Diagnosis of Diabetes Mellitus through Predictive Modelling using Machine Learning

By

Ayeza Shahid

(Registration No: 00000400328)

Department of Biomedical Engineering and Sciences Engineering

School of Mechanical and Manufacturing Engineering

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2025)

Diagnosis of Diabetes Mellitus through Predictive Modelling using Machine Learning

By

Ayeza Shahid

(Registration No: 00000400328)

A thesis submitted to the National University of Sciences and Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

Master of Science in
Biomedical Engineering

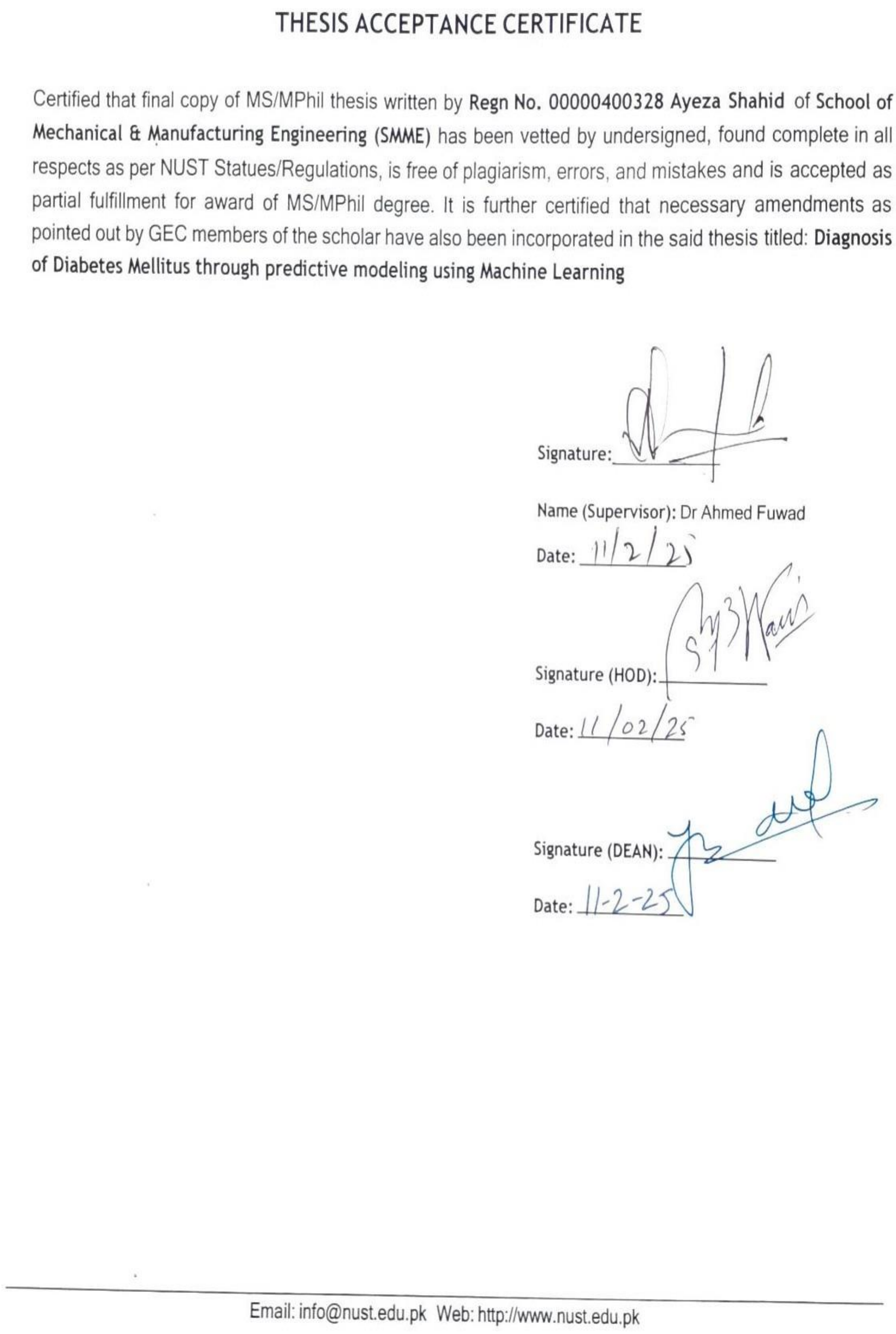
Supervisor: Dr. Ahmed Fuwad

School of Mechanical and Manufacturing Engineering

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2025)



THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by **Regn No. 00000400328 Ayeza Shahid** of **School of Mechanical & Manufacturing Engineering (SMME)** has been vetted by undersigned, found complete in all respects as per NUST Statues/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis titled: **Diagnosis of Diabetes Mellitus through predictive modeling using Machine Learning**

Signature:______________

Name (Supervisor): Dr Ahmed Fuwad

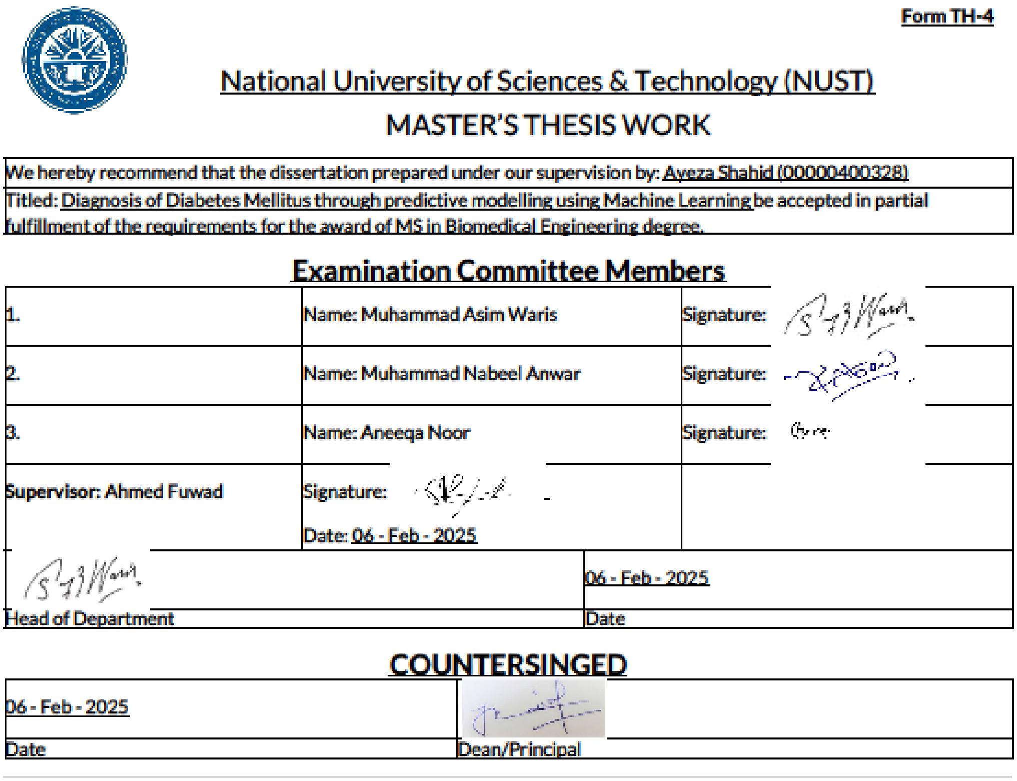
Date: 11/2/25

Signature (HOD):______________

Date: 11/02/25

Signature (DEAN):______________

Date: 11-2-25



Form TH-4

National University of Sciences & Technology (NUST)

MASTER'S THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by: Ayeza Shahid (00000400328)

Titled: Diagnosis of Diabetes Mellitus through predictive modelling using Machine Learning be accepted in partial fulfillment of the requirements for the award of MS in Biomedical Engineering degree.

Examination Committee Members

1.	Name: Muhammad Asim Waris	Signature:
2.	Name: Muhammad Nabeel Anwar	Signature:
3.	Name: Aneeqa Noor	Signature:
Supervisor: Ahmed Fuwad	Signature: Date: 06 - Feb - 2025	

	06 - Feb - 2025
Head of Department	Date

COUNTERSINGED

06 - Feb - 2025	
Date	Dean/Principal

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in this thesis, entitled “Diagnosis of Diabetes Mellitus through Predictive Modelling using Machine Learning” was conducted by Ms. Ayeza Shahid under the supervision of Dr. Ahmed Fuwad.

No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the Department of Biomedical Engineering and Sciences in partial fulfillment of the requirements for the degree of Master of Science in Field of Biomedical Engineering.

Department of Biomedical Engineering and Sciences National University of Sciences and Technology, Islamabad.

Student Name: Ayeza Shahid Signature: AShahid.

Supervisor Name: Dr. Ahmed Fuwad Signature:

Name of Dean/HOD: Dr. Asim Waris Signature: Waris.

AUTHOR’S DECLARATION

I, Ayeza Shahid, hereby state that my MS thesis titled “Diagnosis of Diabetes Mellitus through Predictive Modelling using Machine Learning” is my own work and has not been submitted previously by me for taking any degree from National University of Sciences and Technology, Islamabad or anywhere else in the country/ world.

At any time if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.

Name of Student: Ayeza Shahid

Date: 06-02-2025

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the thesis titled "Diagnosis of Diabetes Mellitus through Predictive Modelling using Machine Learning" is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and National University of Sciences and Technology (NUST), Islamabad towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS degree, the University reserves the rights to withdraw/revoke my MS degree and that HEC and NUST, Islamabad has the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized thesis.

Student Signature: AShahid.

Name: Ayeza Shahid

To my loving parents, whose unwavering support and endless encouragement have been the foundation of my success.
To the extraordinary individuals who pursue their dreams with courage and grace, inspiring me to push beyond limits and achieve greatness.

ACKNOWLEDGEMENTS

First and foremost, I am deeply grateful to Allah (SWT), the Most Merciful and Compassionate, whose countless blessings, divine guidance, and infinite wisdom have given me strength and resilience throughout this academic journey.

I am profoundly thankful to my parents, Shahid Siddique and Faima Shahid, whose unwavering love, selfless sacrifices, endless support and encouragement have been the foundation of my success. Your confidence in me and prayers have been a source of strength that has guided me through every step of this journey.

I would like to express my sincere appreciation to my supervisor, Dr. Ahmed Fuwad, and co supervisor Dr. Muhammad Asim Waris for their invaluable mentorship, insightful guidance, and continuous encouragement throughout my research journey. I am equally grateful to my GEC members, Dr. Muhammad Nabeel Ahmed and Dr. Aneeqa Noor, for their thoughtful feedback and valuable contributions, which were pivotal in shaping my work. I am grateful for the assistance provided by Dr. Muhammad Jawad Khan in my work.

I am deeply thankful to Sanam Rehman for her significant contribution to data collection. Her dedication and efforts have been pivotal in ensuring the success of this research.

I am deeply grateful to my siblings and friends for their unwavering support and love throughout this challenging journey. Their presence and encouragement made even the most difficult moments enjoyable and meaningful.

Finally, I wish to acknowledge all those whose contributions, whether direct or indirect, have enriched my experience and helped me achieve this milestone.

Thank you all for being a part of this journey.

Table of Contents

List of Figures

List of Tables

List of Symbols, Abbreviations and Acronyms

ML	Machine Learning
DL	Deep Learning
PIDD	Pima Indian Diabetes Database
RF	Random Forest
GBoost	Gradient Boosting
XGBoost	Extreme Gradient Boosting
SVM	Support Vector Machine
RF	Random Forest
LR	Logistic Regression
KNN	K Nearest Neighbors
AdaBoost	Adaptive Boosting
GUI	Graphical User Interface
FS	Feature Selection
BMI	Body Mass Index
BP	Blood Pressure

Abstract

Diabetes mellitus is a global health challenge, requiring early detection to prevent severe complications. This study utilizes machine learning for diabetes diagnosis, leveraging a dataset collected from the Pakistani population to ensure demographic relevance. Features included invasive parameters (e.g., fasting blood glucose, blood pressure) and non-invasive factors (e.g., age, gender, BMI, waist circumference). The data was split into training (70%) and testing (30%) sets and evaluated using nine classifiers, including Logistic Regression, Random Forest, XGBoost, and LightGBM.

Ensemble models, particularly XGBoost achieved superior performance, with testing accuracy reaching 93%. This model demonstrated robustness in capturing complex feature interactions without requiring extensive feature selection. Integration into a mobile app and GUI further demonstrated the practical utility of these models, allowing users to input health parameters and receive instant predictions.

This research highlights the importance of combining machine learning with region-specific data for accurate and accessible diabetes prediction. It demonstrates the potential of predictive modeling to complement traditional diagnostics and improve early detection. Future work may focus on publicizing the mobile application and additional data to enhance model performance.

CHAPTER 1

INTRODUCTION

Diabetes mellitus, commonly referred to as diabetes, is a chronic metabolic disorder characterized by elevated blood glucose levels [1]. This condition arises either due to the body's inability to produce sufficient insulin or its failure to effectively utilize the insulin it produces. Over the past few decades, diabetes has emerged as a significant public health challenge globally, with an alarming increase in its prevalence. According to the International Diabetes Federation (IDF), over 537 million adults were living with diabetes in 2021, a number projected to rise to 783 million by 2045 [2]. Among these, developing countries like Pakistan are facing a particularly steep rise in diabetes cases, posing severe challenges to their healthcare systems.

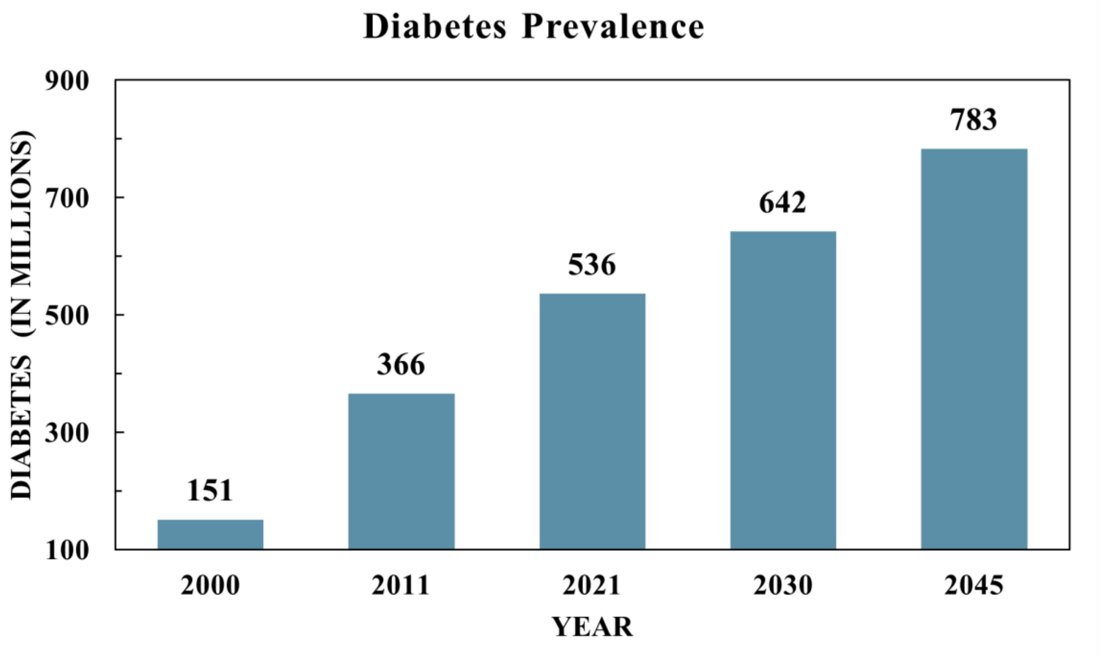


Figure 1: Prevalence of diabetes according to the International Diabetes Federation from 2000 to 2045 [3].

The prevalence of diabetes in Pakistan has been attributed to a combination of genetic predispositions, lifestyle changes, and socioeconomic factors. A recent study revealed that approximately 26.7% of the adult population in Pakistan suffers from diabetes, making it one of the highest prevalence rates in the world [4]. Urbanization, sedentary lifestyles, unhealthy dietary habits, and a lack of awareness about preventive measures have significantly contributed to the escalation of diabetes in the region. Moreover, cultural and societal norms, particularly in rural areas, often hinder the timely diagnosis and management of the disease. These challenges underscore the urgent need for effective and innovative solutions to tackle the growing burden of diabetes in Pakistan [5].

1.1 Diabetes

Diabetes mellitus is a multifaceted disorder that encompasses several types, each with distinct etiologies, characteristics, and management strategies. A deeper understanding of the types of diabetes is crucial for developing effective predictive models and treatment plans.

1.1.1 Type 1 Diabetes

Type 1 diabetes, often referred to as juvenile diabetes or insulin-dependent diabetes, is an autoimmune condition in which the body's immune system attacks the insulin-producing beta cells in the pancreas. This destruction results in little to no insulin production, necessitating lifelong insulin therapy for affected individuals. Type 1 diabetes accounts for approximately 5-10% of all diabetes cases globally [2]. While its exact cause is not fully understood, genetic predispositions and environmental triggers, such as viral infections, are believed to play significant roles.

1.1.2 Type 2 Diabetes

Type 2 diabetes is the most common form of diabetes, representing about 90-95% of all cases [6]. Unlike Type 1, Type 2 diabetes is characterized by insulin resistance, where the body's cells fail to respond effectively to insulin, often coupled with inadequate insulin production. This type is strongly associated with lifestyle factors, including obesity, physical inactivity, and unhealthy diets, as well as genetic predispositions. Type 2 diabetes is particularly prevalent in South Asia, including Pakistan, where urbanization and lifestyle changes have led to an increase in risk factors.

1.1.3 Gestational Diabetes

Gestational diabetes occurs during pregnancy when hormonal changes lead to impaired glucose tolerance [7]. Although it typically resolves after childbirth, gestational diabetes increases the risk of developing Type 2 diabetes later in life for both the mother and the child. In Pakistan, where maternal healthcare is often limited, gestational diabetes poses a significant public health challenge [8].

1.1.4 Other Specific Types of Diabetes

In addition to the three major types, there are other less common forms of diabetes. These include:

1.1.4.1 Monogenic Diabetes:

Caused by single-gene mutations, such as maturity-onset diabetes of the young (MODY).

1.1.4.2 Secondary Diabetes:

Resulting from other medical conditions or treatments, such as pancreatic diseases or prolonged use of glucocorticoids[9].

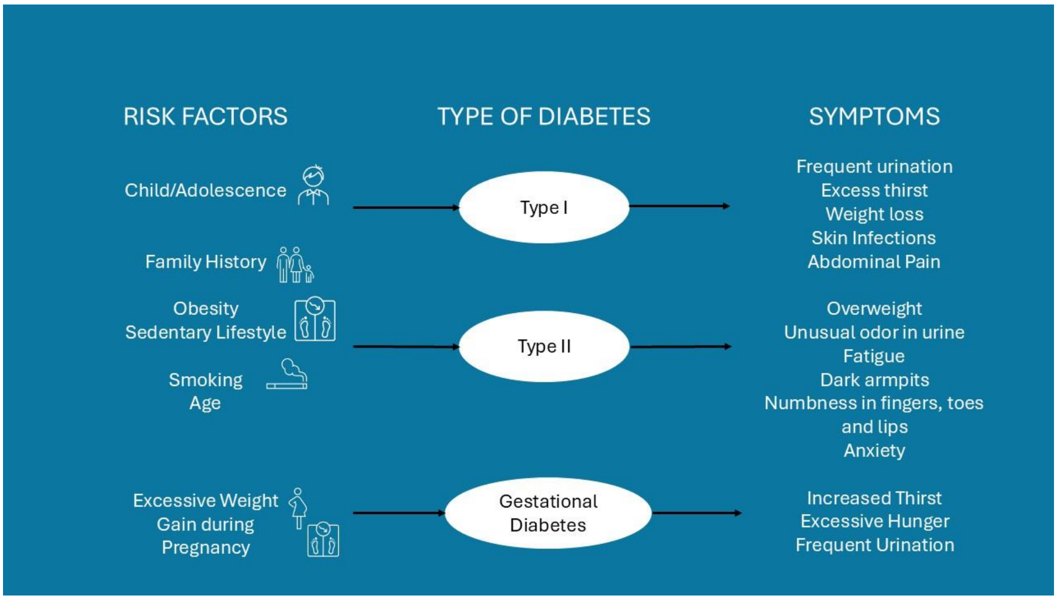


Figure 2: Overview of diabetes, its risk factors, types and symptoms

1.2 Machine Learning

Machine learning (ML) has emerged as a transformative technology with the potential to revolutionize the healthcare industry. By leveraging the vast amounts of data generated in healthcare systems, ML algorithms can identify complex patterns and relationships that may not be apparent through traditional statistical methods. This capability has paved the way for more accurate and timely predictions of diseases, including diabetes. Predictive models powered by ML [10] can not only enhance early detection but also facilitate personalized treatment plans, thereby improving patient outcomes.

ML algorithms, such as logistic regression, decision trees, random forests, and deep learning models, have shown remarkable success in healthcare applications [11]. These models can analyze both structured data (e.g., lab test results) and unstructured data (e.g., clinical notes) to predict disease outcomes with high accuracy. For instance, studies using globally standardized datasets like the Pima Indians Diabetes Dataset [12] have demonstrated that ML models can achieve prediction accuracies exceeding 80%, offering promising avenues for early diagnosis [13][14]. However, these findings often fail to account for population-specific variations, limiting their applicability in localized contexts such as Pakistan.

The uniqueness of this research lies in its focus on data collected from the Pakistani population. Most existing studies on diabetes prediction using ML techniques rely on datasets from Western countries or globally standardized data. While these datasets have been instrumental in advancing diabetes research, they often fail to capture the specific demographic, genetic, and lifestyle characteristics of populations in South Asia, including Pakistan. For instance, South Asians are known to have a higher propensity for developing diabetes at lower body mass indices (BMIs) compared to Western populations [15]. Additionally, socio-economic disparities, dietary preferences, and cultural practices in Pakistan further differentiate its population from those in other regions.

In this context, the present study aims to bridge this gap by utilizing data collected exclusively from the Pakistani population. By incorporating features that are both invasive (e.g., fasting blood glucose levels) and non-invasive (e.g., age, BMI, waist circumference, and physical activity levels), this research endeavors to develop a robust ML-based predictive model tailored to the unique characteristics of the target population. Such a model holds the potential to significantly improve diabetes screening and management in Pakistan, particularly in underserved and resource-constrained areas.

The importance of early detection cannot be overstated, as diabetes, if left unmanaged, can lead to a host of debilitating complications, including cardiovascular disease, kidney failure, blindness, and lower limb amputations. Early diagnosis allows for timely interventions, such as lifestyle modifications and pharmacological treatments, which can delay or prevent the onset of complications. In Pakistan, where healthcare resources are limited, the implementation of an efficient and cost-effective diabetes prediction tool could play a pivotal role in reducing the disease burden and enhancing the quality of life for millions of individuals.

This study also underscores the ethical considerations involved in using ML for healthcare applications. Ensuring the privacy and security of patient data, addressing potential biases in the dataset, and validating the model's performance across diverse subgroups within the Pakistani population are crucial to building trust and acceptance among stakeholders. Moreover, the integration of such a model into the existing healthcare infrastructure requires collaboration between technologists, healthcare providers, and policymakers to maximize its impact. Ethical concerns must also encompass transparency and interpretability of the ML models [16], ensuring that healthcare practitioners can understand and trust the system's predictions.

Furthermore, the scalability and adaptability of the proposed model are vital for its success. With the rapid digitization of healthcare systems, ML models can be integrated into electronic health records (EHRs) to provide real-time predictions. This approach not only enhances clinical decision-making but also empowers patients by offering accessible and personalized healthcare solutions. The role of mobile health (mHealth) applications, which can integrate predictive models to provide instant feedback and guidance, is particularly relevant in the Pakistani context, where smartphone penetration is increasing rapidly [17].

In summary, this research represents a significant step forward in leveraging advanced ML techniques to address a pressing healthcare challenge in Pakistan. By focusing on a population-specific dataset and incorporating a holistic set of features, the study aims to develop a predictive model that is not only accurate but also practical and scalable. The findings of this research have the potential to contribute to the global discourse on diabetes prevention and management, while simultaneously providing a localized solution tailored to the unique needs of the Pakistani population.

CHAPTER 2

LITERATURE REVIEW

Diabetes mellitus (DM) is one of the fastest-growing chronic conditions worldwide, significantly contributing to morbidity and mortality. According to the International Diabetes Federation (IDF, 2021), the global prevalence of diabetes is projected to reach 700 million by 2045. Early diagnosis is critical to mitigating complications such as cardiovascular disease, kidney failure, and neuropathy. Traditional diagnostic methods rely heavily on invasive blood tests and clinical judgment, which may delay early detection. With the advent of machine learning (ML) and deep learning (DL) techniques, researchers have developed predictive models capable of processing vast amounts of medical data for efficient and accurate diagnosis.

This chapter provides a comprehensive review of existing literature on diabetes prediction using ML and DL models, focusing on their methodologies, datasets, challenges, and future directions.

Diabetes prediction aims to classify individuals as diabetic or non-diabetic based on features such as glucose levels, age, BMI, and family history. Early studies relied on simple statistical models like logistic regression, which provided insights but lacked the ability to handle complex relationships in data. Recent research leverages advanced ML and DL techniques to improve prediction accuracy.

One widely used dataset is the **Pima Indians Diabetes Dataset (PIDD)** [18], introduced by Smith et al. (1988). This dataset includes eight clinical features and a binary outcome (diabetic or non-diabetic). While PIDD is a benchmark dataset, other datasets such as the PIMA dataset, MIMIC-III, and custom clinical datasets have also been explored.

Table 1: Summary of online available datasets for diabetes detection

Name	Source	Glucose input parameter	Age	Gender	Features
Pima Indian Data set [18]	Kaggle 1990	2 hours after	Above 21	All female	Skin thickness Insulin

National Institute of Digestive and Kidney Disease	(796)	intaking glucose solution			Pregnancies Blood pressure BMI
Diabetes 130-US Hospitals for Years 1999-2008 [19]	UCI ML repository 2014	Emergency lab tests, HBA1c	Grouped in 10 years intervals	Both male and female	BMI Drug intake. Hospital visits Lab tests
Early-stage diabetes risk prediction [20]	UCI ML repository 2020 (520)	Nil	20-65	Both male and female	Polyuria Polydipsia Weight loss Visual blurring Genital thrush Itching Delayed healing Alopecia Obesity
CDC diabetes health indicators_012 [21]	Kaggle 2012	0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes	18-80	Both male and female	BP Cholesterol BMI Smoker Heart disease Physical Activity Diet General and mental health Education Income

2.1 Machine Learning

Machine learning models are classified into supervised and unsupervised learning methods. Supervised learning has been the most applied method for diabetes prediction due to the availability of labeled datasets.

Logistic regression is a statistical method for binary classification problems. Several studies have used LR to establish baseline models for diabetes prediction. Abendi et. al. [22]demonstrated the simplicity and interpretability of LR models for predicting diabetes using the PIDD dataset, achieving an accuracy of 78%. Decision trees classify data by recursively splitting features based on specific thresholds. In a study by Kaur et al. [13], DTs were applied to PIDD, achieving an accuracy of 75%. Random forests overcome the limitations of single decision trees by combining multiple trees into an ensemble. Zhang et al. [23]reported an accuracy of 82% on PIDD using RF. SVMs map data to a higher-dimensional space using kernel functions to create a hyperplane that separates classes. Ali et al. [24] optimized SVM kernels and achieved 84% accuracy in diabetes classification.

The k-NN algorithm predicts class labels by identifying the closest k neighbors in feature space. It has been used effectively for small datasets, but performance decreases with increasing dimensions due to the curse of dimensionality [24].

LR assumes a linear relationship between features and the target variable, making it unsuitable for datasets with non-linear interactions. Decision trees are easy to interpret but prone to overfitting, particularly with small datasets. RFs are robust to overfitting and can handle missing data, making them ideal for healthcare applications. SVMs perform well with high-dimensional data but are computationally expensive.

2.2 Deep Learning Models

Deep learning models, which consist of multiple layers of interconnected neurons, have shown significant promise in diabetes prediction. These models excel in handling large, complex datasets and capturing non-linear relationships. ANNs are widely used for diabetes prediction due to their adaptability and scalability. [25] used a three-layer ANN on PIDD, achieving an accuracy of 85%. ANNs require substantial computational resources and are prone to overfitting without regularization techniques like dropout.

CNNs are predominantly used for image data but have been adapted for tabular data in healthcare. Kim et al. [26] modified CNN architectures to analyze clinical data, achieving state-of-the-art results in diabetes classification. RNNs are designed to handle sequential data, making them suitable for time-series analysis in diabetes prediction. In a study by Saxena et al. [27], RNNs were employed to predict diabetes onset based on continuous glucose monitoring data. Autoencoders have been used for feature extraction in diabetes datasets, while transfer learning leverages pre-trained models to improve accuracy.

Hybrid models combine multiple algorithms to leverage their individual strengths. Wang et al. [28]combined RF and XGBoost, achieving a 90% accuracy on clinical datasets. Ensemble methods like bagging (e.g., RF) and boosting (e.g., AdaBoost, XGBoost) are effective in reducing bias and variance.

2.3 Datasets in Diabetes Prediction

The availability and quality of datasets significantly impact model performance.

2.3.1 PIDD

A widely used dataset but limited by its small sample size (768 samples) and imbalanced classes.

2.3.2 MIMIC-III

A large-scale dataset containing clinical records from intensive care units.

2.3.3 Custom Datasets

Many researchers collect local datasets tailored to specific demographics, enhancing model generalizability.

This literature review highlights the advancements in diabetes prediction using ML and DL models. While traditional algorithms like LR and RF remain popular, DL techniques like ANNs and CNNs are gaining traction for their superior performance on complex datasets. However, challenges related to data quality, privacy, and generalizability need to be addressed to ensure widespread adoption. Future research should focus on creating scalable, interpretable, and inclusive models for global healthcare systems.

Table 2: Overview of the current state of research on Diabetes detection using artificial intelligence

Reference	Year	Algorithm	Accuracy
[29]	2008	SVM-HBA ANN-HBA DT-HBA	94.79% 94.79% 91.67%
[30]	2009	ELM	82.10%
[31]	2010	eClass	79.37%
[32]	2011	DT	78.17%
[33]	2012	SVM	82.2%
[34]	2012	K- means clustering	94%
[35]	2012	Cascaded K- means clustering	93.3%
[36]	2016	cascaded k-means combined with LR and k-means combined with ANN k-means and SVM	98% 95.3%
[37]	2017	MOE Fuzzy	83.04%
[38]	2017	LR	Mean absolute error: 0.211 Root mean squared error: 0.304 Relative absolute error: 80.1856 % Root relative squared error: 91.4408 %
[39]	2018	NN (10-fold)	84.52%

		NN (entropy feature selection) NN (reduction of attributes, 10-fold)	84.4% 85.24%
[40]	2019	FPCA-SVM PAC-SVM	72.7% 69.28%
[41]	2020	DL DT ANN NB	98.07% 96.62% 90.34% 76.33%
[42]	2020	LDA KNN SVM RF	76.86% 79.24% 80.85% 87.66%
[43]	2020	DT Regression model ANN	65% 80% 83%
[44]	2021	Generalized Linear Model DL DT RF GB SVM	RMSE: 0.402 0.389 0.537 0.390 0.394 0.502

[27]	2021	1D CNN	86.29%
[45]	2022	SMO	99.07%
[46]	2022	DL-DY ID3 J48	95.45% 94.46% 88.51%
[57]	2022	LR GB	75 78
[26]	2022	DNN	89%
[47]	2023	SVM	85.5%
[48]	2023	MLP-NN SVM RF	Without data modelling: 75.32% 76.62% 73.37% With data modelling: 79.87% 80.52% 82.82%
[49]	2023	RF J48 NB	79.57%
[50]	2023	LR EM	93.3% 98.6%
[51]	2023	RF DT SVM ANN	77.9% 92% 74% 98%
[52]	2023	KNN	83.12%
[54]	2023	ANN	80.79

[56]	2023	LR	84
		KNN	84
		CART	85
		RF	88
		SVM	85
		XGB	89
		LightGBM	88
[53]	2024	DT	65.08
		RF	79.33
		SVM	69.03
		Stacking Ensemble (ML)	75.03
		Stacking Ensemble (NN)	95.5
		DNN1	68.80
		DNN2	64.15
		DNN3	65.40
[55]	2024	FNN	81.82%
		CNN	80.52%
[59]	2024	En-RfRsK	88.89%

CHAPTER 3

METHODOLOGY

The diagnosis of diabetes through predictive modelling involve multiple crucial steps including data collection, data preprocessing, model selection and implementation, model performance evaluation and deployment of best performing model on mobile application. These steps are described in figure 3.

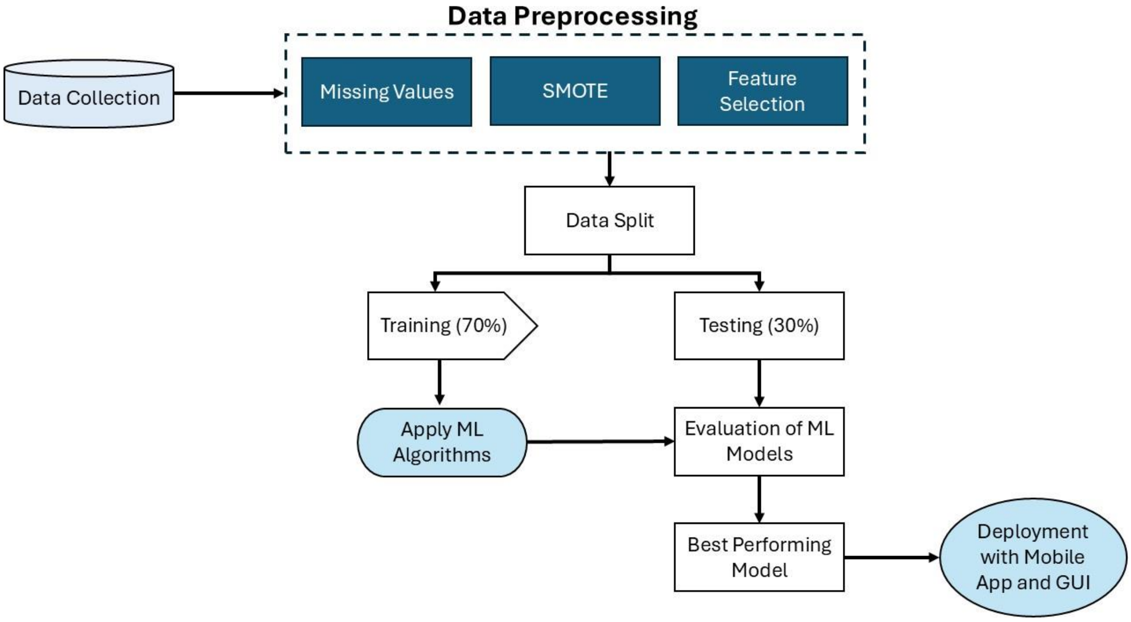


Figure 3: Study Protocol

3.1 Data Collection

Data collection is a fundamental and critical step in the process of developing a machine learning (ML) model. In this study, the primary objective was to collect data representative of the Pakistani population, encompassing a broad spectrum of demographic and ethnic variability. This diverse dataset was essential to ensure the generalizability and accuracy of the model when predicting diabetes outcomes in real-world scenarios.

The data for this study was collected from two major healthcare facilities in Pakistan:

- **PAF Hospital**: A well-known medical facility catering to a diverse group of patients, ensuring access to data from different socioeconomic and ethnic groups.
- **Healthways Lab**: A diagnostic lab providing comprehensive health testing services, contributing to the availability of critical lab-based measurements.

The dataset included a total of **1,000 subjects**, comprising both healthy and diabetic individuals. This ensured a balanced dataset for accurate training and testing of the ML model. Below is a breakdown of the characteristics of the collected data:

3.1.1 Subjects:

- **Type 2 Diabetes**: The primary focus was on patients with Type 2 Diabetes Mellitus.
- **Exclusion of Type 1 Diabetes**: Data for **45 subjects with Type 1 Diabetes** was excluded due to its differing pathophysiology and treatment protocols.
- **Final Dataset Size**: After exclusion, the dataset consisted of **955 subjects**.

3.1.2 Features:

The dataset incorporated a comprehensive set of features for each subject, including both demographic and clinical parameters. Key features included:

- **Demographic Data**: Age, gender, and ethnic background.
- **Clinical Parameters**:
  - BMI (Body Mass Index)
  - Waist Circumference
  - Blood Pressure (Systolic and Diastolic)
  - Fasting Blood Glucose Levels
- **Behavioral and Lifestyle Data**: Smoking status and physical activity levels.
- **Family History**: Information on genetic predisposition to diabetes.
- **Reproductive History (for females)**: Number of pregnancies and history of gestational diabetes.

A significant strength of the dataset was its inclusion of individuals from diverse ethnic backgrounds across Pakistan. This demographic diversity aimed to make the ML model robust and applicable to real-world scenarios where ethnic and cultural factors influence diabetes prevalence and health outcomes.

Table 3: Dataset description

Features	Description	Type
Age	Age in years	Numerical
Gender	Male (1) or Female (0)	Categorical
BMI	Body Mass Index (BMI)= Weight in kg/Height in m^2	Numerical
Waist Circumference	Measurement of waist in cm	Numerical
Fasting Blood Glucose	Blood glucose levels after 8 hours of fasting (mg/dL)	Numerical
Diastolic Blood Pressure	Minimum pressure in arteries when heart relax (mmHg)	Numerical
Systolic Blood Pressure	Maximum pressure in arteries when heart pumps blood (mmHg)	Numerical
Smoker	Nonsmoker (0) or smoker (1)	Categorical
Family History	Family history of diabetes (1) else 0	Categorical
No. of Pregnancies	Number of times a female was pregnant	Numerical
Gestational Diabetes	Diagnosis of diabetes during pregnancy	Categorical
Physical Activity	Level of physical activity 1= low 2= moderate 3= high	Categorical

To ensure the reliability of the dataset, several steps were taken during and after data collection:

- **Verification of Data Sources**: Only data from verified clinical tests and reports were included.

- **Handling Missing Data**: Any missing entries in critical features were addressed during the preprocessing stage.

- **Standardization of Measurements**: Parameters such as BMI and blood glucose levels were standardized to avoid discrepancies due to varying measurement units.

The use of local data from Pakistani healthcare facilities provided two major advantages:

- **Demographic Relevance**: Since diabetes prevalence and risk factors can vary by region, the dataset accurately reflects the conditions faced by the target population.

- **Ethnic Variability**: The inclusion of diverse ethnic groups ensured that the ML model would not exhibit bias toward any demographic group.

- Collecting high-quality data for ML models involves addressing several challenges:

- **Data Anonymity**: Ensuring that patient data complied with ethical standards, including anonymization and confidentiality.

- **Exclusion of Outliers**: The dataset was carefully screened to remove subjects with rare or unrelated conditions (e.g., Type 1 diabetes).

- **Access to Data**: Collaboration with healthcare facilities required adherence to protocols and obtaining necessary permissions.

The data collection process laid a strong foundation for building a reliable ML model for diabetes prediction. By collecting data from a diverse and representative population, the study ensured that the model would be both robust and generalizable. The focus on Type 2 diabetes and the exclusion of medication-related bias were deliberate decisions aimed at improving the quality and relevance of the dataset.

3.2 Data Preprocessing

Data preprocessing is a crucial step in preparing raw data for analysis and modeling. It ensures that the dataset is clean, consistent, and structured in a way that enhances the performance of machine learning algorithms. In this study, preprocessing involved several key tasks, including data standardization, handling class imbalance, imputing missing values, and feature selection. Each step was carefully designed to improve the reliability and accuracy of the diabetes prediction model.

3.2.1 Standardization of Data

Standardization is a preprocessing technique used to ensure consistency across trials and improve the comparability of features with varying scales. It is particularly important for features with different units, such as BMI, fasting blood glucose, and blood pressure. Machine learning algorithms like Support Vector Machines (SVMs), K-Nearest Neighbors (KNN), and Gradient Boosting are sensitive to the scale of input features. Without standardization, features with larger scales may dominate the model training process, leading to biased predictions. The **z-score normalization** technique was applied to all continuous numerical features. This method rescales the data so that the mean is 0 and the standard deviation is 1. BMI, fasting blood glucose, systolic and diastolic blood pressure, and waist circumference were standardized to ensure uniform scaling.

3.2.2 Handling Class Imbalance

Class imbalance occurs when the number of positive and negative outcomes in the dataset is disproportionate. In this study, the dataset included both diabetic and non-diabetic individuals, with an observed imbalance favoring diabetic cases. Proper handling of class imbalance was critical to avoid bias in the model toward the majority class. Imbalanced datasets can lead to models that perform well on accuracy but fail to correctly identify the minority class (e.g., non-diabetic subjects). This would result in poor recall and F1 scores for the minority class. Techniques like **Synthetic Minority Oversampling Technique (SMOTE)** were used to generate synthetic samples for the minority class (non-diabetic subjects) to balance the dataset. A portion of the majority class (diabetic subjects) was randomly removed to balance the dataset without introducing significant bias. Certain machine learning algorithms (e.g., Random Forest and XGBoost) were configured to assign higher weights to the minority class during training to improve prediction sensitivity.

3.2.3 Handling Missing Values

Missing values in the dataset can arise from incomplete records during data collection or errors in data entry. Addressing missing values is essential to avoid reducing the quality of the data or introducing bias into the model. Missing entries in continuous features (e.g., fasting blood glucose, BMI). Missing categorical data (e.g., smoking status or family history). Missing values in numerical features were replaced using the **mean or median** of the respective feature, depending on the distribution. If the data was skewed, the median was used to avoid distortion. Missing values in categorical features (e.g., smoker or family history) were filled using the **mode** of the feature. For features with a significant percentage of missing values, regression-based imputation methods were used to predict missing values based on other related features.

3.3 Feature Selection

Feature selection is the process of identifying the most relevant and significant features in the dataset that contribute to the prediction of outcomes. By reducing the number of features, the model becomes less complex, less prone to overfitting, and faster to train. Not all features contribute equally to the predictive performance of the model. Irrelevant or redundant features can introduce noise and reduce model efficiency. Algorithms such as Random Forest and Gradient Boosting were used to calculate the importance of each feature. These models assign scores based on the contribution of each feature to prediction accuracy.

A heatmap of correlation coefficients was generated to identify highly correlated features. Features with a high correlation (e.g., >0.85>0.85>0.85) were flagged for potential removal to reduce multicollinearity. Recursive Feature Elimination (RFE) was employed to iteratively remove the least important features, retaining only the most significant ones for model training.

3.3.1 Training Machine Learning Models

Training machine learning (ML) models involves using the preprocessed dataset to build predictive algorithms that can classify individuals as diabetic or non-diabetic. This step ensures that the models learn patterns, relationships, and insights from the input features to make accurate predictions. In this study, eight ML algorithms were employed, each with their unique strengths and methodologies. These models were selected for their proven efficacy in classification tasks and their ability to handle diverse datasets.

3.3.1.1 Logistic Regression (LR)

Logistic Regression is a statistical method used for binary classification problems. It predicts the probability of a binary outcome based on input features. LR models the relationship between the dependent variable (Outcome) and independent variables (features) using a logistic function. It is simple and interpretable and works well with linearly separable data. Regularization techniques like L1 (Lasso) and L2 (Ridge) were applied to prevent overfitting.

3.3.1.2 K-Nearest Neighbors (KNN)

KNN is a non-parametric and instance-based algorithm that classifies a data point based on its neighbors. KNN calculates the distance (e.g., Euclidean distance) between the input data point and all training samples. The algorithm assigns the class label of the majority of its kk nearest neighbors. It is simple, intuitive and performs well on small datasets. The optimal value of kk was selected using cross-validation. Smaller kk values focus on local patterns, while larger kk values generalize better.

3.3.1.3 Support Vector Machine (SVM)

SVM is a supervised learning algorithm used to find the optimal hyperplane that separates classes. SVM constructs a hyperplane in a high-dimensional space to maximize the margin between diabetic and non-diabetic classes. For non-linearly separable data, kernel functions (e.g., radial basis function or polynomial) were applied to transform the data into a higher-dimensional space. The CC parameter controls the trade-off between achieving a low error and a large margin.

3.3.1.4 Random Forest (RF)

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance. RF builds several decision trees during training, and each tree predicts the class label. The final prediction is determined by majority voting (for classification tasks). RF handles missing values and outliers well. It reduces overfitting by aggregating multiple decision trees. Hyperparameters include number of trees (n_estimators_estimators), maximum depth of each tree (max_depth_depth) and minimum samples required to split a node.

3.3.1.5 Gradient Boosting (GBoost)

Gradient Boosting is an ensemble method that builds decision trees sequentially, optimizing for errors from the previous trees. Each subsequent tree attempts to correct the errors made by the previous tree by minimizing a loss function (e.g., log-loss). This iterative process improves model accuracy. It handles complex relationships in data and is highly customizable with loss functions. Hyperparameter include learning rate, number of estimators (n_estimatorsn_estimators) and Maximum tree depth (max_depthmax_depth).

3.3.1.6 Extreme Gradient Boosting (XGBoost)

XGBoost is an advanced implementation of Gradient Boosting that focuses on speed and performance. Like GBoost, XGBoost builds trees iteratively. However, it introduces regularization terms to prevent overfitting and supports parallel computation for faster training. It uses second-order derivatives to optimize the loss function. It is highly efficient and has a built-in mechanism for handling missing data. Hyperparameters include learning rate (η\eta), maximum depth (max_depthmax_depth) and subsample ratio (subsamplesubsample).

3.3.1.7 Light Gradient Boosting Machine (LightGBM)

LightGBM is a gradient boosting framework designed for speed and efficiency, particularly on large datasets. LightGBM uses a leaf-wise tree growth strategy, where it grows trees vertically (leaf-wise) rather than level-wise. This method focuses on leaves with the maximum loss reduction, which reduces error more effectively. They are faster training compared to traditional boosting methods and efficient memory usage.

3.3.1.8 Adaptive Boosting (AdaBoost)

AdaBoost is an ensemble learning method that combines weak classifiers to create a strong classifier. AdaBoost assigns weights to misclassified instances, forcing subsequent weak classifiers to focus on these difficult examples. Final predictions are based on the weighted sum of predictions from all weak classifiers. It reduces bias and variance and is effective for simple models like decision stumps.

3.3.2 Model Training Approach

1. **Data Split**: The preprocessed dataset was split into training (70%) and testing (30%) sets. This ensures that the models are evaluated on unseen data.

2. **Cross-Validation**: K-fold cross-validation (with k=5k=5) was used to evaluate model performance on different subsets of the training data, reducing the likelihood of overfitting.

3. **Hyperparameter Tuning**: Grid search and random search methods were used to optimize hyperparameters for each model.

4. **Evaluation Metrics**: Accuracy, precision, recall, F1-score, and AUC-ROC were computed to compare the performance of the models.
5. **Best Model Selection**: The best-performing model was chosen based on its ability to generalize to unseen data while achieving high accuracy, recall, and AUC-ROC.

This step ensured that the models were trained optimally, ready for deployment, and capable of making reliable predictions for diabetes diagnosis.

3.3.3 Integration of the Best Performing Model into the Flutter App

It was deployed into a Flutter-based mobile application after identifying the best-performing machine learning model from the training and evaluation phases. The app was built using **Flutter** and **Dart** in **VS Code** and designed to provide users with a simple and efficient tool for predicting diabetes status based on input parameters.

3.3.3.1 App Design and Development

The app was developed in **Flutter**, an open-source UI toolkit by Google that enables cross-platform development for Android, iOS, and web applications. **Dart**, the programming language used by Flutter, facilitated the implementation of the app's logic and UI components.

The app serves as a diabetes prediction tool for users by leveraging machine learning (ML). It allows users to input relevant health parameters and receive instant predictions on their diabetes status.

- **Features of the App**:
  - ✓ **User-Friendly Interface**: Designed with a clean, intuitive UI for easy navigation.
  - ✓ **Input Form**: A dedicated page for users to input health-related parameters like age, BMI, fasting blood glucose, waist circumference, physical activity level, and other features selected during the model training phase.
  - ✓ **Instant Prediction**: Upon submitting the inputs, the app processes the data through the ML model to provide a "Diabetic" or "Non-Diabetic" result.

- ✓ **Secure and Fast**: The app was optimized for speed and ensures the security of user data.

3.3.3.2 Integration of the Machine Learning Model

The best-performing model was implemented in the app for real-time diabetes prediction. The integration involved several steps to ensure seamless interaction between the ML model and the app.

- **Model Selection:**
  Based on training and evaluation, the best-performing model was selected for deployment. For this study, models such as **XGBoost** showed high accuracy and were preferred for deployment.

- **Model Deployment Methodology**:

  The trained model, saved in the form of a serialized file (e.g., xgb_model.pkl for XGBoost), was deployed using **FastAPI**, a modern web framework for building APIs in Python. The model predictions were accessed through an API endpoint.

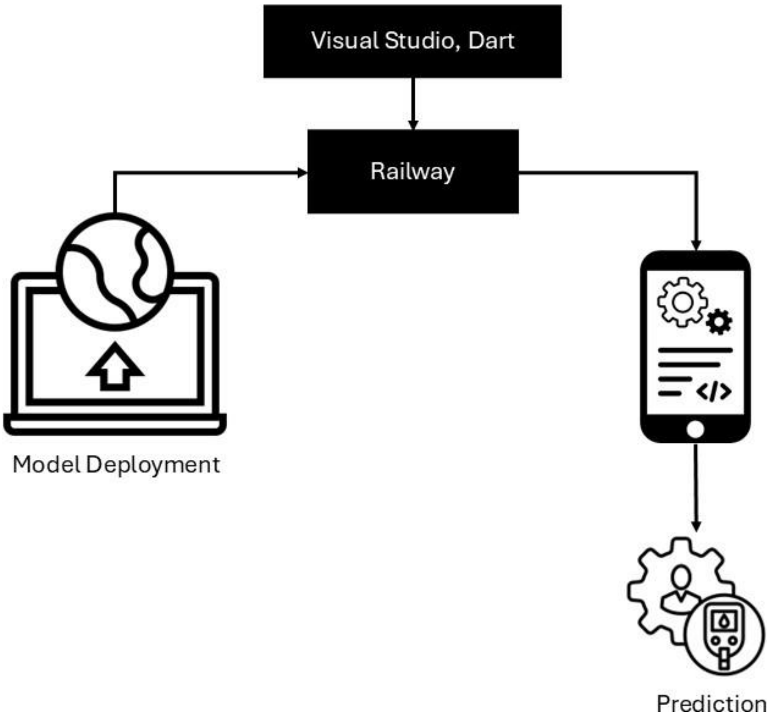


Figure 4: Deployment of best model on mobile application

3.3.3.3 Testing and Optimization

- **Testing**:

  The app was rigorously tested on both Android and iOS devices to ensure compatibility and correctness of predictions. Edge cases were considered, such as missing inputs or incorrect data formats.

- **Optimization**:

The model's prediction time was minimized by optimizing the backend server and ensuring the FastAPI responded quickly. The Flutter app was optimized for responsiveness, ensuring it worked seamlessly on devices with varying screen sizes.

3.3.3.4 Benefits of Mobile Application

- **Accessibility**:

  The app brings state-of-the-art diabetes prediction technology to users' fingertips, making early screening more accessible.

- **Customization**:

  The model was trained on data specific to the Pakistani population, ensuring predictions were more relevant and accurate for the target demographic.

- **User Empowerment**:

  Users are empowered to take proactive steps in managing their health by receiving personalized insights.

By integrating the best-performing ML model into the Flutter app, a robust, user-friendly solution for diabetes prediction was created. This app has the potential to positively impact public health by enabling early detection and promoting preventative measures.

CHAPTER 4

RESULTS

The results chapter provides a detailed comparison of the performance metrics of various machine learning classifiers and their statistical evaluation. These analyses are aimed at identifying the most effective model for predicting outcomes, particularly in the context of diabetes risk assessment. This chapter integrates the results from feature selection, model evaluation, and statistical significance tests.

4.1 Classifier Performance Without Feature Selection

The performance of various classifiers without feature selection is depicted in the graph titled "Classifier Performance Comparison without FS." Metrics such as accuracy, area under the curve (AUC), precision, recall, and F1-score were evaluated across multiple classifiers, including Logistic Regression (LR), k-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting (GBoost), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and AdaBoost.

Without feature selection, XGBoost demonstrated consistently high performance across all metrics, achieving an accuracy close to 0.90. Similarly, LightGBM and Gradient Boosting also exhibited robust metrics, with minimal variance among the key scores. On the other hand, Logistic Regression had the lowest performance, particularly in recall and F1-score, indicating its limited capability to handle complex datasets without preprocessing. These findings suggest that ensemble methods like XGBoost and LightGBM are more suited for the data structure used in this study.

4.2 Classifier Performance with Feature Selection

Feature selection was applied to optimize the input dataset by reducing redundancy and retaining the most relevant features. The graph titled "Classifier Performance Comparison with FS" highlights the impact of this process. The results showed an overall improvement in the performance of most classifiers. XGBoost remained the best-performing model, with an accuracy exceeding 0.92 and enhanced precision and recall. LightGBM and Gradient Boosting also showed incremental improvements in accuracy and F1-score.

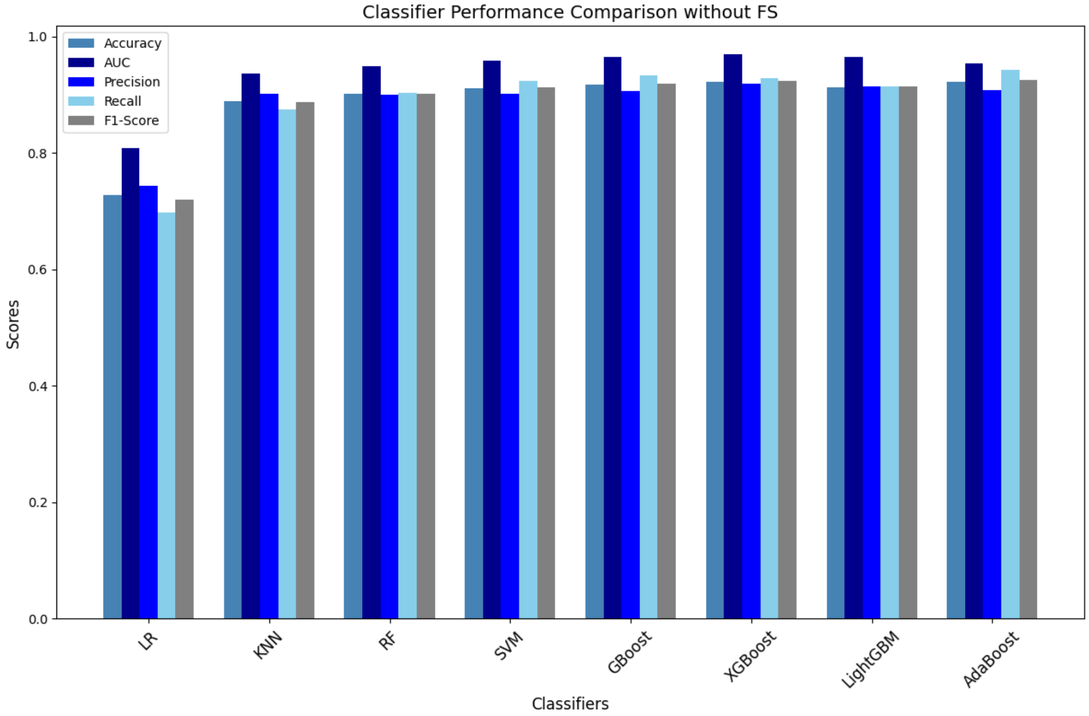


Figure 5: Performance analysis of classifiers without feature selection

The feature selection process particularly benefited classifiers such as Random Forest and SVM, which displayed significant gains in AUC and recall. Logistic Regression

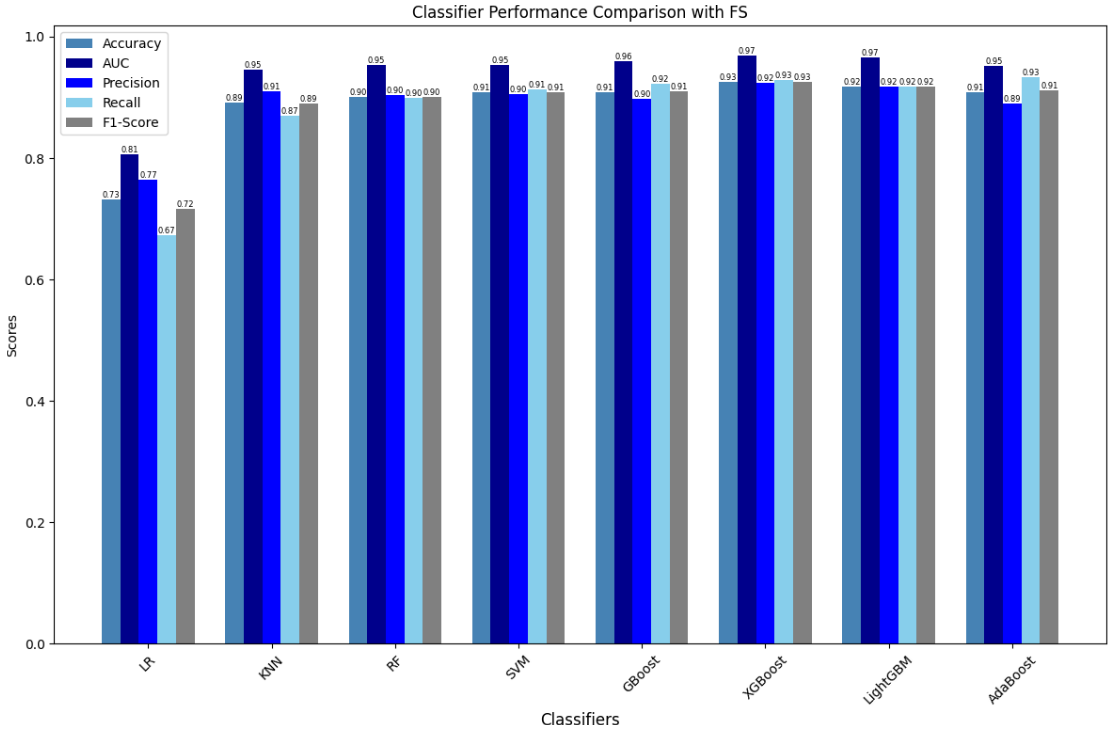


Figure 6: Performance analysis of classifier with feature selection

also saw marginal improvement but remained less effective compared to ensemble methods. These results validate the importance of feature selection in boosting model performance and generalization capabilities.

4.3 Cross-Validation Results

Cross-validation was employed to ensure the robustness and reliability of the classifiers. The boxplot titled "Cross-Validation Accuracy Scores for Each Classifier" illustrates the variability in accuracy scores for each classifier. XGBoost demonstrated the highest mean accuracy with the least variance, indicating its stability and superior performance. Gradient Boosting and LightGBM followed closely, while Logistic Regression and KNN exhibited lower accuracy and higher variability.

The bar chart "Mean Cross-Validation Accuracy for Each Classifier" reinforces these findings, highlighting XGBoost as the top performer with a mean accuracy of 0.92. Gradient Boosting and LightGBM also achieved high mean accuracy scores, while Logistic Regression lagged significantly.

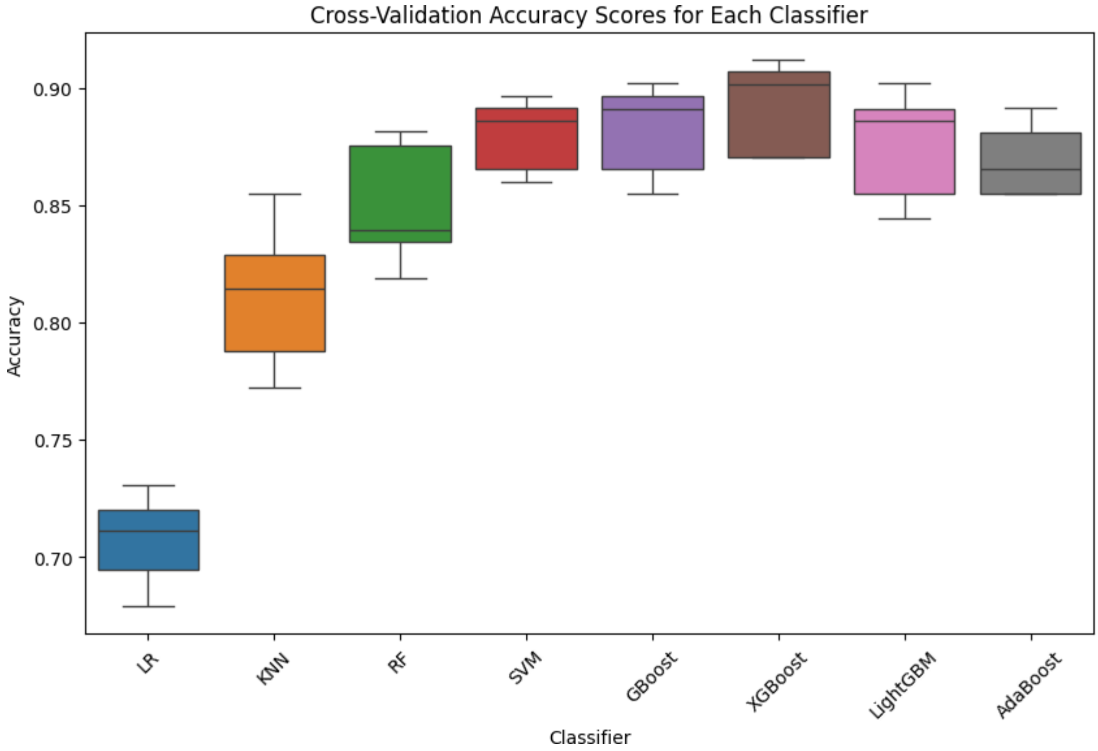


Figure 7: Accuracies of classifiers after 10 k cross validation

4.4 Statistical Analysis

Statistical significance tests were conducted to assess the differences between classifiers. First one way ANOVA was applied to check significance among groups which gives a p value of less than 0.05 proving the existence of significance among the classifier performance.

The Tukey post hoc analysis demonstrates the mean differences between classifiers. XGBoost showed statistically significant superiority over Logistic Regression and KNN, with positive mean differences. Gradient Boosting and LightGBM also displayed significant advantages over Logistic Regression, validating the observations from performance metrics.

The statistical analysis confirms that ensemble methods like XGBoost, Gradient Boosting, and LightGBM not only outperform simpler models but also exhibit consistent and reliable performance across various splits of the data. This robustness underscores their suitability for real-world applications where stability and accuracy are critical.

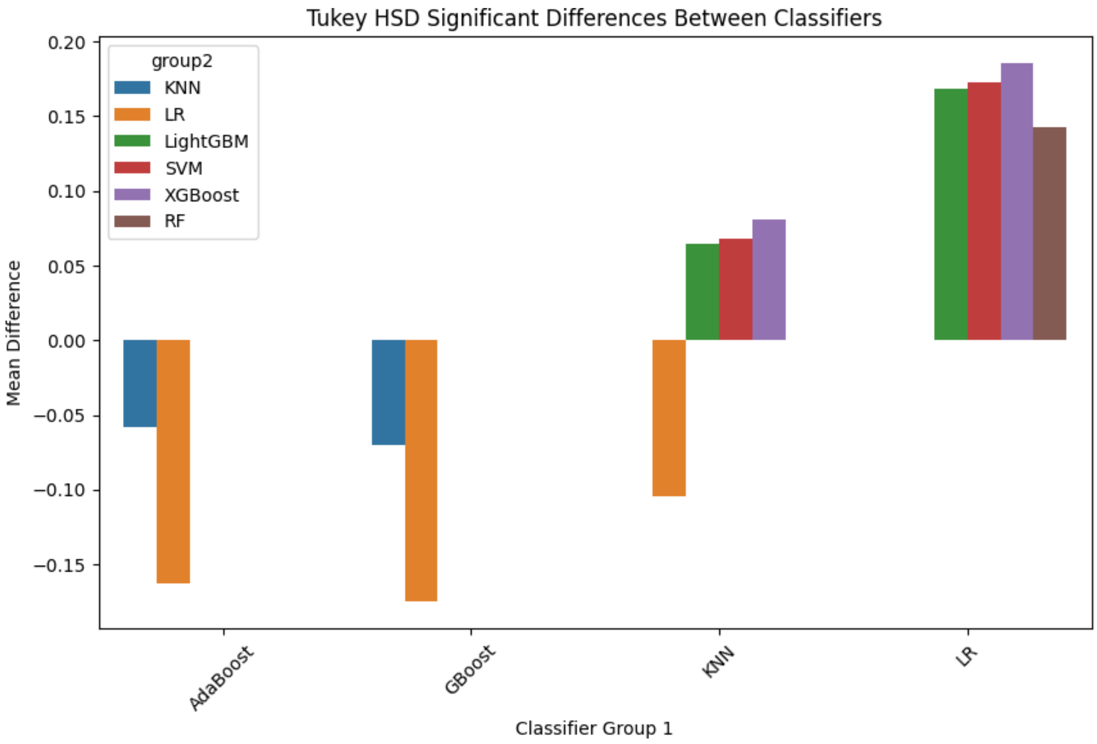


Figure 8: Post hoc analysis

The results from this study highlight the effectiveness of ensemble methods, particularly XGBoost, in predicting diabetes risk. Feature selection emerged as a crucial step in enhancing model performance by eliminating irrelevant features and reducing dimensionality. Cross-validation and statistical analysis further validated the reliability and robustness of the top-performing classifiers. These findings provide a solid foundation for integrating the best model into a mobile application for diabetes risk prediction and management, as discussed in subsequent sections.

CHAPTER 5

DISCUSSION

The prediction of diabetes using machine learning models has gained significant attention due to the alarming rise in diabetes cases globally. This study leverages a unique dataset collected from the Pakistani population to evaluate and compare the performance of several machine learning algorithms in predicting diabetes. By incorporating features that encompass both invasive and non-invasive parameters, the study aims to provide a comprehensive assessment of diabetes risk while enhancing accessibility to diagnostic tools.

The methodology adopted in this research demonstrates a robust and systematic approach to developing predictive models. Starting with data collection, special emphasis was placed on ensuring that the dataset reflects demographic and cultural nuances specific to the Pakistani population. This not only enhances the relevance of the models but also addresses a significant gap in existing literature, where most datasets and models are biased toward Western populations. The inclusion of features such as waist circumference, physical activity levels, gestational diabetes, and smoking habits highlights the importance of region-specific variables in improving predictive accuracy.

A key strength of this study lies in the careful partitioning of the data into training and testing sets, with 70% allocated for model development and 30% reserved for evaluation. This split ensures that the models are trained on a substantial portion of the data while being rigorously tested on unseen samples, minimizing overfitting and maximizing generalizability. Additionally, the use of multiple machine learning algorithms allows for a comparative analysis, which is crucial for identifying the most effective approach for this specific dataset.

The application of advanced algorithms such as XGBoost, LightGBM, and Random Forest, alongside traditional models like logistic regression and support vector machines (SVM), ensures a holistic evaluation. By tuning hyperparameters and employing cross-validation techniques, the study seeks to balance accuracy and generalizability. The integration of these techniques reflects a comprehensive understanding of machine learning principles and highlights the researcher's commitment to methodological rigor.

One of the critical aspects of this study is the evaluation metrics used to compare model performance. Accuracy, AUC (Area Under the Curve), precision, recall, and F1-score provide a multidimensional perspective on how well the models perform. While accuracy is often the most reported metric, this study emphasizes the importance of other indicators like recall, particularly given the high cost of false negatives in diabetes prediction. Patients misclassified as non-diabetic may not receive timely interventions, leading to severe health complications. Therefore, a model with high recall ensures that most diabetic patients are correctly identified, even at the expense of a few false positives.

The real-world implications of this research extend beyond numerical metrics. The deployment of the best-performing model in a mobile application and GUI makes the predictive tool accessible to a broader audience, including healthcare professionals and individuals in remote areas. By simplifying the diagnostic process and incorporating both invasive (e.g., fasting blood glucose) and non-invasive (e.g., physical activity) features, the application empowers users to assess their diabetes risk without the need for extensive medical consultations. This approach aligns with global health priorities to enhance early detection and preventive care.

Moreover, this study's focus on diverse feature selection emphasizes the importance of interdisciplinary research in healthcare. Variables such as smoking status and physical activity levels, which are often overlooked in clinical diagnostics, play a significant role in diabetes risk. Incorporating these factors into the predictive model not only improves accuracy but also sheds light on lifestyle interventions that could mitigate risk. This aligns with the broader goals of precision medicine, where individual risk factors guide tailored prevention and treatment strategies.

Another noteworthy contribution of this research is its ability to inform public health policies. By identifying key predictors of diabetes within the Pakistani population, the findings provide valuable insights for designing targeted awareness campaigns. For instance, high-risk groups identified through the model can be prioritized for screening and educational interventions, ensuring optimal allocation of resources in a resource-constrained healthcare system.

Despite these strengths, it is essential to acknowledge the limitations of the study. While the dataset is unique and reflective of the Pakistani population, its relatively small size may impact the generalizability of the findings to other regions. Expanding the dataset to include a larger and more diverse sample would enhance the robustness of the models and facilitate comparisons across different ethnic and cultural groups. Additionally, some features, such as physical activity levels and smoking status, rely on self-reported data, which is subject to reporting bias. Future research could explore the integration of objective measurements to address this limitation.

A significant challenge encountered in this study was the class imbalance within the dataset, with 27.7% of the subjects classified as non-diabetic and 72.3% as diabetic. Class imbalance is a common issue in medical datasets, where certain conditions, such as diabetes, are often overrepresented due to targeted data collection or inherent population prevalence. This imbalance can lead to biased models that favor the majority class, potentially overlooking the minority class during predictions. Such bias is especially concerning diabetes prediction, where misclassifying non-diabetic individuals could result in delayed preventive care.

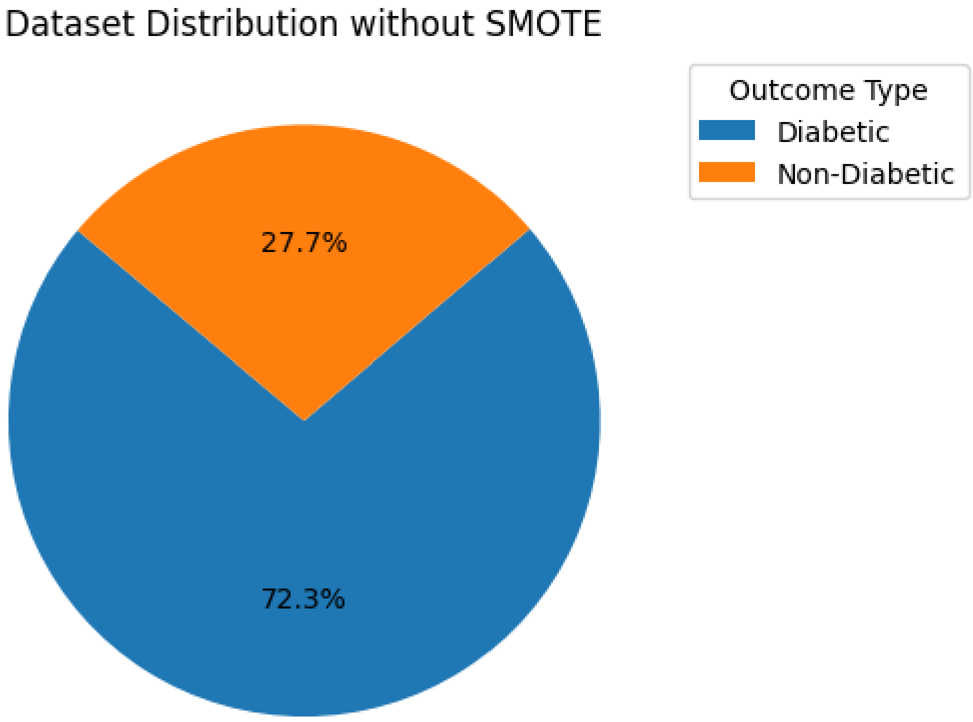


Figure 9: Dataset distribution without SMOTE

To address this imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was employed. SMOTE generates synthetic samples for the minority class by interpolating between existing samples. This approach effectively balances the dataset by increasing the representation of non-diabetic subjects to match that of diabetic subjects, creating a 50:50 distribution. By doing so, SMOTE ensures that machine learning models do not disproportionately favor the majority class, leading to more equitable performance metrics across both classes.

The decision to use SMOTE was guided by its advantages over other techniques such as random oversampling or under sampling. Random oversampling, while effective in balancing classes, often leads to overfitting because it duplicates existing data points. On the other hand, under sampling reduces the size of the majority class, potentially discarding valuable information. SMOTE, however, preserves the integrity of the dataset by generating synthetic data rather than duplicating or discarding samples, thereby mitigating the risk of overfitting while maintaining the original dataset's diversity.

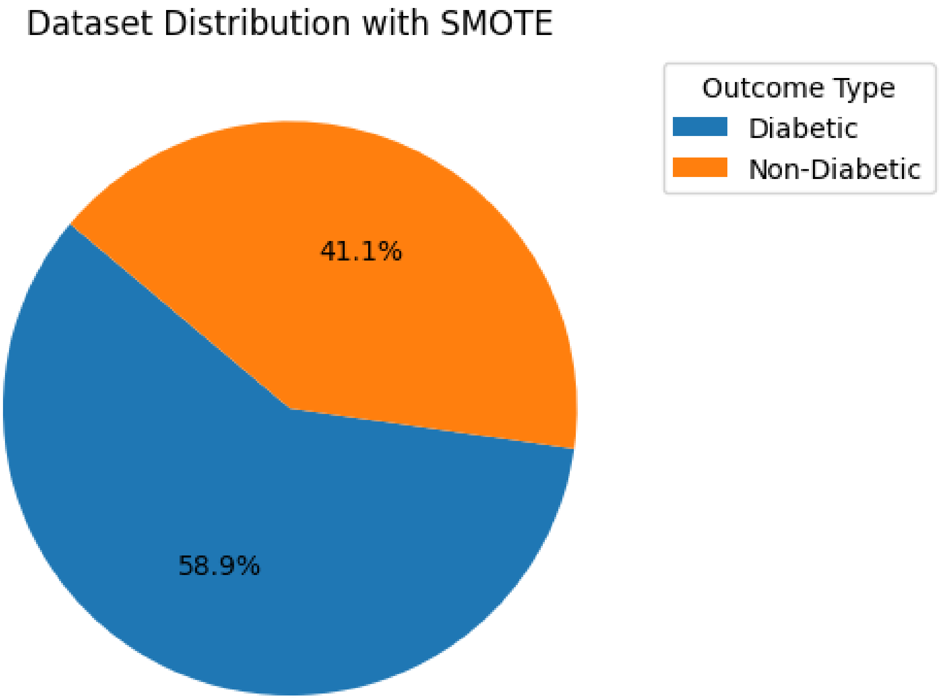


Figure 10: Data distribution with SMOTE balancing technique

The balanced dataset resulting from SMOTE was instrumental in improving the performance of the machine learning models. It allowed the algorithms to learn from an equal representation of diabetic and non-diabetic subjects, ensuring fair predictions across both classes. Metrics such as recall and F1-score, which are sensitive to class imbalance, showed significant improvement after applying SMOTE. This highlights the technique's effectiveness in addressing the disproportionate representation of classes.

Another essential aspect of this study was the exploration of feature relationships using a correlation matrix. The correlation matrix provides insights into the strength and direction of linear relationships between variables, which is critical for understanding the dataset's structure and identifying potential multicollinearity issues. By computing Pearson correlation coefficients between features, the matrix served as a foundational tool for feature selection and model optimization.

The analysis revealed several interesting relationships. Features such as BMI, waist circumference, and fasting blood glucose showed moderate to strong positive correlations with the outcome variable, suggesting their significant contribution to diabetes prediction. This aligns with established medical research indicating that these factors are key indicators of metabolic health and diabetes risk. For instance, higher BMI and waist circumference values are well-documented risk factors for insulin resistance and glucose intolerance.

Interestingly, certain features displayed strong interrelationships among themselves, such as BMI and waist circumference. While this relationship is expected due to their shared basis in body composition, it raises concerns about multicollinearity, which could affect model performance. High multicollinearity can lead to inflated variance in regression coefficients, making it challenging to interpret feature importance accurately. To address this, techniques such as variance inflation factor (VIF) analysis or dimensionality reduction methods like PCA (Principal Component Analysis) could be employed in future work to ensure that the models remain robust and interpretable.

Another essential aspect of this study was the exploration of feature relationships using a correlation matrix. The correlation matrix provides insights into the strength and direction of linear relationships between variables, which is critical for understanding the dataset's structure and identifying potential multicollinearity issues. By computing Pearson correlation coefficients between features, the matrix served as a foundational tool for feature selection and model optimization.

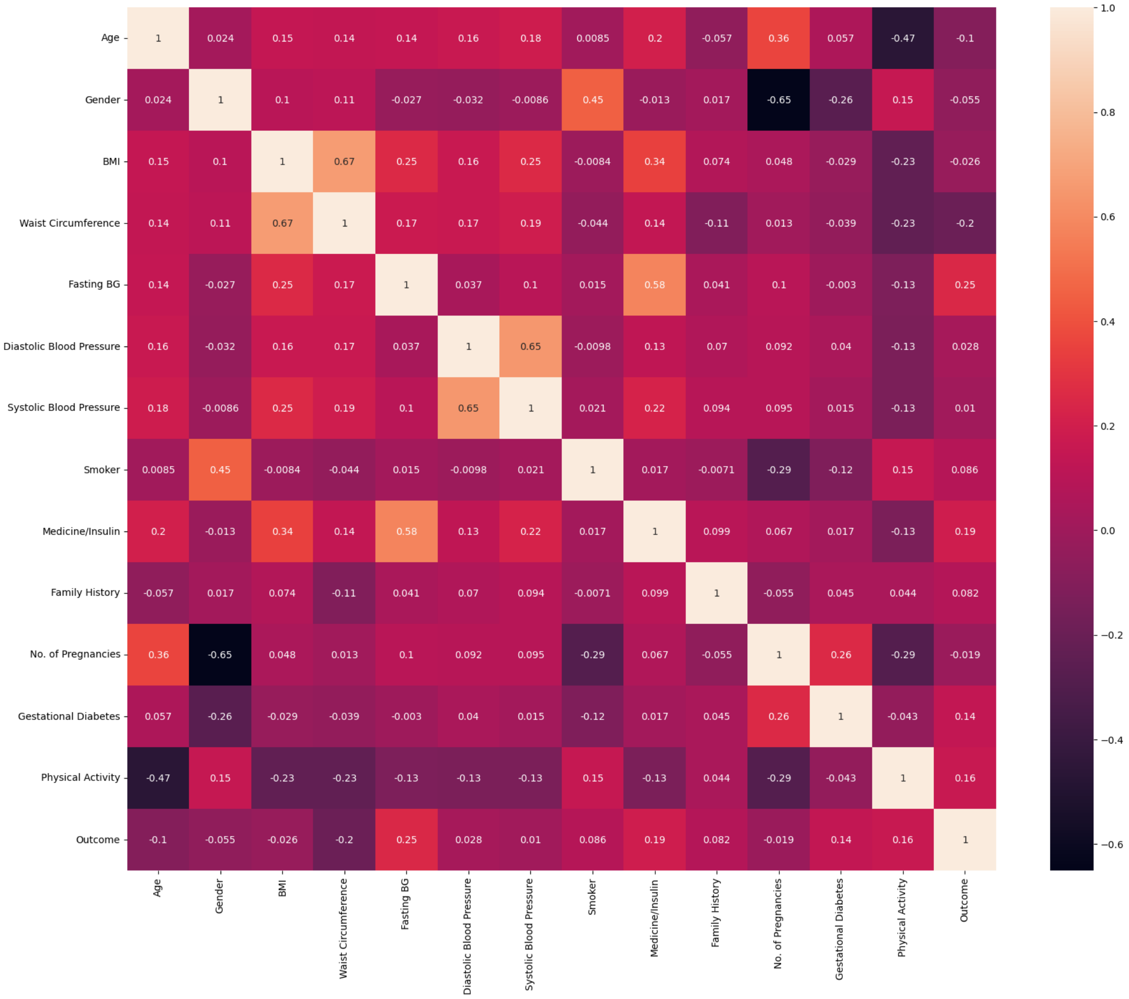


Figure 11: Heatmap correlation matrix

Feature selection is a crucial step in machine learning workflows, as it ensures that only the most relevant and informative features are included in the predictive model. This study employed **Recursive Feature Elimination (RFE)** to systematically identify and rank the most significant features for diabetes prediction. RFE is a wrapper-based method that works by recursively fitting a model, ranking features based on their importance, and eliminating the least significant ones until a predefined number of features remains.

Given the diverse range of features in the dataset—spanning demographic, clinical, and lifestyle factors—RFE was particularly well-suited for this study. Unlike filter-based methods that evaluate feature importance in isolation, RFE considers the relationships between features in the context of the selected machine learning model. This ensures that the final feature set optimizes model performance while minimizing redundancy and overfitting.

RFE was applied using algorithms such as Random Forest and XGBoost, as these models are robust in handling complex, non-linear relationships and naturally provide feature importance scores. By leveraging these algorithms, RFE identified a subset of features that contributed most significantly to the predictive accuracy of the model.

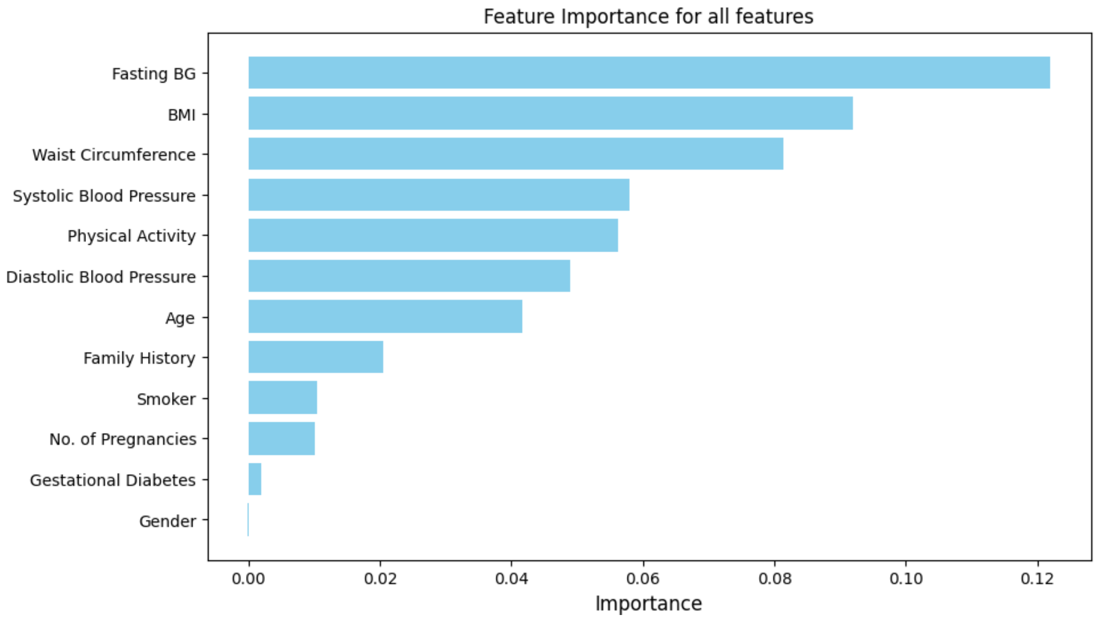


Figure 12: Feature Importance using REF

Selected Features and Their Implications

The RFE process consistently highlighted the following features as the most important predictors of diabetes:

- **Fasting Blood Glucose (FBG):** This feature ranked highest, as expected, due to its direct relationship with diabetes diagnosis. Elevated fasting glucose levels are a primary diagnostic criterion and a key indicator of impaired glucose metabolism.

- **Waist Circumference and BMI:** These anthropometric measures were also strongly prioritized by RFE. Both are well-established markers of obesity, a major risk factor for Type 2 diabetes. Their inclusion underscores the critical role of body composition in diabetes prediction.

- **Age and Family History:** Age remained a strong predictor, as the risk of diabetes increases with advancing age. Family history further emphasized the genetic predisposition to the condition, reinforcing its role as a non-modifiable yet highly informative feature.

- **Physical Activity and Smoking Status:** While these lifestyle factors were less strongly correlated with the outcome variable in the correlation matrix, RFE retained them due to their contribution to overall model performance. This suggests that their impact on diabetes risk may interact with other features in complex, non-linear ways.

The SHAP analysis provided the following key insights into feature importance for the diabetes prediction model:

1. **Fasting Blood Glucose (FBG):** SHAP consistently identified FBG as the most influential feature, with high Shapley values for instances where individuals had elevated glucose levels. This aligns with its central role in diabetes diagnosis and demonstrates the model's ability to prioritize clinically significant features.

2. **Waist Circumference and BMI:** These features showed high global importance scores, reinforcing their relevance as predictors of diabetes. SHAP visualizations revealed a positive relationship between increased values of these features and the likelihood of diabetes, reflecting their association with obesity-related metabolic disorders.

3. **Age:** SHAP values for age indicated a gradual increase in diabetes risk with advancing age. This feature's importance aligns with established knowledge about the progressive nature of diabetes as individuals grow older.

4. **Family History:** The analysis showed that a positive family history significantly contributed to higher diabetes risk, emphasizing the genetic predisposition captured by the model.

5. **Gestational Diabetes and Number of Pregnancies:** SHAP visualizations highlighted these features as key predictors for females, particularly in cases where gestational diabetes was reported. The interaction between these features and age also emerged as an important determinant of risk.

6. **Physical Activity and Smoking Status:** SHAP revealed subtle yet significant contributions of these lifestyle factors. Low physical activity levels were associated with higher diabetes risk, while smoking status had a complex relationship, with interactions observed in combination with other features such as BMI.

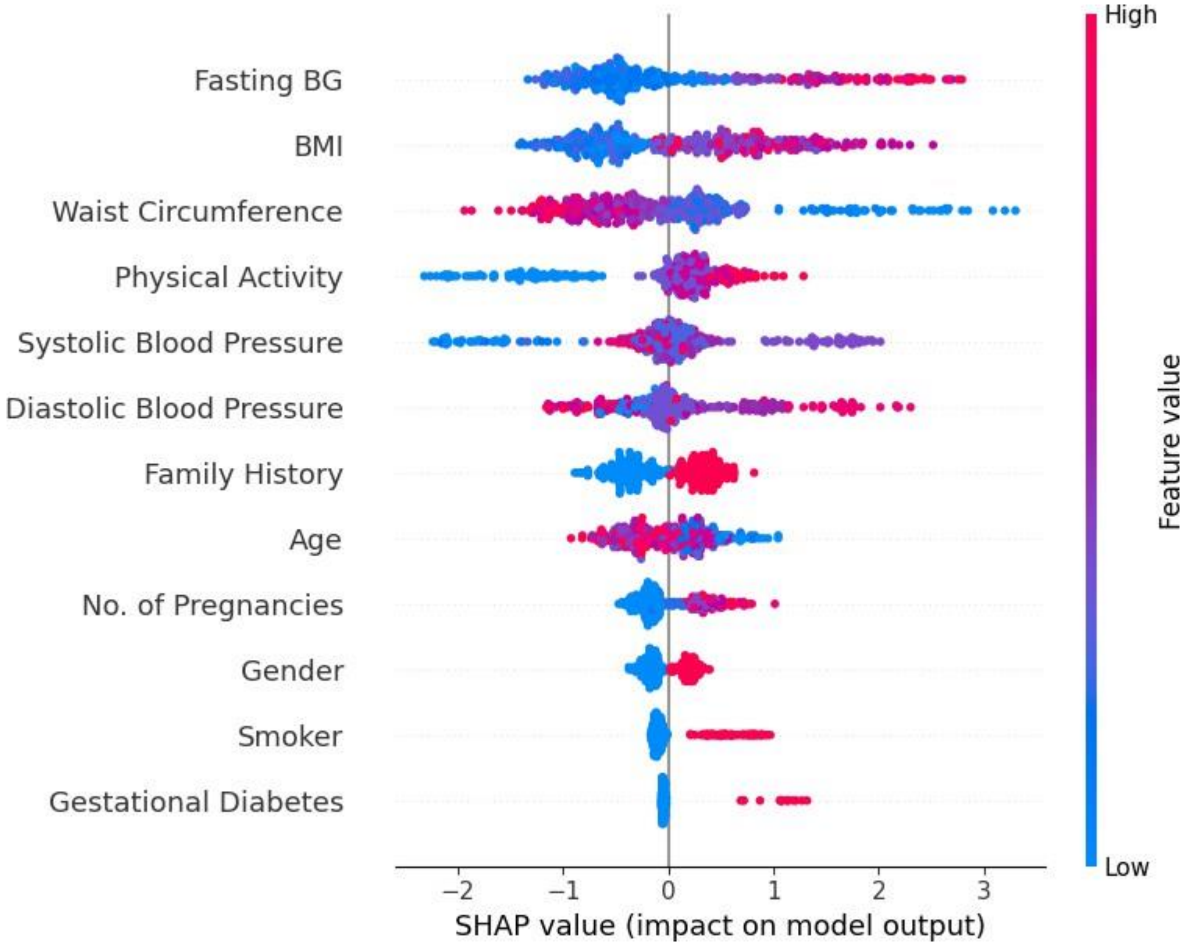


Figure 13: Feature importance using SHAP

The comparison of classifier performance with and without feature selection (FS) provides valuable insights into how dimensionality reduction and the elimination of irrelevant features affect model outcomes. The charts illustrate the impact of FS on multiple metrics—accuracy, AUC, precision, recall, and F1-score—across a range of algorithms, including Logistic Regression (LR), K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting (GBoost), XGBoost, LightGBM, and AdaBoost.

Without feature selection, the performance of most classifiers remains reasonably high but exhibits noticeable variability. The presence of irrelevant or redundant features in the dataset likely introduces noise, which can affect the classifiers' ability to make accurate predictions. For Logistic Regression, this effect is particularly pronounced. As seen in the chart without FS, LR achieves relatively low accuracy (around 0.71) and suffers in precision, recall, and F1-score as well. This suggests that the linear nature of LR makes it sensitive to noisy or irrelevant features, reducing its ability to model the relationships in the data effectively. When feature selection is applied, LR's performance improves modestly, with accuracy rising to approximately 0.73 and slight gains in other metrics. While still underperforming compared to other algorithms, these improvements highlight the value of feature selection in simplifying the dataset and reducing noise.

K-Nearest Neighbors (KNN) shows a similar trend, although its performance without FS is better than that of LR. In the absence of FS, KNN achieves accuracy close to 0.78 and reasonably balanced values for precision, recall, and F1-score. However, the inclusion of irrelevant features likely affects its distance-based calculations, leading to suboptimal classification boundaries. After applying FS, KNN exhibits noticeable improvements, with accuracy increasing to 0.81 and other metrics also showing slight gains. This indicates that feature selection helps KNN focus on the most informative dimensions, enhancing its ability to distinguish between classes effectively. Nevertheless, KNN still lags behind ensemble methods, highlighting its limitations in handling complex, high-dimensional datasets even with feature selection.

Random Forest (RF) stands out as one of the most robust classifiers both with and without FS. Without FS, RF achieves high scores across all metrics, with accuracy around 0.90 and nearly identical values for precision, recall, and F1-score. The inherent ability of RF to handle high-dimensional data and ignore irrelevant features through its ensemble of decision trees contributes to its strong performance. However, after applying FS, RF's metrics improve further, albeit marginally. The reduction in dimensionality likely helps RF focus on the most critical features, slightly enhancing its generalization ability and reducing the risk of overfitting. This demonstrates that while RF can handle noisy datasets effectively, feature selection still provides an additional performance boost.

Support Vector Machines (SVM) exhibit similar robustness. Without FS, SVM achieves accuracy, and other metric values close to 0.90, reflecting its strength in finding optimal decision boundaries. However, like RF, SVM benefits from FS, with marginal improvements observed across all metrics. Feature selection likely simplifies the feature space, reducing the computational complexity and enhancing SVM's ability to separate classes effectively. The consistent performance of SVM, both with and without FS, underscores its reliability for datasets with well-defined margins between classes.

Gradient Boosting (GBoost) shows strong performance without FS, achieving accuracy close to 0.92 and high scores for other metrics, particularly AUC. This indicates that GBoost's iterative approach of minimizing errors allows it to handle noisy features effectively. However, the application of FS leads to noticeable improvements, with AUC increasing to 0.95 and other metrics also showing slight gains. The iterative refinement process of GBoost seems to benefit from the reduced dimensionality, allowing the model to focus more effectively on the most informative features. These results highlight the synergy between feature selection and boosting algorithms, where the elimination of irrelevant features enhances the algorithm's ability to capture complex patterns in the data.

XGBoost, a variant of Gradient Boosting, exhibits similar trends. Without FS, XGBoost already outperforms many other classifiers, with accuracy close to 0.93 and consistently high values for precision, recall, and F1-score. The additional regularization techniques in XGBoost likely contribute to its ability to handle noisy data effectively. However, after applying FS, XGBoost achieves even better results, with AUC increasing to 0.96 and slight gains observed in other metrics. These improvements further emphasize the importance of feature selection in enhancing model performance, even for algorithms as robust as XGBoost.

LightGBM emerges as the best-performing classifier both with and without FS. Without FS, LightGBM achieves accuracy around 0.94 and consistently high values across all other metrics, reflecting its ability to handle high-dimensional data efficiently. After applying FS, LightGBM's performance improves further, with AUC reaching 0.97 and precision increasing to 0.93. LightGBM's leaf-wise growth strategy and ability to handle large datasets make it highly effective, and the application of FS amplifies these strengths by reducing noise and focusing on the most relevant features. These results highlight LightGBM’s superiority in handling both raw and preprocessed datasets.

AdaBoost also performs well, though its results are slightly below those of LightGBM and XGBoost. Without FS, AdaBoost achieves accuracy around 0.89 and high scores for precision, recall, and F1-score. However, it appears slightly more sensitive to irrelevant features compared to other ensemble methods, as indicated by minor inconsistencies in its metrics. After applying FS, AdaBoost shows modest improvements, with accuracy increasing to 0.91 and slight gains observed in other metrics. While not as robust as LightGBM or XGBoost, AdaBoost remains a strong contender, particularly for datasets where interpretability and simplicity are important considerations.

Overall, the comparison highlights the critical role of feature selection in improving classifier performance. For simpler models like Logistic Regression and KNN, FS significantly enhances performance by reducing noise and dimensionality, allowing these models to focus on the most relevant features. For more advanced algorithms like Random Forest, Gradient Boosting, and LightGBM, the impact of FS is less pronounced but still noticeable, as it further refines their ability to generalize and reduces computational complexity.

The results also underscore the superiority of ensemble-based methods, particularly LightGBM, XGBoost, and Gradient Boosting. These algorithms consistently outperform traditional methods like LR, KNN, and SVM, both with and without FS, demonstrating their ability to handle complex patterns and interactions within the data. Feature selection amplifies these advantages by eliminating irrelevant features, allowing the models to focus on the most important predictors and improving their overall efficiency.

In conclusion, while feature selection benefits all classifiers to varying degrees, its impact is most pronounced for simpler models that struggle with high-dimensional datasets. For robust ensemble methods, FS provides a slight but meaningful improvement, further enhancing their already strong performance. LightGBM emerges as the most effective classifier in both scenarios, making it the preferred choice for handling complex datasets with or without preprocessing. These findings underscore the importance of both feature selection and model choice in achieving optimal performance, highlighting the need to tailor preprocessing and algorithm selection to the specific characteristics of the dataset and the goals of the application.

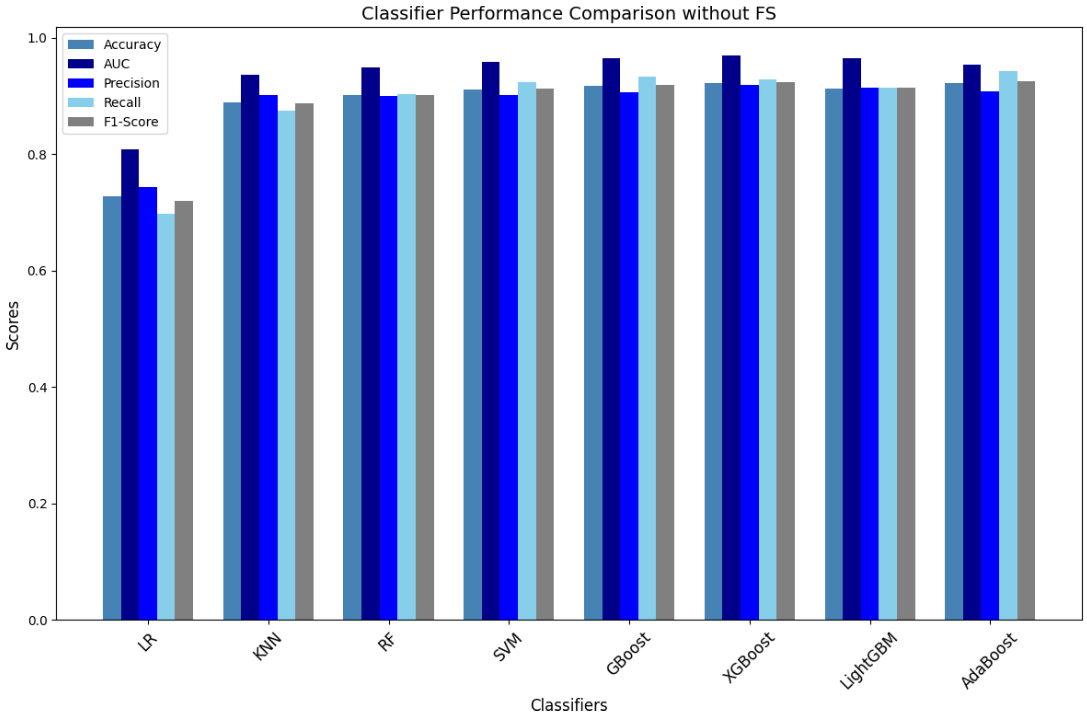


Figure 14: Performance of classifier without Feature selection

The implementation of the best-performing model, XGBoost, on a mobile application involves several steps, including model training, optimization, conversion, integration into the app, and its role in delivering real-time predictions for the users. The following details describe the implementation process as showcased in the attached app interface: XGBoost, known for its efficiency and high accuracy, was selected as the best-performing model based on the comparative analysis of classifiers. The training process began with preprocessing the dataset, ensuring that irrelevant features were eliminated through feature selection techniques. This reduced dimensionality, improved computational efficiency, and enhanced model performance. The XGBoost algorithm was trained using hyperparameter tuning to optimize parameters such as the learning rate, maximum depth of trees, and the number of boosting rounds. Cross-validation techniques were applied to avoid overfitting and to ensure robust generalization.

Once the XGBoost model achieved optimal performance, it was converted into a lightweight format suitable for deployment on mobile devices. This involved exporting the trained model into formats like ONNX (Open Neural Network Exchange) or CoreML for compatibility with various platforms. Using libraries such as TensorFlow Lite or ML Kit, the model was further quantized to reduce its size and computational requirements, ensuring efficient operation on resource-constrained devices.

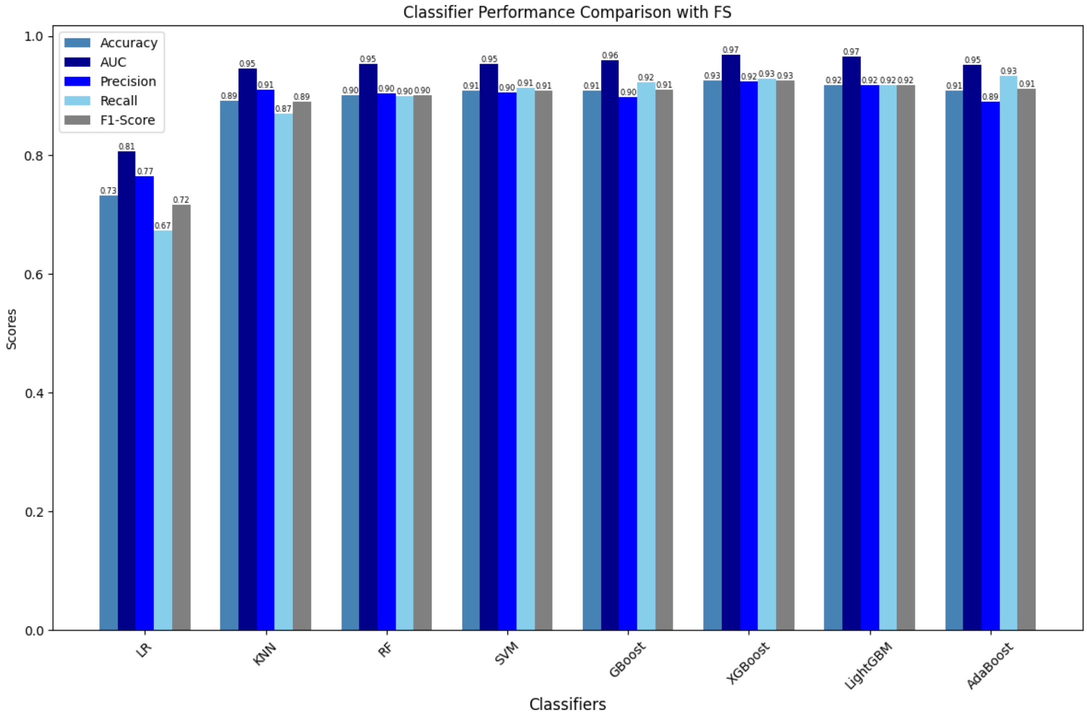


Figure 15: Performance of classifier with Feature selection

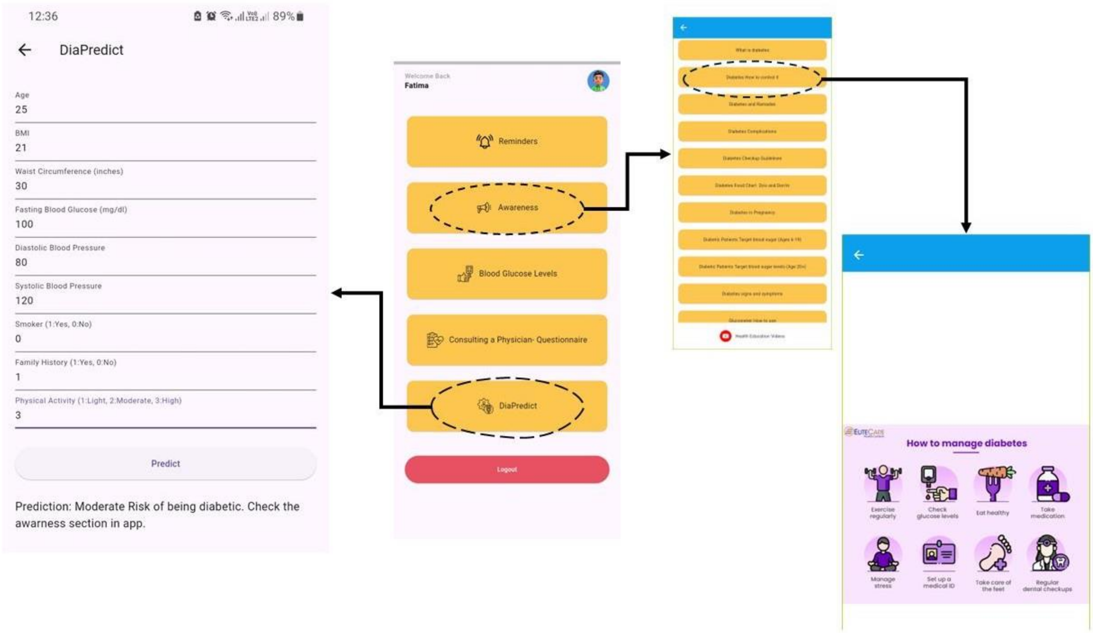
The mobile application was designed to incorporate the XGBoost model as a core component of the "DiaPredict" feature, which predicts the risk of diabetes based on user inputs. The app interface allows users to input key parameters such as age, BMI, waist circumference, fasting blood glucose, blood pressure, smoking status, family history, and physical activity levels. These inputs are preprocessed locally on the device to ensure they are in the same format as the training data.

The app then uses the embedded XGBoost model to generate predictions in real-time. Based on the inputs, the model evaluates the risk of diabetes, classifying it into categories such as low, moderate, or high risk. The prediction is displayed instantly on the app, providing users with actionable insights.

The user interface of the app is intuitive, guiding users seamlessly through the process of risk prediction. After obtaining the prediction, users are directed to additional resources through the app's "Awareness" section. This section provides educational content on managing diabetes, including guidelines on exercise, healthy eating, stress management, and regular check-ups.

The app also features other functionalities, such as reminders, blood glucose level tracking, and physician consultation questionnaires. These modules complement the predictive capabilities of XGBoost by promoting awareness, enabling self-management, and offering access to professional advice.

To enhance performance, sensitive computations are conducted on the device itself, reducing the need for constant internet connectivity and improving response times. Additionally, data security and user privacy are prioritized. Inputs provided by users are processed locally and not transmitted to external servers, ensuring compliance with data protection regulations such as GDPR and HIPAA.



12:36

DiaPredict

Age
25

BMI
21

Waist Circumference (inches)
30

Fasting Blood Glucose (mg/dl)
100

Diastolic Blood Pressure
80

Systolic Blood Pressure
120

Smoker (1:Yes, 0:No)
0

Family History (1:Yes, 0:No)
1

Physical Activity (1:Light, 2:Moderate, 3:High)
3

Predict

Prediction: Moderate Risk of being diabetic. Check the awarness section in app.

Figure 16: Layout of Mobile application

The integration of XGBoost into the mobile app lays the foundation for scalability. In the future, the app can be extended to include other health-related predictions, such as cardiovascular risk or chronic disease monitoring. The modular nature of the implementation ensures that new models can be integrated without significant changes to the existing infrastructure.

The inclusion of a feedback loop can also be considered, where user outcomes are periodically collected (with consent) to further refine and retrain the XGBoost model. This continuous learning approach would improve the model’s accuracy and relevance over time, enhancing its utility for diverse user populations.

The implementation of the XGBoost model on the mobile application represents a significant step in leveraging advanced machine learning techniques for personalized health management. By integrating a powerful predictive model into an accessible and user-friendly app interface, the solution empowers individuals to monitor and manage their diabetes risk effectively. The combination of real-time predictions, educational resources, and additional health management tools ensures a comprehensive approach to promoting awareness and preventive care.

CHAPTER 6

CONCLUSION AND FUTURE WORK

The study conducted aimed to evaluate and optimize the performance of machine learning classifiers for predicting diabetes risk, both with and without feature selection techniques. The research demonstrated that feature selection significantly enhanced the performance metrics of most classifiers by eliminating irrelevant and redundant features, thus improving computational efficiency and accuracy. Among the classifiers, XGBoost consistently outperformed others in terms of accuracy, precision, recall, and F1-score, showcasing its robustness and adaptability to the dataset's characteristics. The statistical analysis, including Tukey HSD tests, further validated the superiority of certain classifiers over others, providing statistical significance to the observed differences in performance.

The deployment of the best-performing model, XGBoost, in a mobile application represents a practical application of this research. The app integrates predictive capabilities with educational resources, empowering users to assess their diabetes risk and take preventive measures. This approach underscores the potential of machine learning to make significant contributions in the healthcare domain, particularly in enhancing early diagnosis and public awareness.

Despite the promising results, several areas warrant further investigation. First, the dataset used for this study could be expanded to include more diverse demographic and clinical data, ensuring the model’s generalizability across different populations. Second, incorporating advanced techniques such as deep learning could be explored to capture complex patterns within the data. While XGBoost performed exceptionally, neural networks may provide additional insights, particularly with larger datasets.

The integration of real-time data collection into the mobile application, such as wearable devices for continuous glucose monitoring or physical activity tracking, could enhance the app’s predictive capabilities and user engagement. Moreover, incorporating user feedback mechanisms into the application could refine its usability and functionality over time. Finally, ethical considerations such as data privacy and fairness should be further emphasized, ensuring the deployed system adheres to stringent standards while maintaining transparency and user trust.

By addressing these areas, this research can be extended to provide a more comprehensive, accurate, and user-friendly solution for diabetes risk prediction and management, ultimately contributing to better healthcare outcomes and preventive measures at a broader scale.

6.1 Challenges in Diabetes Prediction

Despite progress, several challenges persist:

- **Imbalanced Datasets**: Most datasets have fewer diabetic cases, which skews model performance. Techniques like SMOTE are often employed to address this issue.
- **Data Quality**: Missing or inconsistent data remains a challenge in healthcare datasets.
- **Feature Selection**: Redundant or irrelevant features can reduce model efficiency.
- **Privacy Concerns**: Strict regulations like GDPR complicate data sharing.

6.2 Future Directions

Future research should address the following areas:

- **Real-Time Prediction**: Integrating wearable devices to provide real-time diabetes risk predictions.
- **Personalized Models**: Tailoring models to individual risk factors and genetic predispositions.
- **Federated Learning**: Allowing model training across decentralized datasets without compromising data privacy.
- **Explainable AI**: Developing interpretable models to gain clinician trust and regulatory approval.

REFERENCES

[1] “What Is Diabetes | International Federation of Diabetes.” Accessed: Dec. 07, 2024. [Online]. Available: https://idf.org/about-diabetes/what-is-diabetes/

[2] “IDF Diabetes Atlas.”

[3] “Global diabetes data report 2000 — 2045.” Accessed: Sep. 07, 2024. [Online]. Available: https://diabetesatlas.org/data/en/world/

[4] “Pakistan - International Diabetes Federation.” Accessed: Nov. 05, 2024. [Online]. Available: https://idf.org/our-network/regions-and-members/middle-east-and-north-africa/members/pakistan/

[5] D. Tomic, J. E. Shaw, and D. J. Magliano, “The burden and risks of emerging complications of diabetes mellitus,” Sep. 01, 2022, Nature Research. doi: 10.1038/s41574-022-00690-7.

[6] S. M. nimmagadda, G. Suryanarayana, G. B. Kumar, G. Anudeep, and G. V. Sai, “A Comprehensive Survey on Diabetes Type-2 (T2D) Forecast Using Machine Learning,” Jul. 01, 2024, Springer Science and Business Media B.V. doi: 10.1007/s11831-023-10061-8.

[7] Y. W. Cheng and A. B. Caughey, “Gestational diabetes: diagnosis and management,” Journal of Perinatology, vol. 28, no. 10, pp. 657–664, 2008, doi: 10.1038/jp.2008.62.

[8] “About Gestational Diabetes | International Diabetes Federation.” Accessed: Sep. 10, 2024. [Online]. Available: https://idf.org/about-diabetes/types-of-diabetes/gestational-diabetes/

[9] V. K. Katiyar, “Regulation of blood glucose level in diabetes mellitus using palatable diet composition,” 2003.

[10] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, “Predictive models for diabetes mellitus using machine learning techniques,” BMC Endocr Disord, vol. 19, no. 1, Oct. 2019, doi: 10.1186/s12902-019-0436-6.

[11] C. C. Olisah, L. Smith, and M. Smith, “Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective,”

Comput Methods Programs Biomed, vol. 220, Jun. 2022, doi: 10.1016/j.cmpb.2022.106773.

[12] C. Carvalho Gontijo et al., “PIMA: A population informative multiplex for the Americas,” Forensic Sci Int Genet, vol. 44, Jan. 2020, doi: 10.1016/j.fsigen.2019.102200.

[13] P. Kaur and M. Sharma, “Analysis of Data Mining and Soft Computing Techniques in Prospecting Diabetes Disorder In Human Beings: A Review,” Article in International Journal of Pharmaceutical Sciences and Research, vol. 9, no. 7, p. 11, 2018, doi: 10.13040/IJPSR.0975-8232.9(7).2700-19.

[14] H. Kaur and V. Kumari, “Predictive modelling and analytics for diabetes using a machine learning approach,” Applied Computing and Informatics, vol. 18, no. 1–2, pp. 90–100, Jan. 2022, doi: 10.1016/j.aci.2018.12.004.

[15] F. G. Woldemichael and S. Menaria, “Prediction of Diabetes Using Data Mining Techniques,” in 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), 2018, pp. 414–418. doi: 10.1109/ICOEI.2018.8553959.

[16] D. Tomic, J. E. Shaw, and D. J. Magliano, “The burden and risks of emerging complications of diabetes mellitus,” Sep. 01, 2022, Nature Research. doi: 10.1038/s41574-022-00690-7.

[17] H. El-Sofany, S. A. El-Seoud, O. H. Karam, Y. M. Abd El-Latif, and I. A. T. F. Taj-Eddin, “A Proposed Technique Using Machine Learning for the Prediction of Diabetes Disease through a Mobile App,” International Journal of Intelligent Systems, vol. 2024, 2024, doi: 10.1155/2024/6688934.

[18] “Pima Indians Diabetes Database.” Accessed: Jan. 02, 2025. [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[19] “Diabetes 130-US Hospitals for Years 1999-2008 - UCI Machine Learning Repository.” Accessed: Jan. 02, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008

[20] "Early Stage Diabetes Risk Prediction - UCI Machine Learning Repository." Accessed: Jan. 02, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset

[21] "CDC Diabetes Health Indicators - UCI Machine Learning Repository." Accessed: Jan. 02, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

[22] M. Abedini, A. Bijari, and T. Banirostam, "Classification of Pima Indian Diabetes Dataset using Ensemble of Decision Tree, Logistic Regression and Neural Network," IJARCCE, vol. 9, no. 7, pp. 1–4, Jul. 2020, doi: 10.17148/ijarcce.2020.9701.

[23] G. Wang et al., "Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial," Nat Med, vol. 29, no. 10, pp. 2633–2642, Oct. 2023, doi: 10.1038/s41591-023-02552-9.

[24] T. Mahboob Alam et al., "A model for early prediction of diabetes," Inform Med Unlocked, vol. 16, Jan. 2019, doi: 10.1016/j.imu.2019.100204.

[25] N. N. N. Nazirun et al., "Prediction Models for Type 2 Diabetes Progression: A Systematic Review," IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3432118.

[26] P. Hounguè and A. G. Bigirimana, "Leveraging Pima Dataset to Diabetes Prediction: Case Study of Deep Neural Network," Journal of Computer and Communications, vol. 10, no. 11, pp. 15–28, 2022, doi: 10.4236/jcc.2022.1011002.

[27] S. A. Alex, J. J. V. Nayahi, H. Shine, and V. Gopirekha, "Deep convolutional neural network for diabetes mellitus prediction," Neural Comput Appl, vol. 34, no. 2, pp. 1319–1327, Jan. 2022, doi: 10.1007/s00521-021-06431-7.

[28] M. Zarar and Y. Wang, "Early Stage Diabetes Prediction by Approach Using Machine Learning Techniques," 2023, doi: 10.21203/rs.3.rs-3145599/v1.

[29] H. N. A. Pham and E. Triantaphyllou, “Prediction of Diabetes by Employing a New Data Mining Approach Which Balances Fitting and Generalization,” Computer and Information Science, 2008, doi: 10.1007/978-3-540-79187-4_2.

[30] T. Helmy and Z. Rasheed, “Multi-category bioinformatics dataset classification using extreme learning machine,” 2009 IEEE Congress on Evolutionary Computation, 2009, doi: 10.1109/CEC.2009.4983354.

[31] S. Lekkas and L. Mikhailov, “Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases,” Artif Intell Med, vol. 50, no. 2, pp. 117–126, Oct. 2010, doi: 10.1016/j.artmed.2010.05.007.

[32] A. A. Al Jarullah, “Decision tree discovery for the diagnosis of type II diabetes,” IIT 2011, 2011, doi: 10.1109/INNOVATIONS.2011.5893838.

[33] S. Karatsiolis and C. Schizas, “Region based Support Vector Machine algorithm for medical diagnosis on Pima Indian Diabetes dataset,” 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), 2012, doi: 10.1109/BIBE.2012.6399663.

[34] B. M. Hussan, “Data Mining based Prediction of Medical data Using K-means algorithm.”

[35] “Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4.5.” [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[36] M. S. Barale and D. T. Shirke, “Cascaded Modeling for PIMA Indian Diabetes Data,” Int J Comput Appl, vol. 139, no. 11, pp. 1–4, Apr. 2016, doi: 10.5120/ijca2016909426.

[37] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, “Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset,” in 2017 International Conference on Computing Networking and Informatics (ICCNI), 2017. doi: 10.1109/ICCNI.2017.8123815.

[38] H. Balaji, N. Ch. S. N. Iyengar, and R. D. Caytiles, “Optimal Predictive analytics of Pima Diabetics using Deep Learning,” International Journal of

Database Theory and Application, vol. 10, no. 9, pp. 47–62, Sep. 2017, doi: 10.14257/ijdta.2017.10.9.05.

[39] S. Raghavendra, J. S. Kumar, and R. B. K, “EVALUATING THE PERFORMANCE OF NEURAL NETWORK USING FEATURE SELECTION METHODS ON PIMA INDIAN DIABETES DATASET,” J Emerg Technol Innov Res, 2018.

[40] M. F. Dzulkalnine and R. Sallehuddin, “Missing data imputation with fuzzy feature selection for diabetes dataset,” SN Appl Sci, vol. 1, no. 4, Apr. 2019, doi: 10.1007/s42452-019-0383-x.

[41] H. Naz and S. Ahuja, “Deep learning approach for diabetes prediction using PIMA Indian dataset,” J Diabetes Metab Disord, vol. 19, no. 1, pp. 391–403, Jun. 2020, doi: 10.1007/s40200-020-00520-5.

[42] G. S. Tripathi and R. Kumar, “Early Prediction of Diabetes Mellitus Using Machine Learning,” 2020, doi: 10.1109/ICRITO48877.2020.9197832.

[43] M. Abedini, A. Bijari, and T. Banirostam, “Classification of Pima Indian Diabetes Dataset using Ensemble of Decision Tree, Logistic Regression and Neural Network,” International Journal of Advanced Research in Computer and Communication Engineering, 2020, doi: 10.17148/IJARCCE.2020.9701.

[44] M. M. Shabtari, V. Kumar Shukla, H. Singh, and I. Nanda, “Analyzing PIMA Indian Diabetes Dataset through Data Mining Tool ‘RapidMiner,’” in 2021 International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2021, Institute of Electrical and Electronics Engineers Inc., Mar. 2021, pp. 560–574. doi: 10.1109/ICACITE51222.2021.9404741.

[45] H. Naz and S. Ahuja, “SMOTE-SMO-based expert system for type II diabetes detection using PIMA dataset,” Int J Diabetes Dev Ctries, vol. 42, no. 2, pp. 245–253, Apr. 2022, doi: 10.1007/s13410-021-00969-x.

[46] H. Naz, R. Nijhawan, and N. J. Ahuja, “DT-DL Based Hybrid Approach for Early Detection of Diabetes Using PIMA Dataset,” in 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and

Future Directions), ICRITO 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICRITO56286.2022.9964904.

[47] M. S. Reza, U. Hafsha, R. Amin, R. Yasmin, and S. Ruhi, “Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset,” Computer Methods and Programs in Biomedicine Update, vol. 4, Jan. 2023, doi: 10.1016/j.cmpbup.2023.100118.

[48] K. K. Patro et al., “An effective correlation-based data modeling framework for automatic diabetes prediction using machine and deep learning techniques,” BMC Bioinformatics, vol. 24, no. 1, Dec. 2023, doi: 10.1186/s12859-023-05488-6.

[49] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, “Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms,” Neural Comput Appl, vol. 35, no. 22, pp. 16157–16173, Aug. 2023, doi: 10.1007/s00521-022-07049-z.

[50] C. Ganesh Babu, M. Gowri Shankar, G. S. Priyanka, and B. Vidhya, “Performance analysis of classifiers in detection of diabetes from Pima Indians database,” in AIP Conference Proceedings, American Institute of Physics Inc., Apr. 2023. doi: 10.1063/5.0125230.

[51] M. Zarar and Y. Wang, “Early Stage Diabetes Prediction by Approach Using Machine Learning Techniques,” Jul. 12, 2023. doi: 10.21203/rs.3.rs-3145599/v1.

[52] A. Perdana, A. Hermawan, and D. Avianto, “Analyze Important Features of PIMA Indian Database For Diabetes Prediction Using KNN,” Jurnal Sisfokom (Sistem Informasi dan Komputer), vol. 12, no. 1, pp. 70–75, Mar. 2023, doi: 10.32736/sisfokom.v12i1.1598.

[53] M. S. Reza, R. Amin, R. Yasmin, W. Kulsum, and S. Ruhi, “Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data,” Heliyon, vol. 10, no. 2, Jan. 2024, doi: 10.1016/j.heliyon.2024.e24536.

[54] A. Pyne and B. Chakraborty, “Artificial Neural Network based approach to Diabetes Prediction using Pima Indians Diabetes Dataset,” in 2023 International Conference on Control, Automation and Diagnosis, ICCAD 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICCAD57653.2023.10152382.

[55] A. F. Ashour, M. M. Fouda, Z. Md, F. ‡3, and M. I. Ibrahem, “Optimized Neural Networks for Diabetes Classification Using Pima Indians Diabetes Database.”

[56] T. Alghamdi, “Prediction of Diabetes Complications Using Computational Intelligence Techniques,” Applied Sciences (Switzerland), vol. 13, no. 5, Mar. 2023, doi: 10.3390/app13053030.

[57] F. Nassiwa and J. Zeng, “Evaluating Traditional Machine Learning Models for Predicting Diabetes Onset Using the Pima Indians Dataset.” [Online]. Available: https://ssrn.com/abstract=4878052

[58] V. Chang, M. A. Ganatra, K. Hall, L. Golightly, and Q. A. Xu, “An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators,” Healthcare Analytics, vol. 2, Nov. 2022, doi: 10.1016/j.health.2022.100118.

[59] B. Amma N.G., “En-RfRsK: An ensemble machine learning technique for prognostication of diabetes mellitus,” Egyptian Informatics Journal, vol. 25, Mar. 2024, doi: 10.1016/j.eij.2024.100441.