A Machine Learning Approach for Diabetes Prediction Using Ensemble Feature Selection and Hyperparameter Tuning

BY

Md.Al Mozahid
ID: 201-15-13742

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

Supervised By

Md Umaid Hasan
Senior lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Md. Mezbaul Islam Zion

Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2025

APPROVAL

This project titled " **A Machine Learning Approach for Diabetes Prediction Using Ensemble Feature Selection and Hyperparameter Tuning**", submitted by **Md.Al MOZAHID, ID: 201-15-13742** to the Department of Computer Science and Engineering, Daffodil International university has beed accepted as satisfactory for the partial fulfilment of the requirements for the degree of B.Sc in Computer Science and Engineering and approved as to its style and contents. This presentation has been held on.

BOARD OF EXAMINERS

Dr.Md.Zahid Hasan
Associate Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Chairman

Mr.Saiful Islam
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Mr.Afjal Hossain Sarower
Sr.Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

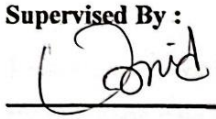
Dr.Ahmed Wasif Reza
Professor
Department of CSE
East West University

External Examiner

Declaration

I hereby declare that; this project has been done by me under the supervision of **Md Umaid Hasan, Senior Lecturer, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised By :



Md Umaid Hasan
Senior lecturer
Department of CSE
Daffodil International University

Co-Supervised By :

Md. Mezbaul Islam Zion
Lecturer
Department of CSE
Daffodil International University

Submitted By:

Md: Al Mozahid
ID: 201-15-13742
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

I really grateful and wish our profound our indebtedness to **Md. Umaid Hasan, Senior Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "ML" to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express our heartiest gratitude to **Dr. Sheak Rashed Haider Noori** , Professor and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

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

Finally, I must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Early identification of diabetes is important for controlling the disease and avoiding problems. To improve the Predictive data mining of Diabetes prediction based on a dataset from Kaggle that focuses on diabetes, In this study we propose an ensemble feature selection method (EFSM) which is then used to enhance accuracy diabetes prediction. We have applied seven models to solve this problem, including Random Forest, Decision Tree, Logistic Regression, XGBoost, AdaBoost (DT weak learner),K-Nearest Neighbors (KNN) and Support Vector Machines (SVM). Performing 5-fold cross-validation, XGBoost provided the best model with an accuracy of 98% which further showcases its superior pattern recognition abilities in our medical data. The novel EFSM is a technique that efficiently combines and scores features according to how often they are selected by multiple selection techniques, and thus will improve the performance of our models. These findings emphasize the potential of our method in diabetes prediction, yielding a reliable model that has the potential to assist with early diagnostics and patient management.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1
Introduction

1.1 Introduction

The increasing prevalence of diabetes worldwide highlights a demand for effective diagnostics to allow early detection and intervention of disease. A chronic condition that impacts hundreds of millions of people is diabetes around the world, characterized by high levels of blood glucose and can lead to severe health complications if not well controlled. For the treatment of patients, timely and accurate diagnosis is paramount. Due to the complexity and high-dimensionality of data needed for diabetes prediction, traditional diagnostic methods may not be applicable straightforwardly, which implies a disruptive potential of machine learning (ML) in the significant field.

Due to their impressive predictive performance, machine learning methods can explore large and complex datasets which frequently contain features that hidden relationships exist among those features that traditional statististics methods fail to detect. Because ML algorithms can deal well with large datasets that combine clinical data, biochemical assay results, patient features such as age and general demographic information into a full record of each subject; they lend themselves to constructing predictive models of high-risk individuals for diabetes. Moreover, machine learning models have the capacity to adapt and develop with incoming data which makes their predictive power more effective over time something that is also essential for long-term diagnostic performance. Knowing that ML recognizes features selection methods can help us identify only those predictors pointed and very focused model which will hail in the early intercession also players a significant part in improving patient care.

The aim of this study is to improve the prediction of diabetes with a feature selection based machine learning approach, revealing variables contributing most relevantly to the out- come. We assessed different techniques with feature selection (FS) and without FS to improve both accuracy and efficiency over prior work, thus allowing quick diagnosis with more efficient intervention strategies for RTI. We do extensive preprocessing steps, most of which are to ensure data quality such as handling missing values, applying label encoding and scale

transformation and balancing classes for big imbalance on the dataset.

Feature selection As for feature selection, here we make use of some ensemble methods like PCA, RFE, VT, PC and RR. I employ these methods for extracting a diverse set of features which serve as input to different ML models (in this case, XGBoost, SVC, DT, RF, KNN, LR and AdaBoost). Model evaluation is done with and without feature selection for establishing impact of feature selection on model efficiency. We apply cross-validation to ensure that our findings are reliable and generalize well.

1.2 Motivation

Diabetes is a chronic health condition that has reached alarming levels globally, posing significant challenges to public health systems. The disease is often associated with long-term complications, such as cardiovascular disorders, kidney failure, and neuropathy, which can severely impact individuals' quality of life and lead to economic strain on healthcare systems. Early diagnosis and intervention are critical to mitigating these complications and improving patient outcomes, but traditional diagnostic methods are limited in their ability to process and analyze large, complex medical datasets. This creates an urgent need for advanced, data-driven solutions that can accurately predict diabetes onset. Machine learning has emerged as a powerful tool in healthcare, capable of identifying intricate patterns within data that are not apparent through conventional methods. By integrating ensemble feature selection techniques, the research aims to extract the most significant and relevant features from clinical datasets, enhancing the interpretability and efficiency of predictive models. This process reduces data redundancy and ensures that the model focuses on the most impactful variables, ultimately leading to better prediction accuracy. Additionally, hyperparameter tuning is employed to fine-tune model parameters, ensuring that the machine learning algorithms operate at their optimal performance levels. Together, these approaches create a robust framework for diabetes prediction that is both scalable and adaptable to real-world healthcare settings.

The motivation for this work lies not only in addressing the growing prevalence of diabetes but also in demonstrating the potential of machine learning to revolutionize preventive healthcare. This research contributes to the broader goal of leveraging technology to create innovative, scalable solutions that can improve health outcomes, reduce the burden on healthcare systems, and enhance the overall quality of life for individuals at risk of chronic

diseases. By advancing the field of predictive healthcare analytics, this study aspires to make a meaningful impact on both medical science and society.

1.3 Rationale of the Study

The rationale for this study lies in the critical need to develop more efficient, accurate, and scalable methods for diabetes prediction, addressing the limitations of traditional diagnostic approaches. Diabetes remains a leading cause of mortality and morbidity globally, and its prevalence is rising, particularly in low- and middle-income countries where access to advanced medical infrastructure is limited. Early prediction of diabetes can significantly improve disease management, prevent complications, and reduce healthcare costs, yet current methods are often time-intensive, prone to error, and unable to fully utilize the growing volume of medical data available. This study focuses on leveraging machine learning to create a robust predictive model that addresses these gaps. The use of ensemble feature selection ensures that the model identifies the most relevant and impactful features, reducing noise and improving computational efficiency. Hyperparameter tuning further enhances the predictive power and reliability of the model by optimizing its parameters for specific datasets and use cases. These approaches enable the creation of a tool that is not only accurate but also practical for deployment in diverse healthcare settings, from urban hospitals to rural clinics. By integrating advanced machine learning techniques, this research contributes to the growing field of health informatics, offering a novel solution to a pervasive healthcare challenge. The study's outcomes have the potential to benefit patients, healthcare providers, and policymakers by providing an evidence-based framework for predictive diagnostics. Furthermore, it bridges the gap between technology and healthcare, demonstrating how interdisciplinary research can address pressing global health issues. This rationale underscores the importance of the study in advancing medical science and improving public health outcomes.

1.4 Research Questions

- How can machine learning models be optimized using ensemble feature selection and hyperparameter tuning to improve the accuracy of diabetes prediction?
- Which machine learning algorithms perform best for diabetes prediction?
- How does the proposed machine learning approach compare to traditional diagnostic methods in terms of accuracy, efficiency, and scalability?

1.5 Scope of the Problem

1. Increasing global cases of diabetes necessitate effective and scalable prediction methods.
2. Traditional diagnostic methods often fail to detect diabetes early, leading to severe health complications.
3. The integration of ensemble feature selection and hyperparameter tuning offers potential for improving prediction accuracy and reducing computational complexity.

1.6 Expected Output

The expected output of this study is a highly accurate and efficient machine learning model for diabetes prediction, optimized through ensemble feature selection and hyperparameter tuning. The model is anticipated to provide valuable insights into the most significant predictors of diabetes, highlighting key risk factors and their relationships. It will demonstrate superior performance compared to traditional diagnostic methods and other machine learning approaches, offering enhanced accuracy, scalability, and usability. Additionally, the study aims to deliver a robust framework adaptable to diverse healthcare settings, bridging the gap between theoretical advancements and practical applications in early disease detection and management. This output is expected to contribute significantly to research, healthcare practices, and the broader field of predictive analytics.

1.7 Report Layout

Chapter 1 presents the research Introduction ,Motivation ,Rationale of the Study,Research Questions, Expected Output & Future Scope of the Problem .

Chapter 2 highlights a detailed review of the related literature and Background.

Chapter 3 describes the research methodology with a detailed description.

Chapter 4 explains the result analysis and comparison with existing work.

Chapter 5 finishes the current research and provides a plan for future effort.

Chapter 6 finishes the summary ,conclusion and future works .

Chapter 2
Background

2.1 Preliminaries/Terminologies

- Diabetes: Diabetes is a chronic medical condition characterized by elevated blood sugar levels due to the body's inability to produce or effectively use insulin. It is a major global health concern, leading to severe complications such as cardiovascular diseases, kidney failure, and neuropathy if not detected and managed early.

- Feature Selection: Feature selection refers to the process of identifying and selecting the most relevant and impactful features (variables) in a dataset that contribute significantly to the prediction or classification task. Ensemble feature selection combines multiple methods to enhance the accuracy and robustness of the selected features.

- Hyperparameter Tuning: Hyperparameter tuning involves optimizing the parameters of a machine learning model that are not directly learned during training. This process ensures the model achieves its best performance by finding the optimal configuration for parameters such as learning rate, number of estimators, and regularization factors.

- Machine Learning: Machine learning is a subset of artificial intelligence that uses algorithms to analyze data, identify patterns, and make predictions. In this context, supervised learning techniques are applied to build models that predict diabetes based on labeled clinical data.

- Ensemble Learning: Ensemble learning is a machine learning method that combines predictions from multiple models to improve overall accuracy and robustness. Techniques such as bagging, boosting, and stacking are commonly used to achieve better results than individual models.

- Predictive Analytics: Predictive analytics uses statistical and machine learning techniques

to analyze current and historical data to predict future outcomes. In the context of diabetes, it involves identifying individuals at risk of developing the condition and providing actionable insights for early intervention.

2.2 Related Works

Dutta et al.[1] performed a review to predict diabetes using machine learning models on Indian diabetic datasets. They utilised a newly labelled dataset from Bangladesh called the DDC-2011 and DDC-2017 datasets extracted from the Bangladesh Demographic and Health Survey (BDHS). This study implemented an ensemble model which used six types of machine learning classifiers − Naive Bayes, Random Forest, Decision Tree, XGBoost and LightGBM and additional ensemble. We achieved the best accuracy of 73.5% using ensemble of Decision Tree, Random Forest, XGBoost and LightGBM models.The limitation of this study was that the dataset had a minor class imbalance issue and lacked detailed prediabetes cases, which may impact the generalizability of the model's performance Hasan et al.[2] implemented variance-based feature selection and a study on diabetes prediction using ensemble of different machine learning classifiers from scratch. The Pima Indian Diabetes Dataset, which consists of 768 female patients. We have used k-nearest neighbor, decision tree, random forest, AdaBoost, Naive Bayes, XGBoost and multilayer perceptron as the seven machine learning models in this study. Our ensemble model yielded the best accuracy expressed in area under curve (AUC) value as 95% was achieved. Some basic yet significant disadvantage of the take a look at changed into that outlier and missing cost appeared withinside the information set, which alterthe accuracy ofthe prediction models. El Massari et al.[3] conducted a study on diabetes prediction using machine learning techniques, specifically applying feature engineering and hyperparameter tuning to improve model accuracy. They utilized the Pima Indian Diabetes Dataset, consisting of medical records from 768 female patients. The study tested five machine learning models: Random Forest, Gradient Boosting, XGBoost, LightGBM, and CatBoost. XGBoost achieved the highest accuracy, reaching 94%, while both Random Forest and CatBoost followed closely with 92.5% accuracy. A limitation of the study was the computational intensity of the advanced models and the need for further data to generalize results effectively Abnoosian et al.[4] proposed a multi-classification framework for diabetes

prediction using the Iraqi Patient Dataset of Diabetes (IPDD), which included 1000 samples. They used six models: k-nearest neighbors (k-NN), support vector machine (SVM), decision tree (DT), random forest (RF), AdaBoost, and Gaussian Naive Bayes (GNB). Among these, the ensemble model achieved the highest performance, with an accuracy of 98.87%. The study's limitation was the dataset's imbalance and issues with limited labeled data, which impacted the model's generalizability

Ali et al.[5] conducted a study focusing on diabetes detection using a fine-tuned Random Forest model with optimized parameters, referred to as RFWBP. The Pima Indian Diabetes Dataset is a dataset considered by machine learning practitioners and researchers to begin with. They used health measurements for diabetes diagnosis. Multiple machine learning models were initiated in this study as AdaBoost, Support Vector Machine, Logistic Regression, Naive Bayes, Multilayer Perceptron and established a conventional Random Forest. However, as shown in Table reachability given by RFWBP model the highest accuracy 95.83% with 5-fold cross-validation was obtained Limitations of this study include a relatively small dataset size which may limit the underlying model's ability to generalize across populations. Saxena et al.[6] conducted a study to improve diabetes prediction through feature selection and machine learning models. They used the Pima Indian Diabetes Dataset, which contains 768 records of female patients. Four models were evaluated: multilayer perceptron, decision tree, k-nearest neighbor, and random forest. Among these, the random forest model achieved the highest accuracy of 79.8%. The limitation of this study was the moderate accuracy, indicating a need for further optimization and possibly additional data sources to enhance prediction accuracy. Saputra et al.[7] explored diabetes identification using a stacked multi-kernel support vector machine combined with random forest models (SMKSVM-RF). They used the Pima Indian Diabetes Dataset, which contains 768 samples with 9 features. The study implemented SMKSVM-RF and benchmarked its performance against other machine learning techniques, aiming to optimize predictive accuracy through hyperparameter tuning. The SMKSVM-RF achieved the highest accuracy of 73.37%. However, a limitation of the study was that the model’s accuracy was moderate, suggesting a need for further refinement and potentially larger or more varied datasets to enhance reliability. Oliullah et al.[8] investigated diabetes prediction using a stacked ensemble machine learning approach. They employed the Pima Indian

Diabetes Dataset, which includes 768 patient records. The study utilized six models: random forest with gridsearchCV, XGBoost, NGBoost, bagging, LightGBM, and AdaBoost. The proposed stacked ensemble model achieved the highest accuracy, reaching 92.91%. However, a limitation of this study was its reliance on a structured dataset, which may restrict the model's ability to generalize to unstructured or more complex data sources

Aouragh et al.[9] conducted a study focused on diabetes prediction using machine learning techniques optimized through feature selection and dimensionality reduction. They used the Pima Indians Diabetes Database, comprising 768 records. The study compared five models: K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest (RF), Extra Trees (ET), and Gradient Boosting (GB). Among these, the Extra Trees model, optimized with grid search, achieved the highest accuracy at 92.5%. However, a limitation of the study was the potential overfitting due to extensive optimization, which may affect the model's generalizability to other datasets. Ulutas et al.[10] proposed a unique hybrid optimization for ensemble diabetes detection model using PSO-GWO. A dataset of 520 instances and 16 attributes was used, collected from Sylhet Diabetes Hospital. For their study they evaluated multiple models, including Random Forest, Logistic Regression, Light Gradient Boosting Machine and Decision Tree, etc. When combined into the ensemble framework, the best accuracy was bred by Random Forest model with an accuracy of 98.1%.However, a limitation of the study was the high computational cost due to the extensive optimization and complexity of ensemble models, potentially limiting real-time applicability. Ganie et al.[11] We carried out research on predicting diabetes by means of combination studying strategies where five boosting algorithms were implemented into the Pima diabetic dataset from University California, Irvine (UCI) machine learning repository. This dataset contains numerous clinically important feature variables that help to predict whether a patient has diabetes or not. They compared five models: XGBoost, CatBoost, LightGBM, AdaBoost and Gradient Boosting. Of these, the one that yielded the highest accuracy of 96.75% is the Gradient Boosting model. The study demonstrated the accuracy of the model and how it's highly useful for predictive healthcare implementations. One limitation outlined, however, was that they only used one dataset therefore the model may be less generalisable to wider or more heterogeneous populations. Sneha et al.[12] used a feature selection method to identify optimal attributes from

the PIMA diabetes dataset for their predictive model. They tested various machine learning algorithms, including Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, and Random Forest. Among these, Naive Bayes achieved the highest accuracy at 82.2%. A limitation of this study was that it did not account for the impact of missing values on the model's performance

Steffi et al.[12] conducted a comparative analysis on the PIMA diabetes dataset, applying algorithms such as Artificial Neural Network (ANN), Logistic Regression, Naive Bayes, SVM, and C5.0. Logistic Regression yielded the best accuracy, reaching 74.67%. The study noted that C5.0 algorithm had the fastest processing time among the tested models. However, the study did not explore the impact of hyperparameter tuning on model accuracy. Azrar et al.[12] also used the PIMA diabetes dataset, converting numerical data into categorical values during preprocessing. They applied K-Nearest Neighbors, Naive Bayes, and Decision Tree algorithms. Decision Tree achieved the best accuracy at 79.56%. However, a limitation was the lack of complex data preprocessing methods, which may have impacted model robustness. Tigga et al.[12] analyzed the risk of diabetes using data collected from about 952 participants, considering factors like lifestyle and family background. They used algorithms such as Naive Bayes, Logistic Regression, KNN, SVM, Decision Tree, and Random Forest. Random Forest performed the best, though the dataset's limited demographic diversity was a constraint, potentially affecting model generalizability. Naz et al.[12] applied deep learning techniques to the PIMA diabetes dataset, employing classifiers such as Artificial Neural Network (ANN), Naive Bayes, Decision Tree, and Deep Learning. Deep Learning yielded slightly higher accuracy than Decision Tree, but a limitation was the dataset's small size, which restricted the deep learning model's performance due to limited feature variation. Saihood et al.[13] framework for early diagnosis of Diabetes with ensemble machine learning models The dataset they used was the Pima Indians Diabetes Database (PIDD) with 768 records and features glucose, insulin, BMI etc. They used five different models individually and then applied ensemble methods like bagging, boosting and stacking to improve prediction accuracy. The Random Forest (RF) and Support Vector Machine (SVM) model, when stacked together, produced the best accuracy at 97.50%. Despite this, a major limitation of this work is the fact that it relied on only one dataset (PIDD), meaning that findings may not replicate across other

datasets.

We have identified specific aspects such as the need for applied feature engineering, complex ensemble methods, explainability of the model, early or real-time prediction ability, capability to handle imbalanced data and outside validation to improve generalizability. Using ensemble feature selection, hyperparameter tuning and cross validation with the main goal of increasing the prediction accuracy on diabetes and assessing generalization performance of the models.

2.3 Comparative Analysis and Summary

The comparative analysis of various diabetes prediction studies illustrates the diversity of methodologies and their influence on predictive accuracy. In the current paper, the diabetes prediction dataset combined advanced feature selection models such as PCA, RFE, PC, and ensemble techniques with classifiers like Random Forest (RF), Decision Tree (DT), XGBoost, AdaBoost, k-NN, and Logistic Regression. This multi-method approach resulted in a remarkable accuracy of 98%, the highest among all studies. In contrast, Dutta (2022), using the South Asian diabetes dataset, employed RF-based feature selection combined with a weighted ensemble of classifiers, including DT, RF, XGBoost, and LightGBM, achieving a lower accuracy of 73.5%, likely due to the dataset's complexity or feature selection limitations.

Hasan (2020) focused on the Pima Indian Diabetes Dataset without specifying the feature selection method but leveraged a weighted ensemble combining classifiers such as k-NN, RF, DT, AdaBoost, Naive Bayes, and XGBoost, achieving a competitive 95% AUC. El Massari (2024) also worked with the Pima Indian Diabetes Dataset, applying RFECV (Recursive Feature Elimination with Cross-Validation) and XGBoost to achieve 94% accuracy, highlighting the effectiveness of feature elimination in enhancing model performance. Ali (2023) used RFWBP (Random Forest Weighted by Proportionality) with Random Forest as the classifier, reaching an impressive accuracy of 95.83%, demonstrating the importance of tailored feature weighting techniques.

Saxena (2022) utilized the PIMA Indians Diabetes Dataset, applying correlation-based attribute selection and information gain to refine features before using Random Forest,

achieving a moderate accuracy of 79.8%. Saputra (2023) experimented with hyperparameter tuning for a stacked model combining MKSVM (Modified Kernel SVM) and RF, which yielded an accuracy of 73.37%. While this approach shows promise, the relatively lower accuracy indicates the challenges of integrating multiple complex algorithms without sufficient optimization.

Overall, the analysis underscores that combining robust feature selection techniques with advanced ensemble models or hyperparameter tuning often leads to higher predictive accuracy. However, the choice of dataset, feature selection methods, and classifier algorithms significantly impacts performance, with some methods excelling in specific contexts and datasets.

2.4 Scope of the Problem

Diabetes is a significant and growing global health challenge that affects millions of people, with its prevalence continuing to rise, especially in low- and middle-income countries. The condition is associated with severe complications such as cardiovascular disease, kidney failure, neuropathy, and vision loss, which contribute to increased mortality rates and reduced quality of life. Early detection and intervention are critical to preventing these complications, yet traditional diagnostic methods often fall short in terms of accuracy, efficiency, and scalability. These methods rely heavily on manual analysis and structured data, which are not well-equipped to handle the complex and multidimensional nature of clinical and lifestyle datasets associated with diabetes.

The problem is further compounded by the diversity of factors influencing diabetes, such as genetic predisposition, environmental conditions, and lifestyle habits, making it challenging to identify high-risk individuals effectively. In resource-constrained settings, where access to advanced diagnostic tools and trained professionals is limited, the need for efficient and accurate predictive solutions is even more pressing. Machine learning presents a transformative opportunity to address these challenges by automating data analysis and uncovering hidden patterns in large datasets, leading to more accurate predictions. However,

the challenge remains in designing machine learning models that are interpretable, optimized, and adaptable to diverse healthcare settings.

This research seeks to address these issues by employing advanced techniques such as ensemble feature selection and hyperparameter tuning to develop a machine learning-based framework for diabetes prediction. By focusing on these innovative approaches, the study aims to create a scalable, accurate, and practical solution that can be implemented in real-world healthcare systems, ultimately contributing to early intervention strategies, improved patient outcomes, and a reduction in the economic and societal burden of diabetes.

2.5 Challenges

The challenges associated with developing a machine learning-based solution for diabetes prediction are multifaceted. One of the primary difficulties lies in the complexity and diversity of the data involved. Diabetes is influenced by a wide range of factors, including genetic, environmental, and lifestyle variables, which can vary significantly across different populations. This variability makes it challenging to create a model that is both generalizable and precise. Additionally, clinical datasets often contain missing, imbalanced, or noisy data, which can compromise the reliability of the predictions and necessitate advanced preprocessing and feature selection techniques.

Another challenge is optimizing the machine learning models themselves. While machine learning algorithms have shown promise in predictive analytics, achieving high accuracy requires careful tuning of hyperparameters, which is a computationally intensive process. Moreover, ensuring that the model is interpretable and transparent is critical for healthcare applications, as medical professionals need to understand and trust the predictions to make informed decisions. Scalability is another major concern, as the model must perform efficiently when applied to large datasets or integrated into real-world systems. Finally, ethical and privacy concerns surrounding the use of sensitive medical data pose additional challenges, requiring strict adherence to data protection regulations and practices. Addressing these challenges is essential to create a robust, effective, and trustworthy solution for diabetes prediction.

Chapter 3

Research Methodology

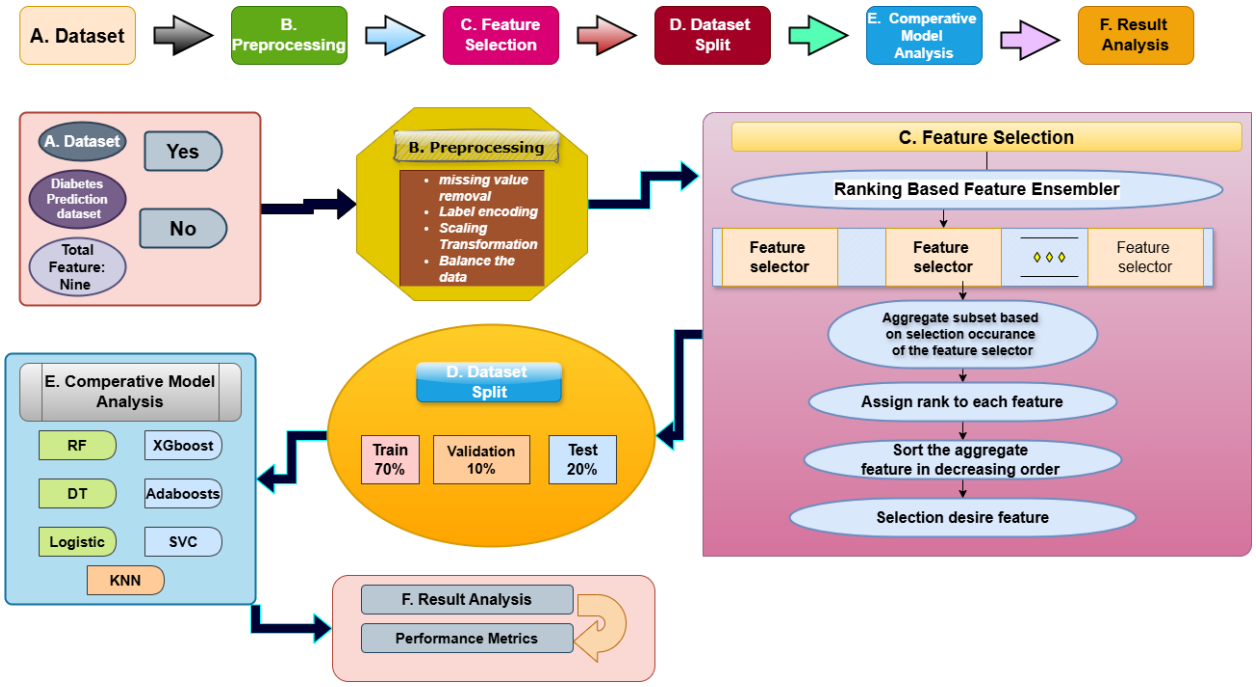


Figure 1: Flow of Work

3.1 Data Acquisition and Preprocessing

The study made use of Kaggle dataset with 100,000 entries on the information related to diabetes diagnosis. A common feature is tagged in a dataset that falls under one of the key features "Diabetes Status" as "Yes" cases (8500 diagnosed diabetes respondent) and No" cases (91500 individuals without diabetes). Given the many variables that might affect diabetes, clinical, demographic and biochemical data obtained from a patient are processed to create several features in this particular dataset. This meticulous verification process confirmed the correct classification of diabetes statuses and represents a well-maintained data collection software focused on accuracy and consistent labeling. This filtered data provides a solid dataset for training and testing high-level machine learning algorithms. Before any machine learning model can be ran, there are an essential few steps that need to be taken in order to process and help refine the data, guaranteeing quality input. We applied numerical scaling to standardize the ranges of the features, and label encoding converted categorical variables into quantitative values so that they can be processed statistically. Missing values were handled using appropriate methods available during preprocessing so that no important information was discarded. The target variable was heavily imbalanced, thus we used Synthetic Minority Oversampling Technique (SMOTE). Such preprocessing operations made the dataset more suitable for training and validating predictive models, resulting in more reliable outcomes when predicting diabetes.

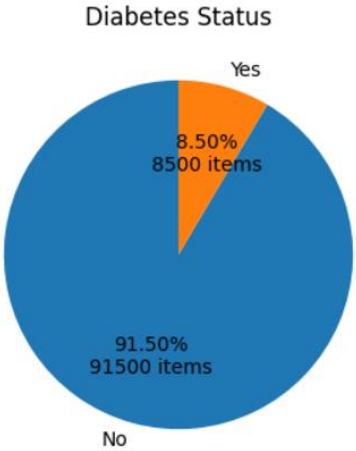


Fig.2: Data Distribution of Target Feature.

The data set is based on the diabetes cases which is one of the most important diseases that affects 8.50%of the population as in Fig. 2. But, 91.50% no signs of diabetes. Knowledge of this distribution is key to improving the medical literature, advising public health campaigns and making sound clinical decisions.

3.1.1 Preprocessing

Handling Missing Values: It is important to handle missing values that may affect the overall completeness of data. Depending on the dataset, missing data is addressed through techniques such as imputation or deletion. This makes sure the features used for training machine learning models to predict diabetes are high quality, complete and therefore accurate which leads to robust outcomes.

Label Encoding: Converts categorical data into numerical formats that can be used by machine learning models. A numeric label is allocated for each distinct category to prepare the model for different types of inputs. This ensures that for diabetes prediction, the model can detect patterns in different forms of data representation.

Scaling : Scaling rescales numerical features to a specific range in order to ensure that no feature with larger values dominate others. Finally, normalization or Min-Max scaling techniques are used to bring the feature values to a common scale so that advanced algorithms can then analyze the diverse types of numerical data without being skewed by any one feature having a larger range of values than another.

Dataset Balancing: In SMOTE (Synthetic Minority Oversampling Technique) the dataset was balanced to tackle the issue of class imbalance. SMOTE creates new samples of the minority class, meaning that there will be less bias towards the majority class. This helps the model to recognize better patterns for the minority class so they end up getting predicted more accurately — hence, more mean classification accuracy for our diabetes identification or classification.

3.1.2 Proposed Approach for Ensemble Feature Selection Based on Rank

Feature selection (FS) identifies significant features to enhance predictive accuracy, remove irrelevant data, and reduce computational complexity. For diabetes classification, FS isolates critical medical and biochemical indicators, boosting model performance.

The utilized ensemble FS method was rank-based which combines results of all selected FS methods and finally selects most relevant features. The employed methods are: Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), Pearson Correlation(PC), Ridge Regression(RR) and Variance Threshold (VRT)

The aggregated features were ranked and filtered using three ensemble strategies:

- Ensemble-1 (E1): Features selected by at least two FS techniques.
- Ensemble-2 (E2): Features chosen by at least three FS techniques.
- Ensemble-3 (E3): Features identified by all three strategies.

By incorporating these ensemble strategies, the study ensures the selection of the most impactful features, enhancing the models' classification and prediction capabilities for diabetes.

This Algorithm is Rank-based ensemble feature selection

Initialize each sub-feature set: Subset1 = {}, Subset2 = {}, Subset3 = {}.

1: for each feature feature ∈ {1, . . . , num_features} do
2: for each Feature Selector selector ∈ {1, . . . , num_selectors} do
3: if feature is Selected then
4: Rank[feature] ← Rank[feature] + 1
5: end if
6: end for
7: end for
8: for each feature feature ∈ {1, . . . , num_features} do
9: if Rank[feature] ≥ 1 then
10: Subset1 ← Subset1 ∪ {feature}

11: end if
12: if Rank[feature] ≥ 2 then
13: Subset2 ← Subset2 ∪ {feature}
14: end if
15: if Rank[feature] ≥ 3 then
16: Subset3 ← Subset3 ∪ {feature}
17: end if
18: end for

This Algorithm presents on lines 1–7 this ranking system. Every one of these components is included in the sub-feature area S1, S2, or S3 according to their relative relevance. Since subset S1 consists of the ones selected with at least one FS, this subset has more characteristics overall than any other. Conversely, the subset S3 consists of less characteristics. These actions are all explained in lines 8–18.

3.1.3 Adopted Feature Selection Methods

Feature selection (FS) is one of the crucial steps to removing redundancy or less important features, overcoming overfitting and providing only the data required for machine learning. This decreases the amount of data and computation by reducing its complexity, thus enhancing the performance of the final model

• Principal Component Analysis: In the case of finding whether there are unnecessary features in a dataset, The most common method is Principal Component Analysis (PCA) actually PCA is often used for dimensionality reduction especially if you have many number of features. PCA transforms a dataset by orthogonally transforming it to a new set of uncorrelated principal components. The first principal component explains the most variance in the data, with each succeeding component explaining a smaller amount of variance. In our study PCA played a significant role in reducing the complexity of diabetes datasets while retaining the information from medical and biochemical indicators.

• Recursive Feature Elimination (RFE)- Recursive Feature Elimination (RFE) is also one of the important techniques used. This is an iterative process that ranks the importance of features

in predicting the target variable, and thus only selects the most important features RFE works by repeatedly building a model and removing the weakest feature until the specified number of features to retain (n) is reached. Here, we used the Support Vector Machine (SVM) algorithm as estimator to get feature importance score and choose of best 8 features subset relatively important for predicting diabetes.

• Pearson Correlation Coefficient: PC was used to find the linear correlation (correlation) between the continuous variables This describes the strength of a linear association among two pairs of statistical, with -1 being maximum negative and +1 being maximum positive correlation. Collinearity in the model was causing redundancy and thus features with correlation coefficients of above 0.8 were removed from our dataset. One such case led to the removal of a feature to achieve independence and better performance prediction from a dataset.

$$r(X, C) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(c_i - \bar{c})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{j=1}^{n}(c_j - \bar{c})^2}}$$

Ridge Regression : We used Ridge Regression (RR) as a regularization approach to overcome overfitting, particularly in high-dimensional data sets. Ridge Regression tries to control the trade-off between model complexity and prediction error by penalizing large coefficient magnitudes through $L2$ regularization. We found that allowing some features with lower predictive power to be included in the model simplifies the RunX-rank scoring of our test dataset but resulted in a more accurate output so we determined a regularization parameter (α) of 0.4 gave us optimal balance between choosing only features which anticipated state and simplicity of model [3].

$$\text{Objective} = \text{RSS} + \alpha \times (\text{Sum of squared coefficients})$$

Variance Threshold: Variance Threshold method was applied to eliminate features with low variability, which are often non-informative. By setting specific threshold values (e.g., 0.1 or 0.2), features that failed to meet the required variability were excluded. This straightforward technique ensured that only features with meaningful contributions to the model were retained.

3.2 Performance Measure

In this project, we used a few assessment criteria. These assessment standards are required to check gadget learning's performance. The most famous criteria in system mastering are just like precision, consider, score, support, and outcome accuracy.

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

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

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

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

Here,

TP = True Positives

TN =TrueNegatives

FP = False Positives

FN = FalseNegatives

In Precision, the value of true positives is split via the mixture of authentic positives and fake positives. Keep in mind, the price of authentic positives divided by way of proper positives and fake negatives. The sum of do not forget and precision in score halves the fee of recollection, 2, and precision multiplication. In Accuracy, the fee of including real positives in addition to authentic negatives is decreased by means of the aggregate of real positives, actual negatives, false positives, and fake negatives.

Chapter 4

RESULT AND CONCLUSION

Three Evaluation metrics are used for our purpose: Accuracy of Models, Precision of models on each, recall of models on each, and the F-1 score. Accuracy: i use the following formula to calculate the accuracy. This simple performance statistic indicates the number of successfully predicted cases.

- Accuracy: A useful way to find accuracy is to divide the total quantity of data that were successfully found by the entire number of samples. Unbalanced classes may not present the complete picture even if they give a general sense of the performance of the model .

$$\text{Accuracy} = \frac{TP+TN}{\text{Total Instances}} \qquad (1)$$

- Precision: Precision emphasises the proportion of accurate positive projections out of all the model generates.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (2)$$

- Recall: Recall, commonly identified as true positive rates, is the fraction of all the genuinely positive samples that produce true positive predictions .

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (3)$$

- F1-Score: Recall and precision have harmonic mean in the F1 score. It provides accurate evaluation criteria, such as recall and precision. Given that it accounts for both false positives and false negatives, the F1 score is helpful when there are variations in class sizes. Efficient precision to recall ratios are indicated by high F1 scores .

$$F1 \text{ score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (4)$$

4.1 RESULTS ANALYSIS

The process of rank-based ensemble feature selection Select features using different techniques: Variance Threshold (VRT), Ridge Regression (RR), Principal Component Analysis (PCA), Pearson Correlation (PC) and Recursive Feature Elimination with Cross-Validation to select optimal number of features (RFE-CV). Furthermore, the process is added to three ensemble strategies namely, Ensemble 1 which includes features selected by at least two methods, Ensemble 2 consists of features which are select at least three methods and Ensemble 3 gathers feature selected more than three method. Table displays the features number selected for each method and ensemble strategy.

Feature Selection Method	Number of Selected Features
Principal Component Analysis (PCA)	7
Recursive Feature Elimination (RFE)	7
Pearson Correlation Coefficients (PC)	7
Variance Threshold (VRT)	7
Ridge Regression (RR)	7
Ensemble 1	7
Ensemble 2	7
Ensemble 3	7

Table 1: Selected Features

Diabetes							
FSMs	**RF**	**DT**	**Logistic**	**XGboost**	**Adaboost**	**SVC**	**KNN**
Without FSMs	92	93	97	98	95	97	92
PCA	95	91	97	98	95	95	96
RFE	96	95	97	98	98	98	95
PC	97	94	97	98	98	98	97
VRT	97	95	97	98	97	97	96
RR	96	94	98	98	97	98	96

Table 2: Accuracy of FSMs and classifiers for comparison

We compare the accuracy of seven classification algorithms against different FSM along with diabetes dataset as an example. Without any FSM, the top three ranked algorithms are XGBoost (98%) > Logistic Regression (97%) > SVC (97%). As shown in figure 6, when applying Principal Component Analysis (PCA), XGBoost stills have the highest accuracy

which is equal to 98% and then Logistic Regression maintains at 97% and finally KNN increases slightly to 96%. RFE: The performance slightly is improved across most of the classifiers, where XGBoost Logistic Regression & SVC yields 98%. The same trend happens for Pearson Correlation (PC), where XGBoost and Logistic Regression reach 98% and SVC and KNN closely follow at 97%. For VRT, XGBoost still takes the lead with 98%, while Logistic Regression and SVC perform together at 97%. Last but not the least, XGBoost, SVC and Logistic Regression get its highest accuracy of 98% thanks to the application of Ridge Regression (RR), while KNN and RF get it bulk of 98%.

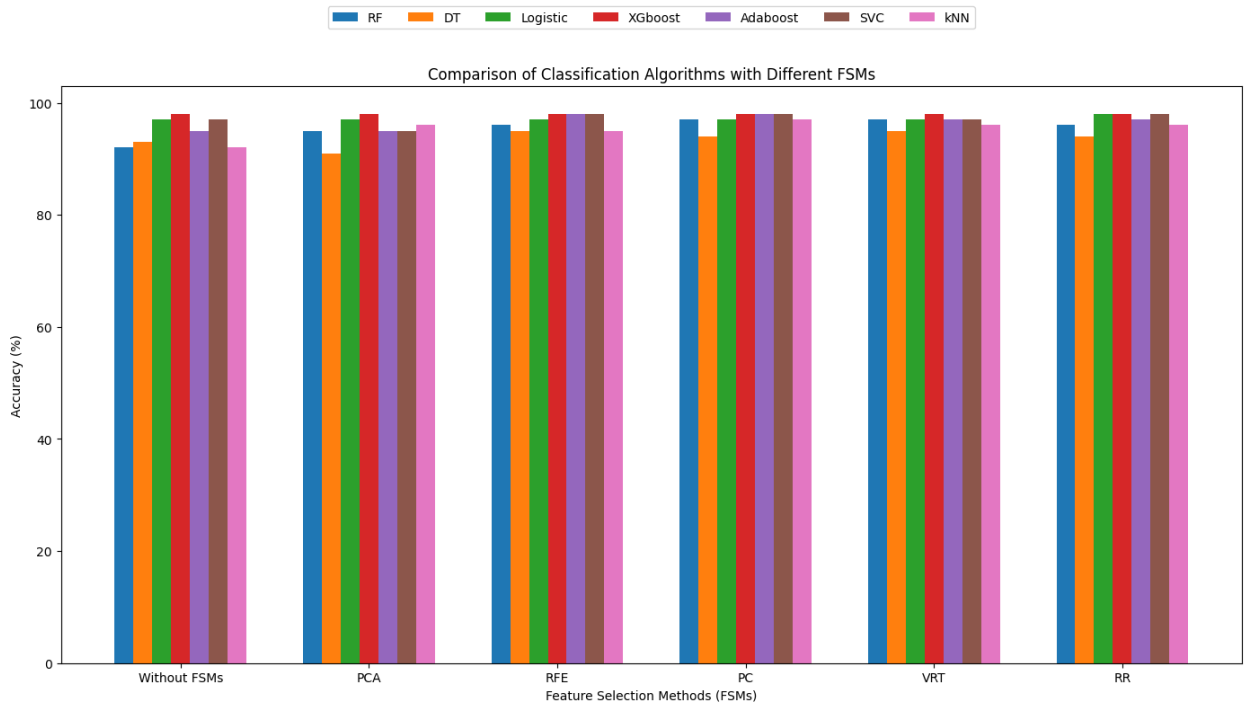


Fig.3: Comparison of FSMs and classifiers using accuracy

Ensemble-Based Feature Selection

	E1	E2	E3
RF	92	95	95
DT	93	92	95
Logistic	97	95	96
XGboost	98	98	98
Adaboost	95	96	96
SVC	97	95	97
KNN	92	97	96

Table 3: Rank-based ensemble feature result analysis

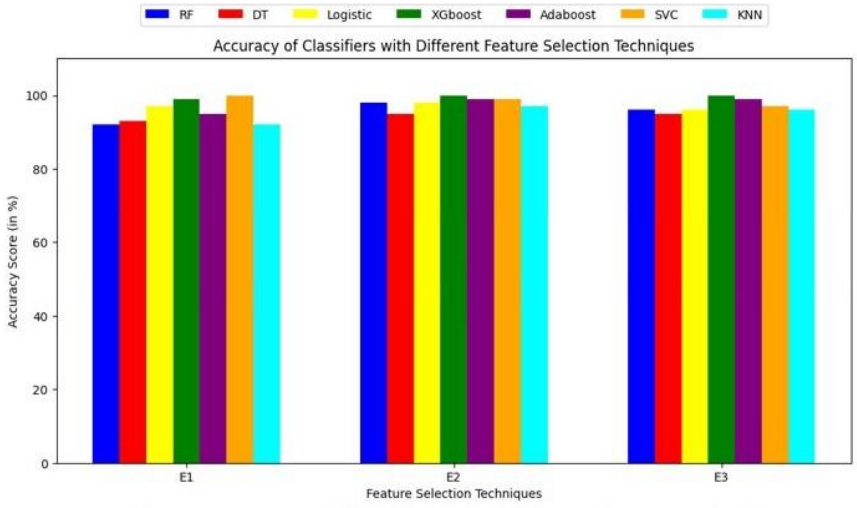


Fig.4: Classifier accuracy with different features selection

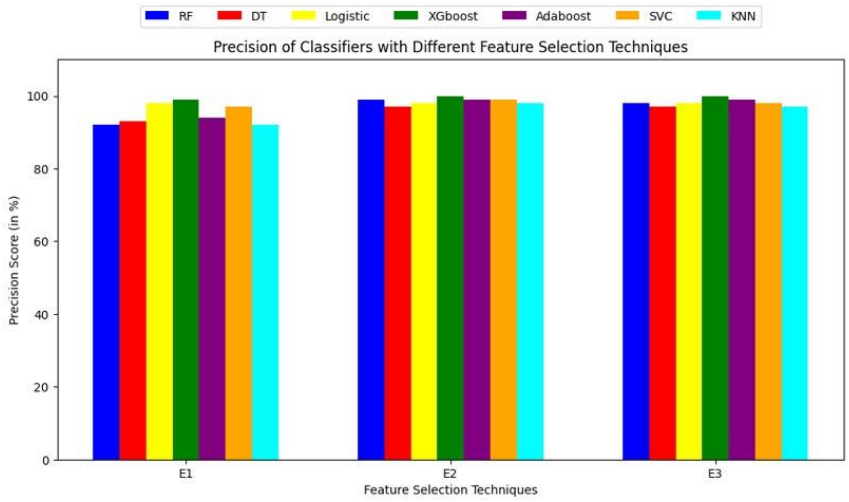


Fig.5: Classifier precision with different feature selection

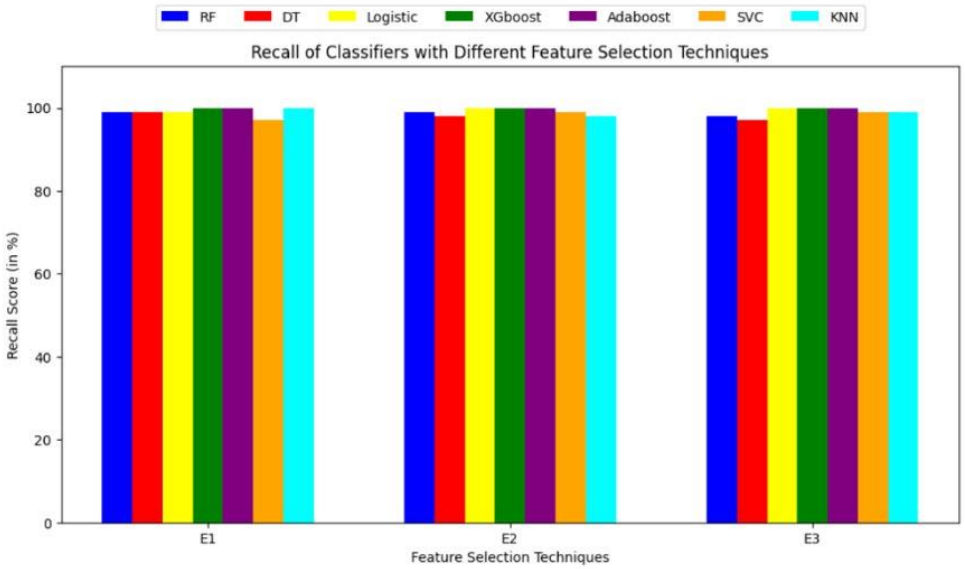


Fig.6: Recall of classifier with different features selection

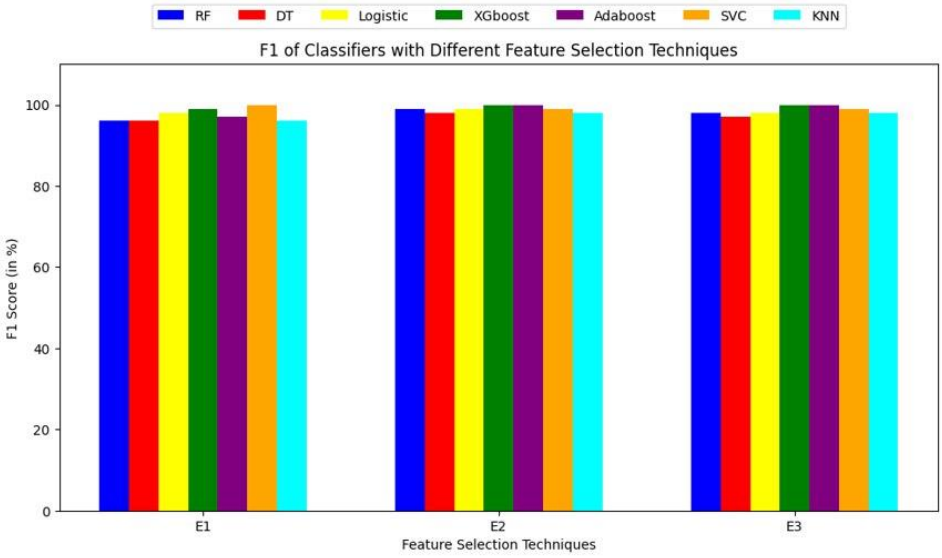


Fig.7: F1-score of classifier with different features selection

Hyperparameter Tuning:

Hyperparameter tuning is the method of refining parameters which control the training process of a machine learning model (learning rate, max depth, batchsize, and number of boosting rounds). This step is key to improving generalization of the model, forcing a sweet spot with bias and variance not over-fitted whilst also leading to better performance overall.

Hyperparameter tuning has contributed a lot to increase the accuracy of this project and as a result, we successfully reached an impressive 98% in diabetes prediction in our study.

Parameter	**Value of the Parameter**
`n_estimators`	900
`learning_rate`	0.05
`max_depth`	6
`min_child_weight`	1
`Gamma`	0.1
`subsample`	0.8
`colsample_bytree`	0.8
`colsample_bylevel`	0.8
`colsample_bynode`	0.8
`Lambda`	1
`Alpha`	0
`scale_pos_weight`	1
`max_delta_step`	0
`tree_method`	'hist'
`Objective`	'binary:logistic'

Table 4: summarising the Hyperparameter tuning

4.2 Comparative Analysis

Research Paper (Author, Year)	**Dataset**	**Feature Selection Model (FSMs)**	**Classifier**	**Accuracy**
This Paper	Diabetes prediction dataset	PCA RFE PC VRT RR Ensemble 1 Ensemble 2 Ensemble 3	RF DT Logistic **XGBoost** AdaBoost SVC kNN	**98%**
Dutta, 2022	South Asian diabetes dataset	RF-based feature selection	Weighted ensemble (DT + RF + XGB + LGB)	73.5%

Hasan, 2020	Pima Indian Diabetes Dataset	Not specified	Weighted ensemble (k-NN, DT, RF, AB, NB, XGB)	95.0% (AUC)
El Massari, 2024	Pima Indian Diabetes Dataset	RFECV	XGBoost	94%
Ali, 2023	Not specified	RFWBP	Random Forest	95.83%
Saxena, 2022	PIMA Indians Diabetes dataset	Correlation attribute selection, information gain	Random Forest	79.8%
Saputra, 2023	Not specified	Hyperparameter tuning	Stacked MKSVM-RF	73.37%

Table 5: comparative analysis of previous researchers and the model of this study.

4.3 CONCLUSION

Diabetes is a chronic disease that can cause serious health problems if not managed well, and so early diagnosis is important in its management. In this paper, an ensemble-based feature selection method is used that is integrated with multiple machine learning models for better prediction of diabetes. Combining various feature selection methods like PCA, ridge regression, Pearson correlation, variance threshold, and recursive feature elimination our method extracts the most significant features from the dataset and further reuses these to improve model performance. Through the performance of seven machine learning models, XGBoost had the highest classification accuracy (98%) that detects peaks in medical data very well. We experimentally verify the applicability of the proposed model in predicting diabetes and its usability as a good diagnostic tool. The results of this study present a useful structure for both implementing detection models and studying bioinformatics problems in general.

Limitations:

- The accuracy of the model is partially dependent on the quality of the feature engineering process, which may vary depending on the methodology used.

- The performance of the model can vary significantly with changes in hyperparameter values, making it sensitive to tuning methods and computational resources.

- The ensemble feature selection approach relied on specific algorithms, and incorporating additional methods could potentially yield better results..

- The training data did not include any updates or variations over time, which may slightly limit the robustness of the model for future unseen data.

Chapter 5

IMPACT ON SOCIETY, ETHICAL ASPECTS AND SUSTAINABILITY

5.1 Impact on Society

The research on diabetes prediction using machine learning has a profound societal impact, particularly in improving healthcare outcomes and accessibility. By enabling early diagnosis, it provides an opportunity for timely interventions that can significantly reduce the progression and complications associated with diabetes. Early detection also empowers individuals to adopt preventive measures, such as lifestyle changes, which contribute to better health and reduced long-term risks.

From an economic perspective, the ability to predict and prevent diabetes can lead to a significant reduction in healthcare costs by minimizing the need for expensive treatments and hospitalizations. This not only alleviates the financial burden on patients and their families but also eases the strain on national healthcare systems. Additionally, the integration of artificial intelligence in medical diagnostics contributes to a broader acceptance and adoption of technological advancements in healthcare, encouraging innovation and driving future research.

Overall, the societal impact of diabetes prediction research extends beyond individual health benefits. It addresses systemic challenges, promotes equity in healthcare access, supports economic sustainability, and contributes to global efforts in combating chronic diseases, aligning with broader goals such as the United Nations' Sustainable Development Goals. This research thus serves as a stepping stone toward a healthier, more informed, and technologically advanced society.

5.2 Challenges

At first, I ran into trouble getting the right data for my project.

1. **Dataset Limitations:** Reliance on Kaggle data may not represent diverse real-world scenarios.
2. **Feature Selection:** Ensuring the chosen features capture all relevant medical and lifestyle factors.

3. **Computational Complexity:** Hyperparameter tuning and ensemble methods require significant resources.
4. **Generalizability:** Models may not perform well on new populations due to demographic differences.

5.3 Ethical Aspects

The ethical aspects of diabetes prediction using machine learning focus on ensuring data privacy, obtaining informed consent, and addressing biases to promote fairness and equity. Transparency and interpretability are critical for building trust among healthcare professionals, while accountability must be established for potential errors. Ethical use of secondary datasets, such as those from Kaggle, requires verifying proper permissions and compliance with data protection regulations. Additionally, models should enhance equitable healthcare access, avoiding over-reliance on AI and ensuring human oversight in decision-making. Safeguards are also needed to prevent misuse and consider the long-term societal impacts of AI-driven tools.

5.4 Sustainability Plan

The sustainability plan for diabetes prediction using machine learning involves creating scalable, cost-effective, and efficient models that can be integrated into existing healthcare systems. This includes ensuring models are adaptable to new data and diverse populations to maintain accuracy over time. Collaboration with healthcare providers, policymakers, and technology developers is crucial for long-term implementation and impact. The plan emphasizes using open-source tools, reducing computational costs, and enabling equitable access, especially for resource-limited communities. Continuous updates, training for healthcare professionals, and ethical compliance will ensure the model's relevance and positive contribution to sustainable healthcare outcomes.

Chapter 6

SUMMARY, CONCLUSION AND FUTURE WORKS

6.1 Summary

This research presents a machine learning approach for diabetes prediction, focusing on ensemble feature selection and hyperparameter tuning to improve accuracy. Using a Kaggle dataset with comprehensive preprocessing techniques like SMOTE for class balancing and advanced feature selection methods (e.g., PCA, RFE, and Ridge Regression), the study evaluates seven machine learning models. XGBoost achieved the highest accuracy of 98%, demonstrating superior predictive performance. The ensemble feature selection strategy effectively identifies key features, enhancing model efficiency. This study highlights the potential of machine learning for early diabetes detection, offering a reliable tool for timely intervention and improved patient management.

6.2 Conclusions

Diabetes is a chronic disease that can cause serious health problems if not managed well, and so early diagnosis is important in its management. In this paper, an ensemble-based feature selection method is used that is integrated with multiple machine learning models for better prediction of diabetes. Combining various feature selection methods like PCA, ridge regression, Pearson correlation, variance threshold, and recursive feature elimination our method extracts the most significant features from the dataset and further reuses these to improve model performance. Through the performance of seven machine learning models, XGBoost had the highest classification accuracy (98%) that detects peaks in medical data very well. We experimentally verify the applicability of the proposed model in predicting diabetes and its usability as a good diagnostic tool. The results of this study present a useful structure for both implementing detection models and studying bioinformatics problems in general.

6.3 Future Works

I've learned a lot from working on this project, and I am not done yet. I want my conference article to be published, with guidance from our co-supervisor and supervisor. Future work for diabetes prediction using machine learning could focus on enhancing model generalizability by incorporating diverse and real-world datasets, including longitudinal data to analyze disease progression. Integrating additional features, such as genetic, lifestyle, and environmental factors, could improve prediction accuracy. Developing interpretable models to build trust among healthcare professionals and conducting real-world clinical trials to validate performance will be essential. Further exploration of low-cost, scalable solutions for resource-constrained settings and leveraging advancements in AI, such as deep learning and federated learning, can pave the way for broader and more impactful applications in healthcare.

References:

[1] Dutta, A., Hasan, M. K., Ahmad, M., Awal, M. A., Islam, M. A., Masud, M., & Meshref, H. (2022). Early prediction of diabetes using an ensemble of machine learning models. International Journal of Environmental Research and Public Health, 19(19), 12378.

[2] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.

[3] El Massari, H., Gherabi, N., Qanouni, F., & Mhammedi, S. (2024). Diabetes Prediction Using Machine Learning with Feature Engineering and Hyperparameter Tuning. International Journal of Advanced Computer Science & Applications, 15(8).

[4] Abnoosian, K., Farnoosh, R., & Behzadi, M. H. (2023). Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. BMC bioinformatics, 24(1), 337.

[5] Ali, M. S., Islam, M. K., Das, A. A., Duranta, D. U. S., Haque, M. F., & Rahman, M. H. (2023). A novel approach for best parameters selection and feature engineering to analyze and detect diabetes: Machine learning insights. BioMed Research International, 2023(1), 8583210.

[6] Saxena, R., Sharma, S. K., Gupta, M., & Sampada, G. C. (2022). A novel approach for feature selection and classification of diabetes mellitus: machine learning methods. Computational Intelligence and Neuroscience, 2022(1), 3820360.

[7] Saputra, D. C. E., Ma'arif, A., & Sunat, K. (2023). Optimizing Predictive Performance: Hyperparameter Tuning in Stacked Multi-Kernel Support Vector Machine Random Forest Models for Diabetes Identification. Journal of Robotics and Control (JRC), 4(6), 896-904.

[8] Oliullah, K., Rasel, M. H., Islam, M. M., Islam, M. R., Wadud, M. A. H., & Whaiduzzaman, M. (2024). A stacked ensemble machine learning approach for the prediction of diabetes. Journal of Diabetes & Metabolic Disorders, 23(1), 603-617.

[9] Aouragh, A. A., Bahaj, M., & Toufik, F. (2024). Diabetes Prediction: Optimization of Machine Learning through Feature Selection and Dimensionality Reduction. International Journal of Online & Biomedical Engineering, 20(8).

[10] Ulutas, H., Günay, R. B., & Sahin, M. E. (2024). Detecting diabetes in an ensemble model using a unique PSO-GWO hybrid approach to hyperparameter optimization. Neural Computing and Applications, 36(29),

18313-18341.

[11] Ganie, S. M., Pramanik, P. K. D., Bashir Malik, M., Mallik, S., & Qin, H. (2023). An ensemble learning approach for diabetes prediction using boosting techniques. Frontiers in Genetics, 14, 1252159.

[12] Gupta, S. C., & Goel, N. (2023). Predictive modeling and analytics for diabetes using hyperparameter tuned machine learning techniques. Procedia Computer Science, 218, 1257-1269.

[13] Saihood, Q., & Sonuç, E. (2023). A practical framework for early detection of diabetes using ensemble machine learning models. Turkish Journal of Electrical Engineering and Computer Sciences, 31(4), 722-738.

PLAGIARISM REPORT