A HYBRID MACHINE LEARNING-BASED TYPE 2 DIABETES MELLITUS RISK PREDICTION TOOL FOR RURAL UGANDAN SETTINGS

NOELA RUGOGAMU

S23B38/016

LORRAINE PAULA ARINAITWE

M23B38/004

ALOYSIOUS SSENDI

S23B38/002

A DISSERTATION SUBMITTED TO THE FACULTY OF ENGINEERING, DESIGN AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF A DEGREE OF BACHELOR OF SCIENCE IN DATA SCIENCE AND ANALYTICS OF UGANDA CHRISTIAN UNIVERSITY

May, 2026

UGANDA CHRISTIAN UNIVERSITY

A Centre of Excellence in the Heart of Africa

Declaration

We, *RUGOGAMU NOELA*, *LORRAINE PAULA ARINAITWE* and *SSENDI ALOYSIOUS*, declare that to the best of our knowledge, the work presented in this report is original and has not been submitted to any university or institution for any academic award. This work is the result of our own effort and investigation.

Signature (*RUGOGAMU NOELA*): ____________________

Signature (*LORRAINE PAULA ARINAITWE*): ____________________

Signature (*SSENDI ALOYSIOUS*): ____________________

Date: 05/06/2026

Ethical Generative AI Usage Declaration

A. **Acknowledgement of AI Usage as an Academic Support Tool**

We, *RUGOGAMU NOELA*, *LORRAINE PAULA ARINAITWE* and *SSENDI ALOYSIOUS*, declare that Generative Artificial Intelligence (AI) tools were used only as academic support tools during the preparation of this report. The AI tools assisted in activities such as brainstorming ideas, improving grammar and sentence structure, refining technical explanations, formatting LaTeX content, and summarising literature sources.

The authors confirm that AI was not used to fabricate research findings, generate experimental results, replace independent analysis, or substitute the intellectual contribution of the research team. All system design, implementation, evaluation, interpretation of results, and final conclusions presented in this project remain the original work of the authors.

Furthermore, all AI-assisted outputs were critically reviewed, edited, and validated by the authors to ensure academic integrity, technical correctness, and compliance with the guidelines of Uganda Christian University.

Signature (*RUGOGAMU NOELA*): ____________________

Signature (*LORRAINE PAULA ARINAITWE*): ____________________

Signature (*SSENDI ALOYSIOUS*): ____________________

Date: 05/06/2026

B. **Appendix Reference for AI Usage**

A detailed explanation of the AI tools used and the nature of assistance provided during this project is included in Appendix C.

Approval

This report has been submitted for examination with the approval of the following advisors/supervisor.

Dr. Terhemba Michael-Ahile (PhD.)
Eng. Ian Raymond Osolo
Department of Computing & Technology
Faculty of Engineering Design & Technology
Uganda Christian University

Signature (Primary Advisor): ______________________

Signature (Co-Advisor): ______________________

Date: 05 / 06 / 2026

Abstract

Type 2 Diabetes Mellitus (T2DM) is becoming a major public health problem in Uganda, especially in rural communities where access to diagnostic equipment and specialised healthcare services is limited. This project presents *PreDia*, a hybrid machine learning-based risk prediction tool developed to support early screening of T2DM in low-resource rural Ugandan settings. The system uses low-cost and clinically accessible features and combines an XGBoost predictive model with rule-based clinical guidelines from the World Health Organization (WHO) to improve the reliability of risk assessment. To improve transparency and user trust, SHAP (SHapley Additive exPlanations) was integrated to provide understandable explanations for each prediction. The system also includes an offline-first bilingual interface in English and Luganda to support accessibility among Village Health Teams (VHTs).

The predictive model was trained and tested using an augmented version of the Pima Indians Diabetes Dataset adapted to reflect selected physiological trends found in Sub-Saharan African populations. Experimental results showed strong predictive performance, achieving an AUC score of 0.9672, an accuracy of 88.96%, a recall of 87.04%, and a precision score of 82.46%. The findings of this study demonstrate that explainable and offline-capable artificial intelligence tools can support community-level diabetes risk screening in resource-constrained healthcare environments.

Dedication

This research report is dedicated to our families for their unwavering support, encouragement, and sacrifices throughout our academic journey. We also dedicate this work to our friends, colleagues, and sponsors whose motivation and assistance greatly contributed to the successful completion of this project.

Special dedication goes to Uganda Christian University and our supervisors, Dr. Terhemba Michael-Ahile and Eng. Ian Raymond Osolo, for their guidance, mentorship, and continuous support during the development of this research. Finally, we dedicate this project to the community health workers in rural Uganda whose commitment to improving healthcare inspired the development of this system.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
APK	Android Package Kit
AUC	Area Under Curve
BMI	Body Mass Index
CDSS	Clinical Decision Support System
CSV	Comma-Separated Values
DM	Diabetes Mellitus
EMR	Electronic Medical Records
FBG	Fasting Blood Glucose
IDF	International Diabetes Federation
ML	Machine Learning
PWA	Progressive Web Application
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SQL	Structured Query Language
SVM	Support Vector Machine
T2DM	Type 2 Diabetes Mellitus
TRL	Technology Readiness Level
VHT	Village Health Team
VHTs	Village Health Teams
WHO	World Health Organization
XAI	Explainable Artificial Intelligence

Acknowledgments

First and foremost, we thank the Almighty God for the gift of life, wisdom, strength, and guidance throughout our academic journey and during the completion of this research project.

We express our sincere gratitude to our supervisors, Dr. Terhemba Michael-Ahile and Eng. Ian Raymond Osolo, for their valuable guidance, technical support, constructive criticism, and continuous encouragement throughout the development of this project. Their expertise and mentorship greatly contributed to the success of this research.

We also extend our appreciation to the Faculty of Engineering, Design and Technology and the Department of Computing and Technology at Uganda Christian University for providing a supportive academic environment and the resources necessary for this study.

Special thanks go to Sunbird AI for their contribution to local language technology, which informed the bilingual accessibility aspect of this project. We are equally grateful to our friends, classmates, and colleagues for their encouragement, teamwork, and support during the various stages of this research.

Finally, we appreciate everyone who directly or indirectly contributed to the successful completion of this project.

Chapter 1

Introduction

1.1 Background

Diabetes Mellitus (DM) is one of the fastest growing non-communicable diseases worldwide and continues to place significant pressure on healthcare systems, especially in low- and middle-income countries [1]. The International Diabetes Federation (IDF) estimated that approximately 537 million adults were living with diabetes globally in 2021, and this number is projected to increase significantly by 2045 [1]. A large proportion of these cases occur in developing countries where access to healthcare services, diagnostic infrastructure, and long-term disease management remains limited [2].

In Sub-Saharan Africa, diabetes has become an important public health concern due to increasing urbanisation, lifestyle changes, and limited access to early screening services [3]. The region also records one of the highest proportions of undiagnosed diabetes cases globally. In Uganda, studies estimate that a large number of adults are affected by Type 2 Diabetes Mellitus (T2DM), many of whom remain undiagnosed until serious complications develop [3].

One of the major challenges in Uganda is the limited availability of affordable diabetes screening services in rural healthcare settings. While urban hospitals may provide Glycated Hemoglobin (HbA1c) and Fasting Blood Glucose (FBG) testing, many rural health facilities lack the equipment, trained personnel, and laboratory infrastructure required for routine screening [4]. Village Health Teams (VHTs), who support community healthcare delivery in Uganda, often rely on basic symptom observation and manual referrals for diabetes-related cases [2].

Recent developments in Machine Learning (ML) and Explainable Artificial Intelligence (XAI) have created opportunities for improving disease risk prediction using low-cost and clinically accessible features such as age, Body Mass Index (BMI), family history, lifestyle-related factors, and glucose measurements where available [5, 6]. However, many machine learning systems operate as black-box models that provide predictions without clear explanations, making them difficult to trust in clinical environments [7]. This project introduces *PreDia*, a hybrid clinical decision support system designed for rural Ugandan settings. The system combines an XGBoost predictive model with clinical rule-based decision logic and SHAP (SHapley Additive exPlanations) to provide explainable and offline-capable diabetes risk screening.

1.2 Problem Statement

Despite the increasing prevalence of Type 2 Diabetes Mellitus (T2DM) in Uganda, access to early and affordable screening services remains limited in many rural communities. Existing diagnostic approaches largely depend on laboratory infrastructure, trained medical personnel, and stable healthcare systems, which are often unavailable at lower-level health centres and community healthcare settings [4].

Many rural healthcare workers, including Village Health Teams (VHTs), lack access to reliable digital tools that can support early diabetes risk identification [2]. In addition, several existing digital health systems require continuous internet connectivity and are designed primarily for English-speaking users, limiting their usability in rural Ugandan communities [8]. The lack of explainability in many machine learning-based healthcare systems also reduces trust and adoption among healthcare workers [7].

Therefore, there is a need for an explainable, offline-capable, and bilingual diabetes risk screening tool suitable for use in low-resource rural healthcare settings.

1.3 Research Objectives

The main objective of this research is to design, implement, and validate *PreDia*, a hybrid machine learning-based risk prediction tool that delivers accessible, explainable, and bilingual (English and Luganda) Type 2 Diabetes Mellitus screening tailored for rural Ugandan settings.

1.3.1 Specific Objectives

1. To identify and prioritise low-cost and clinically accessible features for Type 2 Diabetes Mellitus prediction, with emphasis on features suitable for low-resource healthcare environments.
2. To design and evaluate a hybrid risk prediction engine that integrates high-performing supervised learning algorithms with clinical rule-based overrides aligned to WHO and Ugandan Ministry of Health guidelines.
3. To develop an offline-first Android application featuring a bilingual interface (English and Luganda) and SHAP-based explainability modules for community-level diabetes risk screening.

1.4 Justification

The development of the *PreDia* system is important because diabetes cases in Uganda continue to increase while access to early screening services remains limited, especially in rural healthcare settings [3]. Studies have shown that delayed diagnosis of Type 2 Diabetes Mellitus contributes to preventable complications and increased healthcare costs in low-resource countries [1].

Rural healthcare facilities in Uganda often face challenges such as limited laboratory infrastructure, shortages of trained medical personnel, and inconsistent internet connectivity [4]. As a result, many patients are diagnosed only after symptoms become severe. Existing digital health solutions are also not always suitable for rural environments because they may require stable internet access or lack support for local languages [8].

By using low-cost and clinically accessible features together with offline machine learning technology, the proposed system provides an accessible and cost-effective approach to diabetes risk screening [5]. The integration of SHAP-based explainability improves transparency and supports better understanding of prediction results among healthcare workers [7]. In addition, the bilingual English-Luganda interface improves accessibility and usability for Village Health Teams (VHTs) and community healthcare workers.

The study therefore contributes to ongoing efforts to improve community-level healthcare delivery through explainable and accessible artificial intelligence technologies. The final system was implemented as an offline Android application to support deployment on affordable mobile devices commonly used by community healthcare workers in rural Uganda.

1.5 Scope

1.5.1 Geographical Scope

This study focuses on rural healthcare settings in Uganda, with particular attention given to community-level healthcare delivery and diabetes screening practices among Village Health Teams (VHTs). Stakeholder feedback and contextual considerations were mainly based on rural communities within Mukono District.

1.5.2 Subject Scope

The research focuses on the early risk screening of Type 2 Diabetes Mellitus (T2DM) among adults using low-cost and clinically accessible features. The study covers machine learning model development, explainable artificial intelligence (XAI), bilingual healthcare accessibility, and offline mobile deployment for community healthcare support.

1.5.3 Technical Scope

The technical scope of this study includes the development of a hybrid machine learning-based prediction system using Python, XGBoost, Scikit-learn, and SHAP for model explainability. The mobile application was developed using Kotlin/Java and designed to operate offline on Android devices without requiring continuous internet connectivity or cloud-based infrastructure.

Chapter 2

Literature Review

2.1 Introduction

The literature review provides a detailed examination of existing research and technological solutions aimed at diabetes risk prediction. This chapter follows a thematic and chronological approach to understanding how machine learning (ML) has been applied to healthcare, specifically in low-resource settings [5, 6]. The review was conducted using a systematic search of the Scopus database, identifying studies related to clinically accessible diabetes screening approaches. By analyzing these related works, the study identifies the technical strengths and the remaining gaps that the *PreDia* system addresses for the Ugandan context.

Table 2.1: Diabetes Statistics in Uganda (2000–2050)

Indicator	**2000**	**2011**	**2024**	**2050**
Adults (20–79) with diabetes (thousands)	88.7	307.9	369.1	1,152
Age-standardised prevalence of diabetes (%)	–	2.8	2.2	2.7
Proportion of people with diabetes who are undiagnosed (%)	–	–	61.2	–
Adults (20–79) with undiagnosed diabetes (thousands)	–	–	225.9	–
Adult population (20–79 years, thousands)	8,405	13,932	22,022	50,963
Impaired fasting glucose (IFG) (%)	–	–	2.0	2.3
Impaired glucose tolerance (IGT) (%)	–	8.6	18.1	23.7
Deaths attributable to diabetes (20–79)	–	1,913	2,750	–
Diabetes-related health expenditure per person (USD)	–	84.0	222.6	198.9

Table 2.1 presents key diabetes indicators for Uganda from 2000 to 2050 based on estimates from the International Diabetes Federation (IDF) Diabetes Atlas. The statistics highlight a

growing diabetes burden, with the number of adults living with diabetes projected to increase from 369,100 in 2024 to approximately 1.15 million by 2050. The data further indicate that 61.2% of diabetes cases remain undiagnosed, emphasizing the need for accessible early screening and prevention interventions.

Source: International Diabetes Federation (IDF) Diabetes Atlas, 11th Edition, 2024.

Diabetes is a major non-communicable disease and an increasing public health concern in Sub-Saharan Africa [3]. Studies estimate that Uganda has a growing number of adults living with Type 2 Diabetes Mellitus (T2DM), many of whom remain undiagnosed until complications occur [3]. Research shows that the Ugandan healthcare system is primarily designed for infectious diseases, leaving a gap in chronic disease management [2].

Literature further highlights that rural populations are at higher risk because they often lack access to affordable diagnostic services and healthcare infrastructure [4]. The cost and availability of laboratory-based diabetes testing continue to limit access to early screening in many rural healthcare environments. This literature confirms the need for a diabetes screening approach that does not depend heavily on expensive laboratory infrastructure or continuous access to urban healthcare facilities.

2.2 Predictive Clinical Features

This section reviews clinically accessible features commonly used in Type 2 Diabetes Mellitus (T2DM) prediction systems. Several studies have shown that features such as age, Body Mass Index (BMI), waist circumference, family history, blood pressure, diet, and physical activity levels can be effective predictors of diabetes risk [9, 10].

2.2.1 Clinically Accessible Predictors for Low-Cost Screening

Research conducted in several developing countries demonstrated that low-cost and clinically accessible variables can achieve strong predictive performance while reducing dependence on expensive laboratory testing and specialised diagnostic infrastructure [5]. Age and BMI are among the most commonly used predictors because they are easy to collect in low-resource healthcare environments.

Family history and lifestyle-related factors such as smoking, alcohol consumption, physical inactivity, and unhealthy diet have also been identified as important contributors to diabetes risk [11]. These predictors are particularly suitable for rural healthcare settings where access to laboratory infrastructure is limited.

2.2.2 Limitations in Existing Predictive Feature Selection

Although many studies have successfully used clinically accessible predictors, most existing datasets are derived from non-African populations such as the Pima Indians Diabetes Dataset [12]. This creates challenges when applying such systems directly to Ugandan populations due to differences in lifestyle, genetics, and healthcare conditions.

In addition, several predictive systems rely heavily on laboratory variables such as glucose concentration and HbA1c values, reducing their suitability for rural community healthcare deployment [4].

2.3 Hybrid Machine Learning Models

Several supervised machine learning algorithms have been applied in diabetes prediction studies, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and XGBoost [5, 13]. Among these, ensemble learning methods such as XGBoost have demonstrated strong predictive performance due to their ability to handle non-linear relationships and missing data.

2.3.1 Supervised Learning Algorithms in Diabetes Prediction

Studies comparing machine learning techniques for diabetes prediction consistently report that XGBoost and Random Forest achieve higher predictive accuracy and recall than traditional statistical models [6, 14]. XGBoost is particularly effective in handling structured clinical datasets and reducing false negatives in healthcare prediction tasks.

2.3.2 Hybrid Clinical Decision Models

Recent studies have explored combining machine learning models with rule-based clinical guidelines to improve reliability and safety in healthcare systems [15]. Hybrid approaches integrate predictive analytics with predefined medical rules derived from established healthcare standards.

In the context of diabetes screening, healthcare guidelines provide important thresholds and clinical recommendations that can improve the interpretability and safety of machine learning predictions [16].

Such guidelines are particularly useful in low-resource healthcare settings where decision support tools must balance predictive accuracy with clinical safety and interpretability [7, 15].

2.3.3 Strengths and Limitations of Hybrid Models

Hybrid machine learning systems improve prediction reliability by combining the adaptability of machine learning with the consistency of clinical decision rules [15]. However, many existing systems remain difficult to deploy in low-resource settings because they depend on internet connectivity or complex computational infrastructure.

2.4 Mobile Health Applications in T2DM

Mobile health (mHealth) technologies are increasingly being used to support disease screening, monitoring, and healthcare delivery in low- and middle-income countries [17]. Android-based healthcare applications provide an affordable platform for delivering digital health services in rural communities.

2.4.1 Offline-First Mobile Systems

Several studies have shown that internet connectivity remains unreliable in many rural healthcare environments in Sub-Saharan Africa [2]. Offline-first healthcare applications are therefore important because they allow healthcare workers to continue operating without continuous internet access.

Offline deployment also improves response speed, reliability, and accessibility during community outreach activities [8].

2.4.2 Bilingual and Local Language Interfaces

Language barriers continue to affect healthcare communication in multilingual communities. Studies on healthcare accessibility have shown that bilingual interfaces improve usability and patient engagement among community health workers [18].

For this reason, integrating English and Luganda support improves accessibility and allows Village Health Teams (VHTs) to communicate screening outcomes more effectively.

2.4.3 Mobile Applications for Community Health Workers

Several healthcare studies have reported successful use of Android applications by community health workers for disease screening and patient monitoring [17]. However, many existing applications focus primarily on data collection rather than explainable clinical decision support.

2.5 Explainability in Machine Learning Models

Explainability is an important requirement in healthcare machine learning systems because healthcare workers need to understand how predictions are generated before trusting automated recommendations [7].

2.5.1 SHAP and LIME

SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) are among the most widely used explainable artificial intelligence techniques [7]. SHAP is particularly suitable for healthcare applications because it provides consistent feature importance explanations for individual predictions.

2.6 Explainable AI (SHAP) Integration

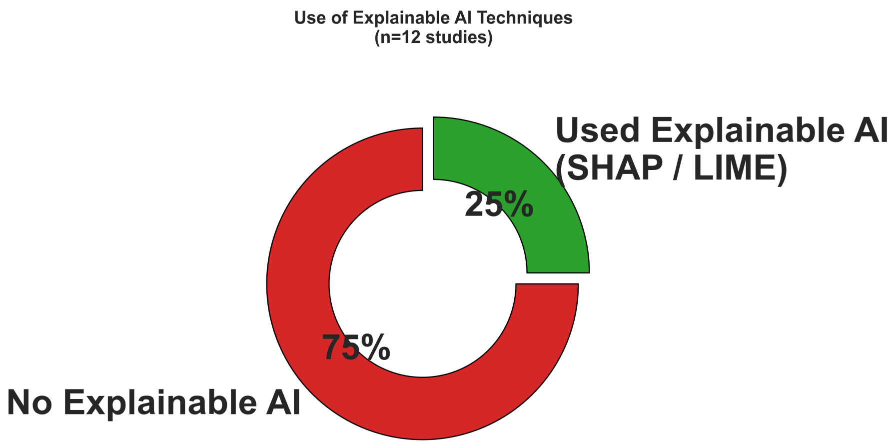


Figure 2.1: Use of Explainable AI Techniques in Reviewed Studies

2.6.1 Importance of Explainability in Healthcare

Explainability improves trust, transparency, and accountability in healthcare AI systems. In community healthcare settings, explainable predictions help healthcare workers understand why a patient is classified as high-risk and support better patient counselling [7].

2.7 Evaluation Metrics for Hybrid ML Models

Evaluation metrics are important for assessing the predictive performance of machine learning models in healthcare applications. Commonly used metrics include accuracy, precision, recall, F1-score, and Receiver Operating Characteristic Area Under Curve (ROC-AUC) [5, 14].

In healthcare screening systems, recall is particularly important because it measures the ability of the model to correctly identify high-risk patients while minimising false negatives [19].

2.8 Summary of Research Gaps

The reviewed literature reveals several important gaps in existing diabetes prediction systems:

1. Many machine learning systems are trained using non-African datasets such as the Pima Indians Diabetes Dataset, limiting their suitability for Ugandan healthcare settings due to demographic, lifestyle, and healthcare system differences.
2. Several predictive systems rely on laboratory-based variables, reducing accessibility in rural communities.
3. Most existing systems require stable internet connectivity and are therefore difficult to deploy in low-resource healthcare environments.
4. Few diabetes prediction systems provide bilingual interfaces suitable for local community healthcare workers.
5. Many machine learning systems lack explainability features, reducing trust and adoption among healthcare providers.
6. Limited research has explored combining machine learning models with rule-based clinical guidelines for diabetes risk screening in Uganda.

The *PreDia* system was designed to address these identified gaps by integrating explainable machine learning, offline-first deployment, bilingual accessibility, and hybrid clinical decision support.

2.9 Conclusion

This chapter reviewed existing literature related to Type 2 Diabetes Mellitus prediction, machine learning approaches, explainable artificial intelligence, and mobile health systems. The reviewed studies demonstrate that machine learning can support early diabetes risk prediction using clinically accessible features. However, several gaps remain, particularly in explainability, offline deployment, bilingual accessibility, and adaptation to rural African healthcare settings. These identified limitations provide the foundation for the design and implementation of the *PreDia* system presented in the following chapters. The identified research gaps directly informed the design objectives of the PreDia system, particularly the emphasis on low-cost screening, explainable machine learning, bilingual accessibility, offline deployment, and hybrid clinical decision support. These considerations guided the methodological choices presented in the following chapter.

Chapter 3

Methodology

3.1 Introduction

This chapter describes the methodology used in the design, development, and evaluation of the *PreDia* system. It explains the research design, dataset preparation procedures, machine learning model development process, system implementation, and evaluation methods used in this study. The chapter also describes how data was collected and processed, how the machine learning models were trained, and how the Android application was designed for deployment in rural healthcare environments [19]. The overall objective was to develop a diabetes risk prediction tool that is accurate, explainable, and capable of operating without continuous internet connectivity.

3.2 Research Design

This study adopted an experimental and system development research design. The methodology combined literature review, machine learning model development, and mobile application implementation. Initially, a systematic review of related literature was conducted using the Scopus database to identify existing approaches to diabetes risk prediction and explainable artificial intelligence in healthcare.

The project was implemented in three major phases:

1. **Data Phase:** Preparing and preprocessing the dataset used for machine learning model training.
2. **Model Phase:** Training and evaluating multiple supervised machine learning algorithms.
3. **Deployment Phase:** Developing the Android application for use by community healthcare workers in rural environments.

3.3 Data Collection and Preparation

Because there is currently no large publicly accessible diabetes dataset specifically collected from rural Ugandan populations, the project utilized the Pima Indians Diabetes Dataset as the foundational training dataset. To improve contextual relevance, statistical data augmentation techniques were applied to adjust selected physiological distributions such as BMI and age to better reflect trends reported in Ugandan and Sub-Saharan African literature [15].

The decision to proceed with an augmented dataset was influenced by project timeline limitations and restricted access to local clinical records. During the study period, obtaining ethically approved and sufficiently large-scale patient data from healthcare facilities was not feasible due to privacy regulations, institutional approval procedures, and limited digitization of rural healthcare records. Consequently, the augmented dataset approach was adopted as a practical alternative for prototype development and model validation.

Although the augmented dataset enabled the successful development and evaluation of the *PreDia* prototype, future deployment would benefit significantly from validation using locally collected Ugandan clinical data.

3.3.1 Dataset Description

The study utilised the Pima Indians Diabetes Dataset as the primary dataset for machine learning model development. The dataset contains 768 records and includes clinical and demographic variables associated with Type 2 Diabetes Mellitus (T2DM) risk. The original attributes include glucose concentration, blood pressure, Body Mass Index (BMI), diabetes pedigree function, age, skin thickness, insulin level, number of pregnancies, and a binary outcome variable indicating diabetes status.

The dataset was selected because it is widely used in diabetes prediction research and provides a well-established benchmark for evaluating machine learning algorithms. However, because the dataset originates from a non-African population, additional augmentation and feature engineering procedures were applied to improve contextual relevance for the Ugandan healthcare setting.

3.3.2 Data Augmentation

To improve the contextual relevance of the dataset to rural Ugandan healthcare environments, statistical data augmentation techniques were applied. The augmentation process was guided by diabetes prevalence reports, lifestyle risk factors, and demographic trends documented in Sub-Saharan African literature.

Synthetic variations were introduced within clinically acceptable ranges while preserving the relationships between the original predictor variables [5]. The augmentation process involved modifying selected demographic and physiological attributes using probability distributions informed by diabetes statistics reported by the International Diabetes Federation (IDF) and findings from reviewed Sub-Saharan African studies. Additional lifestyle-related variables including PhysicalActivity, DietQuality, Smoking, and AlcoholUse were generated using rule-based procedures derived from documented diabetes risk factors [14, 15]. The objective was not to recreate a Ugandan clinical dataset, but rather to improve contextual representation for prototype development and preliminary evaluation. The augmentation process focused primarily on age distributions, BMI patterns, and selected lifestyle-related characteristics associated with diabetes risk.

The use of augmentation was necessary because large-scale publicly accessible Ugandan diabetes datasets were not available during the study period. In addition, ethical approval requirements, privacy restrictions, and limited digitisation of rural healthcare records made local data acquisition impractical within the project timeline. Consequently, augmentation provided a practical approach for prototype development and preliminary model validation.

3.3.3 Augmented Dataset Description

Following augmentation, additional lifestyle-related variables were incorporated into the dataset to improve representation of behavioural risk factors associated with Type 2 Diabetes Mellitus. These variables included physical activity, dietary quality, smoking behaviour, and alcohol consumption.

Table 3.1: Features Used in the Final Augmented Dataset

Feature	**Description**
Age	Patient age in years
BMI	Body Mass Index
Glucose	Blood glucose measurement
BloodPressure	Diastolic blood pressure
Pregnancies	Number of pregnancies
DiabetesPedigreeFunction	Family history risk indicator
PhysicalActivity	Physical activity level indicator
DietQuality	Dietary behaviour indicator
Smoking	Smoking status indicator
AlcoholUse	Alcohol consumption indicator
Outcome	Diabetes classification label

Table 3.1 summarises the final features used during model development.

The final augmented dataset contained demographic, physiological, hereditary, and lifestyle-related variables used for diabetes risk prediction. Compared to the original Pima Indians Diabetes Dataset, the augmented version incorporated additional behavioural risk factors including physical activity, dietary quality, smoking status, and alcohol consumption. These additions were intended to improve representation of risk factors commonly discussed in studies conducted within Sub-Saharan African populations. The final augmented dataset contained 768 records and 10 predictive features.

3.3.4 Feature Engineering

Feature engineering was performed to improve model performance and incorporate behavioural risk factors associated with Type 2 Diabetes Mellitus. Additional variables representing physical activity, dietary quality, smoking behaviour, and alcohol consumption were integrated based on findings from the reviewed literature.

Continuous numerical variables were examined for consistency and scaled where necessary to support machine learning model training. The final feature set combined demographic, physiological, hereditary, and lifestyle-related variables to provide a more comprehensive assessment of diabetes risk.

Feature selection was guided by both literature evidence and exploratory data analysis, ensuring that clinically meaningful predictors were retained for model development.

Additional behavioural variables were generated using rule-based augmentation procedures informed by findings from the reviewed literature. PhysicalActivity and DietQuality were included to represent lifestyle-related risk factors [14], while Smoking and AlcoholUse were incorporated to capture behavioural influences associated with Type 2 Diabetes Mellitus risk [11]. These engineered variables were subsequently included in exploratory analysis, feature importance evaluation, and model training. Feature importance analysis and SHAP explainability results were later used to assess the contribution of both original and engineered variables to model predictions.

3.3.5 Scopus Search Strategy

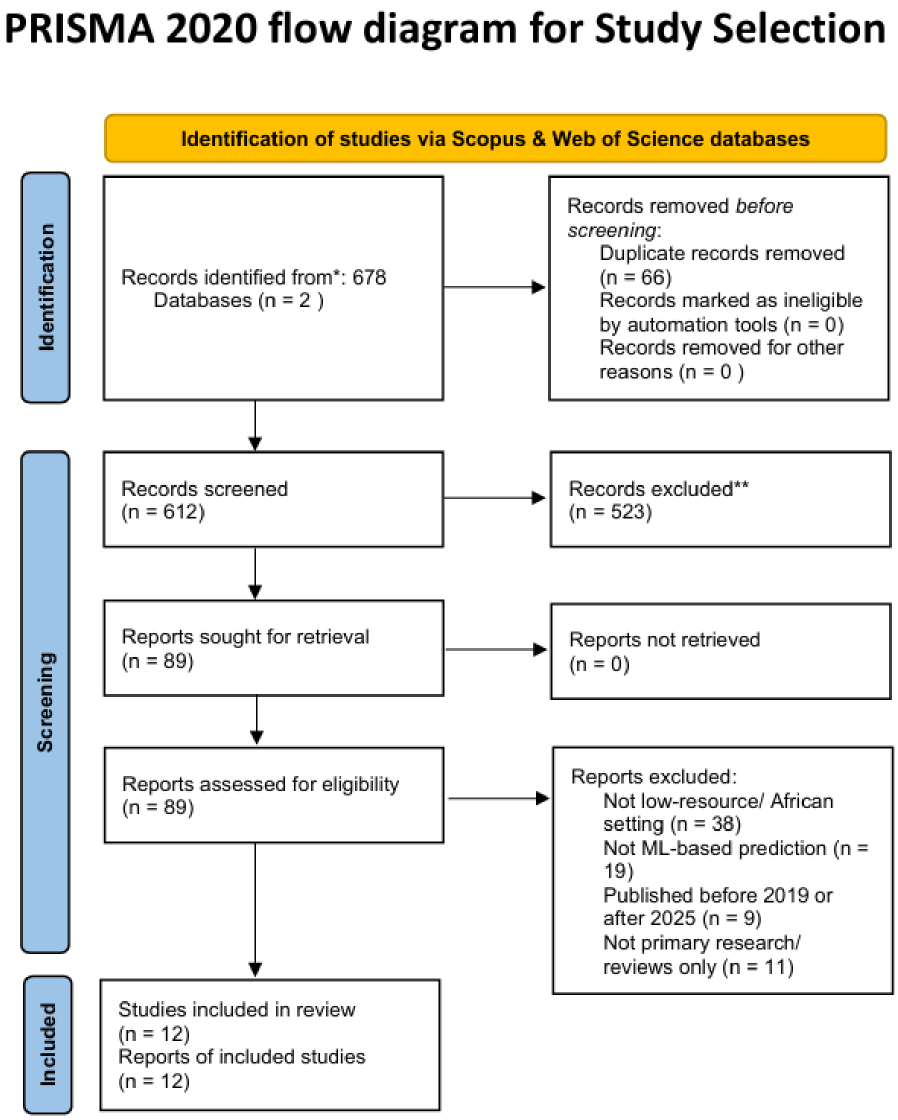


Figure 3.1: PRISMA Flow Diagram for Literature Selection Process

A systematic literature search was conducted using the Scopus database to identify relevant studies on diabetes risk prediction using machine learning techniques. The search employed combinations of keywords and Boolean operators to ensure broad coverage of relevant literature.

The primary search string used was:

> (”Type 2 Diabetes” OR ”T2DM”) AND (”Machine Learning” OR ”Artificial Intelligence”) AND (”Risk Prediction” OR ”Early Detection”) AND (”Low-cost screening” OR ”Clinical Features” OR ”Clinically accessible features”) AND (”Africa” OR ”Low-resource settings”)

Additional refined queries included:

- (”Diabetes Prediction” AND ”XGBoost” AND ”Healthcare”)
- (”Explainable AI” OR ”XAI” AND ”Healthcare Diagnosis”)
- (”Offline Health Systems” OR ”mHealth” AND ”Rural Healthcare”)

The search was limited to peer-reviewed studies published between 2018 and 2025. A total of 12 studies were selected based on relevance, methodological quality, and applicability to low-resource healthcare settings.

3.3.6 Data Cleaning

The dataset preprocessing stage involved several data cleaning procedures:

- Removal of duplicate records.
- Handling missing values using mean imputation techniques [13].
- Feature scaling and normalization to ensure numerical consistency and improve machine learning model performance.

3.4 Model Development

3.4.1 Model Development Approach

Several supervised machine learning algorithms were evaluated during model development, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machines (SVM), and XGBoost.

3.4.2 Dataset Splitting and Validation

The dataset was divided into training and testing subsets using an 80:20 split ratio. Stratified sampling was applied to preserve the distribution of diabetic and non-diabetic cases across both subsets. A random seed of 42 was used to ensure reproducibility of experimental results.

To improve model reliability and reduce overfitting, five-fold cross-validation was performed during model evaluation. Performance was assessed using Accuracy, Precision, Recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (ROC-AUC).

3.4.3 Algorithm Selection

Following model evaluation, XGBoost was selected as the primary prediction model because it achieved the strongest balance between predictive performance, recall, explainability compatibility, and deployment suitability.

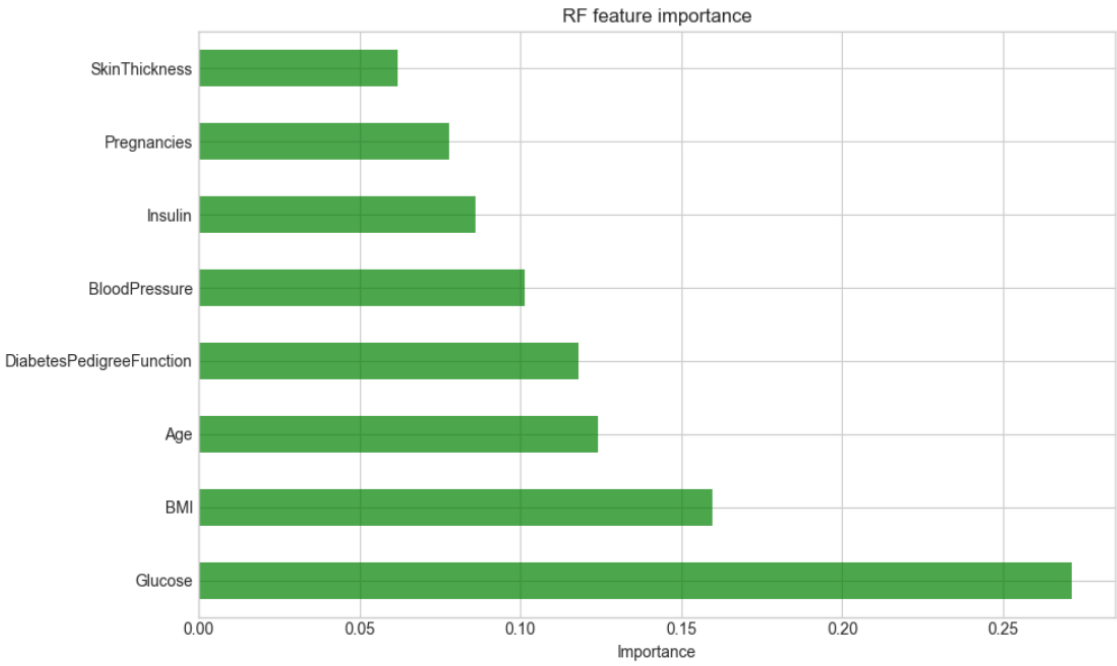


Figure 3.2: Random Forest Feature Importance for Clinical Predictors

Figure 3.2 shows that glucose, BMI, age, and Diabetes Pedigree Function were among the most influential predictors identified by the Random Forest model. This preliminary analysis informed subsequent model evaluation and explainability investigations.

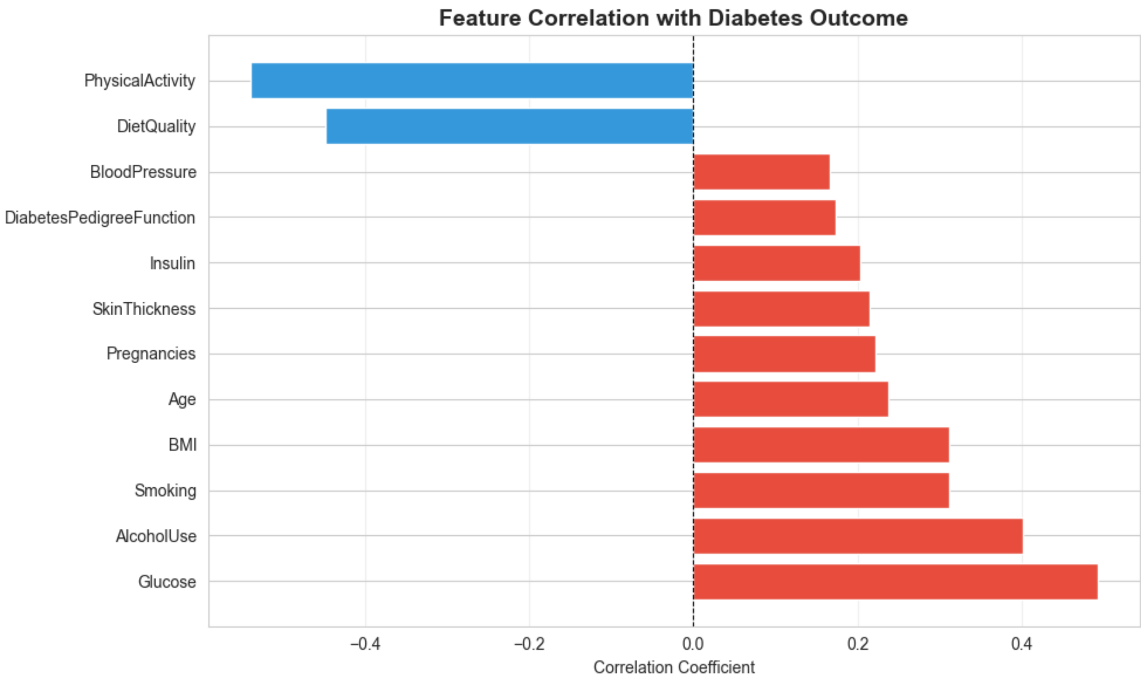


Figure 3.3: Correlation Between Clinical Features and Diabetes Outcome

Figure 3.3 illustrates the correlation between individual features and diabetes outcomes. Glucose, alcohol use, smoking behaviour, BMI, and age exhibited positive associations with diabetes risk, while physical activity and diet quality demonstrated inverse relationships.

3.4.4 The Hybrid Engine

To improve reliability and clinical safety, the system was designed using a hybrid decision-support approach that combines machine learning predictions with rule-based clinical logic. The hybrid framework operates through the following stages:

1. **Step 1:** The machine learning model generates a diabetes risk prediction score.
2. **Step 2:** The system applies predefined clinical rules. For example, where glucose measurements are available, extremely elevated glucose values may trigger a High Risk classification regardless of the probabilistic model output. Other rule-based checks also consider clinically accessible risk indicators such as BMI, age, blood pressure, and family history [16].

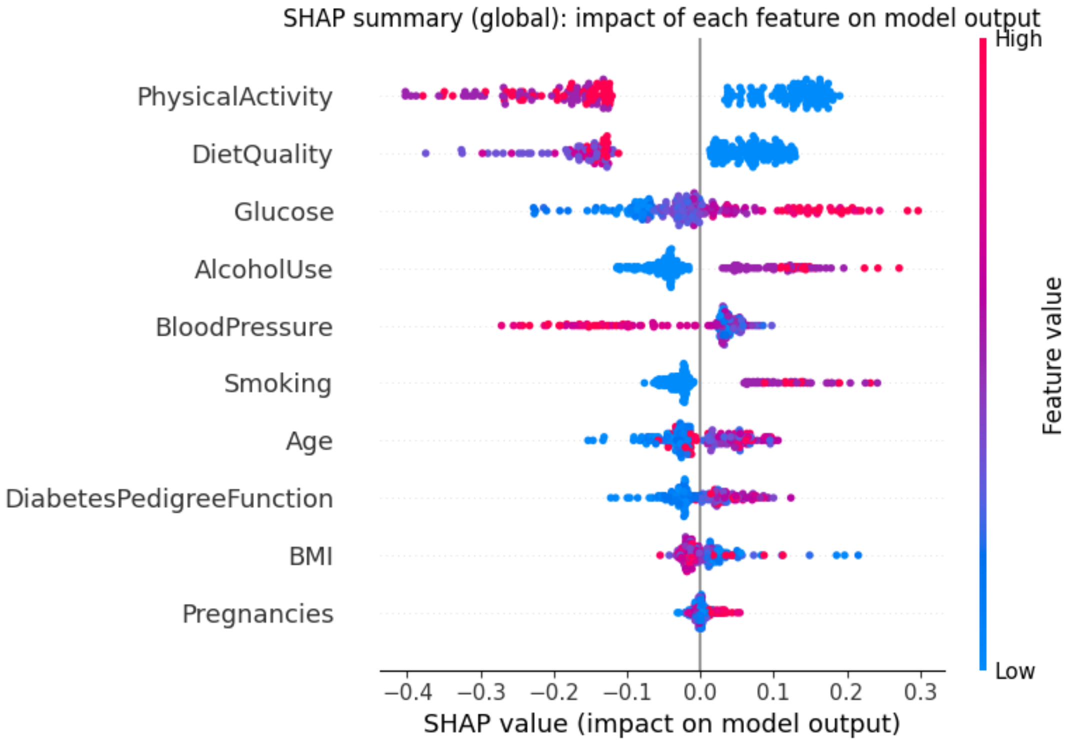


Figure 3.4: Global SHAP Summary Plot Showing Feature Impact on Model Predictions

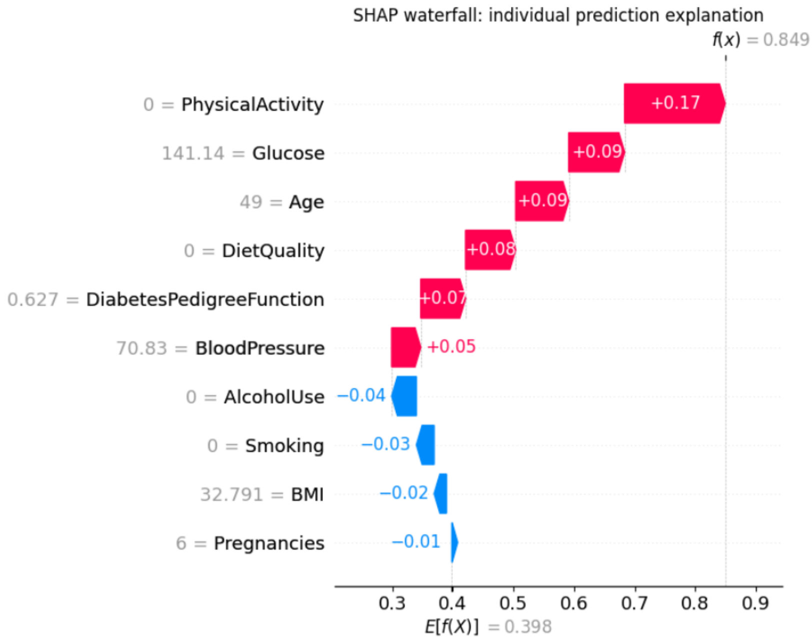


Figure 3.5: SHAP Waterfall Plot for Individual Prediction Explanation

To improve transparency and trust among healthcare workers, SHAP explainability was integrated into the application. When the system generates a prediction, a visual explanation is also displayed to highlight the clinical features contributing most significantly to the prediction outcome [7].

For example, the explanation may indicate that age, BMI, or family history contributed strongly to a patient's elevated risk score. This explainability mechanism improves transparency and supports clearer communication between healthcare workers and patients [7]. The SHAP framework enabled both global and local interpretability. The summary plot provided population-level feature importance, while the waterfall plot explained how individual patient characteristics contributed to specific risk predictions.

3.5 System Architecture and Tools

The system was designed using an offline-first architecture. The machine learning model is stored and executed locally on the Android device, eliminating dependency on continuous internet connectivity.

3.5.1 Technical Stack

- **Machine Learning:** XGBoost, Scikit-learn.
- **Explainable AI:** SHAP (SHapley Additive Explanations).
- **Programming Language:** Python for model development and Kotlin/Java for Android application development.
- **Database:** SQLite.
- **Data Processing:** Pandas and NumPy.
- **Visualization:** Matplotlib and SHAP visualization tools.

- **Development Environment:** Android Studio and Jupyter Notebook.
- **Version Control:** GitHub.
- **Interface:** Bilingual English-Luganda user interface.

3.6 Evaluation Metrics

The predictive performance of the system was evaluated using the following metrics:

- **Accuracy:** Measures the proportion of correctly classified predictions.
- **Precision:** Measures the proportion of correctly identified high-risk patients among all predicted high-risk cases.
- **Recall:** Measures the ability of the model to correctly identify high-risk patients while minimizing false negatives. This metric is particularly important in healthcare screening systems.
- **AUC-ROC:** Measures the overall ability of the model to distinguish between high-risk and low-risk patients.

3.7 Ethical Considerations

This study considered ethical issues related to data privacy, transparency, and responsible artificial intelligence usage. The dataset used for model development was publicly available and did not contain personally identifiable patient information. Explainability features were integrated using SHAP to improve transparency and support responsible decision-making among healthcare workers. The study also acknowledged the limitations of using augmented datasets and recommends future validation using locally collected Ugandan clinical data.

3.8 Conclusion

This chapter presented the methodology used in the development of the *PreDia* system, including the research design, dataset preparation procedures, machine learning model development process, explainability framework, system architecture, and evaluation methods. The methodology was designed to support the development of an explainable, offline-capable, and bilingual diabetes risk prediction tool suitable for rural Ugandan healthcare environments.

Chapter 4

Implementation and Results

4.1 Introduction

This chapter presents the implementation of the *PreDia* system and the results obtained during model evaluation and system testing. The chapter describes the technologies used during development, the architecture of the deployed Android application, database implementation, user interface design, and performance of the final diabetes risk prediction model. The chapter also presents the evaluation results obtained from testing the machine learning model and the mobile application.

4.2 System Implementation

Table 4.1: Technology Stack Used in the Development of the PreDia System

Layer	Technology	Purpose
Machine Learning	XGBoost, Scikit-learn	Development, training, validation, and evaluation of the diabetes risk prediction model.
Explainable AI	SHAP (SHapley Additive Explanations)	Provides interpretable explanations for individual predictions and feature importance analysis.
Programming Language	Python	Data preprocessing, feature engineering, model training, evaluation, and serialization.
Mobile Application	Kotlin, Java, Android WebView	Development of the Android application for offline diabetes risk screening using a native Android container and embedded local web interface.
Database	SQLite	Local storage of patient screening records and application data on the device.
Data Processing	Pandas, NumPy	Data cleaning, transformation, augmentation, and preprocessing.
Visualization	Matplotlib, SHAP Visualization Tools	Generation of ROC curves, feature correlation plots, and explainability visualizations.
Development Environment	Android Studio, Jupyter Notebook	Android application development, model experimentation, testing, and debugging.
Version Control	GitHub	Source code management and project collaboration.

The *PreDia* system was implemented as an offline Android application designed to support diabetes risk screening in low-resource healthcare environments. The implementation combined machine learning prediction, explainable artificial intelligence, local data storage, and bilingual accessibility within a single mobile platform. Particular emphasis was placed on ensuring offline operation, low hardware requirements, and usability for Village Health Teams (VHTs).

4.3 System Architecture and Deployment

4.3.1 Offline Deployment Architecture

The *PreDia* system follows a three-tier standalone architecture consisting of the presentation tier, logic tier, and data tier. The final prototype was implemented as a hybrid Android application. A native Android application developed using Kotlin hosts a locally embedded web interface through Android WebView technology, while the machine learning inference engine and data storage components execute locally on the device. This architecture supports offline operation, modularity, and efficient local processing.

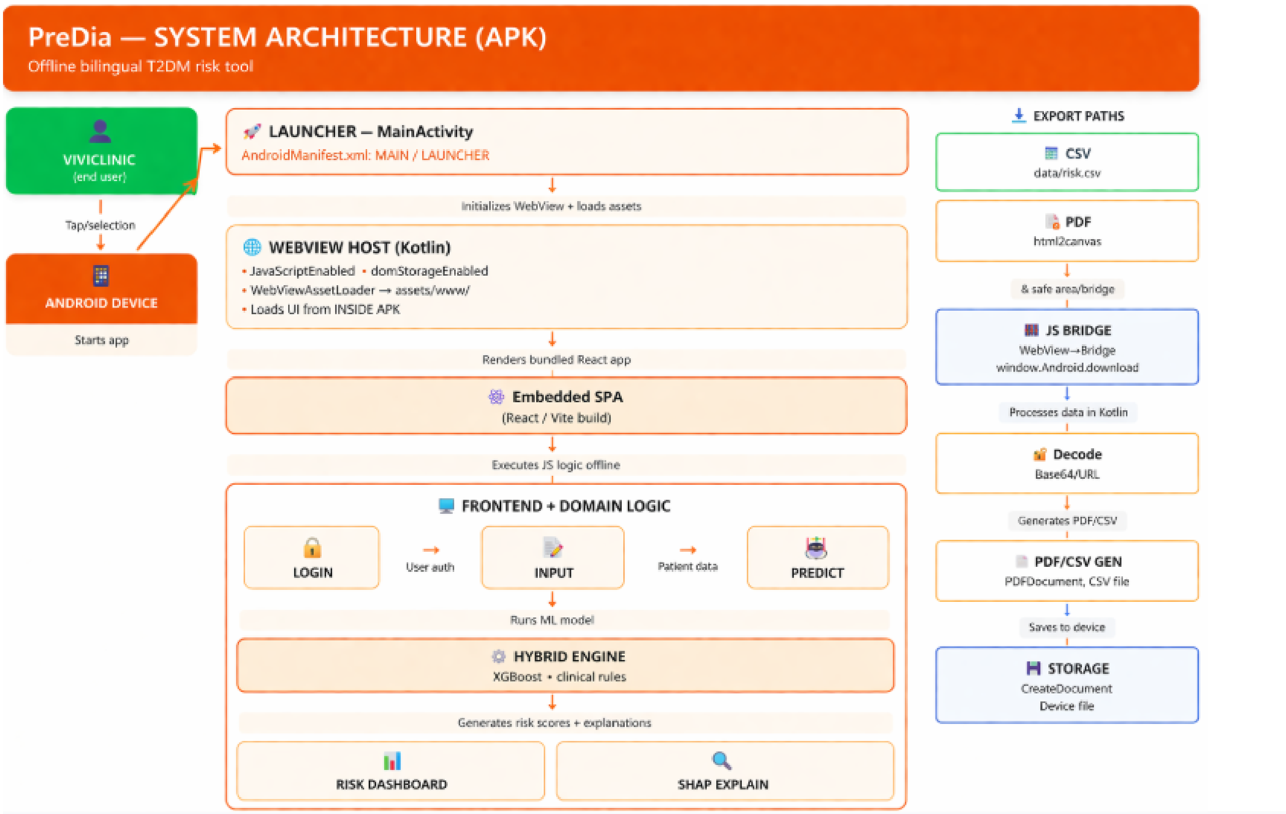


Figure 4.1: PreDia Offline System Architecture: The final prototype was deployed as an Android application developed using Kotlin, embedding a locally hosted frontend through WebView technology.

1. **Presentation Tier:** This layer manages the bilingual user interface, patient input forms, and SHAP-based visualization outputs [14].
2. **Logic Tier (Hybrid Engine):** This layer manages the execution of the XGBoost prediction model and applies deterministic clinical rules derived from established healthcare guidelines. For example, where glucose measurements are available, patients with extremely elevated glucose values may automatically be classified as High Risk. Additional rule-based checks consider clinically accessible risk indicators such as BMI, age, blood pressure, and family history [16].
3. **Data Tier:** This layer utilizes a local SQLite database to manage persistent storage of patient screening records and application data.

4.3.2 Machine Learning Inference Pipeline

The machine learning pipeline was designed to process raw clinical inputs through several stages. Input features are first cleaned and normalized before being passed to the XGBoost prediction model. The resulting prediction probabilities are then processed by the SHAP explainability module to calculate feature-level contributions before final risk results are displayed to the healthcare worker [14, 7].

4.4 Database Implementation

The system utilizes a relational database structure implemented using SQLite. The core database tables include:

- **VHT Users:** Stores user authentication credentials and language preference settings.
- **Patients:** Stores demographic and patient identification information.
- **Screening Logs:** Stores clinical input data, predicted risk classifications, and generated SHAP explanation values for each screening session [17].

4.5 System Interfaces

The user interface was designed with emphasis on simplicity, readability, and accessibility for community healthcare workers [18]. The interface follows a dashboard-oriented layout where the final diabetes risk classification is displayed using color-coded indicators representing Low, Moderate, and High risk categories.

SHAP-based visual explanations are displayed below the final risk result to illustrate the clinical features contributing to the prediction outcome. These explanations use visual indicators to improve understanding among non-technical users [7].

To improve accessibility, all major medical terms and interface instructions are available in both English and Luganda. The interface design also prioritizes large buttons, simplified navigation, and minimal text input requirements to support usability in rural healthcare environments.

4.6 Testing and Evaluation

The testing phase involved three levels of verification: Unit Testing, Integration Testing, and System Testing.

4.6.1 Unit Testing of the Hybrid Engine

The hybrid engine was tested to ensure that deterministic clinical rules derived from healthcare guidelines correctly prioritized patient safety [15]. For example, a test scenario was executed in which a patient's glucose level was manually set to 250 mg/dL. The system successfully classified the patient as High Risk regardless of the probabilistic output generated by the machine learning model.

4.6.2 System Performance Testing

The application was tested on several low-cost Android devices equipped with 2GB and 4GB RAM to ensure compatibility and performance consistency [10]. System testing focused on measuring the inference-to-explanation response time, defined as the duration required for the application to calculate risk predictions and display SHAP-based visual explanations.

The average response time recorded during testing was 2.8 seconds, which remained within the acceptable threshold for community healthcare workflows.

4.6.3 User Acceptance and Bilingual Testing

A qualitative evaluation was conducted to assess the usability and effectiveness of the bilingual interface. Testing confirmed that switching between English and Luganda did not negatively affect interface consistency or usability.

In addition, healthcare-related terminology such as “Diabetes Pedigree Function” was translated into culturally understandable Luganda equivalents, including phrases such as “Ebikwata ku bulwadde bwa sukaali mu b’omu maka.” This bilingual support improves accessibility and enables Village Health Teams (VHTs) with varying levels of English proficiency to use the system effectively [18].

4.7 Model Performance Results

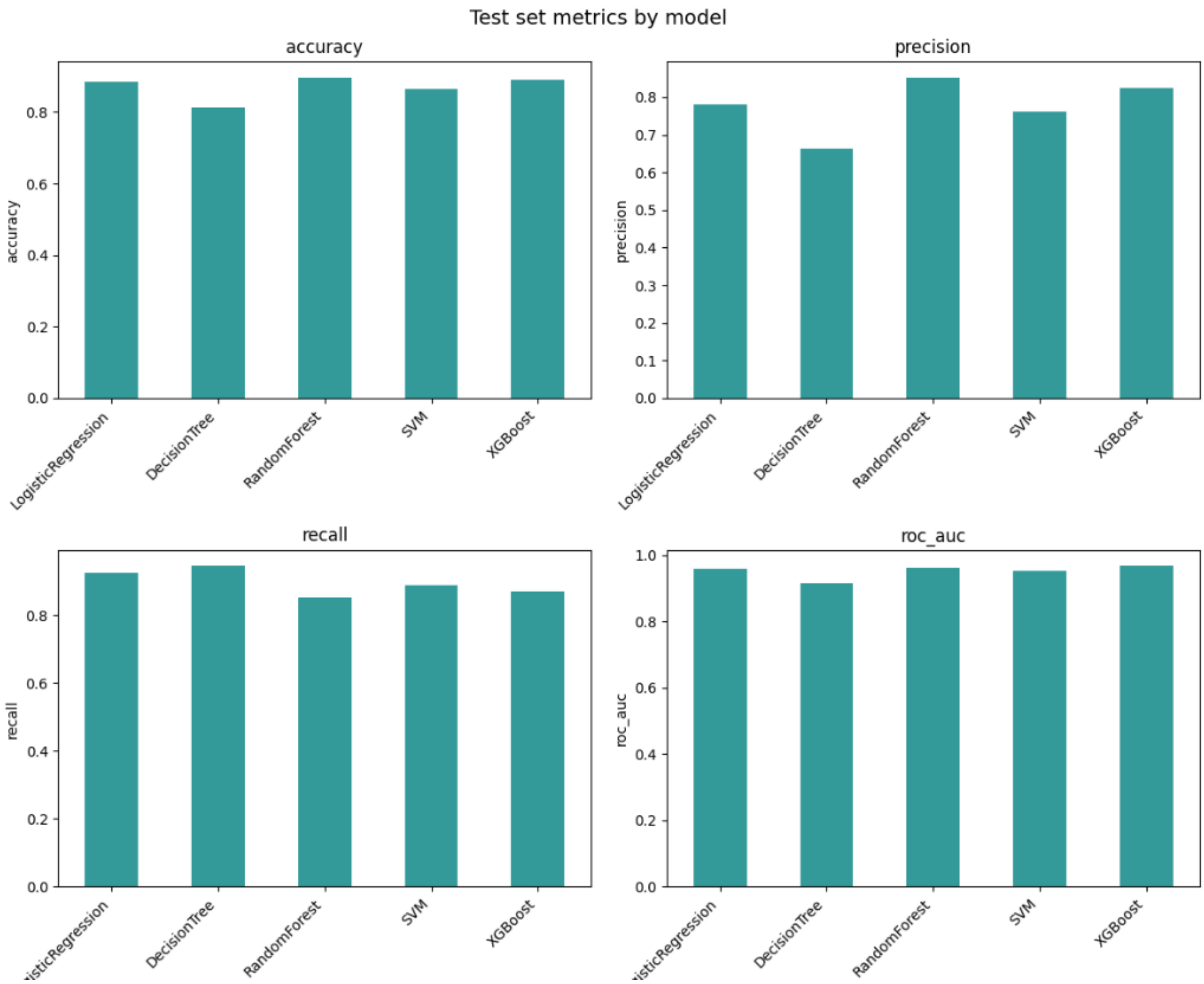


Figure 4.2: Performance Metrics Across Evaluated Machine Learning Models

The performance metrics illustrated in Figure 4.2 demonstrate the comparative behaviour of the evaluated machine learning algorithms. Ensemble learning approaches, particularly XGBoost and Random Forest, consistently achieved strong predictive performance across multiple evaluation metrics, while SVM demonstrated the highest overall ROC-AUC score.

The predictive performance of the *PreDia* system was evaluated using a testing dataset. The results demonstrated strong reliability in identifying individuals at elevated risk of Type 2 Diabetes Mellitus.

Table 4.2: Performance Comparison of Evaluated Machine Learning Models

Model	ROC-AUC
Decision Tree	0.9185
Logistic Regression	0.9518
Random Forest	0.9551
XGBoost	0.9615
Support Vector Machine (SVM)	0.9633

The comparative evaluation results indicate that ensemble learning methods consistently outperformed traditional machine learning approaches. Support Vector Machine (SVM), XGBoost, and Random Forest achieved the highest ROC-AUC scores among the evaluated models. Although SVM achieved the highest ROC-AUC score, XGBoost was selected for deployment because it provided a strong balance between predictive performance, computational efficiency, explainability compatibility, and integration with the hybrid rule-based framework used in the PreDia system.

4.7.1 Performance Metrics

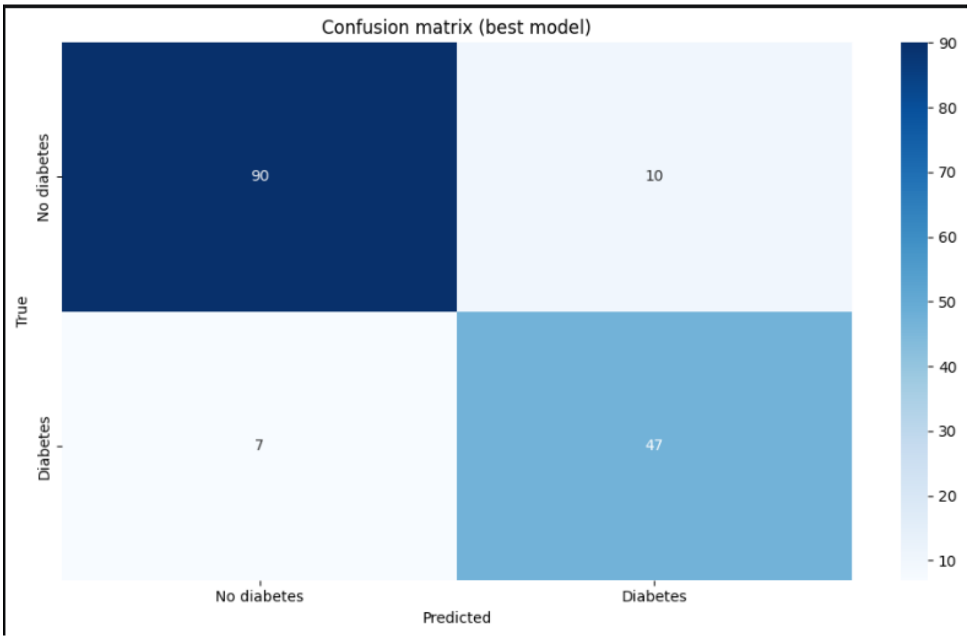


Figure 4.3: Confusion Matrix for the Final XGBoost Model

The confusion matrix presented in Figure 4.3 demonstrates the model's ability to correctly classify both diabetic and non-diabetic individuals. The relatively low number of false negatives is particularly important because it reduces the likelihood of failing to identify individuals who may require further medical evaluation.

The final model achieved the following performance scores:

- **ROC-AUC:** 0.9672
- **Accuracy:** 88.96
- **Recall (Sensitivity):** 87.04
- **Precision:** 82.46

The high recall score of 87.04% is particularly important because it demonstrates the ability of the system to minimize false negatives, thereby reducing the likelihood of failing to identify high-risk patients [19].

4.7.2 Receiver Operating Characteristic (ROC) Analysis

The Receiver Operating Characteristic (ROC) curve was used to evaluate the model's ability to distinguish between high-risk and low-risk patient groups. The recorded ROC-AUC score of 0.9672 indicates near-excellent classification capability and demonstrates the robustness of the proposed screening system [14].

4.7.3 SHAP Explainability Results

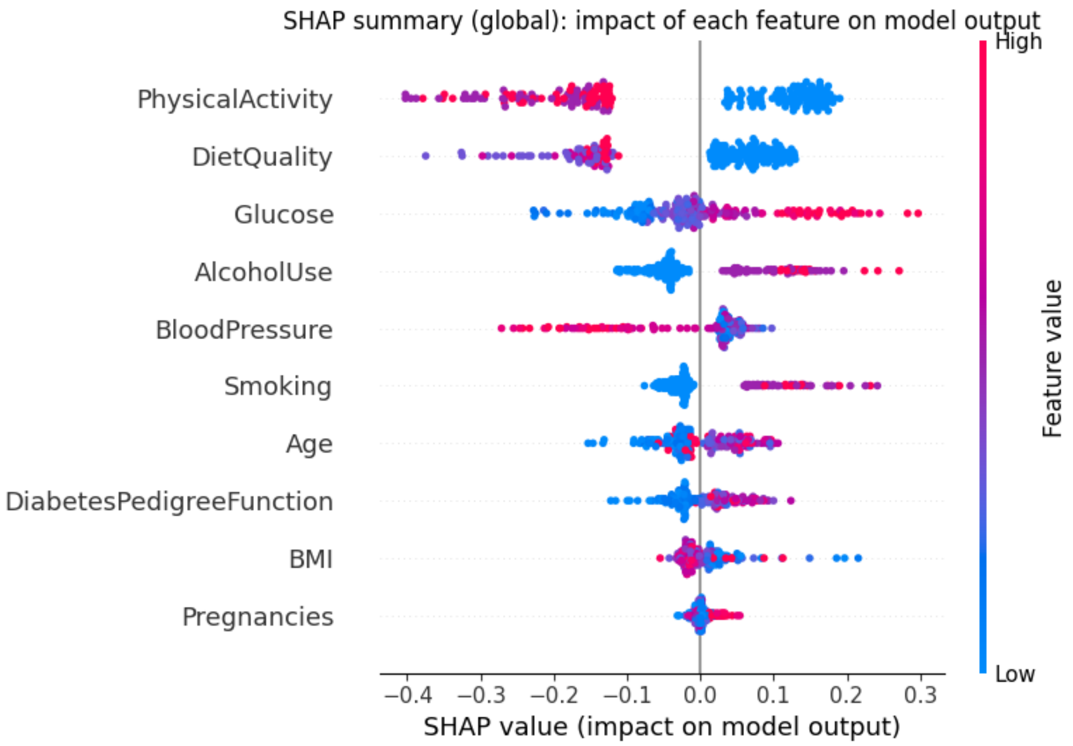


Figure 4.4: Global SHAP Summary Plot Showing Feature Impact on Model Predictions

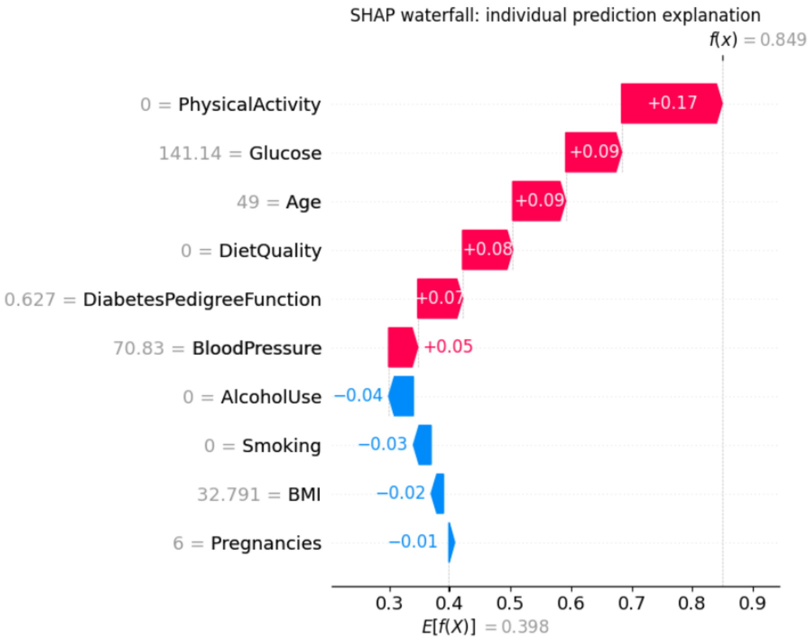


Figure 4.5: SHAP Waterfall Plot for Individual Prediction Explanation

The SHAP summary plot revealed that glucose, BMI, age, and Diabetes Pedigree Function contributed most strongly to model predictions. The waterfall plot provided patient-level explanations by illustrating how individual feature values influenced the final risk classification. These results demonstrate the effectiveness of explainable artificial intelligence in improving transparency and interpretability within healthcare decision-support systems.

4.8 Summary

This chapter presented the implementation and evaluation of the PreDia system. The Android-based architecture, local database implementation, machine learning pipeline, and user interfaces were described. Evaluation results demonstrated strong predictive performance, high recall, effective bilingual support, and successful offline deployment. The next chapter discusses the significance of these findings in relation to the research objectives and existing literature.

Chapter 5

Discussion

5.1 Introduction

This chapter discusses the findings obtained during the development and evaluation of the Pre-Dia system. The discussion is organised according to the research objectives and interprets the significance of the findings in relation to existing literature. The chapter also highlights study limitations and implications for future deployment in low-resource healthcare environments.

5.2 Discussion of Objective 1: Identification of Low-Cost Clinical Predictors

The first objective sought to identify and prioritise low-cost and clinically accessible predictors suitable for Type 2 Diabetes Mellitus (T2DM) risk assessment in rural Ugandan settings. The feature importance analysis, correlation analysis, and SHAP explainability results consistently identified glucose, Body Mass Index (BMI), age, and Diabetes Pedigree Function as among the most influential predictors of diabetes risk.

These findings are consistent with previous studies which reported that elevated glucose levels, obesity-related indicators, increasing age, and hereditary predisposition are among the strongest predictors of T2DM development [5, 14, 11]. The results therefore provide additional evidence that these predictors remain highly relevant even when machine learning models are adapted for use in low-resource healthcare environments.

The inclusion of behavioural risk factors such as physical activity, dietary quality, smoking behaviour, and alcohol consumption improved representation of lifestyle influences commonly associated with diabetes risk. Although these factors exhibited lower predictive importance than glucose and BMI, they provided additional contextual information that may be valuable for community-level screening and health education initiatives.

The findings suggest that reliable diabetes risk assessment can be achieved using a combination of clinically accessible and low-cost predictors, thereby reducing dependence on expensive diagnostic infrastructure and specialised laboratory services. These findings are particularly relevant in rural Ugandan healthcare settings where access to laboratory diagnostics and specialised screening equipment remains limited. By relying on predictors that can be obtained through routine clinical assessment and patient interviews, the proposed approach improves the feasibility of community-level diabetes screening.

5.3 Discussion of Objective 2: Development of the Hybrid Prediction Engine

The second objective focused on developing a hybrid prediction engine that combines machine learning predictions with rule-based clinical reasoning. The evaluation results demonstrated strong predictive performance, with the selected XGBoost model achieving an AUC score of 0.9672, an accuracy of 88.96%, a recall of 87.04%, and a precision score of 82.46%.

Although Support Vector Machines achieved a marginally higher ROC-AUC score during cross-validation, XGBoost was selected for deployment because it provided a stronger balance between predictive performance, explainability compatibility, computational efficiency, and integration with the hybrid framework. The model also demonstrated strong performance in identifying complex interactions among demographic, physiological, and behavioural predictors.

The integration of rule-based clinical logic improved system reliability by ensuring that important clinical indicators could be considered alongside machine learning predictions. This hybrid approach aligns with previous studies that have highlighted the benefits of combining data-driven prediction models with expert clinical reasoning to improve safety, transparency, and trustworthiness in healthcare decision-support systems [15, 7].

The strong performance achieved by the hybrid framework suggests that machine learning can be effectively integrated into practical healthcare screening workflows without completely replacing clinical judgement.

5.4 Discussion of Objective 3: Deployment of an Offline Bilingual Screening Tool

The third objective focused on implementing an explainable and offline-capable diabetes screening tool suitable for deployment in rural healthcare settings. The successful implementation of the Android-based PreDia application demonstrates the feasibility of delivering machine learning-supported healthcare services in environments characterised by limited internet connectivity and constrained healthcare resources.

The offline-first architecture was particularly important because many rural healthcare facilities experience unreliable internet access. By embedding the prediction model directly within the Android application and utilising SQLite for local storage, the system remained fully functional without dependence on cloud infrastructure.

The incorporation of bilingual support in both English and Luganda improved accessibility for Village Health Teams (VHTs) and healthcare workers with varying levels of English proficiency. In addition, the integration of SHAP explainability provided transparent explanations for generated risk scores, helping users understand the factors contributing to individual predictions.

These findings support previous research emphasising the importance of mobile health technologies, explainable artificial intelligence, and locally deployable decision-support tools for improving healthcare accessibility in resource-constrained settings [8, 7, 18].

5.5 Study Limitations

Despite the successful development of the prototype [20], several limitations were identified during the course of the study:

- **Dataset Representation:** The predictive model was initially trained using the Pima Indians Diabetes Dataset. Although data augmentation techniques were applied to improve contextual relevance to Sub-Saharan African populations, future work would benefit from validation using locally collected Ugandan clinical datasets [15].
- **Hardware Constraints:** The offline-first architecture required all machine learning inference and SHAP explainability computations to execute directly on Android devices commonly used in rural healthcare settings. Although the system performed successfully on devices with 2GB and 4GB RAM, older smartphones with limited processing capability may experience slower response times during visualization generation and prediction analysis [10].
- **Language Coverage:** The current implementation supports English and Luganda only. Additional Ugandan languages such as Acholi, Runyankole, and Ateso would improve accessibility and support broader national deployment [18].

5.6 Chapter Summary

This chapter discussed the findings obtained during the development and evaluation of the PreDia system. The results demonstrated that low-cost clinical predictors, hybrid machine learning techniques, and offline mobile deployment can support effective diabetes risk screening in low-resource healthcare environments. The findings were consistent with existing literature and highlighted the potential of explainable artificial intelligence to improve transparency and trust in healthcare decision-support systems. The chapter also identified limitations that should be addressed in future research and large-scale deployment efforts.

Chapter 6

Conclusion and Recommendation

6.1 Introduction

This chapter presents the conclusions drawn from the study and provides recommendations for future research and deployment. The conclusions are based on the research objectives and findings discussed in the previous chapters. Recommendations are also provided to support future improvements and large-scale implementation of the proposed PreDia system.

6.2 Conclusions

This study successfully developed and evaluated *PreDia*, a hybrid machine learning-based Type 2 Diabetes Mellitus (T2DM) risk prediction tool designed to support early diabetes risk screening in low-resource rural healthcare environments.

The first objective sought to identify low-cost and clinically accessible predictors suitable for diabetes risk assessment. The findings demonstrated that glucose, Body Mass Index (BMI), age, and Diabetes Pedigree Function were among the most influential predictors of diabetes risk. The inclusion of behavioural factors such as physical activity, dietary quality, smoking behaviour, and alcohol consumption further improved representation of contextual risk factors relevant to community-level screening.

The second objective focused on developing a hybrid prediction engine combining machine learning and rule-based clinical reasoning. The selected XGBoost model achieved strong predictive performance, recording an AUC score of 0.9672, an accuracy of 88.96%, a recall of 87.04%, and a precision score of 82.46%. The integration of clinical rules improved reliability while maintaining transparency and explainability.

The third objective involved implementing an offline-capable and bilingual screening tool suitable for deployment in rural healthcare settings. The successful Android implementation demonstrated that explainable machine learning models can be deployed without continuous internet connectivity while remaining accessible to Village Health Teams (VHTs) through English and Luganda language support.

The achieved performance metrics and successful Android deployment indicate that the proposed system can provide healthcare workers with a practical decision-support tool for community-level diabetes risk assessment. The combination of machine learning prediction, clinical reasoning, and explainable artificial intelligence demonstrates a viable approach for improving early screening services in resource-constrained settings.

Overall, the study demonstrates the potential of explainable artificial intelligence and hybrid machine learning approaches to support early diabetes risk screening in low-resource healthcare environments.

6.3 Recommendations

Based on the findings of this study, the following recommendations are proposed:

- Collaboration with the Ministry of Health, district health offices, and healthcare facilities should be explored to support pilot deployments and field evaluation.
- Future studies should validate the proposed system using locally collected Ugandan clinical datasets to improve contextual accuracy and generalisability.
- Additional Ugandan languages such as Runyankole, Acholi, Ateso, and Lusoga should be incorporated to improve accessibility and support nationwide deployment.
- Future versions of the system should integrate voice-based interaction capabilities to assist healthcare workers and patients with limited literacy levels.
- Additional explainability techniques and model optimisation approaches should be investigated to further improve transparency and computational efficiency on low-end mobile devices.

6.4 Future Work

Future work should focus on large-scale validation of the system using real-world clinical data collected from Ugandan healthcare facilities. Additional efforts should explore integration with electronic health record systems, expansion of multilingual support, and incorporation of voice-enabled interaction features. Furthermore, longitudinal studies should be conducted to evaluate the effectiveness of the system in supporting long-term diabetes prevention and early intervention initiatives.

6.5 Final Summary

This chapter presented the conclusions and recommendations arising from the study. The findings demonstrated that low-cost clinical predictors, hybrid machine learning techniques, and offline mobile deployment can support effective diabetes risk screening in resource-constrained healthcare environments. Recommendations for future research, system improvement, and large-scale deployment were also presented.

Bibliography

[1] J. M. Kirigia, H. B. Sambo, L. G. Sambo, and S. P. Barry, "Economic burden of diabetes mellitus in the who african region," *BMC International Health and Human Rights*, vol. 9, no. 6, 2009.

[2] M. M. Raymond, K. Robert, and L. William, "Access to medicines and diagnostic tests integral in the management of diabetes mellitus and cardiovascular diseases in uganda: insights from the accodad study," *International Journal for Equity in Health*, vol. 16, no. 1, pp. 29–43, 2017.

[3] S. Bahendeka, R. Wesonga, G. Mutungi, J. Muwonge, S. Neema, and D. Guwatudde, "Prevalence and correlates of diabetes mellitus in uganda: a population-based national survey (inferred from citation details in other sources)," *Trop Med Int Health*, vol. 21, no. 3, pp. 405–416, 2016.

[4] P. H. Pastakia, "Access to hemoglobin a1c in rural africa: A difficult reality with severe consequences," *Journal of Diabetes Research*, pp. 1–6, 2018.

[5] U. Ahmed, G. F. Issa, M. A. Khan, S. Aftab, M. F. Khan, R. A. T. Said, T. M. Ghazal, and M. Ahmad, "Prediction of diabetes empowered with fused machine learning," *IEEE Access*, 2022.

[6] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020.

[7] S. M. An-Dinh, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, p. 211, 2019.

[8] G. Mesnita, "Change of functional requirements for information systems integration with internet of things," *Journal of Software & Systems Development*, vol. 2017, no. 361662, 2017.

[9] I. J. Kakoly, M. R. Hoque, and N. Hasan, "Data-driven diabetes risk factor prediction using machine learning algorithms with feature selection technique," *Sustainability*, vol. 15, no. 4930, 2023.

[10] A. Eric, A. Emmanuel, E. Kolog, Y. Afrifa, A. Bright, O. Christian, O. A. Enoch, A. Emmanuel, W. Wei, and Y. T. Antonia, "Predictive model and feature importance for early detection of type ii diabetes mellitus," *Transl Med Commun*, vol. 6, no. 17, 2021.

[11] X. Lian, J. Qi, M. Yuan, X. Li, M. Wang, G. Li, T. Yang, and J. Zhong, "Study on risk factors of diabetic peripheral neuropathy and establishment of a prediction model by machine learning," *BMC Medical Informatics and Decision Making*, vol. 23, no. 146, pp. 1–12, 2023.

[12] A. Mujumdar and D. V. Vb, "Diabetes prediction using machine learning algorithms," in *Procedia Computer Science, INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019, ICRTAC 2019*, vol. 165, pp. 292–299, 2019.

[13] M. F. Faruque, Asaduzzaman, and I. H. Sarker, "Performance analysis of machine learning techniques to predict diabetes mellitus," in *2019 International Conference on Electrical, Computer and Communication Engineering*, (Cox'sBazar, Bangladesh), 2019.

[14] A. Dutta, M. K. Hasan, M. Ahmad, M. A. Awal, M. A. Islam, M. Masud, and H. Meshref, "Early prediction of diabetes using an ensemble of machine learning models," *International Journal of Environmental Research and Public Health*, vol. 19, no. 12378, 2022.

[15] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "Development of disease prediction model based on ensemble learning approach for diabetes and hypertension," *IEEE Access*, vol. 7, pp. 144777–144789, 2019.

[16] M. W. Eliana, R. T. Maria, F. Maicon, T. Janet, A. D. Maria, A. C. Maria, B. D. Bruce, and I. S. Maria, "Gestational diabetes and pregnancy outcomes - a systematic review of the world health organization (who) and the international association of diabetes in pregnancy study groups (iadpsg) diagnostic criteria," *BMC Pregnancy and Childbirth*, vol. 12, no. 23, pp. 22–30, 2012.

[17] N. Ahmed, R. Ahammed, M. M. Islam, M. A. Uddin, A. Akhter, M. A. Talukder, and B. K. Paul, "Machine learning based diabetes prediction and development of smart web application," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 229–241, 2021.

[18] H. N. Shilubane, "Knowledge of patients and family members regarding diabetes mellitus and its treatment," Master's thesis, UNISA, Pretoria, 2003. MA Dissertation.

[19] A. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad, G. A. Khan, and A. Ali, "Intelligent machine learning approach for effective recognition of diabetes in e-healthcare using clinical data," *Sensors*, vol. 20, no. 2649, 2020.

[20] D. E. and M. T., "Data driven machine learning methods for diabetes prediction (title inferred from document name)," *Sensors*, vol. 22, no. 5304, 2022.

Appendix A

Budget

A.1 Projected Budget for Pilot Deployment of PreDia

The *PreDia* system was developed as an offline-first Android-based diabetes risk screening solution designed for deployment in rural healthcare environments. In the event of sponsorship or institutional support, the following budget is proposed for a pilot deployment involving approximately 50 Village Health Teams (VHTs) across selected districts in Uganda.

Table A.1: Projected Pilot Deployment Budget

Item	**Quantity**	**Cost (UGX)**
Android Smartphones for VHTs	50	10,000,000
VHT Training and Capacity Building Workshops	2 Sessions	3,000,000
Application Deployment and Device Configuration	1 Exercise	1,500,000
Field Testing and Clinical Validation Activities	1 Exercise	4,000,000
Model Refinement and Local Dataset Integration	1 Exercise	2,000,000
Translation and Additional Language Support	1 Exercise	1,500,000
Technical Support, Maintenance and APK Updates (12 Months)	1 Year	2,500,000
Monitoring and Evaluation Activities	1 Exercise	2,000,000
Contingency (10 **Total Estimated Budget**		**29,150,000**

A.2 Budget Justification

The proposed budget reflects the technical and operational requirements necessary for pilot deployment of the *PreDia* system in rural healthcare environments. The largest expenditure is associated with procurement of Android devices for Village Health Teams (VHTs), who serve as the primary end users of the application. Additional costs support healthcare worker training, field validation exercises, application deployment, and system maintenance.

The inclusion of model refinement and local dataset integration activities recognises the need for future validation using Ugandan clinical data. Resources have also been allocated for language expansion, application updates, and long-term technical support to ensure sustainability beyond the pilot phase.

Because the application operates offline and relies primarily on open-source technologies including XGBoost, SHAP, SQLite, FastAPI, and Android development frameworks, deployment costs remain substantially lower than those associated with laboratory-based diabetes screening programmes. The proposed pilot would support early diabetes risk screening, strengthen preventive healthcare initiatives, and provide evidence for potential nationwide deployment.

Appendix B

Research Project timelines

B.1 Project Management Timeline

The development of the *PreDia* system was planned and executed over a six-month period. The project was divided into six major phases, each producing specific deliverables that contributed to the successful completion of the system.

Table B.1: Research Project Timeline

Phase	**Duration**	**Deliverable**	**Dependency**
Project Proposal and Planning	Month 1	Approved Proposal	None
Literature Review and Requirements Analysis	Month 1–2	Literature Review Report	Proposal Approval
Dataset Preparation and Feature Engineering	Month 2	Prepared Dataset	Literature Review
Machine Learning Model Development and Evaluation	Month 2–3	Trained and Evaluated Models	Dataset Preparation
System Design and Mobile Application Development	Month 3–4	Functional Application Prototype	Requirements Analysis
SHAP Integration and Offline Deployment	Month 4–5	Explainable Offline System	Application Development
System Testing and Validation	Month 5	Evaluation Results	System Integration
Documentation, Review and Final Submission	Month 6	Final Project Report	Testing and Validation

B.2 Project Phases

B.2.1 Phase 1: Project Planning

This phase involved project selection, problem identification, proposal development, and supervisor approval.

B.2.2 Phase 2: Literature Review and Data Preparation

Relevant literature was reviewed, research gaps were identified, and the dataset was prepared through preprocessing and feature engineering.

B.2.3 Phase 3: Model Development

Multiple machine learning algorithms were trained and evaluated. XGBoost was selected as the final predictive model based on performance metrics.

B.2.4 Phase 4: System Development

The Android application, bilingual interface, and offline deployment architecture were developed and integrated with the trained machine learning model.

B.2.5 Phase 5: Testing and Validation

The integrated system was tested for functionality, performance, usability, explainability, and offline operation.

B.2.6 Phase 6: Documentation and Submission

The final report was prepared, reviewed, corrected, and submitted for academic assessment.

Table B.2: Gantt Chart for the PreDia Project

Activity	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**
Project Proposal Development	X					
Literature Review and Requirements Analysis	X	X				
Dataset Preparation and Feature Engineering		X				
Machine Learning Model Development		X	X			
System Design and Architecture Development			X	X		
Android Application Development			X	X	X	
SHAP Integration and Explainability Module				X	X	
System Integration and Offline Deployment				X	X	
Testing and Performance Evaluation					X	
Documentation and Report Writing					X	X
Final Review and Submission						X

Appendix C

Declaration of AI use

This appendix provides a declaration of the ethical use of Generative Artificial Intelligence (AI) tools as academic support tools during the development of this project. The AI tools were used only to support learning, documentation refinement, and technical guidance, while all core research activities, implementation, evaluation, and conclusions remained the original work of the authors.

Declaration Statement:

We, *RUGOGAMU NOELA*, *LORRAINE PAULA ARINAITWE* and *SSENDI ALOYSIOUS*, declare that Generative Artificial Intelligence (AI) tools were ethically used as academic support tools during the development of this report. The AI tools assisted in improving understanding, refining documentation, and supporting technical communication. All generated outputs were critically reviewed, verified, and rewritten where necessary to reflect the understanding and original contribution of the authors.

The AI tools were used in the following ways:

i. **ChatGPT (OpenAI)** was used to support brainstorming of research ideas, refinement of technical explanations, and improvement of academic writing structure. Example prompts included *"Explain SHAP explainability in simple academic language"*, *"Improve the academic tone of this methodology section"*, and *"Generate a professional conclusion for a machine learning healthcare project."* The outputs generated were reviewed critically, modified extensively, and aligned to the project objectives and implementation details.

ii. **ChatGPT** was also used to assist in debugging and refining LaTeX formatting, chapter organization, table insertion, figure referencing, and bibliography management. Example prompts included *"Generate LaTeX code for inserting system architecture diagrams"* and *"Correct formatting issues in a thesis chapter."*

iii. **Google Gemini** was used to assist in summarising literature concepts related to machine learning, explainable artificial intelligence, mobile health systems, and offline healthcare technologies. Example prompts included *"Summarize the role of explainable AI in healthcare systems"* and *"Explain offline-first mobile architecture for healthcare applications."* All generated summaries were verified against academic literature and rewritten appropriately.

iv. **AI-assisted writing tools** were used to improve grammar, sentence flow, readability, and document consistency. These tools were used strictly for language refinement and not for generating experimental findings, datasets, evaluation results, or final research conclusions.

v. **Sunbird AI** was used to support the local language accessibility component of the project, particularly in understanding how Luganda language support could improve communication between Village Health Teams (VHTs) and community members. It informed the bilingual interface direction of the system but was not used to generate research findings, model results, or evaluation outcomes.

vi. The authors confirm that Generative AI tools were **not** used to fabricate data, manipulate results, generate system evaluation outcomes, or replace independent implementation and analysis. All machine learning model development, Android application development, testing procedures, evaluation metrics, interpretation of findings, and final conclusions presented in this project remain the original intellectual work of the authors.

Signature (*RUGOGAMU NOELA*): ______________________________

Signature (*LORRAINE PAULA ARINAITWE*): ______________________________

Signature (*SSENDI ALOYSIOUS*): ______________________________

Date: ________________________

Appendix D

List of 10 Major Journal Publications reviewed

D.1 Major Publications Reviewed

The publications listed in Table D.1 were among the most influential sources reviewed during the development of the *PreDia* system. They provided the theoretical, methodological, and technical foundation for the study, particularly in the areas of diabetes risk prediction, machine learning, explainable artificial intelligence, mobile health systems, and healthcare deployment in low-resource settings.

Table D.1: Major Publications Reviewed

No.	Publication Title	Type	Journal / Conference	Year	Impact Factor*
1	Gestational Diabetes and Pregnancy Outcomes: A Systematic Review of WHO and IADPSG Diagnostic Criteria (Wendland et al.)	Review Paper	BMC Pregnancy and Childbirth	2012	3.1
2	Explainable AI Techniques in Healthcare Decision Support Systems (Reviewed through Dinh et al.)	Review Paper	BMC Medical Informatics and Decision Making	2019	3.4
3	Machine Learning Approaches for Diabetes Prediction: Literature Synthesis from Recent Studies	Review Paper	Multiple Sources Reviewed	2018–2023	N/A
4	Prediction of Diabetes Empowered With Fused Machine Learning (Ahmed et al.)	Journal Article	IEEE Access	2022	3.9
5	Early Prediction of Diabetes Using an Ensemble of Machine Learning Models (Dutta et al.)	Journal Article	International Journal of Environmental Research and Public Health	2022	4.5
6	Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers (Hasan et al.)	Journal Article	IEEE Access	2020	3.9
7	Data-Driven Diabetes Risk Factor Prediction Using Machine Learning Algorithms with Feature Selection Technique (Kakoly et al.)	Journal Article	Sustainability	2023	3.9
8	Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension (Fitriyani et al.)	Journal Article	IEEE Access	2019	3.9
9	Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus (Faruque et al.)	Conference Proceeding	International Conference on Electrical, Computer and Communication Engineering (ECCE)	2019	N/A
10	Diabetes Prediction Using Machine Learning Algorithms (Mujumdar and Vaidehi)	Conference Proceeding	International Conference on Recent Trends in Advanced Computing (ICRTAC)	2019	N/A
11	Machine Learning-Based Diabetes Prediction Research Presented in International Computing Conferences Reviewed During the Study	Conference Proceeding	Various Proceedings	2019–2023	N/A

D.2 Summary

The reviewed publications provided the conceptual and technical foundation for the development of the *PreDia* system. Particular emphasis was placed on studies involving non-invasive diabetes prediction, ensemble and hybrid machine learning approaches, explainable artificial intelligence, mobile health technologies, and healthcare solutions suitable for low-resource environments. These publications informed the selection of predictive features, model development strategies, explainability mechanisms, and deployment considerations adopted in this project.

Appendix E

Scopus Search Strings used during the literature review

E.1 Introduction

This appendix presents the search strings and filtering strategy used during the systematic literature review conducted for this study. The literature search was performed primarily using the Scopus database to identify studies related to machine learning, explainable artificial intelligence, and Type 2 Diabetes Mellitus (T2DM) prediction in low-resource healthcare settings.

E.2 Primary Search String

> (”Type 2 Diabetes” OR ”T2DM”) AND (”Machine Learning” OR ”Artificial Intelligence”) AND (”Risk Prediction” OR ”Early Detection”) AND (”Non-invasive” OR ”Clinical Features”) AND (”Africa” OR ”Low-resource settings”)

E.3 Additional Refined Queries

1. (”Diabetes Prediction” AND ”XGBoost” AND ”Healthcare”)
2. (”Explainable AI” OR ”XAI”) AND (”Healthcare Diagnosis”)
3. (”Offline Health Systems” OR ”mHealth”) AND (”Rural Healthcare”)
4. (”SHAP” OR ”LIME”) AND (”Diabetes Prediction”)
5. (”Community Health Workers”) AND (”Artificial Intelligence”) AND (”Uganda”)

E.4 Inclusion Criteria

- Peer-reviewed journal articles and conference papers.
- Studies published between 2018 and 2025.
- Studies focusing on diabetes prediction using machine learning approaches.
- Studies involving explainable AI techniques.

- Studies related to healthcare systems in low-resource or African settings.

E.5 Exclusion Criteria

- Studies unrelated to diabetes prediction.
- Studies without machine learning or predictive modeling approaches.
- Non-English publications.
- Review papers without primary experimental findings.

E.6 PRISMA Literature Selection Diagram

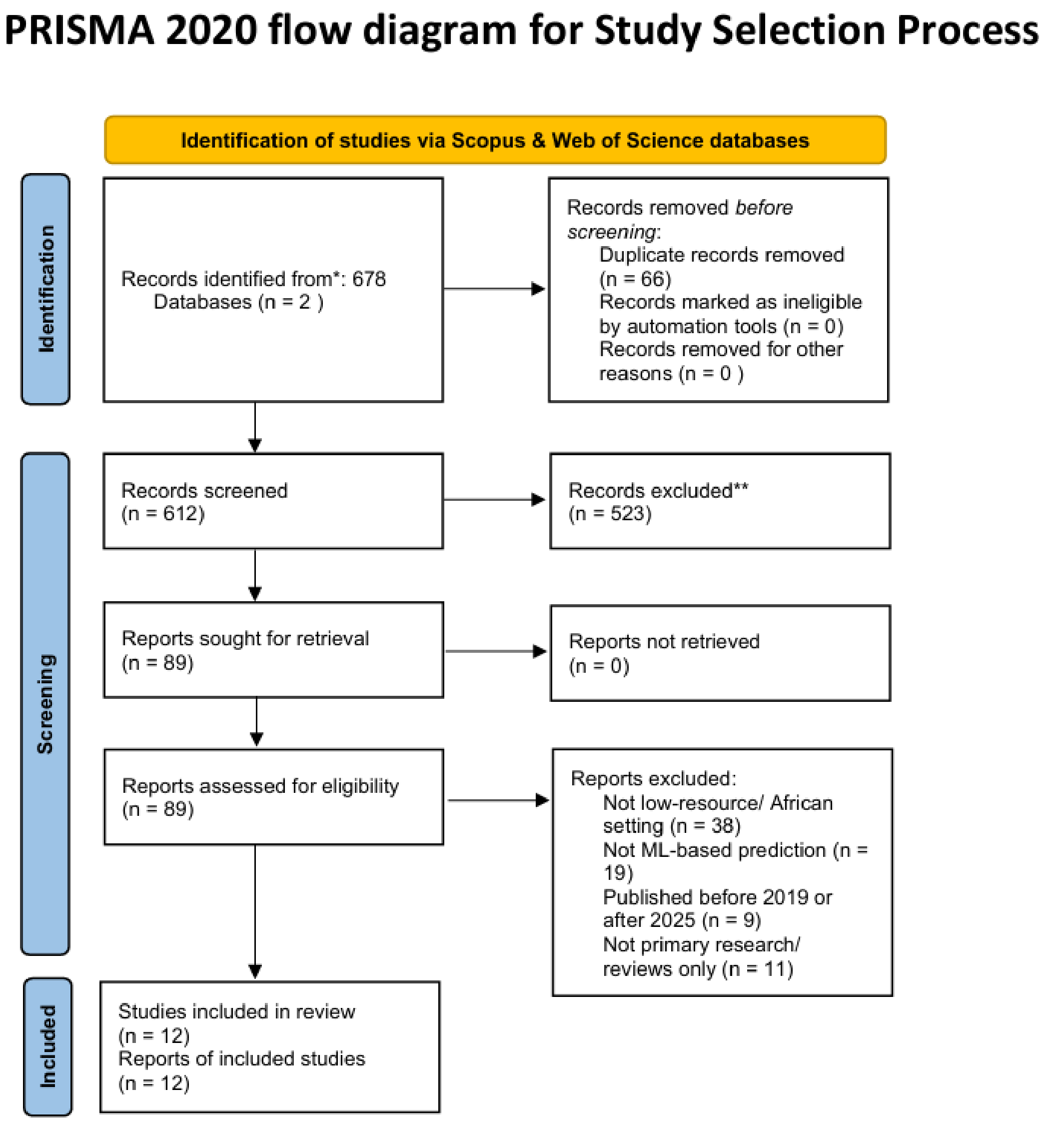


Figure E.1: PRISMA Flow Diagram for Literature Selection

Appendix F

Research alignment with specialization, SDG, NDPIV & Vision 2040

F.1 Alignment with Specialization

The *PreDia* project aligns closely with the Artificial Intelligence and Data Science specialization track. The project applies machine learning techniques, predictive analytics, explainable artificial intelligence (XAI), and mobile computing technologies to address a real-world healthcare challenge. Specifically, the project demonstrates the application of supervised machine learning algorithms, feature engineering, model evaluation, explainability techniques using SHAP, and deployment of AI solutions within low-resource environments.

The project further integrates software engineering principles, mobile application development, database management, and healthcare informatics, thereby reflecting the multidisciplinary nature of Artificial Intelligence and Data Science.

F.2 Alignment with the United Nations Sustainable Development Goals (SDGs)

The project contributes directly to several Sustainable Development Goals (SDGs):

SDG 3: Good Health and Well-Being

The primary objective of the *PreDia* system is to support early screening and risk prediction of Type 2 Diabetes Mellitus (T2DM). Early identification of high-risk individuals can improve preventive healthcare interventions, reduce complications, and strengthen community health outcomes.

SDG 9: Industry, Innovation and Infrastructure

The project promotes innovation through the development of an explainable artificial intelligence system tailored to low-resource healthcare environments. The offline-first architecture enables deployment even in areas with limited digital infrastructure.

SDG 10: Reduced Inequalities

By supporting bilingual communication in both English and Luganda and enabling deployment in underserved rural communities, the system contributes toward reducing inequalities in access to healthcare technologies and information.

F.3 Alignment with Uganda's National Development Plan IV (NDP IV)

Uganda's National Development Plan IV emphasizes the use of science, technology, innovation, and digital transformation to improve service delivery and strengthen human capital development.

The *PreDia* system aligns with NDP IV through:

- Supporting the adoption of digital health technologies within Uganda's healthcare system.
- Promoting innovation-driven healthcare solutions for non-communicable diseases.
- Strengthening community healthcare services through technology-assisted screening and decision support.
- Supporting evidence-based healthcare interventions through data-driven risk prediction.

The project contributes to NDP IV's objective of improving healthcare accessibility, efficiency, and quality through technology-enabled service delivery.

F.4 Alignment with Uganda Vision 2040

Uganda Vision 2040 identifies Information and Communication Technology (ICT), science, technology, and innovation as key drivers of socio-economic transformation.

The *PreDia* system supports Vision 2040 by:

- Leveraging artificial intelligence and digital technologies to improve healthcare service delivery.
- Promoting innovation in healthcare and community health management.
- Supporting the modernization of healthcare systems through intelligent decision-support tools.
- Enhancing access to healthcare services in rural and underserved communities.
- Contributing to the development of a knowledge-based economy through practical application of emerging technologies.

The project demonstrates how locally developed AI solutions can contribute toward achieving Uganda's long-term vision of a modern, healthy, and technologically empowered society.

F.5 Summary

The *PreDia* project demonstrates strong alignment with the Artificial Intelligence and Data Science specialization, the United Nations Sustainable Development Goals, Uganda's National

Development Plan IV, and Uganda Vision 2040. Through the application of explainable artificial intelligence, mobile health technologies, and offline healthcare deployment, the project contributes to improved healthcare accessibility, innovation, and sustainable socio-economic development.

Appendix G

Algorithms, Code of interest

This appendix presents actual Python code components of interest for the proposed AI/ML healthcare research. The code covers data loading, cleaning, preprocessing, model training, model evaluation, and prediction.

G.1 Importing Required Libraries

```
import pandas as pd
import numpy as np

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer

from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from xgboost import XGBClassifier

from sklearn.metrics import accuracy_score, precision_score
from sklearn.metrics import recall_score, f1_score
from sklearn.metrics import confusion_matrix, classification_report
```

G.2 Loading the Healthcare Dataset

```
# Load dataset
data = pd.read_csv("healthcare_dataset.csv")

# Display first five records
print(data.head())

# Display dataset information
print(data.info())
```

```
# Check for missing values
print(data.isnull().sum())
```

G.3 Data Cleaning and Preprocessing

```
# Remove duplicate records
data = data.drop_duplicates()

# Define target column
target_column = "target"

# Separate features and target
X = data.drop(target_column, axis=1)
y = data[target_column]

# Identify numerical and categorical columns
numerical_columns = X.select_dtypes(include=["int64", "float64"]).columns
categorical_columns = X.select_dtypes(include=["object"]).columns

# Preprocessing for numerical data
numerical_transformer = Pipeline(steps=[
    ("imputer", SimpleImputer(strategy="median")),
    ("scaler", StandardScaler())
])

# Preprocessing for categorical data
categorical_transformer = Pipeline(steps=[
    ("imputer", SimpleImputer(strategy="most_frequent")),
    ("encoder", LabelEncoder())
])
```

Note: Since `LabelEncoder` does not work directly inside a `ColumnTransformer` for multiple categorical columns, the categorical encoding can alternatively be handled using `OneHotEncoder`, as shown below.

```
from sklearn.preprocessing import OneHotEncoder

categorical_transformer = Pipeline(steps=[
    ("imputer", SimpleImputer(strategy="most_frequent")),
    ("encoder", OneHotEncoder(handle_unknown="ignore"))
])

preprocessor = ColumnTransformer(
    transformers=[
        ("num", numerical_transformer, numerical_columns),
        ("cat", categorical_transformer, categorical_columns)
    ]
)
```

G.4 Splitting the Dataset

```
X_train, X_test, y_train, y_test = train_test_split(
    X,
    y,
    test_size=0.2,
    random_state=42,
    stratify=y
)
```

G.5 Logistic Regression Model

```
logistic_model = Pipeline(steps=[
    ("preprocessor", preprocessor),
    ("classifier", LogisticRegression(max_iter=1000))
])

logistic_model.fit(X_train, y_train)

logistic_predictions = logistic_model.predict(X_test)

print("Logistic Regression Results")
print("Accuracy:", accuracy_score(y_test, logistic_predictions))
print("Precision:", precision_score(y_test, logistic_predictions, average="weighted"))
print("Recall:", recall_score(y_test, logistic_predictions, average="weighted"))
print("F1 Score:", f1_score(y_test, logistic_predictions, average="weighted"))
print(classification_report(y_test, logistic_predictions))
```

G.6 Decision Tree Model

```
decision_tree_model = Pipeline(steps=[
    ("preprocessor", preprocessor),
    ("classifier", DecisionTreeClassifier(
        max_depth=5,
        random_state=42
    ))
])

decision_tree_model.fit(X_train, y_train)

decision_tree_predictions = decision_tree_model.predict(X_test)

print("Decision Tree Results")
print("Accuracy:", accuracy_score(y_test, decision_tree_predictions))
print("Precision:", precision_score(y_test, decision_tree_predictions, average="weighted
print("Recall:", recall_score(y_test, decision_tree_predictions, average="weighted"))
print("F1 Score:", f1_score(y_test, decision_tree_predictions, average="weighted"))
print(classification_report(y_test, decision_tree_predictions))
```

G.7 Random Forest Model

```
random_forest_model = Pipeline(steps=[
    ("preprocessor", preprocessor),
    ("classifier", RandomForestClassifier(
        n_estimators=100,
        random_state=42
    ))
])

random_forest_model.fit(X_train, y_train)

random_forest_predictions = random_forest_model.predict(X_test)

print("Random Forest Results")
print("Accuracy:", accuracy_score(y_test, random_forest_predictions))
print("Precision:", precision_score(y_test, random_forest_predictions, average="weightec
print("Recall:", recall_score(y_test, random_forest_predictions, average="weighted"))
print("F1 Score:", f1_score(y_test, random_forest_predictions, average="weighted"))
print(classification_report(y_test, random_forest_predictions))
```

G.8 Support Vector Machine Model

```
svm_model = Pipeline(steps=[
    ("preprocessor", preprocessor),
    ("classifier", SVC(kernel="rbf", probability=True))
])

svm_model.fit(X_train, y_train)

svm_predictions = svm_model.predict(X_test)

print("Support Vector Machine Results")
print("Accuracy:", accuracy_score(y_test, svm_predictions))
print("Precision:", precision_score(y_test, svm_predictions, average="weighted"))
print("Recall:", recall_score(y_test, svm_predictions, average="weighted"))
print("F1 Score:", f1_score(y_test, svm_predictions, average="weighted"))
print(classification_report(y_test, svm_predictions))
```

G.9 XG Boost Model

```
xgboost_model = Pipeline(steps=[
    ("preprocessor", preprocessor),
    ("classifier", XGBClassifier(
        n_estimators=100,
        learning_rate=0.1,
        max_depth=4,
        random_state=42,
        eval_metric='logloss'
    ))
])
```

```
xgboost_model.fit(X_train, y_train)

xgboost_predictions = xgboost_model.predict(X_test)

print("XGBoost Results")
print("Accuracy:", accuracy_score(y_test, xgboost_predictions))
print("Precision:", precision_score(y_test, xgboost_predictions, average="weighted"))
print("Recall:", recall_score(y_test, xgboost_predictions, average="weighted"))
print("F1 Score:", f1_score(y_test, xgboost_predictions, average="weighted"))
print(classification_report(y_test, xgboost_predictions))
```

G.10 Confusion Matrix Code

```
import matplotlib.pyplot as plt
from sklearn.metrics import ConfusionMatrixDisplay

cm = confusion_matrix(y_test, random_forest_predictions)

disp = ConfusionMatrixDisplay(confusion_matrix=cm)
disp.plot()

plt.title("Confusion Matrix for Random Forest Model")
plt.show()
```

G.11 Comparing Model Performance

```
results = pd.DataFrame({
    "Model": [
        "Logistic Regression",
        "Decision Tree",
        "Random Forest",
        "Support Vector Machine",
        "XGBoost"
    ],
    "Accuracy": [
        accuracy_score(y_test, logistic_predictions),
        accuracy_score(y_test, decision_tree_predictions),
        accuracy_score(y_test, random_forest_predictions),
        accuracy_score(y_test, svm_predictions),
        accuracy_score(y_test, xgboost_predictions)
    ],
    "Precision": [
        precision_score(y_test, logistic_predictions, average="weighted"),
        precision_score(y_test, decision_tree_predictions, average="weighted"),
        precision_score(y_test, random_forest_predictions, average="weighted"),
        precision_score(y_test, svm_predictions, average="weighted"),
        precision_score(y_test, xgboost_predictions, average="weighted")
    ],
    "Recall": [
```

```
        recall_score(y_test, logistic_predictions, average="weighted"),
        recall_score(y_test, decision_tree_predictions, average="weighted"),
        recall_score(y_test, random_forest_predictions, average="weighted"),
        recall_score(y_test, svm_predictions, average="weighted"),
        recall_score(y_test, xgboost_predictions, average="weighted")
    ],
    "F1 Score": [
        f1_score(y_test, logistic_predictions, average="weighted"),
        f1_score(y_test, decision_tree_predictions, average="weighted"),
        f1_score(y_test, random_forest_predictions, average="weighted"),
        f1_score(y_test, svm_predictions, average="weighted"),
        f1_score(y_test, xgboost_predictions, average="weighted")
    ]
})

print(results)
```

G.12 Saving the Best Model

```
import joblib

# Save the trained XGBoost model
joblib.dump(xgboost_model, "best_healthcare_model.pkl")

print("Model saved successfully.")
```

G.13 Loading the Saved Model

```
loaded_model = joblib.load("best_healthcare_model.pkl")

print("Model loaded successfully.")
```

G.14 Making Predictions on New Patient Data

```
new_patient_data = pd.DataFrame({
    "Age": [45],
    "BMI": [28.5],
    "Glucose": [140],
    "BloodPressure": [130],
    "Pregnancies": [2],
    "DiabetesPedigreeFunction": [0.65],
    "PhysicalActivity": [3],
    "DietQuality": [4],
    "Smoking": [0],
    "AlcoholUse": [1]
})

prediction = loaded_model.predict(new_patient_data)

prediction_probability = loaded_model.predict_proba(new_patient_data)
```

```
print("Predicted Class:", prediction[0])
print("Prediction Probability:", prediction_probability)
```

G.14 Summary

The code presented in this appendix demonstrates the main implementation logic for the proposed AI/ML healthcare system. It includes data preprocessing, machine learning model training, evaluation, performance comparison, model saving, and prediction using new patient data.

Appendix H

Other useful figures

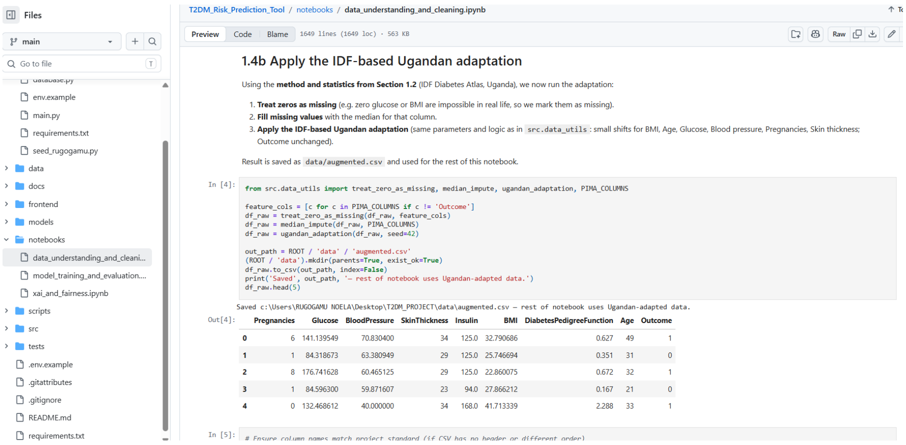


Figure H.1: Screenshot showing how data was augmented using IDF Ugandan diabetes statistics

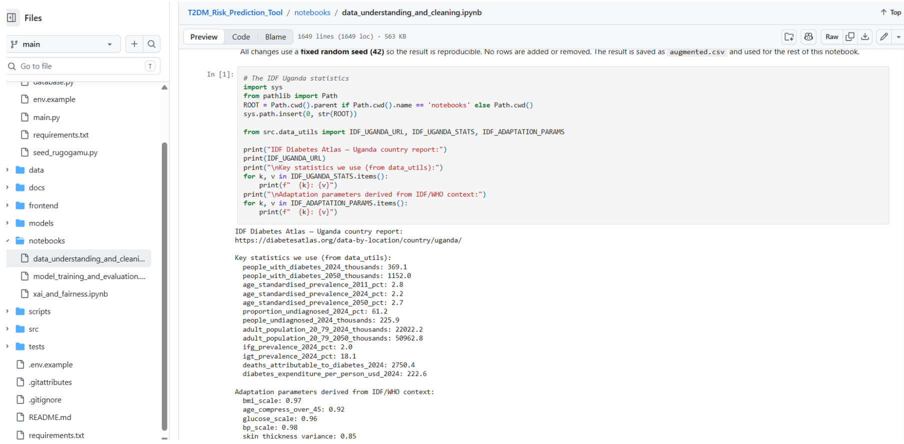
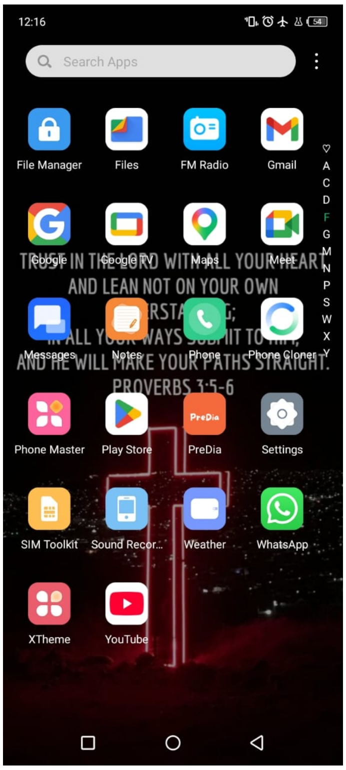
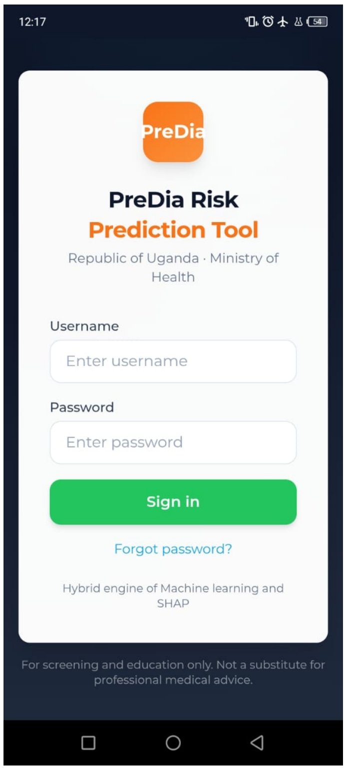


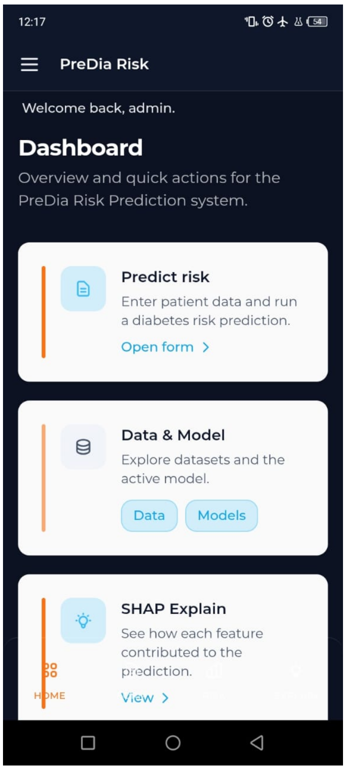
Figure H.2: Screenshot showing the code saved onto the GitHub platform



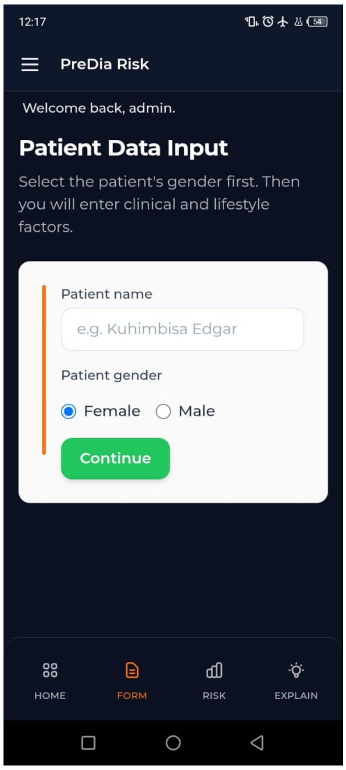
(a) PreDia running offline



(b) Sign In page

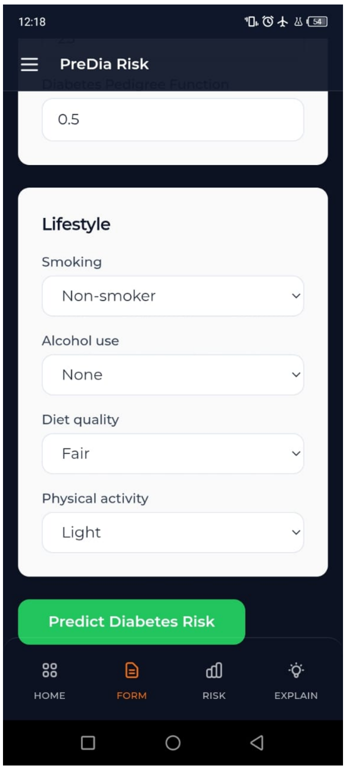


(c) Dashboard

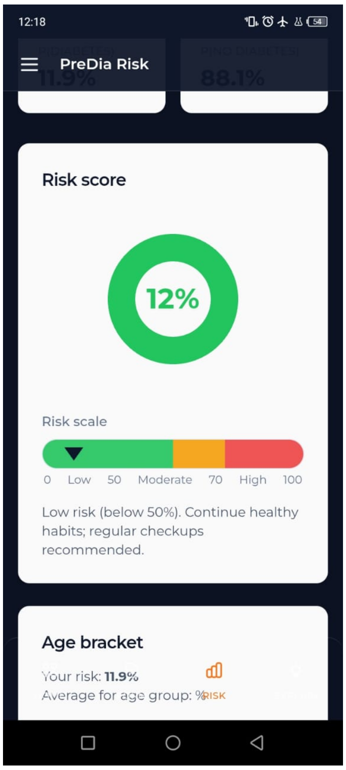


(d) Patient Registration: Enter Gender

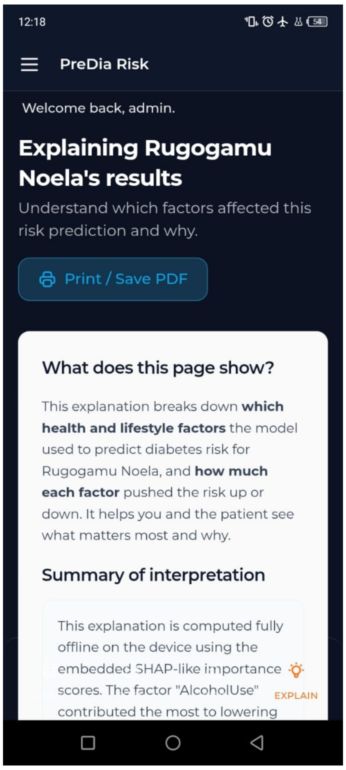
Figure H.3: Key interfaces of the PreDia Android application



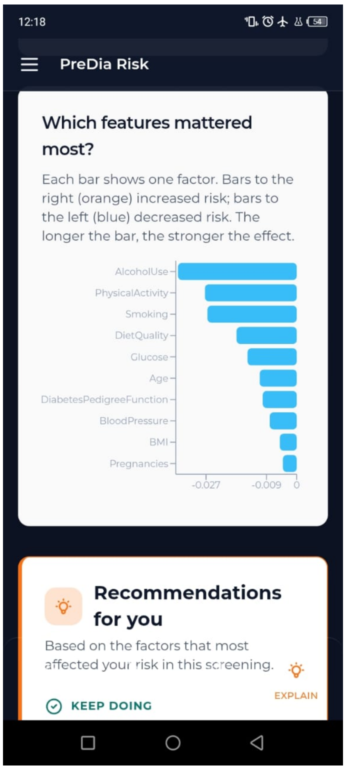
(a) Clinical Data Entry



(b) Prediction Results

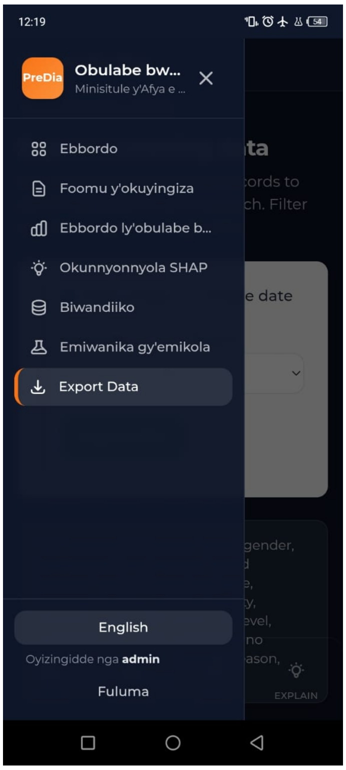


(c) SHAP explanations of results

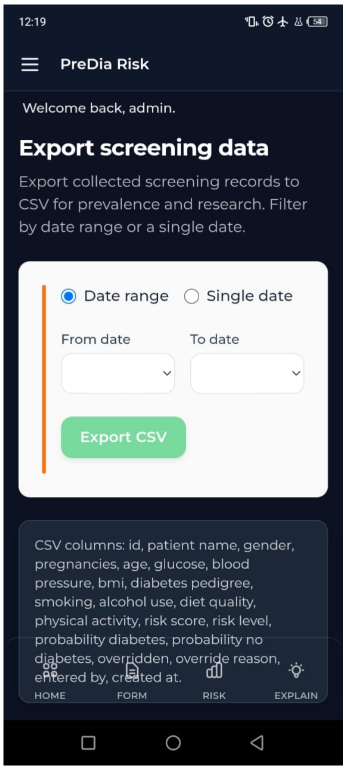


(d) Feature contributions to Risk

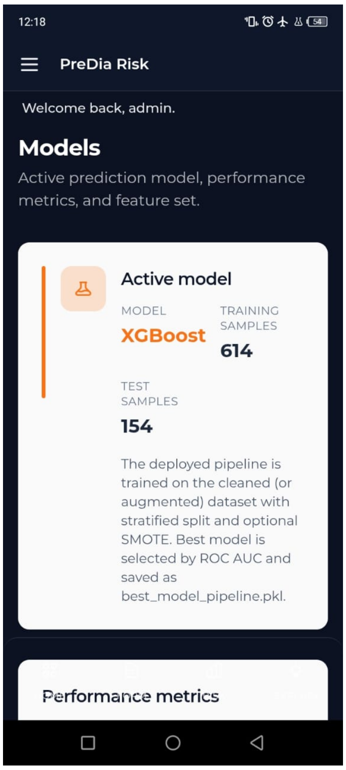
Figure H.4: Key interfaces of the PreDia Android application(Continuation)



(a) English to Luganda Translation



(b) Save or Export Results (Optional)



(c) Model Performance (for Admin Access)

Figure H.5: Key interfaces of the PreDia Android application (Final Continuation)

Appendix I

Research poster layout

I.1 Research Poster

Figure I.1 presents the research poster developed for the *PreDia* project. The poster summarizes the research concept, problem statement, objectives, methodology, results, and expected impact of the proposed diabetes risk prediction system.

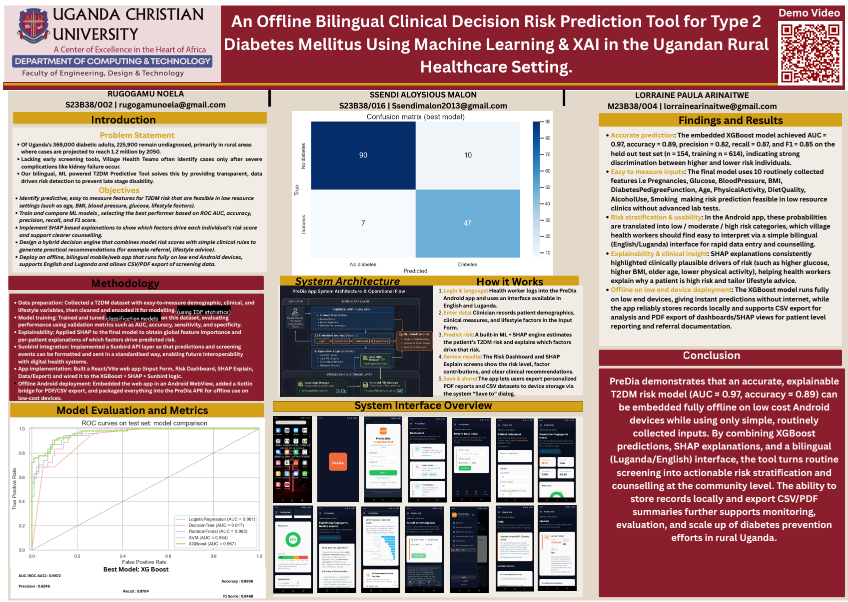


Figure I.1: Research Poster for the PreDia Project

I.2 Poster Summary

Research Concept

PreDia is an offline bilingual clinical decision support system developed to support early screening and risk prediction of Type 2 Diabetes Mellitus (T2DM) in rural Uganda. The system combines machine learning and explainable artificial intelligence (XAI) to provide transparent risk assessments using routinely collected demographic, clinical, and lifestyle information.

Problem Being Addressed

A large proportion of diabetes cases in Uganda remain undiagnosed, particularly in rural communities with limited access to diagnostic services and specialist healthcare workers. This often results in delayed diagnosis and severe complications. PreDia addresses this challenge by providing an accessible and explainable screening tool that operates fully offline.

Research Objectives

1. Identify predictive and easy-to-measure factors associated with T2DM risk.
2. Develop and evaluate machine learning models for diabetes risk prediction.
3. Integrate SHAP-based explainability to improve transparency and clinical understanding.
4. Deploy an offline bilingual Android application for community-level screening.

Methodology

A diabetes dataset containing demographic, clinical, and lifestyle variables was cleaned and prepared for modelling. Multiple machine learning models were evaluated, with XGBoost selected as the best-performing model. SHAP explanations were integrated to provide interpretable predictions, and the final system was deployed as an offline Android application with English and Luganda support.

Results

The XGBoost model achieved strong predictive performance with an AUC of 0.9672, accuracy of 0.8896, precision of 0.8246, recall of 0.8704, and F1-score of 0.8468. The system successfully provides explainable risk predictions and supports PDF/CSV report generation for healthcare workers.

Impact

PreDia demonstrates that accurate and explainable diabetes risk prediction can be delivered on low-cost Android devices without internet connectivity. The system supports early detection, counselling, referral decisions, and community-level diabetes prevention efforts in rural Uganda.