A WEB-BASED DIABETES RISK PREDICTION SYSTEM USING MACHINE LEARNING FOR ENUGU STATE, NIGERIA.

BY

NWAOHA PRECIOUS MMASICHUKWU
GOU/U22/CSC/926

DEPARTMENT OF COMPUTER SCIENCE AND MATHEMATICS, FACULTY OF NATURAL SCIENCES AND ENVIRONMENTAL STUDIES,
GODFREY OKOYE UNIVERSITY, ENUGU STATE. **IN PARTIAL FULFILLMENT OF THE AWARD OF BACHELOR OF SCIENCE (B.SC), DEGREE IN COMPUTER SCIENCE.**

SUPERVISOR: DR LUCY I. EZIGBO.

JUNE, 2026.

APPROVAL PAGE

This research, titled **A Web-Based Diabetes Risk Prediction System Using Machine Learning for Enugu State, Nigeria**, has been assessed and approved by the Department of Computer Science and Mathematics Committees in the Faculty of Natural Sciences and Environmental Studies, Godfrey Okoye University, Enugu, Nigeria.

DR LUCY I. EZIGBO **SUPERVISOR**	**DATE**
DR. FRANK OKEBANAMA **HOD, DEPARTMENT OF COMPUTER SCIENCE AND MATHEMATICS**	**DATE**
PROF. I. A. AJAH **DEAN, FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY.**	**DATE**
EXTERNAL EXAMINER	**DATE**

CERTIFICATION

I certify that I completed the research included therein and that it was primarily based on my observations and investigation. Consequently, I will fully accept the University's decision if I am found to have cheated on the assignment.

NWAOHA PRECIOUS MMASICHUKWU
GOU/U22/CSC/926.

DEDICATION

This project is dedicated to the women and girls around the world who are denied the chance to learn, dream, and be heard. May this stand as a reminder that your voices matter and your stories deserve to exist in every room they were told they could never enter. I see you.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisor, **Dr. Lucy I. Ezigbo**, whose guidance, patience, and invaluable expertise shaped this project from inception to completion. Your mentorship has been a light throughout this journey, and I am truly grateful for every piece of advice, correction, and encouragement you offered.

My sincere appreciation also goes to **Dr. Frank Okebanama**, Head of the Department of Computer Science and Mathematics, Godfrey Okoye University, for his leadership and for fostering an academic environment that encouraged growth and excellence. To all the dedicated lecturers of the Department of Computer Science and Mathematics, your teachings laid the foundation upon which this work was built. I am grateful for your time, knowledge, and commitment to shaping the next generation.

To my mother, **Mrs. Angela Nwaoha**, the strongest woman I know. To the woman whose sacrifices were quiet but endless, whose love carried me even on the days I could not carry myself. Everything I am becoming has your fingerprints on it. I hope one day I can return even a fraction of the grace, strength, and softness you have given me. I pray this work keeps a smile on your face and pride in your heart.

To **Mrs. Okwurumgbe**, my second mother. Thank you for opening your doors to me, for creating a space where I felt safe enough to open up, and for taking me in as a daughter. Your warmth and kindness gave me a sense of belonging that carried me through some of my hardest days, and I am deeply grateful.

To **Miss Doris**, my guardian, who became a friend. Thank you for the words of encouragement you poured into me, often at the moments I needed them most. Your belief in me meant more than you may ever know, and I carry it with me as I step into the next chapter.

To my dear friends, Victory, Khloe, Tessa, Anita, Christabel, Daniel, Pete, Angel, Sopulu, Akay, Febe, and Dim Dim, thank you for the laughter, the late nights, the pep talks, and for making this season of life one I will always look back on with warmth. You were my people through it all.

To my sisters, Lilian, Chisom, Uzoma, Chigozie, Mmesoma, Abigail, and Jessica, you are my heart walking outside my body.

To every single one of you, my mother, my second mother, my guardian, my friends, my sisters, the love I have for you will transcend space and time, and even when I leave this earth, may it echo in everything I leave behind. Thank you for being home.

To everyone who believed in me before I believed in myself — this is for you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The problem of diabetes mellitus continues to be a public health problem in Enugu State, Nigeria, because late diagnosis is common in the region due to structural factors such as lack of laboratory facilities, expensive testing, and non-availability of self-service risk screening instruments, with a lot of cases being diagnosed at later stages when complications have already set in. This research project, therefore, aimed to create a web-based diabetes risk predictor that allows people in Enugu State to determine their individual risk of having diabetes using health and lifestyle data that do not require laboratory tests and a visit to a hospital. Logistic regression, decision tree, support vector machine, and random forest, four supervised machine learning classifiers, were used in this work and tested against each other using Pima Indians Diabetes Database (768 instances) as the main training data and 40 patient records (anonymized) obtained from UNTH, Ituku-Ozalla, and Parklane as the supplementary training data after preprocessing which involved median imputation for missing clinical data, z-score normalization, and stratified 80/20 training/testing partitioning.

The model was implemented in a Flask based web app where users enter seven parameters (age, weight, height, blood pressure, family history, physical activity level and diet type), which goes through the rule-based feature estimation process to determine the clinical variables required by the model, which then categorizes the probability score as Low, Medium or High risk levels with corresponding health suggestions and a medical disclaimer. Random Forest Classifier performed the best among the models used, with an accuracy score of 79.87%, a precision of 76.47%, and an F1-Score of 68.75%. This model was chosen for deployment as a result of the above scores. All eight functional test cases worked perfectly, while the layout appeared properly on both desktop and mobile versions. These results provide proof of concept that shows that a machine learning-enabled diabetes testing web app can be developed without the use of a lab as a preventive awareness solution for people in Enugu State. It is recommended that the model be retrained using more local data in Nigeria in the future.

CHAPTER ONE

INTRODUCTION

1.1 Background to the Study

Diabetes mellitus is a chronic condition characterized by metabolic disorders that has become a serious public health problem of the twenty-first century. According to data collected by the International Diabetes Federation, the number of people with diabetes has increased from 108 million in 1980 to more than 537 million in 2021, with a projected figure of 783 million in 2045. The rapid increase in diabetes cases puts serious pressure on the global healthcare sector. Nevertheless, the effects of this health problem are especially devastating for low- and middle-income countries, where healthcare infrastructure is still under development and cannot provide patients with all necessary care.

Nigeria, being the most populous nation in Africa with a population of over 200 million citizens, has experienced a considerable growth in the number of people with diabetes. Prevalence rates of diabetes in Nigeria vary depending on the region, from 2% to 6%, with urban areas demonstrating higher rates compared to rural communities. This difference may be explained by the impact of urbanization on people's health due to changes in people's diet and occupation: traditional nutrition with whole foods and vegetables is substituted by highly processed and high-calorie food, while agricultural jobs turn into desk-based ones. Enugu State is a good example of such a process.

Another key feature of the diabetes epidemic in Nigeria is the high percentage of undiagnosed cases. Indeed, empirical data reveal that about fifty percent of patients with diabetes in Nigeria do not know that they have diabetes. The reasons for the undiagnosed cases are rooted in the slow nature of the development of Type 2 diabetes. People may assume that they suffer from some other diseases due to minor symptoms like exhaustion or excessive thirst. Therefore, people will go to a doctor when the disease develops to the point where it causes such complications as nephropathy, vision impairment, cardiovascular

complications, and diabetic foot ulcers. At this stage, treatment will be difficult, expensive, and damaging for the patient's well-being.

However, the importance of detecting diabetes at an early stage lies in the fact that this disease can be managed and prevented. Through changes in lifestyle (better nutrition, physical exercises, etc.), people can avoid developing diabetes if they have been suffering from pre-diabetes. Similarly, the chances for proper treatment and recovery are much higher when a person gets to the doctor on time. Nevertheless, there is only a possibility for this to happen when the screening is done prior to any complications. However, in Enugu State, such possibilities are significantly limited by a number of issues. First, laboratory screening tests like fasting plasma glucose and HbA1c can be conducted only in cities. Second, the price of the tests is too high for ordinary families. Third, the lack of self-screening tools prevents people from evaluating their condition without a visit to a healthcare institution.

Recent advancements in machine learning and the widespread use of internet-connected devices create the possibility of addressing these problems. Indeed, machine learning algorithms can analyze combinations of different non-laboratory health indicators (age, body mass index, blood pressure, genetic predisposition, and others) to give accurate predictions of diabetes risk for a certain individual. If this technology is used in a web-based application, the algorithm can provide people with a rapid, free service of estimating the risk without the need for any lab equipment or a visit to the doctor. Thus, this project promotes creating a similar web-based predictive algorithm for the Enugu State residents as an instrument of preventive awareness.

1.2 Statement of the Problem

Type 2 diabetes has become a growing health problem in Enugu State, with the incidence of the disease becoming more pronounced each year. A significant number of people in the population suffer from diabetes; however, a significant percentage of those individuals go

undiagnosed until they develop more severe and expensive complications from diabetes. The delay in the diagnosis of diabetes is due mainly to structural obstacles to making regular, proactive screenings available to everyone in the community.

In its current state, the screening of type 2 diabetes follows a clinical and reactive process. People who want to get screened for diabetes will have to go to a hospital or clinic, get tested through a blood test performed by healthcare workers, and then get their results. Such a screening system does not offer any built-in conveniences that encourage proactive risk assessment on the part of individuals. People usually only get tested for diabetes after symptoms have appeared. The obstacles to earlier testing include the following:

- Poor accessibility of diagnostic facilities, more so in rural and peri-urban areas, where hospitals or clinics with properly functioning laboratory facilities are few and far between.
- Expensive tests that discourage people without any apparent signs from coming forward to undergo screening tests, especially those in poor families.
- The distance to the health facility means that one will have to incur expenses to travel, especially in the case of combining travel costs with the need to take time off work.
- Poor self-assessment instruments, as there is no online portal that allows users to input information about their known health history and receive an informal evaluation of their risk level without consulting a healthcare practitioner first.
- Ignorance about risk factors associated with diabetes and the need for preventive screenings means a lack of motivation for the said tests.

  All of these factors result in the identification of diabetes cases in Enugu State being always too late, at a point where it becomes very hard to reverse the damage that has been done by the condition. Thus, there is a pressing need for a simple yet effective self-testing tool for early risk assessment.

1.3 Aim and Objectives of the Study

1.3.1 Aim

This research aims to design a web-based system for predicting the risks of diabetes among individuals living in Enugu State, Nigeria, using machine learning, allowing people to evaluate their own risk for developing diabetes using freely available data, without needing lab tests or going to any medical facilities.

1.3.2 Objectives

To accomplish the set objectives, the following four specific objectives were formulated:

1. To conduct a literature review of the available studies concerning the risk factors of diabetes, the use of machine learning techniques in medicine, and previous diabetes prediction models to identify suitable approaches and pinpoint the research gap that will be tackled by this study.
2. To obtain, preprocess, and analyze an existing public diabetes dataset and apply different machine learning classification algorithms, which will be tested and compared to identify the most efficient algorithm for further implementation.
3. To develop a convenient web application for obtaining health and lifestyle-related inputs from the user, running the selected algorithm, predicting the user's risk of developing diabetes, and providing health advice along with a medical disclaimer.
4. To test and verify the developed system, as well as provide recommendations for future development of the study.

1.4 Significance of the Study

The relevance of the study lies in the fact that it attempts to meet a clear-cut and underserved public health requirement in Enugu State by providing an actual technological solution that is available, affordable, and operational without the involvement of medical professionals and laboratories. The relevance can be gauged in multiple ways.

From a public health point of view, the system is an effective measure for engaging people in the problem of diabetes risk before the manifestation of symptoms. This gives people a chance to change their way of life and make clinical consultations when necessary, which cannot be provided now in a self-service form via digital channels.

In terms of accessibility, the fact that the system is delivered over the web allows everyone with an internet-connected device to have access to it. Nowadays, it has become possible through smartphones.

In terms of resource efficiency, the system helps allocate healthcare resources to people who can benefit most from the follow-up care, instead of having to make the healthcare system screen everybody. People who are assessed as being high-risk through the system should seek professional advice, whereas low-risk people will be provided with reassurance and preventive information.

On the academic and technical side, the project presents an example of how machine learning and web development skills can be applied to address a particular health issue in one's local community.

1.5 Scope of the Study

This research seeks to investigate the design, development, and testing of a web-based diabetic risk predictor that will be used by the general adult population of Enugu State, Nigeria. The scope can be clearly articulated as:

- The health and lifestyle parameters of the individual are the inputs for the system. They include age, weight, height, blood pressure, history of diabetes in the family, physical activities, and dieting. No laboratory values, such as blood sugar level or HbA1C, are needed from the user.
- The system is a web-based one that will work in any modern web browser in desktop, tablets, and mobile devices.

- The system is designed for the purpose of screening only, and it is not a diagnostic nor replacement for the clinical examination.

1.6 Limitations of the Study

Nevertheless, there are several limitations of this research that need to be considered when evaluating the results of the study and its applications.

Firstly, the machine learning algorithm was trained using the Pima Indians Diabetes Database, which includes only the medical data of female patients who belong to the group of Pima Indians and have a genetic predisposition to Type 2 diabetes. Hence, the obtained prediction model cannot be generalized since the health indicators of this population are not identical to those of the Nigerian population. The database of the local population is needed to create an effective predictive model.

Secondly, all the data used for the prediction will come from the users themselves. The system does not include any verification procedures; therefore, the prediction accuracy will directly depend on the accuracy of user-provided values.

Finally, some of the input features needed for the trained machine learning algorithm, such as plasma glucose level, serum insulin, and triceps skinfold thickness, can be obtained only after conducting some tests in the lab.

This is accomplished by means of rule-based estimation algorithms used to estimate these values based on the information provided through the questionnaires. Estimation creates controlled approximation errors that cannot be completely avoided without conducting clinical tests.

Since this is an online system, it requires the existence of internet connectivity, something that is not available in the rural parts of Enugu State due to weak and unreliable connectivity.

Moreover, people who lack proficiency in using computers might find it challenging to use this system.

This is accomplished by means of rule-based estimation algorithms used to estimate these values based on the information provided through the questionnaires. Estimation creates controlled approximation errors that cannot be completely avoided without conducting clinical tests.

Since this is an online system, it requires the existence of internet connectivity, something that is not available in the rural parts of Enugu State due to weak and unreliable connectivity.

Moreover, people who lack proficiency in using computers might find it challenging to use this system.

1.7 Operational Definition of Terms

Diabetes Mellitus: A persistent metabolic ailment where the patient's body cannot control the blood glucose levels because of a deficiency in insulin production or malfunction of the insulin process. This research project will be centered on Type 2 diabetes, which is the common form and is greatly linked to obesity, sedentary lifestyles, and genetic inheritance.

Machine Learning: A field of Artificial Intelligence where computer systems use a pattern learned from data to make predictions or decisions on future inputs without having to be programmed explicitly for every possibility.

Risk Prediction: Using mathematical/statistical methods or computational methods to determine the probability of an individual developing certain diseases/conditions using a number of identified risk factors or predictors.

Classification Algorithm: A kind of machine learning technique where the algorithm uses input variables to categorize the data into several predefined classes. In this case, the binary classification algorithm used will determine whether a certain input profile will result in a diabetic or non-diabetic output.

Body Mass Index (BMI): It is a numerical value derived by dividing an individual's weight in kilograms by the square of their height in meters. Body mass index is a proxy measure of body fatness and is classified into underweight (BMI below 18.5), normal weight (BMI 18.5-24.9), overweight (BMI 25-29.9), and obese (BMI more than 30).

Web-based Application: Software systems that run on servers remotely and are accessed by users via a web browser without installation on the local computer. For this project, Python Flask is utilized in the development of the back-end, while the front-end is developed using HTML/CSS/JavaScript.

Random Forest: An ensemble machine learning method that involves developing several decision trees in random subsets of training data, which combine their results via majority voting. It is the chosen deployment model in this project owing to its high accuracy on the testing dataset.

Feature Estimation: A method of estimating clinical measurements that are not obtainable from the input provided by the users. In this project, feature estimation was employed to fill the gap between the inputs provided by users and the feature vector required for inference by the trained model.

Stratified Split: An approach used to partition the dataset in a way that the distribution of classes in the entire dataset is maintained proportionately in the train and test partitions. It was used in this particular project considering the imbalanced nature of the training dataset.

Risk Factor	Description	Effect on Diabetes Risk
Age	Individuals aged 45 years and above are more susceptible to Type 2 diabetes.	Increases diabetes risk.
Family History	Having a parent or sibling	Significantly increases the

	with diabetes.	likelihood of developing diabetes.
Obesity (BMI ≥25 kg/m²)	Excess body weight, especially abdominal obesity.	Causes insulin resistance.
Physical Inactivity	Lack of regular exercise.	Increases the risk of diabetes.
Unhealthy Diet	Frequent intake of sugary and processed foods.	Raises blood glucose and obesity risk.
High Blood Pressure	Blood pressure of 140/90 mmHg or above.	Associated with increased diabetes risk.
Prediabetes	Blood sugar levels above normal but below diabetic levels.	Strong predictor of Type 2 diabetes.
Smoking	Regular tobacco use.	Increases insulin resistance.
Excessive Alcohol Intake	Frequent heavy alcohol consumption.	Contributes to obesity and poor glucose regulation.
Gestational Diabetes	Diabetes developed during pregnancy.	Increases future Type 2 diabetes risk.

1.8 Organization of the Study

This project report is divided into five chapters, which include different stages of the research.

Chapter One contains the background and justification of the study, formulation of the problem, statement of aim and objectives, importance and scope, limitations, and definitions of terminologies used in the whole report.

Chapter Two includes an extensive literature review, including the theoretical background concerning diabetes, machine learning algorithms, and their applications in predicting the health status. It also reviews the related works and concludes with the identification of the research gap that the project will cover.

Chapter Three covers the system analysis and design with an ethical statement, analysis of the current screening method, requirements of the system, system design approach, UML diagrams, and design of the system input and output, database design, and architecture.

Chapter Four explains the implementation of the system, which includes the development environment, data set preparation, training and evaluation of the machine learning models, system integration, testing results, interface, and discussion of findings.

Chapter Five includes the conclusion of the entire research and recommendations for future work.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

In this chapter, an extensive literature review will be presented about the existing body of knowledge in connection with the development of a web-based diabetes risk prediction system. The objective of this literature review is to provide the theoretical and empirical basis of the study, to identify approaches and methodologies that are proven to be successful in their related research, and to point out the gap in the current literature that this research aims to fill.

The structure of this chapter is as follows. Section 2.2 covers the theoretical foundations, which include the definition, types of diabetes mellitus, risk factors and complications of diabetes mellitus, machine learning and its use in healthcare classification problems, and the utilization of web-based applications in health information. Section 2.3 discusses previous research related to diabetes prediction through machine learning. Section 2.4 summarizes the previous research in table form. Section 2.5 discusses the research gap and how this research fills it.

2.2 Theoretical Background

2.2.1 Diabetes Mellitus: Types, Causes, and Complications

Diabetes mellitus is a medical condition that is associated with hyperglycemia, a term used to refer to high concentrations of glucose in the blood as a result of a defect in insulin production, insulin action, or both. The hormone insulin is secreted by the beta cells in the pancreas. This hormone helps the body cells in absorbing glucose for use in energy production. When there is a problem with the insulin, the glucose accumulates in the bloodstream because the cells do not use it. There are various problems that may arise from this condition.

There are three major types of diabetes. Type 1 diabetes is the type of diabetes that is autoimmune. This means that the body's immune system attacks the beta cells in the pancreas responsible for the production of the hormone insulin. As a result of this, the body fails to

produce enough insulin. This type is common among children and adolescents. Type 2 diabetes makes up about 90 to 95 percent of all diabetes cases in the world. It is common in adults, but recently, it has been seen in young people.

It is highly linked to modifiable risk factors like obesity, lack of exercise, and poor nutrition, and also non-modifiable risk factors like age, family history, and race. Gestational diabetes is a form of diabetes that occurs during pregnancy, but ceases after childbirth. However, women who suffer from it have an increased risk of getting Type 2 diabetes in the future.

Untreated diabetes results in high blood glucose levels, which damage blood vessels and nerves over time and cause severe complications. The microvascular complications of diabetes include retinopathy, a condition whereby the blood vessels of the retina are affected and is one of the major causes of blindness; nephropathy, which means kidney failure that could progress to end-stage renal disease and requires dialysis or transplantation; and neuropathy, which involves damage to the nerves, particularly those of the feet and legs, causing numbness and pain. Macrovascular complications include coronary artery disease, stroke, and peripheral artery disease. Individuals with diabetes have a two- to four-times greater risk of suffering from cardiovascular complications than people without diabetes.

These problems emphasize the importance of early detection and treatment. Early detection of diabetes can help greatly reduce the likelihood of developing any complications through lifestyle changes and treatment. Also, for those who have pre-diabetes, they can avoid getting diabetes entirely by making such changes.

2.2.2 Risk Factors for Type 2 Diabetes

The risk factors of Type 2 diabetes have been established by epidemiological research, and those factors include non-modifiable risk factors that cannot be changed and modifiable risk factors that can be corrected by changing lifestyle and medical interventions.

Non-modifiable risk factors include age because the risk increases dramatically starting from the age of 45; a family history because if there are any first-degree relatives having diabetes, then the chances of having diabetes increase because of shared genetics; ethnicity because people having African, Asian, Hispanic, and Native American descent have more chances to develop diabetes than those with European descent, despite other differences.

As for the modifiable risk factors, the most important one is obesity. Body Mass Index is an indicator of having excessive body fat, and if a person's BMI is 25 or more, then his/her chances of developing diabetes are high. Abdominal obesity is particularly connected with insulin resistance, regardless of total body mass index. An inactive lifestyle causes overweight and worsens insulin resistance, compounding the risk of metabolic disorders. The risk of developing diabetes is heightened through high intake of refined carbohydrates, sugar, and processed foods, while diets consisting of plenty of fruits and vegetables and low-fat foods decrease this risk. High blood pressure not only increases the risk of developing diabetes but also is one of the complications that occur frequently due to diabetes, because of shared metabolic abnormalities – resistance to insulin and dysfunction of vascular endothelium. Prediabetes refers to the condition when the level of glucose in the blood is elevated, but not enough to be diagnosed as diabetes; however, prediabetes is a marker of the highest risk of developing diabetes. The significance of these risk factors for the current project is obvious: the system’s input form includes questions concerning the patient's age, BMI, blood pressure, family history, physical activity, and dietary pattern.

2.2.3 Machine Learning for Healthcare Classification

Machine learning is a type of artificial intelligence that uses algorithms to learn from a set of data and uses the learned patterns to predict or classify new data. To perform disease prediction tasks, only supervised learning algorithms are useful, since the algorithm needs to be trained on labeled data where the outcome of each observation is known and then learn the mapping from input to output.

Some of the supervised classification algorithms used to predict diabetes include Logistic Regression, which is a statistical linear model that calculates the probability of a binary variable as a function of other variables. Although this algorithm is simple, it is quite efficient on tabular structured data and provides an easily interpretable coefficient.

Another algorithm is Decision Tree, which creates a tree with a sequence of binary decisions on the values of certain attributes. Each split of the tree uses a certain attribute value that maximally separates two classes in the dataset. Although this method is very easy to understand and interpret, it is highly prone to overfitting on small data sets.

Random Forest is another type of algorithm that solves this issue by building an ensemble of decision trees that use bootstrap samples and different sets of attributes while making decisions, and then averaging out the results by majority voting. Such an approach helps minimize variance and provides a better generalization performance than a single tree. SVM is capable of finding the best possible separating hyperplane in the feature space, which provides the maximum margin of separation between the two classes. It can also deal with non-linear boundaries using kernels.

These parameters measure the quality of prediction: accuracy, precision, recall, and F1-score. Accuracy is defined as the ratio of all correct predictions. Precision measures how selective a model is; it is defined as the ratio of true positive predictions among all true positive predictions. Recall is also called sensitivity, and it measures how well the model can detect a condition – this is the ratio of all true positive predictions. F1-score is the harmonic mean of

precision and recall. It is especially useful if there is a class imbalance, because it punishes models that have good scores in one aspect but poor scores in another.

2.2.4 Web-Based Health Applications

There is a class of health apps called web-based health apps, which provide the service of delivering health information, risks, and symptoms assessments via web browsers with no requirement of downloading any application software. The major advantages of web-based health apps over existing methods in healthcare provision are that they are accessible anywhere as long as there is an internet connection, they deliver the results instantly, are scalable because one instance of an app can serve many users simultaneously, and are cost-effective as many such apps are free of charge.

Currently, there exist many online tools assessing the risk of developing diabetes, such as the American Diabetes Association's Risk Test and risk assessment tools designed by the national health authorities in the UK, Finland, and Australia. However, these tools are tailored to fit the specific population within a country and do not take into account the risk factors of developing diabetes among Nigerians/West Africans. In addition, the accuracy of most of these tools is poor as they are based on self-scorable questionnaires.This project’s system is based on the combination of the convenience provided by a web application together with the predictive capacity of a trained machine learning model, specifically designed to be used without any laboratory equipment, which is an approach that sets it apart from many other research diabetes prediction models.

2.3 Review of Related Studies

Much research effort has been put into the development of machine learning models for diabetes prediction using the Pima Indians Diabetes Database available in the UCI Machine Learning Repository. The works listed below vary greatly in terms of methodology, data set

used, and application, but all show the progress made in the field and the particular areas where research is still needed.

In contrast to the traditional approach of using laboratory features to predict disease onset, Tigga and Garg (2020) created a prediction model using only the information collected from questionnaires. Their model was a Random Forest one, which reached an accuracy of 75.9 percent without relying on any laboratory data. The current work is closely related to the paper mentioned above, as it shows the possibility of reaching accurate predictions while using only questionnaires. The dataset used was rather limited and population-specific, and no web application was developed.

Khanam & Foo (2021) reviewed machine learning models used in predicting diabetes, and compared Logistic Regression, SVM, Random Forest, and Gradient Boosting models in several research works. In their research, they concluded that ensemble models, especially Random Forest and Gradient Boosting models, performed better than single classifiers. They also indicated the commonality among several research projects where results were obtained only using the Pima dataset without validating on other population groups, which is also a feature of the current research.

Another example of a relatively unique study carried out in Nigeria is that of Olanrewaju et al. (2022). The authors applied three models, Random Forest, Decision Tree, and SVM, to predict chronic disease, one of them being diabetes. Although they indicated that the Random Forest model performed better than the other two models, they noted that the dataset used for the analysis came from a single institution, the sample size was small, and no web application was created based on the models used.

2.4 Summary of Literature Review

Table 2.1, presented below, compares some of the key studies carried out in this chapter in terms of their methodologies and limitations.

S/N	Author(s)	Contribution / Work Done	Technique / Methodology	Dataset	Limitations
1	Islam et al. (2020)	Developed a prediction model using KNN, LR, and RF; Random Forest achieved 83% accuracy after feature selection.	K-Nearest Neighbors, Logistic Regression, Random Forest	Pima Indians Dataset	No web deployment; laboratory features required.
2	Tigga & Garg (2020)	Built a diabetes prediction model using questionnaire-based data without lab tests; achieved 75.9% accuracy.	Random Forest on self-reported inputs	Custom questionnaire dataset (Indian population)	Limited sample size; single population.
3	Khanam & Foo (2021)	Reviewed and compared ML algorithms; ensemble methods consistently outperformed single classifiers.	Logistic Regression, SVM, Random Forest, Gradient Boosting	Pima Indians Dataset	Purely comparative; no deployed system.
4	Olanrewaju et al. (2022)	Applied ML to Nigerian health data; Random Forest outperformed other algorithms for chronic disease prediction.	Random Forest, Decision Tree, SVM	Nigerian clinical records	Small dataset; single facility; no web interface.

2.5 Research Gap

The review of related literature identifies several consistent gaps that the present research aims to close.

First, most of the studied papers were experimental, training and evaluating the proposed models on a benchmark dataset without developing a user-friendly application. The output of these research efforts would be the trained model along with performance metrics but there was no application available for members of the public to apply the model and get their risk estimate. This gap is closed by integrating the proposed algorithm into a complete web application.

Second, many of the diabetes risk assessment models currently in use, all models based on the Pima Indians Diabetes Database, use lab results such as plasma glucose concentration, serum insulin level, and triceps skinfold thickness, which means that these models are not suitable for application in community settings where there is no laboratory available. This research effort takes care of this gap by developing the user input form to accept data, which could be provided by people based on what they know about themselves, and feature estimation logic to connect the user-provided data and the feature vector.

Thirdly, research on populations from Nigeria or West Africa is very scarce, and there are no models created that have led to the development of any deployable solutions. Although our solution uses a globally available dataset to train the model, it is tailored for the local population of the state of Enugu, using an input interface suitable for this environment.

CHAPTER THREE

SYSTEM ANALYSIS AND DESIGN

3.0 Introduction

The system analysis and design in this chapter describes the whole diabetes risk prediction web-based system designed in this project. First of all, the chapter discusses the ethics behind the design and implementation of the system. It goes on to discuss the currently existing way of diabetes detection, along with the problems the system is meant to solve. In addition, the chapter explains the system requirements, the development methodology chosen, and the UML models that depict the system structure and behavior. Finally, this chapter includes the discussion of the input design of the system, the output design, the design of data storage, and the general architecture of the system, which is followed by an overview of the system, the dataset, and the limitations of the system.

All the analysis and design performed in this chapter were the basis for the implementation discussed in the next chapter. All the choices made during the analysis and design process, such as the input design of the system, the feature estimation process, the risk classification thresholds, etc., were passed to the implementation process.

3.1 Ethical Statement

The emergence of an information collection system capable of making predictions based on information related to health poses certain moral questions. In this part, we will consider some of the main moral issues in our project from four different angles: user privacy, responsible communication of results, data set ethics and consent to conduct research, and algorithm fairness.

3.1.1 User Privacy and Data Confidentiality

User privacy is one of the key design features of this system. All information entered by the user concerning his/her health condition is not stored, logged, shared with third-party

companies or organizations, or saved for more than the current HTTP request response exchange. User registration in the system, account creation, and submission of any personal information, including name and/or email address, are not required to make use of the system. The input form collects only those health and lifestyle parameters needed to produce risk estimation results, which are then immediately deleted from memory once the result is sent back to the user's browser.

3.1.2 Non-Diagnostic Purpose and Informed Communication

User privacy is one of the key design features of this system. All information entered by the user concerning his/her health condition is not stored, logged, shared with third-party companies or organizations, or saved for more than the current HTTP request response exchange. User registration in the system, account creation, and submission of any personal information, including name and/or email address, are not required to make use of the system. The input form collects only those health and lifestyle parameters needed to produce risk estimation results, which are then immediately deleted from memory once the result is sent back to the user's browser.

3.1.3 Dataset Ethics and Research Consent

The research was done using both primary and secondary datasets. The primary dataset, which comprised 40 patient records, was collected from the University of Nigeria Teaching Hospital (UNTH), Ituku-Ozalla, Enugu, and Enugu State University Teaching Hospital (Parklane), Enugu. All the data used had their identities stripped before usage. Due to the small number of records contained in the primary dataset, it was not adequate enough to train the machine learning model. Therefore, the Pima Indians Diabetes Database was selected as the secondary dataset for model training.

3.2 System Analysis

System analysis entails an examination of the current problem situation being faced, the weaknesses of such situations, and also the improvements the proposed system is designed to make. In this section, an analysis will be done of the current way of doing diabetes screening in Enugu State, and also the weaknesses in the current approach.

3.2.1 Analysis of the Existing System

In the case of Nigeria, and especially in Enugu State, the screening for diabetes is a clinical exercise. Any individual interested in having his/her risk of contracting diabetes checked will have to go to a hospital, clinic, or diagnostic lab where various clinical methods like fasting blood glucose, oral glucose tolerance test, and HbA1c are conducted by trained healthcare professionals. Although clinically accurate and medically proven, this method cannot be used by anyone for his/her own risk assessment.

Most individuals get themselves screened for diabetes only when they start developing symptoms like extreme thirst, excessive urination, unexpected weight loss, or constant tiredness. In this scenario, by the time the individual gets himself/herself screened, the condition might have reached an advanced stage. There is little preventive action taken by people in terms of risk checking because there is no tool that enables one to check his/her risk profile without visiting a health center. There is no doubt that healthcare professionals from clinics carry out various awareness exercises periodically, but it is not enough to form a screening mechanism.

3.2.2 Limitations of the Existing System

Analysis of the current approach with an emphasis on structure identifies several weaknesses of the current system, each of which represents a unique area for application of technology:

- Inaccessibility: The process of diabetes screening involves physically visiting a healthcare provider who possesses the ability to conduct laboratory testing. Such visits are not easily accessible geographically and financially for everybody, especially for people living in rural areas or those who cannot afford to spend both time and money on the prevention
- Costliness: Any laboratory testing, including fasting blood glucose level check and HbA1c test, entails certain costs, both direct and indirect ones (transportation, missed work time), which lead many people without any symptoms to refrain from conducting these tests.
- Reactive Approach: The current system is reactive by definition; most people undergo testing after some symptoms emerge, which does not allow detecting diabetes at its earlier stages.
- Absence of Self-service Risk Assessment Platform: There is no digital platform that allows an individual living in Enugu State to input his/her health information and then immediately get the informal assessment of his/her diabetes risk levels as the first step toward awareness.
- Lack of Health Care Professionals: Due to the lack of health care workers in Enugu State, preventive care often gets overlooked, and treatment of those who are already sick takes precedence.

3.2.3 Analysis of the Proposed System

This system overcomes the drawbacks discussed above by offering a web-based application through which any person with access to the internet can input simple data regarding his/her health and lifestyle and get an instant evaluation of the risk of developing diabetes. The system does not require any lab work, any clinical knowledge, or any prior experience or education of the user – it uses only data that any adult can easily obtain himself/herself: age,

body mass, height, measurement of the blood pressure measurements, family history, physical activities, and eating habits.

The system is created to be a screening and raising of awareness tool and is not intended to serve as a diagnostic one. It should help users determine whether their health status puts them into low, medium, or high risk of getting diabetes and encourage people who are in medium and high-risk groups to consult a doctor about their health. The advantages of the proposed system over the current one are the following: accessibility from any device connected to the Internet (including smartphones); it is free and does not require any payment or appointments; it provides immediate results; it promotes a proactive approach to raising awareness; it gives personal health advice.

3.3 System Requirements Specification

The System Requirements Specification describes exactly what is required of the system in terms of functionality. The above-mentioned requirements were obtained from project goals, current system analysis, and desired system behavior. The requirements can be classified into functional requirements, which specify particular system abilities, and non-functional requirements, which define system qualities.

3.3.1 Functional Requirements

S/N	Requirement	Description	Priority	Status
FR1	Input Form	The system provides a form for users to enter age, weight, height, blood pressure, family history, activity level, and diet type.	High	Implemented
FR2	Input Validation	System validates all entered values before processing and displays descriptive error messages for invalid or out-of-range inputs.	High	Implemented
FR3	BMI Calculation	The system automatically calculates BMI from weight and height and categorises the result.	High	Implemented
FR4	Feature Estimation	System internally derives approximate values for clinical model features (Glucose, Insulin, SkinThickness, DiabetesPedigreeFunction) from questionnaire responses.	High	Implemented
FR5	Risk Prediction	System passes the processed feature vector to the trained Random Forest model and retrieves a probability score.	High	Implemented
FR6	Risk Classification	The system classifies the adjusted probability into Low Risk, Medium Risk, or High Risk categories with corresponding colour coding.	High	Implemented
FR7	Result Display	System displays risk level, probability score, BMI, model used, recommendations, and medical disclaimer on the result page.	High	Implemented
FR8	Model Loading	System loads the trained model and metrics files automatically at application startup without requiring user intervention.	Medium	Implemented
FR9	No Registration	The system allows users to access and use the platform without creating an account or providing personal identifying information.	High	Implemented

Table 3.1: Functional Requirements of the Diabetes Risk Prediction System

3.3.2 Non-Functional Requirements

S/N	Attribute	Requirement	Rationale
NFR 1	Performance	The system must return a prediction result within 3 seconds of form submission under normal network conditions.	Ensures a responsive and frustration-free user experience.
NFR 2	Usability	The interface must be navigable by users with no technical or medical background. All labels and instructions must use plain, non-medical language.	The target users are members of the general public, not healthcare professionals.
NFR 3	Responsiveness	The system must render correctly on both desktop and mobile devices using a fully responsive web layout.	Mobile access is critical in Nigeria, where most internet users browse via smartphones.
NFR 4	Reliability	The system must handle unexpected or out-of-range inputs gracefully without crashing or producing undefined behaviour.	Robust error handling is essential for maintaining trust in a health-related tool.
NFR 5	Security	Client-side and server-side validation must be in place. No personal health data should be stored beyond the active session.	User privacy must be protected,d given the sensitive nature of health information.
NFR 6	Maintainability	The codebase must be modular and documented so that the model or interface components can be updated independently in the future.	Supports future improvement without requiring a complete system rewrite.
NFR 7	Portability	The system must run on any device with a modern browser and an internet connection, with no software installation required.	Maximises accessibility across diverse user environments and device types.

Table 3.2: Non-Functional Requirements of the Diabetes Risk Prediction System

3.4 System Design Methodology

The system was implemented using an iterative and milestone-based development approach.

In this approach, the system is built incrementally through a set of predetermined stages with

each of them including a test and review process, followed by a move on to the next stage. The choice of this approach was informed by the fact that the project consisted of two components that are very much inter-related but technologically different. These were the machine learning predictive model and the web application, which had to be developed and tested systematically.

An iterative methodology suits best projects where some design decisions are not known at the time of beginning but will appear through the process of development. For instance, in our case, the final deployment model would be chosen after training and testing of all the four possible algorithms. Therefore, the choices made during the design would be influenced by the results of that test. A Waterfall methodology would have involved making those decisions before the implementation process started, which was impossible in this case.

Key developmental phases undertaken in this research work were: identification of the problem and its objective; review of literature and the current approaches to solving this problem; analysis of the data set; data preprocessing; training and comparing four machine learning models; selection of the best performing model; designing the interface of the web application; implementing the Flask backend and frontend; integrating the trained model; and testing.

3.4.1 Justification for Methodology

The iterative method was used due to the dual aspect of the project, implying that the two parts could not always be worked on sequentially. For example, designing the web application interface required information about the input parameters for the model, which became known only during the model training process. Thus, iterating allowed for making necessary changes at certain points without having to undo all previous work. Also, testing at different stages was more convenient than doing so exclusively at the very end since it facilitated finding and solving the arising problems.

3.4.2 Method of Data Collection

The research design involved a mixture of approaches to data gathering. Primary data, which includes 40 records of patients, was gathered from UNTH and Parklane Hospital, Enugu. Nonetheless, the sample size was too small for training and testing the machine learning models, and therefore, the Pima Indians Diabetes Database was used as secondary data for developing, training, and evaluating the machine learning models.

The selection of such a database was based on the following considerations: it has been commonly used in previous studies to predict diabetes and, thus, offers a reliable basis for comparison of results with other findings; it can be used in binary classification problems; and the characteristics contained in the database have medical value related to the risk factors of Type 2 diabetes. The database does not reflect the population of Enugu State but is sufficient for showing the feasibility of the selected approach within the scope of an undergraduate research project.

3.5 System Modeling Using UML

The Unified Modeling Language (UML) was applied for the modeling of the system's structure and behavior using diagrams. The three diagrams created using UML include a Use Case Diagram, which specifies actors in the system and their interaction with the system; an Activity Diagram, which follows the order of actions in the process of inputting and displaying results; and a Class Diagram, which depicts logical components of the system and their interrelations.

3.5.1 Use Case Diagram

Use Case Diagram

This diagram shows the overall picture of the interaction between actors and use cases of the system. Two actors were found. First is the User, which is described as any person from the general public who uses the web application to evaluate his/her risks for diabetes. The second

actor is the Admin, who is described as a technically educated person who can update the model or check system performance metrics.

The following use cases of the User were found: access the web application, fill the health input form, submit the form for prediction, and receive the risk assessment result. Many of the use cases have background activities performed by the system, which are included in the use cases, for example, validation of the user input, evaluation of derived model features, and calculation of the machine learning prediction.

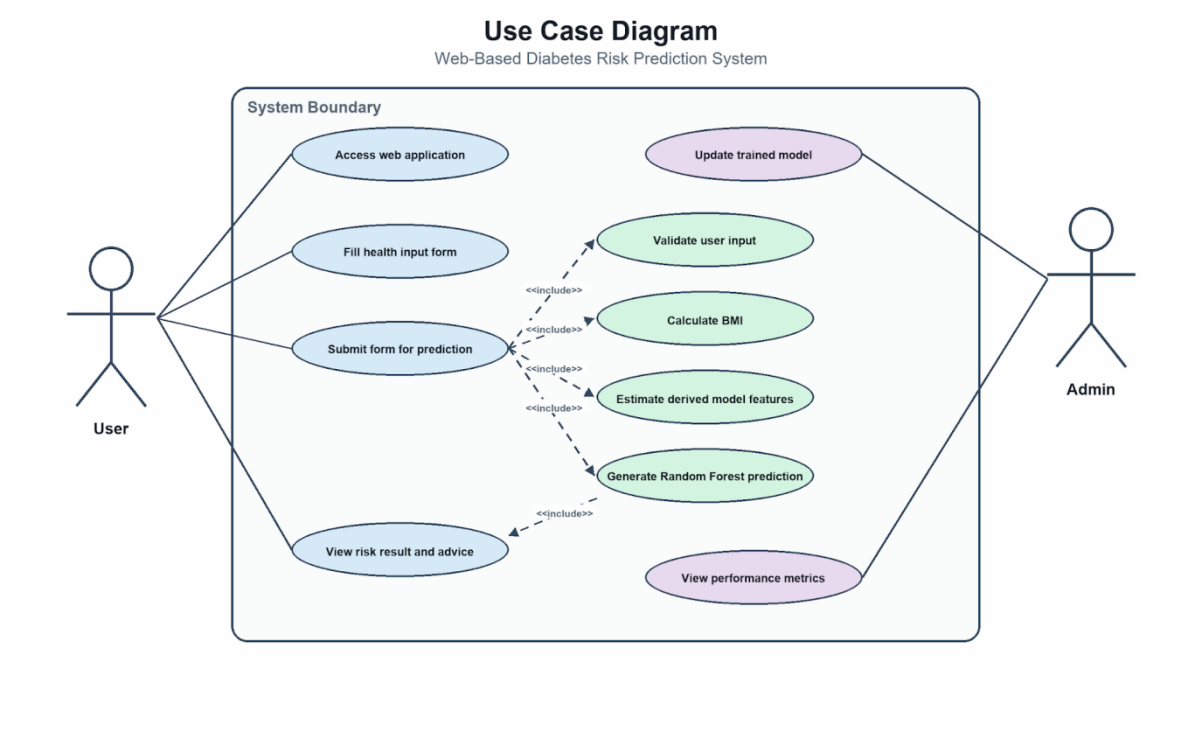


Figure 3.1: Use Case Diagram for the Web-Based Diabetes Risk Prediction System

3.5.2 Activity Diagram

An activity diagram shows the sequence of activities beginning when the user starts using the application till the time they get the result of their risk assessment. Activity diagram shows

that the sequence of activities performed will be: the user opens the homepage through a browser; the user fills all seven fields with information about their health; client-side JavaScript checks all the fields entered and gives error message for any invalid entry; the user clicks submit; the browser sends an HTTP POST request to the Flask server; server-side validation happens; if the validation is successful then the server computes BMI and estimates derived model features from the questions answered by the user; the feature vector is input to the pre-loaded random forest classifier; the classifier gives the probability value; the system applies risk adjustment based on the questionnaire; the probability value after risk adjustment is classified as Low, Medium or High Risk; and finally the Flask server returns the result page to the browser.

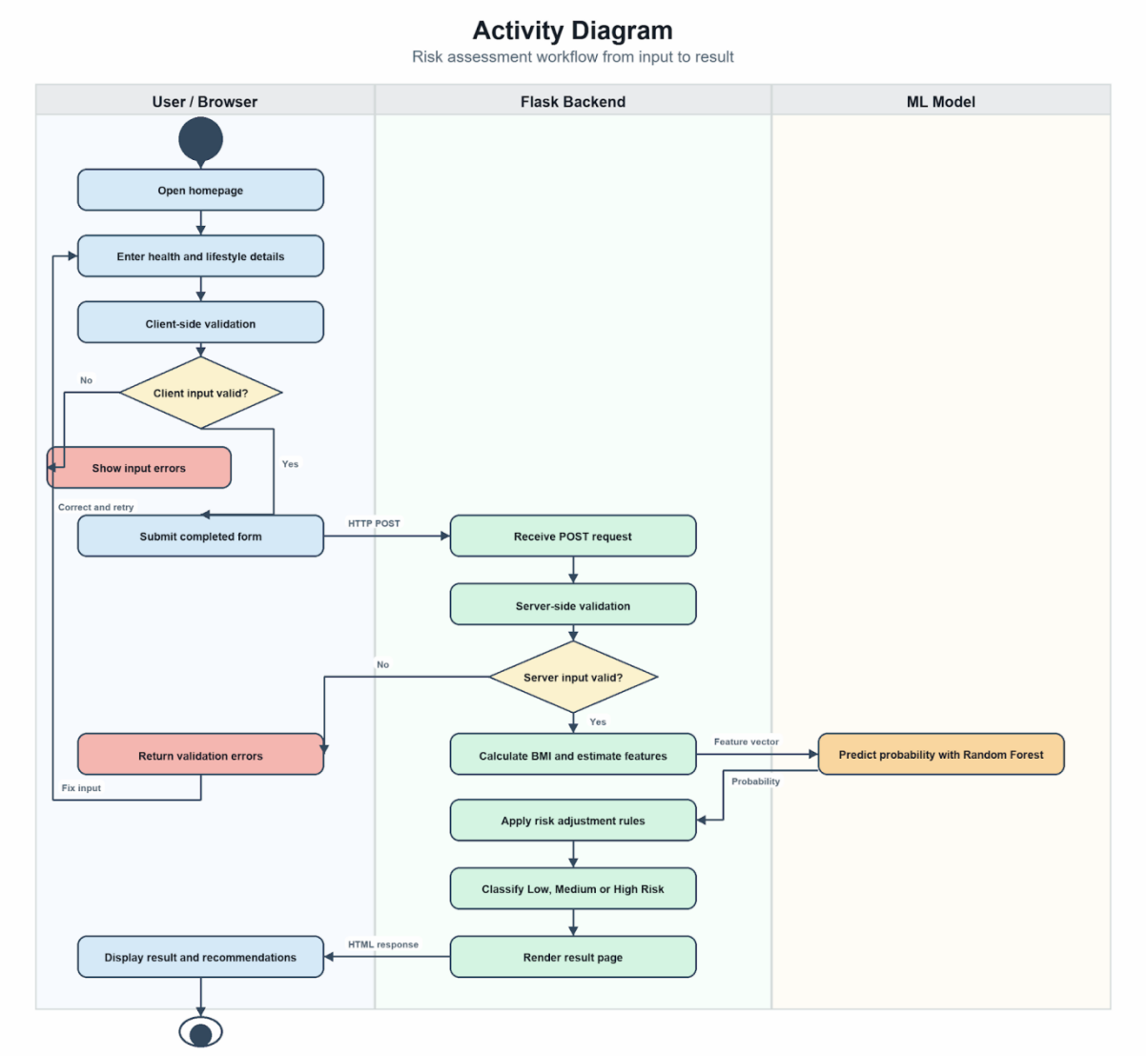


Figure 3.2: Activity Diagram for the Web-Based Diabetes Risk Prediction System

3.5.3 Class Diagram

The Class Diagram outlines the main components of the system from the software design perspective. There are five main components in the diagram. The User is the end user who provides inputs using the form and gets the output results. The InputProcessor component deals with unprocessed form input, validates it, calculates BMI, estimates features, and creates the model-ready feature vector. The DiabetesModel component contains the trained Random Forest classifier and loads the serialized model on start and returns the probability score for each feature vector submitted. The RiskClassifier component uses questionnaire-based rules to adjust the raw probability score and classify it into one of the three risk levels – Low, Medium or High. The ResultPresenter collects all the output information, including the risk level, probability, BMI, recommendations, and disclaimer information, and displays it in the result page.

This set of components works in a pipeline manner: the User sends data to the InputProcessor, which sends processed data to the DiabetesModel. The output from the DiabetesModel is passed to the RiskClassifier, and the classified result is collected by the ResultPresenter.

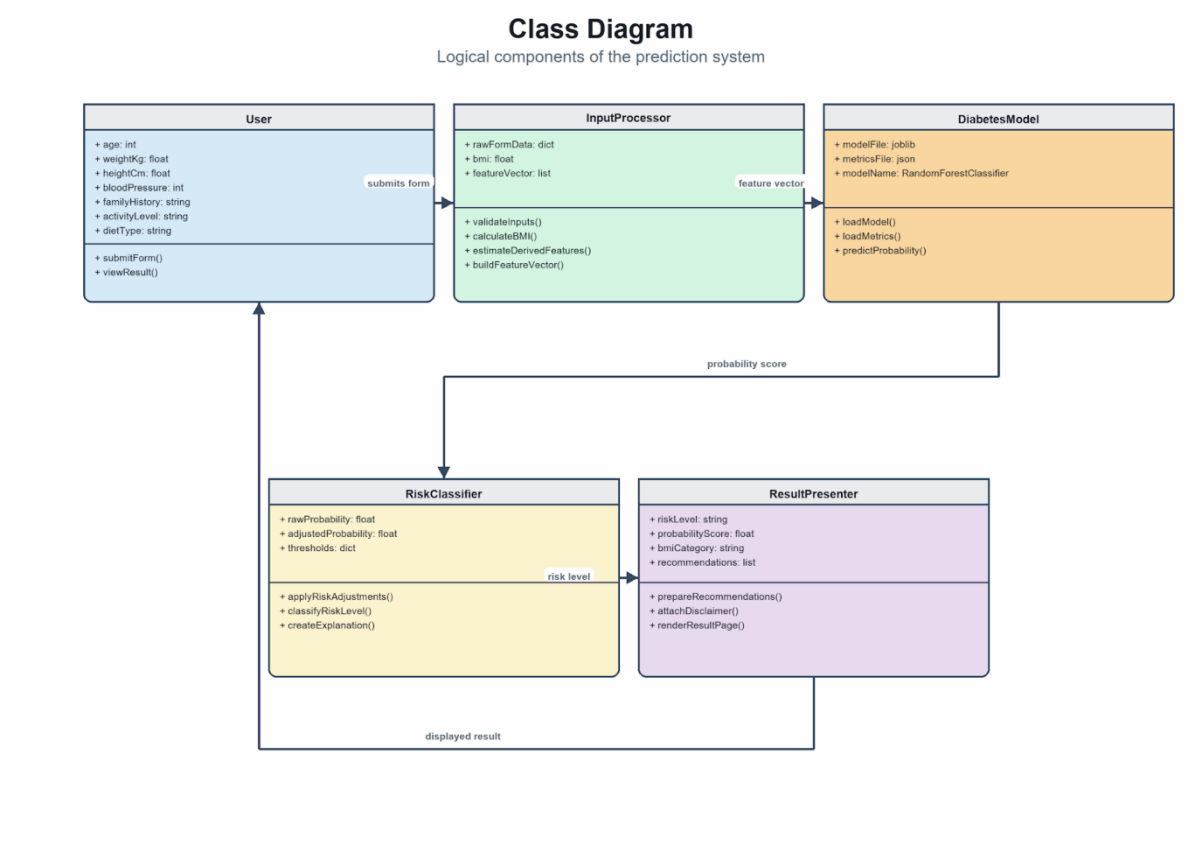


Figure 3.3: Class Diagram for the Web-Based Diabetes Risk Prediction System

3.6 System Design

In this section, we take the system requirements and UML model into an actual design plan, which involves the design of inputs, outputs, data storage, and overall architecture of the system.

3.6.1 Input Design

In the design of the input form, the major issues kept in mind were the user-friendliness and accessibility of the form. As our target audience does not have any medical background, only data that is easy to collect for an ordinary person. There are no l• Age: an integer type field that accepts values between 1 and 120, indicating the age of the person in years.

• Weight: an integer type field that accepts values in kilograms to calculate BMI.

• Height: an integer-type field that accepts values in centimeters to calculate BMI.

• Systolic Blood Pressure: an integer type field that accepts values in mm Hg.

• Family History of Diabetes: a drop-down box with four choices – no family history, one parent affected, two parents affected, and siblings or other relatives affected.

• Physical Activity Level: a drop-down box with four choices – sedentary, lightly active, moderately active, and very active.

• Diet Type: a drop-down box with three choices – healthy, mixed, and unhealthy.

All fields are validated on the client as well as the server side. The numerical fields are validated to check if they contain reasonable values. Drop-down fields are validated for the selection of a valid choice. When there is an error in any field, an appropriate message will be shown below the field.

3.6.2 Output Design

The output page shows the results of the prediction in a legible and non-threatening manner. The information provided includes: Risk Level displayed in colored texts according to levels (Low Risk - green, Medium Risk - amber, High Risk – red); Probability Score presented as a percentage; the computed body mass index (BMI) together with its classification; Model Name with a short description of its accuracy; Explanation of the meaning of the result in simple terms; Health Recommendations relevant to the predicted risk; and a Disclaimer indicating that the output is only an estimated screening tool and should not be considered as a diagnosis.

3.6.3 Database Design

Currently, the system does not use any relational database. The entire data that goes into the process is processed in the context of a single HTTP request-response cycle and is lost immediately after sending a response. Two files are used as data sources in this system. First, artifacts/diabetes_model.joblib contains a serialized Random Forest model that is loaded into

application memory upon initialization; second, artifacts/metrics.json contains all the evaluation metrics of all four models and is used for presenting comparative statistics on the results page.

If the system were further developed to allow user accounts or saved prediction history, a relational database may be considered. The following tables should be included in the proposed schema: Users (user_id, full_name, email_address, password_hash, date_registered) and Predictions (prediction_id, user_id foreign key, submission_timestamp, age, weight_kg, height_cm, bmi, blood_pressure, family_history, activity_level, diet_type, risk_level, probability_score). In the context of this project, this design is perfectly fine.

3.6.4 System Architecture Design

The framework is based on a three-tier web application architecture, which divides the structure into the presentation layer, application logic layer, and data layer. The Presentation Layer is the component where the user interacts. It is developed using HTML5, CSS3, and JavaScript and is composed of two primary components, namely the input form page and the result page. It is responsive to suit both desktop and mobile display. The Application Logic Layer is the backend that uses the Flask framework and is written in Python. Flask performs all routing, validation, BMI calculation, feature calculation, querying of the model, and formatting of the results. The server-side script receives the POST request from the user, processes it through the whole process flow, and generates the required HTML output. The Data Layer is composed of two artefacts found in the artefacts folder: the trained model and the metrics file.

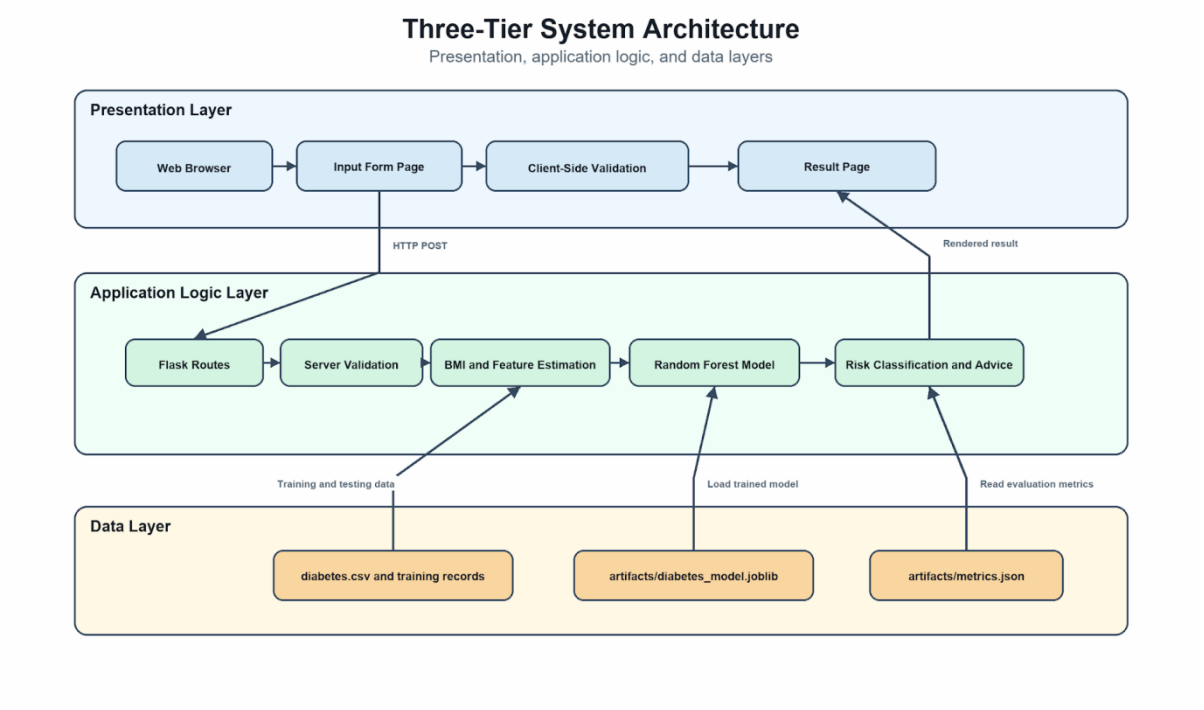


Figure 3.4: Three-Tier System Architecture for the Web-Based Diabetes Risk Prediction System

3.7 Overview of the System

A web-based diabetes risk prediction system is an information technology application used for raising health awareness, whereby people can know their chances of having diabetes without the need for a medical test or even a visit to a health facility. The application uses seven inputs from the user in a form, such as age, weight, height, systolic blood pressure, diabetes family history, physical activity level, and dietary type. The inputs are validated, and the application calculates the body mass index of the individual and then estimates the values of the four derived clinical features: plasma glucose level, serum insulin level, triceps skinfold thickness, and diabetes pedigree function.

This vector of features is fed into the trained random forest classifier, chosen from four machine learning models compared based on accuracy against the Pima Indians Diabetes Database dataset. This model gives a risk probability score, and using the risk rule generated from the questionnaires, this score is categorized as either low risk (less than 30%), medium

risk (30%-60%), or high risk (over 60%). The result page will display the risk level, probability, BMI, and its categorization, the model used, the interpretation of the result, health advice for the user, and a medical disclaimer.

3.8 Dataset Description

Two datasets were used in this experiment. The main dataset contained 40 anonymized patient records collected from UNTH and Parklane Hospitals, Enugu. Since this dataset was small and not enough to train the model using machine learning, the second dataset was the Pima Indians Diabetes Database.

3.8.1 Composition of Dataset

The dataset comprises 768 records of patients consisting only of female Pima Indians aged more than 21 years. There are eight predictor variables and one binary target variable for each record. Out of 768 records, 500 (65.1%) fall under the non-diabetic category (Outcome = 0) and 268 (34.9%) fall under the diabetic category (Outcome = 1). The eight predictor variables include:

Pregnancies (range 0 to 17, mean 3.85)

Glucose, plasma glucose concentration in mg/dL two hours post oral glucose tolerance test (range 0 to 199, mean 120.9); 5 records have a zero value for missing data.

Blood pressure in mmHg (range 0 to 122, mean 69.1); 35 records have zero value

SkinThickness, triceps skinfold thickness in mm (range 0 to 99, mean 20.5); 227 records have a zero value

Insulin, two-hour serum insulin in microU/mL (range 0 to 846, mean 79.8); 374 records have a zero value

BMI (range 0 to 67.1, mean 32.0); 11 records have a zero value

DiabetesPedigreeFunction, a diabetes pedigree function which denotes the probability of disease due to genetic makeup from family history (range 0.078 to 2.42, mean 0.47)

Age in years (range 21 to 81, mean 33.2)

3.8.2 Data Preprocessing

The zeroes in the Glucose, BloodPressure, SkinThickness, Insulin, and BMI fields were treated as missing values because it is physiologically impossible to have such zero values in a live patient's data for these parameters. The missing values were filled in by taking the median value of the non-zero elements along each column, as the median is more robust than the mean due to right-skewed data, especially in the case of the Insulin and SkinThickness columns. After this step, all eight parameters were scaled via z-score normalization, and the entire dataset was divided into training (80%, 614 rows) and testing (20%, 154 rows) sets.

3.8.3 Suitability of the Dataset

Pima Indians Diabetes Database was chosen for the project as it has a binary classification structure, its attributes are based on the scientifically proven risk factors for type 2 diabetes, and there have been numerous academic papers published that support the validity of the dataset. Although the dataset is not based on people from Nigeria or any other part of Africa, this drawback was acceptable in the scope of a feasibility undergraduate project. A user-friendly data entry form was created separately from the dataset for local data collection purposes, and a feature estimation algorithm was written to make up for it.

3.9 Limitations of the System

The limitations below have been observed in this system and must be kept in mind while interpreting its output and developing it further.

Population specificity: While 40 entries have been collected for hospitals in Enugu, the predictive algorithm is based on the Pima Indians Diabetes Dataset because the local dataset is too small. Hence, the algorithm may not represent the attributes of the local population.

Self-reporting of inputs: The system depends completely on user entry without any outside validation, so erroneous inputs would yield erroneous predictions.

Prediction errors: Glucose, Insulin, SkinThickness, and DiabetesPedigreeFunction are estimates from the questionnaire entries and not measurements, hence adding an element of error in the prediction.

Small size and moderate class imbalance: There are only 768 entries in the dataset with a 65:35 class split, which is considered too small and imbalanced in today's context.

High missingness: Around 48.7% of the Insulin and 29.6% of SkinThickness are missing in the dataset and need to be imputed.

Lack of persistent storage: Prediction history is not kept in this system and therefore cannot be used to track risks longitudinally for repeat visitors.

Dependence on Internet: Being a web-based application, this system cannot be accessed where there is no proper Internet connectivity.

3.10 Chapter Summary

The current chapter has presented all the analysis and design aspects of the web-based diabetes risk prediction system. In the first place, the chapter presented the ethical considerations guiding the development of the project. In addition to that, the chapter highlighted the weaknesses of the currently utilized diabetes screening process in Enugu State as well as stated the requirements that the new system needs to fulfill. The design process included modeling the system using the UML diagrams, such as the use case diagram, activity diagram, and class diagram. Besides, the design included specifications of input, output, data storage strategy, and architecture of the system. Furthermore, the dataset used in training the models was presented, including the components of the dataset and its appropriateness for the project.

CHAPTER FOUR

SYSTEM IMPLEMENTATION

4.0 Introduction

This chapter describes the deployment of the diabetes risk prediction system using the web platform as described in Chapter Three. First, the chapter will describe the development

environment, development tools, and libraries employed in developing the system. This is followed by the description of data preprocessing and training of the machine learning algorithms. Then the chapter will describe how the selected model was embedded in the Flask application for deployment. It will discuss how the system was tested for validation, correctness, and usability. Finally, it will give details of the interface and the results of the evaluation of the model and functionality testing.

4.1 Development Environment and Tools

4.1.1 Programming Language and Libraries

Python 3.10 was the programming language that was mainly used for the machine learning algorithm and the web application back-end. The decision to use Python was based on its large number of scientific computing and web development frameworks, readability of the code, and the widespread use of this programming language in both machine learning research projects and web development.

The following software libraries were used during the development:

• pandas: used for loading, analyzing, and processing the dataset, which included such tasks as reading the CSV file, replacing zeros in clinical columns, and calculating the statistics.

• NumPy: used for numerical calculations, including median imputation and construction of feature vectors.

• scikit-learn: the main library for machine learning, which was used for training of all classifiers, a stratified 80/20 train/test split, feature scaling using StandardScaler, and calculating the accuracy, precision, recall, and F1-score for each classifier.

• joblib: employed to serialize the trained random forest model on disk to the artifact/diabetes_model.joblib file so that it can be loaded back and reused in the web application without training again.

• json: used for saving/loading the performance metrics file (artifact/metrics.json), which contains the evaluation metrics for all four algorithms.

• Flask (v. 2.3): the Python web framework utilized for constructing the backend HTTP server, processing the URL routing, handling form POST requests, and rendering the Jinja2 HTML templates.

• Jinja2, the template engine that comes with Flask, is utilized for populating the result page with the risk level, probability value, BMI, recommendations, and other output variables.

• HTML5, CSS3, and JavaScript: used for constructing the frontend interface that includes the input form and the result page.

4.1.2 Development Platform and Hardware Requirements

The software was built on a personal computer that had the Windows 11 operating system. The development and testing were done entirely on one machine with an Intel Core i5 processor and 8 GB of RAM. No GPU was needed since the dataset size for training is relatively small (768 samples), and all four models were trained in a matter of seconds without the need for any GPU. The editor Visual Studio Code was used for writing the code, and the Python extension was active for highlighting and linting purposes.

4.2 Data Preprocessing

In preparation for applying classification models to the diabetes.csv file, several processing steps were applied to the initial data set to address issues with data quality and to prepare the feature matrix for further analysis.

Zero values in Glucose, BloodPressure, SkinThickness, Insulin, and BMI variables were considered as missing values, as there are no possible physiological zero values for these

clinical tests for living patients. Zero values were substituted with medians calculated across non-zero samples for each variable separately. The median value was chosen as a substitution method due to the right-skewed distribution of values in the Insulin and SkinThickness features, where the mean value is shifted toward large outliers.

After imputation, all eight predictors were normalized via StandardScaler implemented in scikit-learn, as it makes each feature have a mean of zero and a standard deviation of one. Normalization was performed to prevent a possible disproportionate impact on distance calculations by Support Vector Machine and Logistic Regression due to features with a very wide range of values, particularly Insulin (0 to 846) and Glucose (0 to 199).

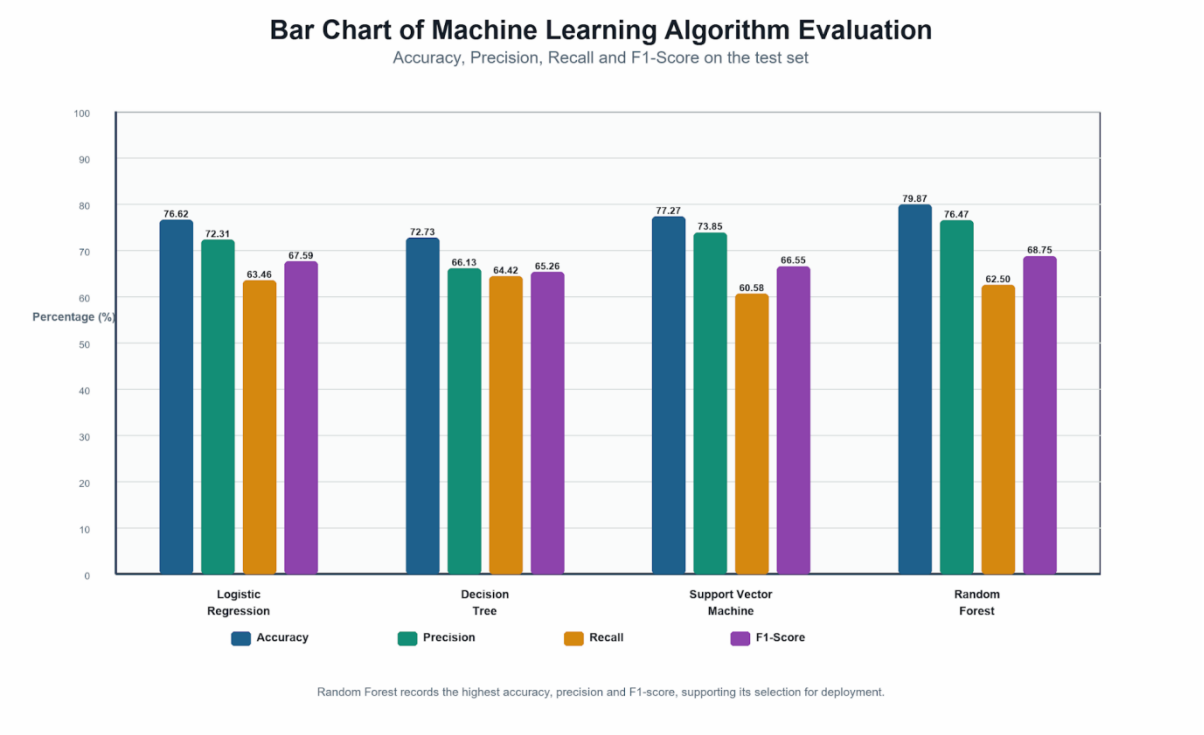
The pre-processed data set was split into training data that consisted of 80 percent of the data (614 entries) and test data consisting of 20 percent (154 entries) by using a stratified split method. This ensured that there were almost the same proportions (65 percent: 35 percent) in both data sets due to the class imbalance problem.

4.3 Model Training and Evaluation

Four supervised classifiers were trained and tested on the processed dataset. Each classifier was trained using one trial on the 614 samples in the training set and then tested on the 154 samples in the test set, for a total of four trials overall. The four classifiers used are: Logistic Regression, Decision Tree Classifier, Support Vector Machine with Radial Basis Function Kernel, and Random Forest Classifier (n_estimators=100, random_state=42).

All classifiers were trained using scikit-learn's default hyperparameters. Such a choice was intentional, as it was needed to provide a consistent point of comparison across classifiers that had not been tuned differently from each other, which makes sense in this context of an undergraduate feasibility study, in which extensive hyperparameter tuning would be out of the scope. The performance evaluation metrics calculated for each classifier are accuracy,

precision, recall, and F1-Score, since the four metrics together provide a more comprehensive evaluation than just the accuracy metric in the presence of the data class imbalance.



Graph 4.1: Bar Chart of Machine Learning Algorithm Evaluation Results

From Table 4.1 above, the Random Forest Classifier was the algorithm that had the highest level of accuracy (79.87%) and the highest level of precision (76.47%). The algorithm achieved the highest F1-score of 68.75%, indicating it performed best among the four models in balancing the prediction of diabetic cases and avoiding false positives. Although Logistic Regression had a lower F1-score, it gave a high level of accuracy (76.62%), making it competitive. SVM had slightly higher accuracy than Logistic Regression but lower recall, resulting in a lower F1-score. On the other hand, the Decision Tree algorithm had the lowest performance compared to the rest.

Based on the performance levels of the algorithms above, the Random Forest Classifier was chosen to deploy the project. It was serialized using joblib and saved in

artifacts/diabetes_model.joblib. Performance levels of all four algorithms were stored in artifacts/metrics.json and will be presented on the result page.

4.4 System Implementation

In this section, we discuss how the developed model was incorporated into the Flask web application to achieve the fully developed system. This implementation uses the three-tier architecture explained in Chapter Three.

4.4.1 Project Structure

The project had an organized folder structure. The root folder included app.py, the main Flask application file that included all the routes and functionalities of the application. The templates/ folder contains the Jinja2 HTML templates used in index.html for input and in result.html for output display. The static/ folder includes the CSS stylesheet and the JavaScript file for validating the input. The artifacts/ folder includes the trained machine learning model (diabetes_model.joblib) and metrics.json. The data/ folder includes diabetes.csv as the dataset and train_model.py, which was executed independently to generate the artifacts.

4.4.2 Flask Routing and Processing Pipeline

Two main routes were implemented in the Flask application. In the GET route for the root URL, the application would load the trained model and its metrics into memory and then send the input form to the user's browser. The POST route would handle the form submissions and run the entire process chain in the following order: server-side input validation, calculation of BMI based on user weight and height, estimation of model-derived features (Glucose, Insulin, SkinThickness and DiabetesPedigreeFunction) via rule-based mapping of answers to the questionnaire, standardisation of the collected feature vector using the same StandardScaler used during training, prediction of diabetes probability via the loaded Random Forest model, risk score adjustment via the questionnaire, risk level classification

into Low, Medium or High, creation of personalised recommendations, and rendering of the result template with all the output values sent as template variables.

4.4.3 Feature Estimation Logic

Since the user input screen does not capture any direct measurements for the Glucose, Insulin, Skin Thickness, and Diabetes Pedigree Function variables, the system adopts estimation strategies to compute the variables using information provided by the questionnaire. The Glucose variable is estimated through stratified estimates using the user's recorded blood pressure and diet, where more Glucose readings corresponded to higher levels of blood pressure and an unhealthy diet. The Insulin value was estimated from the BMI and the diet type. The skin thickness value was computed as a linear function of BMI. The Diabetes Pedigree Function variable is estimated from the family history drop-down box, where more diabetes pedigree readings were assigned to those users whose parent had diabetes.

4.4.4 Risk Adjustment and Classification

Once the model produced a raw probability score, the algorithm used the results of the questionnaires to make further adjustments to the estimate. For instance, a participant who indicated that both parents had diabetes, a physically inactive lifestyle, and an unhealthy diet would have been awarded a positive adjustment as a result of compounded risks. On the other hand, a participant who indicated an active life and healthy eating habits without any family history was given a slight negative adjustment. The resulting probability would then be categorized as either Low Risk, which is below 0.30, Medium Risk, which is 0.30 to 0.60, or High Risk, above 0.60.

4.5 Testing and Evaluation

4.5.1 Evaluation Metrics

The main criteria that were considered to measure the performance of the model were accuracy, precision, recall, and F1-Score, and they were calculated on the hold-out test dataset of 154 records. All these criteria are explained in Chapter Three (Section 3.3.1). They give a complete picture of how well our classifiers performed for the binary classification problem with an imbalanced distribution of classes. These results are shown in Table 4.1 above. No cross-validation technique has been used. Each of the models was trained once and then tested once.

4.5.2 System Testing

Manual functional testing was done by providing different sets of valid, invalid, and boundary values as input to the software and ensuring that the actual output matches the specified behavior as per the test cases. The eight test cases run, and their respective results are listed in Table 4.2.

TC#	Test Case	Input	Expected Output	Priority	Result
TC1	Valid high-risk submission	Age: 55, Weight: 102 kg, Height: 162 cm, BP: 150, Family: Both Parents,	High Risk displayed; probability above 60%; recommendations include clinical referral.	High	Pass

		Activity: Sedentary, Diet: Unhealthy			
TC2	Valid low-risk submission	Age: 24, Weight: 58 kg, Height: 168 cm, BP: 110, Family: None, Activity: Very Active, Diet: Healthy	Low Risk displayed; probability below 30%; maintenance recommendations shown.	High	Pass
TC3	Empty required field	All fields empty; form submitted	Validation error messages displayed beneath each empty field; form not submitted.	High	Pass
TC4	Out-of-range age	Age: -5	Error: "Age must be between 1 and 120" displayed beneath the field.	High	Pass
TC5	Non-numeric weight	Weight: "eighty"	Error: "Please enter a valid number" displayed.	High	Pass
TC6	Model loading at startup	Flask application started	Model and metrics files loaded without errors; no exceptions raised.	Medium	Pass
TC7	Mobile responsiveness	Form loaded on 375 px viewport (iPhone SE equivalent)	All fields, labels, and buttons render correctly without horizontal scrolling.	Medium	Pass
TC8	Result page completeness	Valid submission processed	Result page displays: risk level, probability, BMI, model name, recommendations, and disclaimer.	High	Pass

Table 4.2: Functional Test Cases and Results

None of the eight functional test cases encountered any errors. These include valid input validation, rejection of invalid inputs with an appropriate error message, correct risk

classification output of both the high-risk and low-risk inputs, proper initialization and loading of the model and metrics file, and verification that no submission survives beyond the current HTTP session.

4.5.3 User Interface Testing and Walkthrough

Testing of the user interface was done through the deployment of the system in three different environments, namely: Google Chrome on the desktop computer with a screen width of 1440px, Mozilla Firefox on the desktop computer, and finally Google Chrome on the mobile phone with a width of 375px (equivalent to iPhone SE). All seven fields in the input form were presented together with correct labels, the dropdown menu operated normally, there were validation messages presented below corresponding fields after entering invalid data, and finally, all necessary output fields were presented on the result page. Form fields were presented in a vertical arrangement in the mobile environment, and horizontal scroll was not needed at all. Risk level indication was color-coded and was clearly visible in all tested screens, while the disclaimer was present on the result tpage without any additional scrolling.

4.6 Results Presentation and Analysis

4.6.1 Output Samples and Interface Screens

The user interface of the system comprises two pages. On the input form page, there is a single-column form comprising an introduction to the purpose of the system, seven labelled input fields accompanied by a description, and a submit button. On the result page, the structure includes the result card and recommendations. First, the risk level is displayed in big colored letters. Afterward, the probability score is displayed as a percentage, followed by the

body mass index (BMI) and its corresponding classification. Next is an interpretation of the result, followed by the recommendations related to the risk level in a bulleted list format. A disclaimer box is included in the lower part of the page, which indicates that the result is an estimate for screening purposes and not a clinical diagnosis. The screenshots of the system interface are provided below.

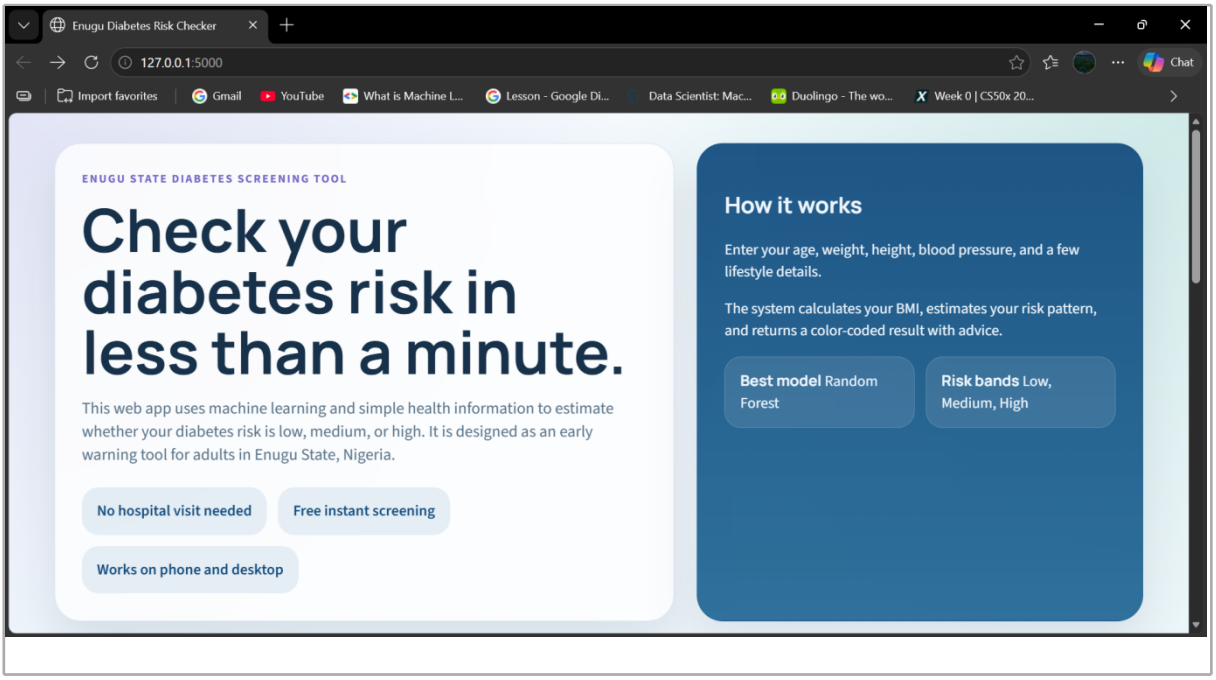


Figure 4.1: Input Form Page of the Web-Based Diabetes Risk Prediction System

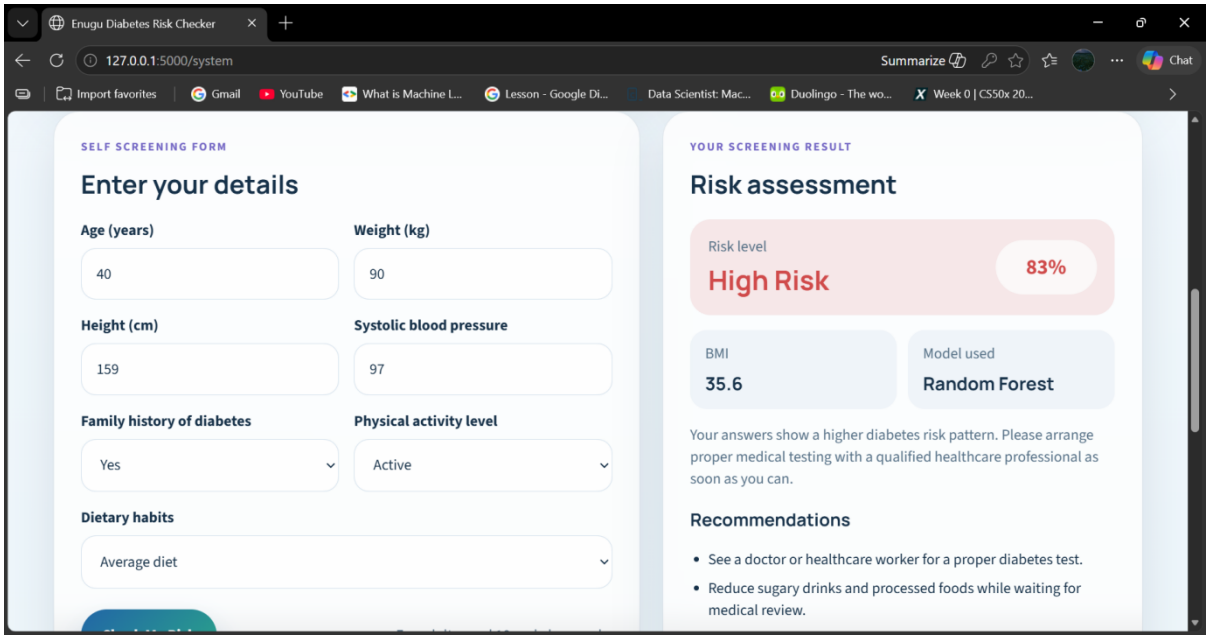


Figure 4.2: Result Page Showing a High Risk Assessment

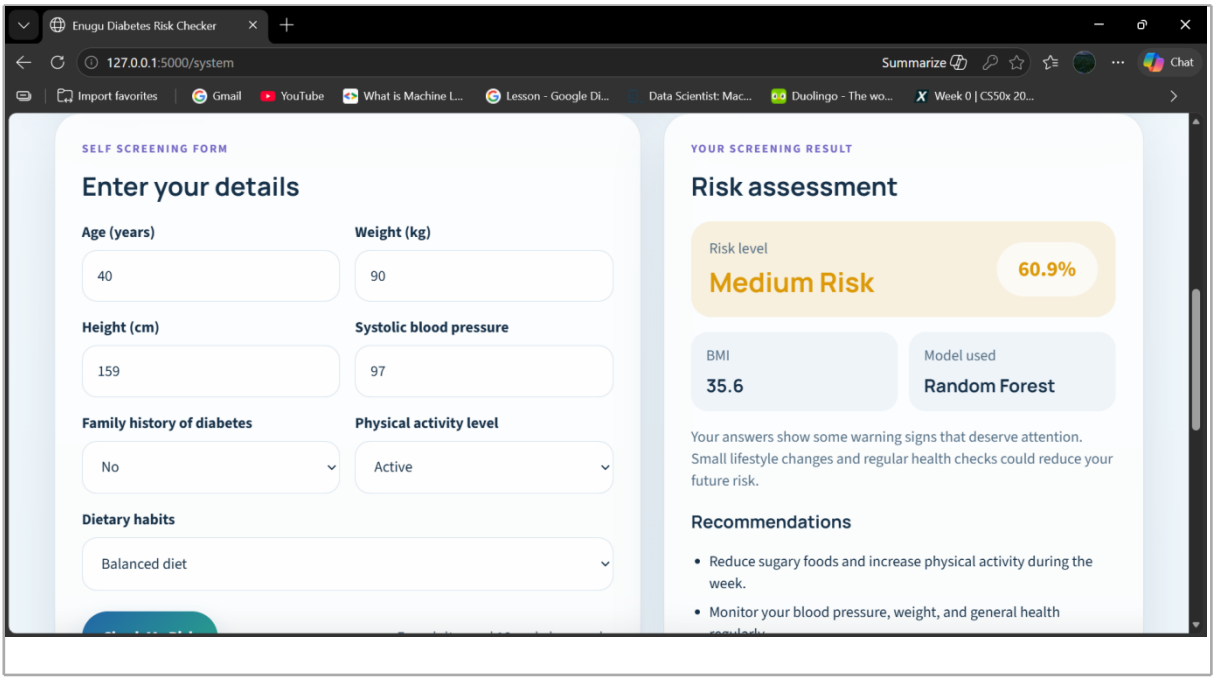


Figure 4.3: Result Page Showing a Low Risk Assessment

4.6.2 Discussion of Results and System Performance

According to the results of both evaluations – the model and the functional testing – the system works as it is expected to work within its intended scope. The most accurate algorithm used in the system was the Random Forest Classifier, which reached an accuracy of 79.87% and an F1-score of 68.75% on the unseen test data. Such outcomes coincide with the ones stated in scientific literature concerning the use of the Pima Indians Diabetes Database, where the accuracy rates for Random Forest models without class balancing are around 75-82%.

The difference between the model's accuracy and recall stems from the class imbalance of the training data and the conservative threshold used. In the case of health risk assessment application, high recall is more desirable than high precision since false negatives, i.e., the failure to alert an actually at-risk individual, are more damaging than false positives. Risk adjustment rules based on the questionnaire used after the model's prediction take care of the moderate recall by increasing the probability score for individuals possessing several risk factors. It decreases the rate of false reassurance among the final results for users.

All eight functional tests passed successfully, indicating that the web application has been developed according to the specified pipeline, invalid and boundary-value input is properly handled, and the interface is rendered correctly for both desktop and mobile environments. The stateless architecture was confirmed by the absence of submission data persistence during the processing cycle.

In summary, the system proves to be technically feasible as a diabetes risk prediction application utilizing machine learning technology and running through a web interface without the necessity for any clinical tests and expert assistance from the user side. The principal limitation of the system is the specificity of the population in the training dataset, thus making the model predictions applicable only to this particular population group.

CHAPTER FIVE

SUMMARY, CONCLUSION, AND RECOMMENDATIONS

5.1 Introduction

The present chapter provides a short synopsis of what was done in each chapter, completes the results given in Chapter Four, and gives recommendations for future work. There is also a part about directions for future work that have been determined during the course of the project; the directions concern how to improve and extend the system in order to enhance its precision, coverage, and applicability. Last but not least, this chapter discusses to what extent the objectives outlined in Chapter One have been fulfilled.

5.2 Summary

This study aimed at developing a web-based diabetic disease risk prediction system utilizing machine learning techniques such that people living in Enugu State in Nigeria could be able to use health data about themselves to predict the risk of diabetes without the need for laboratory tests.

Chapter One provided the basis of the motivation of this study by discussing the prevalence of diabetes in the state, the barriers to conducting regular screening of diabetes in the state, and how the web-based machine learning techniques could help overcome the problems. In addition, the chapter stated the following four research objectives: conducting a literature review; training and testing the machine learning models; designing and developing the web system; and validating and documenting the system.

The second chapter outlined the theoretical and practical background of the project. In particular, it considered the nature of Type 2 diabetes, the risk factors of the disease, the theory behind the supervised classification algorithm of machine learning, and the state-of-the-art diabetes prediction projects. The literature review showed that there were several shortcomings in the previous diabetes prediction research, which included training the

algorithms using benchmark datasets, a lack of a practical system that can be accessed by users, reliance on laboratory results not available to the average user, and the focus of the research mainly on the non-Nigerian population.

In Chapter Three, the complete system analysis and design were discussed. First, the ethical guidelines of the project were stated, and then the limitations of the current approach to clinical diabetes screening in Enugu State were analyzed. Next, the requirements for the system were defined, including functional and non-functional requirements. In terms of system analysis, UML diagrams for use cases, activities, and classes were developed. About system design, it included the definition of specific requirements for the input form, the output page, data storage, and the three-tier web application.

The process of system implementation was described in Chapter Four. In order to implement four machine learning models – Logistic Regression, Decision Tree, Support Vector Machine, and Random Forest – each of the algorithms was trained once on the prepared 614-sample training dataset and validated on the 154-sample test dataset. The Random Forest Classifier was chosen as the model for deployment due to its higher accuracy (79.87%) and F1-score (68.75%). The developed model was incorporated into the Flask-based web application that takes user input, calculates clinical features, computes risk probability, applies adjustment rules, makes a classification, and provides feedback to the user. All functional test cases were successfully passed, and the interface was confirmed to be displayed properly both on desktop and mobile devices.

5.3 Conclusion

All of the four objectives of the project were attained. A web application that predicts the probability of developing diabetes was developed, implemented, tested, and well-documented. The application uses seven easily available factors as input, performs classification using the trained random forest algorithm, and provides the output in the form

of categorized risk level, useful medical advice, and disclaimers without the need to register for any accounts, pay anything, and not requiring any pre-existing medical knowledge on the user's part.

The comparative analysis of the four machine learning algorithms proved that the random forest classifier provided the best performance metrics when applied to the Pima Indian Diabetes Database, achieving an accuracy of 79.87% and an F1-score of 68.75%. The obtained results correspond to those found in other sources and prove that the random forest algorithm trained on the database in question is a valid core of a first-line screening tool when used in a well-thought-out pipeline.

The functional tests validated the correctness and effectiveness of the web application on all kinds of valid, invalid, and boundary value inputs. The software met its intended role of being non-diagnostic and preventative by estimating outcomes, prominently displaying the medical disclaimer, and prompting high-risk individuals to seek medical advice.

Even though the limitations include demographic-specificity of the training data set (since it only consisted of female patients with Pima Indian ethnicity), which prevents the generalization of model predictions in other groups of people (particularly, the Nigerians targeted in the system), it still proves that the system can be done and works, and serves as a good basis for future work outlined below.

5.4 Recommendations

Recommendations for adoption and further study based on the findings and lessons learned from this project include:

• Deploy the tool further: The system needs to be hosted in the cloud and marketed via community health resources in the state of Enugu. This can be achieved through hospitals and community centers, as well as via social media marketing.

• Perform formal usability testing with actual end-users: This would consist of conducting a user evaluation exercise using actual members of the target group, including those who have lower levels of technological knowledge, and finding out what problems users encounter with the user interface.

• Adding more clinically significant fields to the tool: Future iterations of the form could include additional fields such as waist circumference, whether the person is a smoker or not, frequency of drinking, sleep habits, and gestational diabetes history, which are all proven risk factors for diabetes.

• Use class rebalancing approaches while retraining the algorithm: The class imbalance in the training data was moderate, which led to the discrepancy between the precision and recall scores of the predictive model. The future training sessions could be approached with techniques like SMOTE (Synthetic Minority Over-sampling Technique) or class weighting of the training data.

• Get clinical approval of the system: Before launching the tool for use by everyone other than being used just as an academic demonstration, the logic of risk estimation, feature estimation rules, health advice, and the disclaimer of the website should be approved by medical professionals. In Nigeria, the Nigerian Medical Association and NAFDAC should be approached.

5.5 Future Work

The following topics identify specific research and development directions for the future, which would significantly improve the functionality and value of the system. The following are especially important considerations for researchers, supervisors, and collaborators who may wish to take up the project.

• Retraining of the model using a locally sourced Nigerian or West African dataset: This represents the single most important step forward. Gathering data in hospitals, family doctors' offices, or other health surveys conducted in Enugu State or any similar locality in Nigeria would result in a trained model that reflects the health and risk factors of the intended users.

• Explore more sophisticated machine learning models: Although Random Forest showed the best results out of the four models tested in this research work, other, more complex models can potentially perform even better, for example, gradient boosting techniques (such as XGBoost or LightGBM) or artificial neural networks. Testing those on the local dataset is a logical continuation of this work.

• Create a mobile application with offline support: The creation of a mobile application that works both online and offline would expand the scope of the target audience to include people living in remote parts of Nigeria without internet access.

• Add optional account creation and long-term risk tracking features: In the future version of the web application, users should have the ability to create an account and store their information, thus allowing tracking of the changes in the user's risk profile over time and detecting any positive trends due to the changes in their lifestyles.

• Connect the system to pre-existing telemedicine or community health software: The system can be integrated into existing digital healthcare systems used by the community health practitioners in Nigeria. The connection of the system to pre-existing tools utilized by the community health extension practitioners, for instance, would enable the tool to be used during screening drives rather than being left to depend upon user discretion.

• Conduct a prospective clinical trial study: The clinical validity and reliability of the system can be confirmed through a study where the risk stratification made through the system is validated using a confirmed diagnosis in a group of subjects. Such a study will be necessary before the tool can be used for any public health project.

REFERENCES

American Diabetes Association. (2024). *Standards of care in diabetes—2024*. *Diabetes Care, 47*(Supplement_1), S1–S350.

Atlas, I. D. F. (2021). *IDF diabetes atlas* (10th ed.). International Diabetes Federation.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Islam, M. M., Yang, H. C., Poly, T. N., Jian, W. S., Li, Y. C., & Jack Li, Y. C. (2020). Machine learning algorithms for diabetes prediction: A systematic review. *Journal of Healthcare Engineering*, 2020, 1–12.

Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express, 7*(4), 432–439.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Olanrewaju, R. F., Adekunle, A. A., & Yusuf, M. O. (2022). Machine learning techniques for chronic disease prediction using Nigerian clinical data. *International Journal of Advanced Computer Science and Applications, 13*(5), 233–242.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D.,

Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Powers, A. C., & Stafford, J. M. (2022). Diabetes mellitus: Diagnosis, classification, and pathophysiology. In J. Loscalzo (Ed.), *Harrison's principles of internal medicine* (21st ed.). McGraw-Hill.

Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science, 167*, 706–716.

World Health Organization. (2023). *Diabetes*. World Health Organization.

World Health Organization. (2024). *Noncommunicable diseases country profiles 2024*. World Health Organization.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.). Springer.

Aggarwal, C. C. (2018). *Machine learning for text*. Springer.

Geron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.

Tan, P. N., Steinbach, M., & Kumar, V. (2019). *Introduction to data mining* (2nd ed.). Pearson.