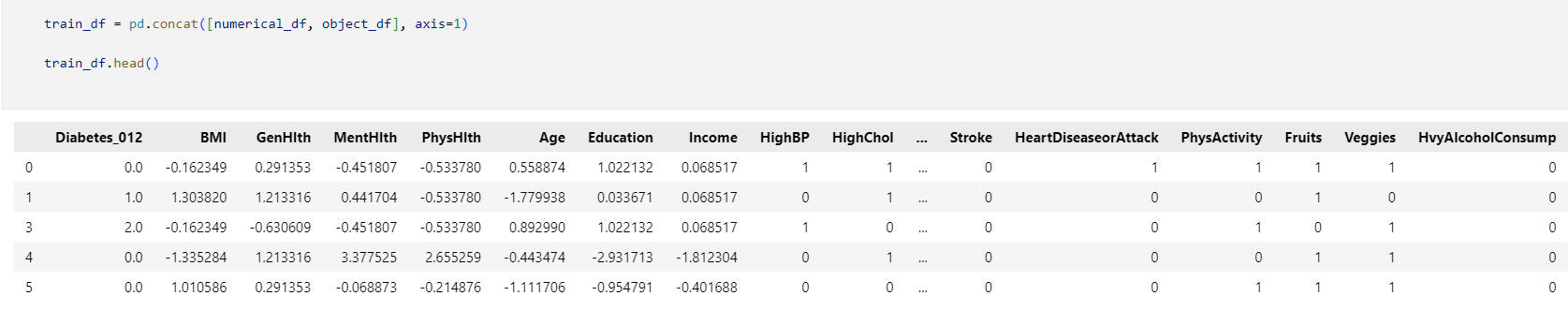


Figure : Training of the numerical data frame

Here two parameters figure, axes are used for creating subplots and a pl.tight\_layout() is used for automatically adjusting subplots to set in all figure areas (Zach, 2023).

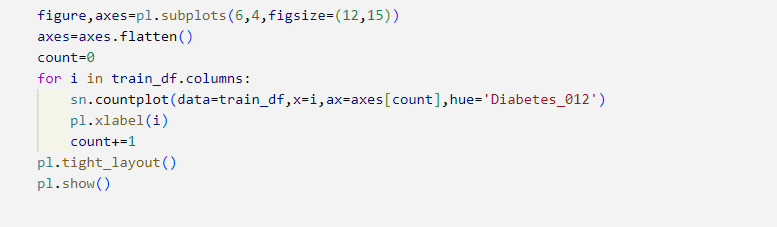


Figure : Plotting of the figures

The below-illustrated figure depicts the created subplots for the considered attributes of the selected dataset.

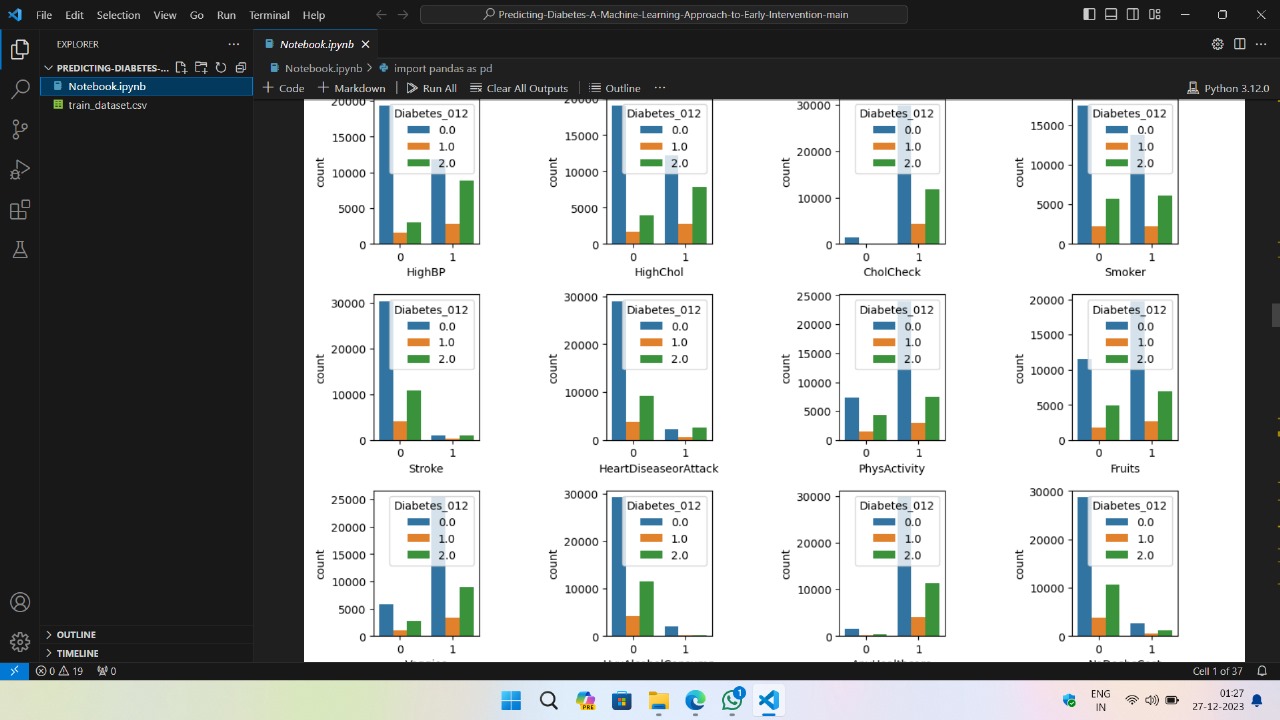


Figure : Subplots for dataset attributes

The below-illustrated figure depicts a correlation matrix heatmap for visualization of the correlation between different metrics to provide an easier understanding. Here seaborn lbrary is used for plotting the heatmap.

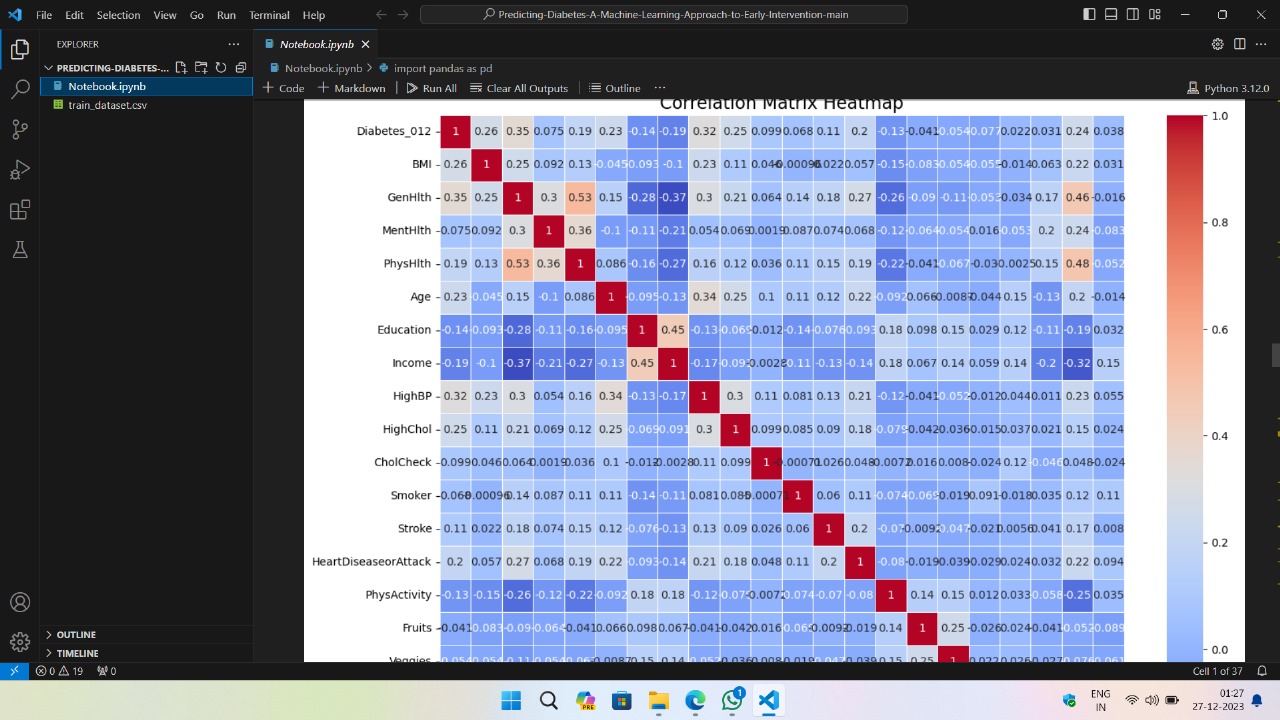


Figure : Correlation matrix heatmap

In these data attributes, some of the attributes contain missing or null values. So here I used the train\_df function for handling those null or missing values from these attributes (Reader, 2021).



Figure : Handling missing values

The below-illustrated diagram depicts the details of the defined ML classifiers, such as decision tree, logistic regression, KNN, random forest, XGBoost, and AdaBoost (Iparraguirre-Villanueva et al., 2023). Then they are trained and evaluated based on their accuracy of the diabetes detection and classification.



Figure : Define, train, and evaluate the selected ML classifiers

The below-illustrated figure depicts the details of the defined hyperparameters from GridSearchCV using the scikit-learn library.



Figure : Defining hyperparameters

The below-illustrated figure depicts the creation of an efficient logistic regression model using the best hyperparameters, including precision, accuracy, recall, and F1-score.



Figure : Creating an LR model using the best hyperparameters

Here the hyperparameter grid is defined including a few critical parameters. Then a RandomizedSearchCV is created to print the best hyperparameters.



Figure : Defining hyperparameter grid and creating RandomizedSearchCV

Then the best hyperparameters, such as criterion, bootstrap, max\_depth, min\_samples\_leaf, min\_samples\_split, and n\_estimators are taken from the RandomizedSearchCV and used to create an accuracy score, classification report, and confusion matrix (Bhattacharya & Roy, 2022).



Figure : Creating classification report, accuracy score, and confusion matrix

Then all the used ML classifiers, such as KNN, decision tree, random forest, logistic regression, XGBoost, and AdaBoost are represented.



Figure : Representing the used classifiers

Chapter 5: Testing, Results, and Discussion

# Introduction

This is the fifth chapter of this research thesis, which is based on the discussion of selected testing methods, obtained research results, and discussion of those results. Here I first discuss the suitable testing method to test the effectiveness of the proposed system for an effective and accurate prediction of diabetes and ensure whether it operates or not as intended. Then I will discuss the obtained results and research findings. Then I will evaluate the performance of the proposed system while comparing the effectiveness of the used machine learning methods. Then I will deliberate on the economic and commercial feasibility of this proposed system.

# Testing, Results, and Discussion

## Testing

Software testing is very important to test the effectiveness of any software system to ensure that it does not have any errors or bugs and it behaves as expected (Yasar & Black, 2023). Multi-level testing helps in the early recognition of any errors or bugs to timely resolve them to ensure an efficient quality of the system. For the proposed system, I conducted a software testing process in four stages, namely unit testing, integration testing, system testing, and user acceptance testing.

* **Unit testing**: This is the first level of testing, which is performed by the developers. Here an individual component or module of this proposed system is tested. This testing is limited to only a single component or module (Hamilton, 2023). It provides an early detection of any bugs or errors to save the required time and money for fixing them. It cannot address the integration issues, so it might be possible that the perfect component in isolation does not work well after integration.
* **Integration testing**: This is the second level of testing in which a group of relevant modules is tested for identifying interfacing issues between the modules. This testing ensures the compliance of different system components after integration for simulating the module functioning.
* **System testing**: This is the third level of testing, where the whole integrated application or system is tested. This testing mainly intends to determine the ability of the proposed system to adapt to the business requirements. This testing can be carried out in an atmosphere that is similar to the production atmosphere.
* **User acceptance testing**: This is the final testing level after which the developed application or system can be released in real-world settings. It intends to ensure that the proposed system meets or does not the intended business requirements following the specified quality standards (Abhisek et al., 2023).

## Results

The proposed diabetes detection and prediction mode is tested and evaluated using the selected machine learning methods, such as Decision Tree, Naïve Bayes, Random Forest, K-Nearest Neighbor, and Support Vector Machine algorithms. To address the effectiveness, I further used a Kaggle dataset that contain about 768 instances and 9 attributes.

### Dataset

This dataset comprises about 768 instances and 9 attributes for the detection of diabetes in formulating binary classification problems. This dataset comprises the details about only female patients. Here table 1 represents the attribute’s description and table 2 represents the dataset’s statistical values (Akturk, 2020).

Table : Description of the available attributes in the dataset

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Type of attribute** | **Description of attribute** |
| Pregnancies | Numeric | Number of times a patient pregnant |
| Blood pressure | Numeric | Diastolic pressure of blood in mmHg |
| Glucose | Numeric | Concentration of plasma glucose in mmol/L |
| Insulin | Numeric | 2-hour serum insulin in muU/ml |
| Skin thickness | Numeric | Skin fold thickness of triceps in mm |
| Age | Numeric | Age in years |
| BMI | Numeric | Body Mass Index = Weight in Kg/Height in Meter |
| Diabetes pedigree function | Numeric | The diabetes pedigree functions represent a function for measuring the diabetes chances following the family history. |

Table : Statistical description of the dataset

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Pregnancies | Blood pressure | Glucose | Insulin | Skin thickness | Age | BMI | Diabetes Pedigree Function | Outcome |
| count | 768 | 768 | 768 | 768 | 768 | 768 | 768 | 768 | 768 |
| std | 3.36 | 12.2 | 30.49 | 98.38 | 9.42 | 11.76 | 6.88 | 0.33 | 0.47 |
| mean | 3.84 | 72.37 | 121.59 | 153.18 | 29.11 | 33.24 | 32.42 | 0.47 | 0.34 |
| max | 17 | 122 | 199 | 846 | 99 | 81 | 67.1 | 2.42 | 1 |
| 75% | 6 | 80 | 140.25 | 190.15 | 35 | 41 | 36.6 | 0.2 | 1 |
| 50% | 3 | 72 | 117 | 133.7 | 29 | 29 | 32.09 | 0.37 | 0 |
| 25% | 1 | 64 | 99 | 87.9 | 23 | 24 | 27.5 | 0.24 | 0 |
| min | 0 | 24 | 44 | 14 | 7 | 21 | 18.2 | 0.07 | 0 |

### Performance metrics

The performance measures of the proposed model can be measured by the confusion matrix illustrated in Table 4. This confusion matrix comprises 4 diverse outcomes, such as TP (true positive), FP (false positive), TN (true negative), and FN (false negative).

Table : Cnfusion matrix

|  |  |  |
| --- | --- | --- |
| **Predicted Outcomes** | **Actual Positive** | **Actual Negative** |
| NO | FN | TN |
| Yes | TP | FP |

The following metrics are considered for analyzing the proposed model.

* **Precision**: This performance metric represents the proportion of the correct and positive predictions with positive predictions. It can be presented as follows:
* **Accuracy**: This performance metric helps in measuring the correctly made predictions and could be described as the ratio of the no. of correct predictions with the total no. of all test cases (Ahmed et al., 2021). It can be presented as follows:
* **Recall**: This performance metric represents the ratio of overall positive predictions with active positive values. It can be presented as follows:
* **F1-score**: This performance metric focuses on both recall and precision and could be described as follows,

### Experimental results

The experimental results obtained from this research study based on the Kaggle and Pima Indian Diabetes dataset are presented in Table 5. The proposed method attained an accuracy of 98.7% in the detection and prediction of the risk of diabetes.

Table : Performance evaluation of different ML methods

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Used ML methods** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| Decision Tree | 70.75 | 68 | 69 | 68.5 |
| Random Forest | 72.50 | 72 | 70.5 | 72.5 |
| Naïve Bayes | 75.50 | 74 | 72.5 | 73.9 |
| Logistic Regression | 79.30 | 75.8 | 74.5 | 74.5 |
| KNN | 79.97 | 76.8 | 75.5 | 75.9 |
| SVM | 82.50 | 80 | 77 | 79 |
| Proposed method | 98.7 | 98.5 | 96.9 | 97.5 |

### Research findings

Here I will present the key findings obtained from this research.

* The increased consumption of processed or fast foods and inactive lifestyles are the main causes behind the increased cases of diabetes across the world (Qi et al., 2023).
* There is a lack of an efficient method or system that can correctly predict and detect diabetes. The emergence of machine learning methods largely revolutionized the medical field, especially in the accurate prediction and detection of diabetes.
* This research addressed the effectiveness of the ML methods in the detection of diabetes at an early stage to prevent the severe health impacts of this disease.
* This research addressed the importance of features of ML models to enhance the detection of diabetes by emphasizing performance metrics.
* This research reveals the abilities of the proposed system to accurately and consistently predict and detect diabetes across different populations (Aslan & Sabanci, 2023).
* This research also addresses the key ethical considerations, such as informed consent, data privacy, and data biases during the development and implementation of this proposed model.
* This research also discusses the practical implications of this research, such as system interface design, integration with current healthcare infrastructure, and ease of use for end-users and healthcare professionals.

## Discussion

Accurate detection and classification of the risk of diabetes is a major problem in medical science. There are several machine learning methods or algorithms that could be used to address this problem. In this research, I used the five most effective and relevant ML algorithms, including DT, RF, NB, KNN, and SVM for identifying and classifying the associated problems with diabetes detection. All of the ML algorithms are very useful for the accurate detection and classification of diabetes, but each of them has its individual strengths and limitations. For instance, KNN is easy to use and implement but is very sensitive to the outlier data, whereas NB is less sensitive to the outlier data and better performs with noisy & large datasets. The RF, DT, and SVM algorithms behave in similar ways and are comprehensively used in the recognition and classification of diabetes. RF is simple and faster, DT is easy to understand & use, and SVM provides higher accuracy in the classification of non-linear data. The obtained results in this work are acceptable for only KNN and SVM algorithms.

Table : Comparison of two best ML algorithms

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Used ML methods** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| Logistic Regression | 79.30 | 75.8 | 74.5 | 74.5 |
| KNN | 79.97 | 76.8 | 75.5 | 75.9 |
| SVM | 82.50 | 80 | 77 | 79 |

According to Table 6, the used KNN ML algorithm attained an accuracy of 79.97%, which can better detect and classify diabetes, whereas SVM poses about 82.50% accuracy in the detection and classification of diabetes. These results depend on the used dataset and its characteristics. However, Logistic regression also poses a good accuracy of 79.97, which is significantly different from the attained accuracy in other works (Aslan & Sabanci, 2023). It can be efficiently used in categorizing peripheral diseases in patients using ML models and the attained accuracy is about 92% and the sensitivity is 91.85%. This difference is due to the type of dataset in which it is used and its characteristics. The DT and RF algorithms did not attain the expected outcomes, only achieving 70.75% and 72.50% accuracy, which are below the threshold value. However, in another research work, they attained about 85% and 89% accuracy, respectively. Thus, it cannot be concluded that the ML methods are not efficient, instead, the effectiveness of these methods depends on the used dataset and its characteristics.

### Comparison of the used ML algorithms

Here I will provide a graphical representation of the comparison of various used ML algorithms, such as DT, RF, NB, KNN, and SVM with respect to precision, accuracy, recall, and F1-score.

Figure : Comparison of ML algorithms based on their Accuracy

Figure : Comparison of ML algorithms based on their Precision

Figure : Comparison of ML algorithms based on their Recall

Figure : Comparison of ML algorithms based on their F1-score

# Conclusion

This chapter is based on the discussion of the used testing method to test the effectiveness of the proposed system, the obtained results, and a discussion of the research results. Here I first discussed the four-level testing method, including unit testing, integration testing, system testing, and user acceptance testing. Then I discussed the used dataset by describing the dataset’s attributes. Then I discussed the used performance metrics, such as precision, accuracy, recall, and F1-score. Then I discussed the obtained results by providing a comparison of used ML methods based on the considered performance metrics.

Chapter 6: Conclusion and Recommendations

# Conclusion and Recommendations

## Conclusion

This research thesis was based on the development of an efficient diabetic detection system for making an accurate detection, prediction, and classification of diabetes. The increased consumption of processed foods, inactive lifestyle, and lack of physical activities are the main causes behind the increased number of diabetic patients across the world. The existing methods used for the detection of diabetes are not effective and may provide wrong detection results. The increased evolution of machine learning methods largely enhances the medical field, especially in the prediction and detection of diabetes. In this research, we used six machine learning methods, including decision tree, random forest, Naïve Bayes, logistic regression, k-nearest neighbor, and support vector machine for the detection of diabetes. Among all of these methods support vector machine provides the highest accuracy in the prediction of diabetes.

Here we developed an efficient and robust diabetic detection system based on machine learning methods and an ensemble learning & stacking method to accurately predict and detect diabetes. We used Python programming language to create the necessary for real-time development and implementation of the proposed system. Further, we used different machine learning and Python libraries, such as scikit-learn, pandas, NumPy, seaborn, and matplotlib libraries. The conducted experimental results demonstrate its efficacy in the recognition and prediction of diabetes. This proposed system provides about 98.7% accuracy, 98.5% precision, 96.9% recall, and 97.5% F1 score.

In this thesis, I first provided a detailed introduction of this research project, including background information, problem statement, research questions, aim & objectives rationale, and thesis structure to provide an overview of this thesis. Then in the second chapter, I critically reviewed and discussed the existing research papers or journal articles to analyze other authors’ works and gain a better understanding of this research. In the third chapter, I discussed the chosen mixed research methodology to conduct this research thesis, including the transformative research approach, pragmatic research philosophy, and Agile project management methods. In the fourth chapter, I created the architecture of the proposed system along with its implementation and evidence of created codes. In the fifth, chapter, I provided the obtained results, testing methods, and a discussion of the results. In the sixth chapter, I concluded this research thesis by summarizing this research and key insights and providing recommendations to further proceed with this research.

## Recommendations

The obtained research findings in this research provide valuable insights to further proceed with this research in the future to explore the applicability of ML methods for better medical diagnosis and research. The research can adopt the following recommendations to further proceed with this research and conduct future research.

* Future research should focus on using more efficient feature engineering methods to improve the overall accuracy of this proposed model to enhance its performance.
* The availability of more data can enhance system accuracy by enabling ML models to capture underlying patterns and insights from data, which can enhance the performance measures of these models.
* The researchers should collaborate with healthcare organizations or institutions to attain access to independent datasets to confirm the robustness and effectiveness of the diabetic detection system.
* Future research should focus on using hyperparameter tuning to optimize the proposed system’s parameters to enhance its performance while decreasing the system's complexity and increasing generalizability.
* Further rigorous cross-validation strategies should be used for assessing the performance of the proposed system to increase its generalizability beyond training data.
* Future research should collaborate with healthcare experts and professionals and collect their insights and feedback for further refining this system and ensuring its usability and practicality.

# References

Abdulhadi, N. and Al-Mousa, A. (2021) ‘Diabetes detection using machine learning classification methods’, *2021 International Conference on Information Technology (ICIT)* [Preprint]. doi:10.1109/icit52682.2021.9491788.

Abhisek *et al.* (2023) *Levels of testing: Unit, integration, system and acceptance testing*, *ArtOfTesting*. Available at: https://artoftesting.com/levels-of-software-testing (Accessed: 31 December 2023).

Ahmed, N. *et al.* (2021) ‘Machine learning based diabetes prediction and development of Smart Web Application’, *International Journal of Cognitive Computing in Engineering*, 2, pp. 229–241. doi:10.1016/j.ijcce.2021.12.001.

Akturk, M. (2020) *Diabetes dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/mathchi/diabetes-data-set (Accessed: 30 December 2023).

Alfian, G. *et al.* (2020) ‘Blood glucose prediction model for type 1 diabetes based on artificial neural network with time-domain features’, *Biocybernetics and Biomedical Engineering*, 40(4), pp. 1586–1599. doi:10.1016/j.bbe.2020.10.004.

Allemang, B., Sitter, K. and Dimitropoulos, G. (2021) ‘Pragmatism as a paradigm for patient‐oriented research’, *Health Expectations*, 25(1), pp. 38–47. doi:10.1111/hex.13384.

AlmaBetter (2023) *Popular python libraries - numpy, Pandas, Seaborn, Sklearn*, *AlmaBetter*. Available at: https://www.almabetter.com/bytes/tutorials/python/popular-python-libraries (Accessed: 30 December 2023).

Amalia, R., Bustamam, A. and Sarwinda, D. (2021) ‘Detection and description generation of diabetic retinopathy using convolutional neural network and long short-term memory’, *Journal of Physics: Conference Series*, 1722(1), p. 012010. doi:10.1088/1742-6596/1722/1/012010.

Arora, N. *et al.* (2022) ‘A novel architecture for diabetes patients’ prediction using K-means clustering and SVM’, *Mathematical Problems in Engineering*, 2022, pp. 1–9. doi:10.1155/2022/4815521.

Aslan, M.F. and Sabanci, K. (2023) ‘A novel proposal for deep learning-based diabetes prediction: Converting clinical data to Image Data’, *Diagnostics*, 13(4), p. 796. doi:10.3390/diagnostics13040796.

Bhattacharya, A. and Roy, R.K. (2022) ‘Analysis of machine learning algorithms for prediction of diabetes’, *Swarm Intelligence and Machine Learning*, pp. 138–149. doi:10.1201/9781003240037-8.

Butt, U.M. *et al.* (2021) ‘Machine learning based diabetes classification and prediction for Healthcare Applications’, *Journal of Healthcare Engineering*, 2021, pp. 1–17. doi:10.1155/2021/9930985.

Chen, H. *et al.* (2014) ‘The diagnostics of diabetes mellitus based on ensemble modeling and hair/urine element level analysis’, *Computers in Biology and Medicine*, 50, pp. 70–75. doi:10.1016/j.compbiomed.2014.04.012.

Das, D., Biswas, S.Kr. and Bandyopadhyay, S. (2022) ‘A critical review on diagnosis of diabetic retinopathy using machine learning and Deep Learning’, *Multimedia Tools and Applications*, 81(18), pp. 25613–25655. doi:10.1007/s11042-022-12642-4.

Dritsas, E. and Trigka, M. (2022) ‘Data-driven machine-learning methods for diabetes risk prediction’, *Sensors*, 22(14), p. 5304. doi:10.3390/s22145304.

G., S.S. and K., M. (2019) ‘Diagnosis of diabetes diseases using optimized fuzzy rule set by grey wolf optimization’, *Pattern Recognition Letters*, 125, pp. 432–438. doi:10.1016/j.patrec.2019.06.005.

George, T. (2023) *Mixed methods research: Definition, guide & examples*, *Scribbr*. Available at: https://www.scribbr.com/methodology/mixed-methods-research/ (Accessed: 31 December 2023).

Gillis, A.S., Torode, C. and Pratt, M.K. (2023) *What is Agile Project Management (APM)?: Definition from TechTarget*, *CIO*. Available at: https://www.techtarget.com/searchcio/definition/Agile-project-management (Accessed: 31 December 2023).

Gollapalli, M. *et al.* (2022) ‘A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM’, *Computers in Biology and Medicine*, 147, p. 105757. doi:10.1016/j.compbiomed.2022.105757.

Hamilton, T. (2023) *Levels of testing in software testing*, *Guru99*. Available at: https://www.guru99.com/levels-of-testing.html#:~:text=Unit%20Testing%20%3A%20checks%20if%20software,functional%20needs%20for%20the%20testing. (Accessed: 31 December 2023).

Howsalya Devi, R.D., Bai, A. and Nagarajan, N. (2020) ‘A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms’, *Obesity Medicine*, 17, p. 100152. doi:10.1016/j.obmed.2019.100152.

Iparraguirre-Villanueva, O. *et al.* (2023) ‘Application of machine learning models for early detection and accurate classification of type 2 diabetes’, *Diagnostics*, 13(14), p. 2383. doi:10.3390/diagnostics13142383

Jaloli, M. and Cescon, M. (2022) ‘Long-term prediction of blood glucose levels in type 1 diabetes using a CNN-LSTM-based deep neural network’, *Journal of Diabetes Science and Technology*, 17(6), pp. 1590–1601. doi:10.1177/19322968221092785.

Kibria, H.B. *et al.* (2022) ‘An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI’, *Sensors*, 22(19), p. 7268. doi:10.3390/s22197268.

Kumari, S., Kumar, D. and Mittal, M. (2021) ‘An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier’, *International Journal of Cognitive Computing in Engineering*, 2, pp. 40–46. doi:10.1016/j.ijcce.2021.01.001.

Liu, J., Peng, B. and Yin, Z. (2023) ‘A hybrid machine learning method for diabetes detection based on unsupervised clustering’, *2023 The 7th International Conference on Machine Learning and Soft Computing (ICMLSC)* [Preprint]. doi:10.1145/3583788.3583809.

Lukmanto, R.B. *et al.* (2019) ‘Early detection of diabetes mellitus using feature selection and fuzzy support vector machine’, *Procedia Computer Science*, 157, pp. 46–54. doi:10.1016/j.procs.2019.08.140.

M, R. (2021) *What is Matplotlib in python? how to use it for plotting?*, *ActiveState*. Available at: https://www.activestate.com/resources/quick-reads/what-is-matplotlib-in-python-how-to-use-it-for-plotting/ (Accessed: 30 December 2023).

Mertens, D.M. (2021) ‘Transformative research methods to increase social impact for vulnerable groups and cultural minorities’, *International Journal of Qualitative Methods*, 20, p. 160940692110515. doi:10.1177/16094069211051563.

Perveen, S. *et al.* (2016) ‘Performance analysis of data mining classification techniques to predict diabetes’, *Procedia Computer Science*, 82, pp. 115–121. doi:10.1016/j.procs.2016.04.016.

Qi, H. *et al.* (2023) ‘KFPREDICT: An ensemble learning prediction framework for diabetes based on fusion of key features’, *Computer Methods and Programs in Biomedicine*, 231, p. 107378. doi:10.1016/j.cmpb.2023.107378.

Rabhi, S. *et al.* (2022) ‘Temporal Deep Learning Framework for retinopathy prediction in patients with type 1 diabetes’, *Artificial Intelligence in Medicine*, 133, p. 102408. doi:10.1016/j.artmed.2022.102408.

Rajagopal, A. *et al.* (2022) ‘A novel hybrid machine learning framework for the prediction of diabetes with context-customized regularization and prediction procedures’, *Mathematics and Computers in Simulation*, 198, pp. 388–406. doi:10.1016/j.matcom.2022.03.003.

Raja, J.B. and Pandian, S.C. (2020) ‘PSO-FCM based data mining model to predict diabetic disease’, *Computer Methods and Programs in Biomedicine*, 196, p. 105659. doi:10.1016/j.cmpb.2020.105659.

Reader, T.C. (2021) *Data preprocessing in python - handling missing data*, *Medium*. Available at: https://medium.com/@theclickreader/data-preprocessing-in-python-handling-missing-data-b717bcd4a264 (Accessed: 31 December 2023).

Sharma, S. and Sharma, B. (2020) ‘EDAS based selection of machine learning algorithm for diabetes detection’, *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)* [Preprint]. doi:10.1109/smart50582.2020.9337118.

Su, Y. *et al.* (2023) ‘Multi-party diabetes mellitus risk prediction based on secure Federated Learning’, *Biomedical Signal Processing and Control*, 85, p. 104881. doi:10.1016/j.bspc.2023.104881.

Talukder, Md.A. (2023) *Machine learning based diabetes detection model for false negative reduction* [Preprint]. doi:10.31224/3044.

Team, D.E. (2023) *Mixed Methods Research Guide with examples*, *Mixed Methods Research Guide With Examples*. Available at: https://dovetail.com/research/mixed-methods-research/#:~:text=Mixed%20methods%20research%20designs%20such,limitations%20of%20either%20approach%20alone. (Accessed: 30 December 2023).

Wu, H. *et al.* (2018) ‘Type 2 diabetes mellitus prediction model based on data mining’, *Informatics in Medicine Unlocked*, 10, pp. 100–107. doi:10.1016/j.imu.2017.12.006.

Yasar, K. and Black, R. (2023) *What is system testing? – TechTarget definition*, *Software Quality*. Available at: https://www.techtarget.com/searchsoftwarequality/definition/system-testing (Accessed: 31 December 2023).

Zach (2023) *How to use tight\_layout() in Matplotlib*, *Statology*. Available at: https://www.statology.org/matplotilb-tight\_layout/ (Accessed: 31 December 2023).

Zhou, H., Xin, Y. and Li, S. (2023) ‘A diabetes prediction model based on Boruta feature selection and Ensemble Learning’, *BMC Bioinformatics*, 24(1). doi:10.1186/s12859-023-05300-5.

Zhu, C., Idemudia, C.U. and Feng, W. (2019) ‘Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques’, *Informatics in Medicine Unlocked*, 17, p. 100179. doi:10.1016/j.imu.2019.100179.

.