UGANDA CHRISTIAN UNIVERSITY

Faculty of Engineering, Design and Technology
Department of Computing & Technology

A Clinical Decision Risk Prediction Tool for Type 2 Diabetes Mellitus in the Ugandan Rural Setting Using Machine Learning and Explainable Artificial Intelligence (SHAP)

Final Year Project Proposal Document

Submitted in Partial Fulfilment of the Requirements for the Award of
Bachelor of Science in Data Science & Analytics

Authors

Rugogamu Noela S23B38/016
Lorraine P. Arinaitwe M23B38/004
Ssendi Aloysious M. S23B38/002

Date: 29 November 2025

Chapter 1

INTRODUCTION

Background of the Study

Diabetes mellitus is a chronic metabolic disease that has become one of the largest public health challenges of the 21st century. The International Diabetes Federation reported in 2021 that 537 million adults worldwide were living with diabetes6, and this number is projected to rise to 783 million by 2045. More worryingly, 79% of these adults live in low- and middle-income countries where health systems are already overstretched. In sub-Saharan Africa, approximately 24 million adults have diabetes, but up to 75% remain undiagnosed because of limited access to diagnostic facilities.

In Uganda specifically, the Ministry of Health and several epidemiological surveys estimate that between 1.2 and 1.8 million adults have diabetes, yet the majority are unaware of their condition. Urban prevalence is estimated at 3.5–4.0%, while rural prevalence appears much lower at 1.4%7. However, health experts agree that this apparent lower rural prevalence is largely an artefact of under-diagnosis rather than true lower risk. More than 75% of Uganda's population lives in rural areas where the nearest health centre with laboratory services may be 20–50 km away, reachable only by foot or motorcycle taxi. Electricity and internet connectivity are unreliable or non-existent in many villages.

The primary level of healthcare in rural Uganda is delivered by approximately 25,000 Village Health Team members (VHTs). These are volunteer community members, usually with only primary-level education, who receive two to four weeks of basic training. They visit households, treat malaria and diarrhoea, distribute family planning commodities, and provide health education. Despite their critical role, VHTs currently have no simple, reliable, and affordable way to identify individuals at high risk of developing type 2 diabetes in the next five to ten years. Random capillary glucose testing using a glucometer and test strips is possible but expensive (approximately UGX 700 per test) and supply chains are unreliable. When diabetes is eventually diagnosed, patients already present with advanced complications such as retinopathy, nephropathy, or foot ulcers that require costly specialist care that most families cannot afford.

Machine learning has revolutionised disease risk prediction in high-income countries. Models trained on simple clinical variables routinely achieve AUC values above 0.90 and are integrated into electronic health records. Unfortunately, almost all publicly available high-performing models were trained on datasets from the United States, Europe, India, or the original Pima Indian community. African populations differ in genetics, diet (high carbohydrate, low processed food), body composition (lower visceral fat at same BMI), and socio-economic determinants of health. Directly applying foreign models in Uganda therefore risks poor calibration, reduced accuracy, and hidden bias.

This project was conceived to fill this exact gap: to build a lightweight, fully offline, explainable, and trustworthy diabetes risk prediction tool that can be used today by Ugandan Village Health Team members on cheap Android smartphones.

Problem Statement

Despite the rapid increase in type 2 diabetes mellitus in Uganda, rural communities continue to experience late diagnosis and preventable complications because Village Health Teams lack practical, low-cost, laboratory-free screening instruments1. Existing machine-learning-based risk prediction models perform excellently in high-income settings but suffer from four critical shortcomings when considered for rural Ugandan deployment:
(i) they were trained on non-African datasets and may not generalise well,
(ii) they are almost always "black boxes" that provide no understandable justification for a high-risk prediction, eroding trust among non-technical health workers,
(iii) many published models exhibit gender and socio-economic bias.
(iv) they require constant internet connection or powerful hardware that is unavailable in villages.
These limitations prevent safe, equitable, and sustainable adoption in Ugandan primary healthcare.

General Objective

To design, develop, test, and prepare for deployment of an offline clinical decision support mobile tool that accurately predicts individual risk of type 2 diabetes mellitus for rural Ugandan adults using only low-cost measurements, while providing transparent, plain-language explanations that Village Health Team members can understand and trust.

Specific Objectives

1. To systematically identify the smallest set of predictive clinical and socio-demographic features that can be collected without laboratory equipment in rural Ugandan settings.

2. To train, optimise, and rigorously compare Logistic Regression, Random Forest, XGBoost, and LightGBM models and select the algorithm offering the best trade-off between predictive performance, inference speed, and model size on low-end Android devices.

3. To integrate SHapley Additive exPlanations (SHAP) into the selected model so that every risk prediction is accompanied by an intuitive visual and textual explanation.

4. To develop a fully offline Streamlit web application that can be packaged and installed as an Android APK on inexpensive smartphones (2 GB RAM, Android 9 or higher) commonly owned by VHTs(Village Health Teams).

5. To conduct usability testing with real Village Health Team members and prepare detailed protocols for Phase II data collection of more than 1,200 Ugandan patient records in Mukono district in 2026.

Research Questions

1. Which combination of non-laboratory features (age, BMI, random glucose, family history, blood pressure, etc.) provides the highest predictive power in publicly available datasets that are closest to African physiology3?

2. Among modern gradient boosting algorithms, does LightGBM continue to outperform XGBoost and Random Forest in both AUC and inference speed when constrained to less than 300 KB model size?

3. How can SHAP force plots and summary explanations be translated into simple English (and later Luganda) sentences that a health worker can confidently explain to a villager?

4. Is it technically feasible to package a LightGBM model + SHAP explainer + Streamlit interface into a single offline Android APK that works without internet on devices costing less than UGX 250,000?

Significance of the Study

This project will deliver the first known diabetes risk prediction tool specifically designed for offline use by Ugandan Village Health Teams. Early identification of high-risk individuals will enable timely lifestyle counselling and referral before irreversible complications develop, thereby reducing morbidity, mortality, and healthcare expenditure. The explainable nature of the tool will build trust between community health workers and the technology, encouraging

sustained adoption. The prototype produced during this project will be immediately deployable while we collect local data for the definitive Uganda-validated model in 2026. The work directly contributes to Sustainable Development Goals 3, 9, and 106.

Scope and Delimitations

The current phase of the project focuses on building a functional prototype using the Pima Indians Diabetes Dataset, as no publicly available Ugandan diabetes dataset exists. The system will have its usability assessed in controlled demonstrations and simulated rural environments but will not yet be deployed clinically. Clinical validation involving Ugandan patients is reserved for Phase II of the project in 2026. The study is therefore limited by its reliance on a non-African dataset, although this limitation will be addressed through local data collection in the next phase.

Limitations

The primary limitation of this project is the use of the Pima dataset, which consists exclusively of female Native American participants. This introduces demographic and physiological differences that may affect the generalizability of the model to Ugandan adults. Another limitation is the use of random glucose measurements, which may introduce noise compared to fasting glucose or OGTT values. Additionally, although insulin levels are included in the Pima dataset, they cannot feasibly be collected by VHTs due to the costs and logistical constraints associated with glucometer test strips. Despite these limitations, the dataset remains useful for prototyping and methodological development, with the understanding that Phase II will mitigate these issues through real Ugandan data.

Chapter 2

LITERATURE REVIEW

Historical Development of Diabetes Risk Prediction

The eaeved AUCs of 0.70–0.85. From 2015 onward, the availability of large electronic health record datasets and powerful ensemble algorithms dramaticalrliest diabetes risk scores (FINDRISC, QDScore, Framingham) were simple logistic regression or point-based systems developed in Europe and North America between 2001 and 2010. They used 7–10 clinical variables and achily improved performance2. Studies using Random Forest, XGBoost, and LightGBM began reporting AUCs of 0.90–0.98 on datasets such as NHANES and the original Pima Indians dataset.

A large body of evidence now confirms that plasma glucose is by far the strongest single predictor, followed by BMI, age, family history, and hypertension. Gradient boosting machines consistently outperform neural networks on tabular health data because they naturally handle missing values, mixed data types, and non-linear interactions. LightGBM and XGBoost are particularly favoured for deployment because they produce tiny models (often ¡500 KB) with millisecond inference times.

In November 2025, we conducted a PRISMA 2020-compliant systematic literature review titled "Machine Learning for Type 2 Diabetes Risk Prediction in Ugandan Healthcare: Bridging Data to Decision Gaps". Two databases (Scopus and Web of Science) were searched on 7 November 2025. The initial search returned 678 records. After duplicate removal, title/abstract screening and full-text assessment, only 12 studies remained.

Key findings were: ensemble methods dominated with AUCs of 0.88–0.96; glucose, BMI, and age were the top three features in every study; only 25% incorporated explainable AI; gender bias was common; zero studies used Ugandan data; no study discussed offline deployment or usability by community health workers.

Chapter 3

METHODOLOGY

3.1 Research Paradigm and Design

This study follows the Design Science Research Methodology (DSRM), which is particularly suitable for projects that involve the creation of technological artefacts aimed at solving real-world problems. DSRM emphasizes iterative development through problem identification, definition of objectives, artefact design, evaluation, and communication. This approach ensures that the tool developed in this project is both scientifically sound and practically relevant to its intended users—Village Health Teams operating in rural Ugandan contexts.

3.2 Data Source and Ethical Considerations

The study uses the Pima Indians Diabetes Dataset obtained from the UCI Machine Learning Repository. The dataset contains 768 anonymized patient records with eight clinical predictors and a binary diabetes outcome. Since the data is publicly available and fully anonymized, no ethical approval is required for its use. However, the future Phase II of this research8, which will involve real Ugandan patient data, will strictly adhere to ethical guidelines, including informed consent, approval from institutional review boards, and compliance with patient confidentiality standards11.

3.3 Data Pre-processing Pipeline

Pre-processing is a crucial step in building reliable machine learning models. An initial exploratory data analysis revealed that several features contain physiologically impossible zero values, such as zero blood pressure or zero BMI. Such values were treated as missing and imputed using K-Nearest Neighbours imputation stratified by outcome class to preserve predictive relationships. Outliers were reduced through winsorization to prevent distortion of model parameters. Features were then scaled using RobustScaler to minimize the influence of extreme values. The dataset was split into an 80

3.4 Feature Selection and Engineering

To identify the most predictive and context-appropriate features, we employed Pearson correlation, Mutual Information, and Recursive Feature Elimination. While all eight features were initially considered, the anticipated minimal set suitable for rural deployment included glucose, BMI, age, blood pressure, pregnancies, and diabetes pedigree function ,eliminating Skin thickness because it's difficult to attain11.

```
FEATURE SELECTION: STATISTICAL METHODS
============================================================

Features with significant correlation (>0.1) with Outcome:
1. Glucose                  :  0.498
2. BMI                      :  0.325
3. Age                      :  0.248
4. DiabetesPedigreeFunction :  0.197
5. Pregnancies              :  0.195
6. SkinThickness            :  0.184
7. BloodPressure            :  0.158
8. Insulin                  :  0.153
```

Figure 3.1: Feature Selection Method 1

3.5 Model Development, Training, and Hyper-parameter Optimisation

Four machine learning algorithms—Logistic Regression, Random Forest, XGBoost, and LightGBM—were developed and evaluated. Hyperparameter optimization was performed using five-fold stratified cross-validation to ensure that results were robust and unbiased. The primary evaluation metric was ROC-AUC, given its suitability for binary classification problems with potential class imbalance. Additional considerations included model interpretability, inference speed, and model size9, all of which influence the practical deployability on low-end Android devices. LightGBM is expected to provide the best balance of these factors due to its gradient boosting framework and efficient leaf-wise tree growth.

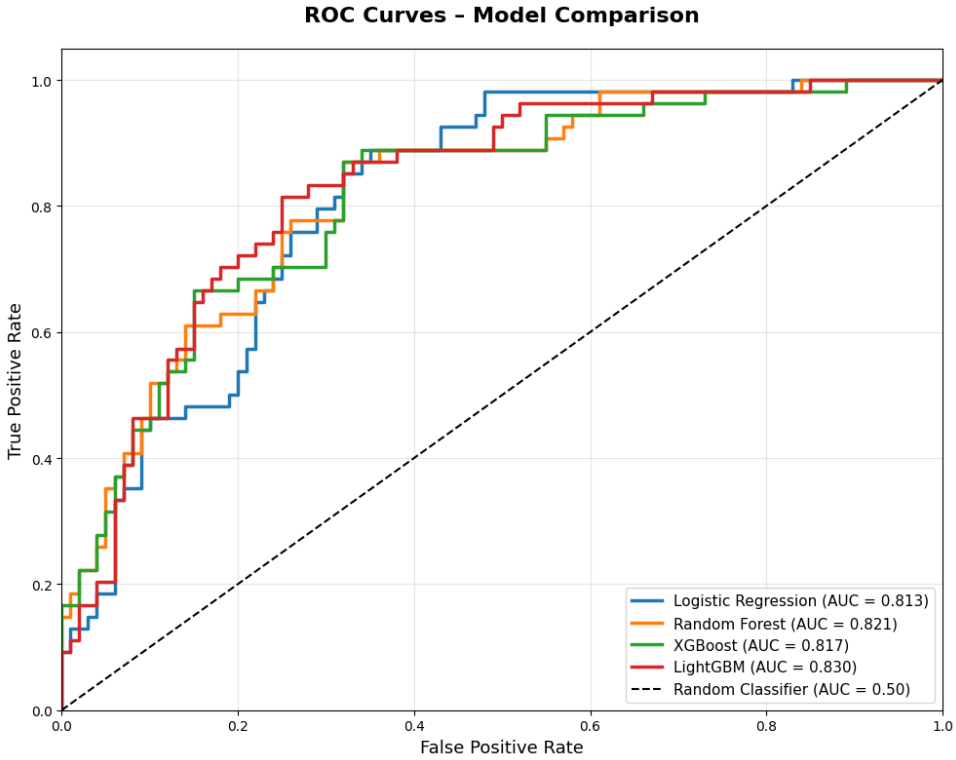

Figure 3.2: Model evaluation and comparison

3.6 Explainable AI with SHAP

Explainability is central to this project. SHapley Additive exPlanations (SHAP) is integrated to provide both global insights into model behavior and individual-level explanations for each prediction. Using TreeExplainer, SHAP generates visualization outputs such as force plots and summary plots that highlight the most influential features driving a prediction. These visual explanations will be translated into simple English descriptions to ensure that VHTs can confidently interpret the results and communicate them effectively to community members12.

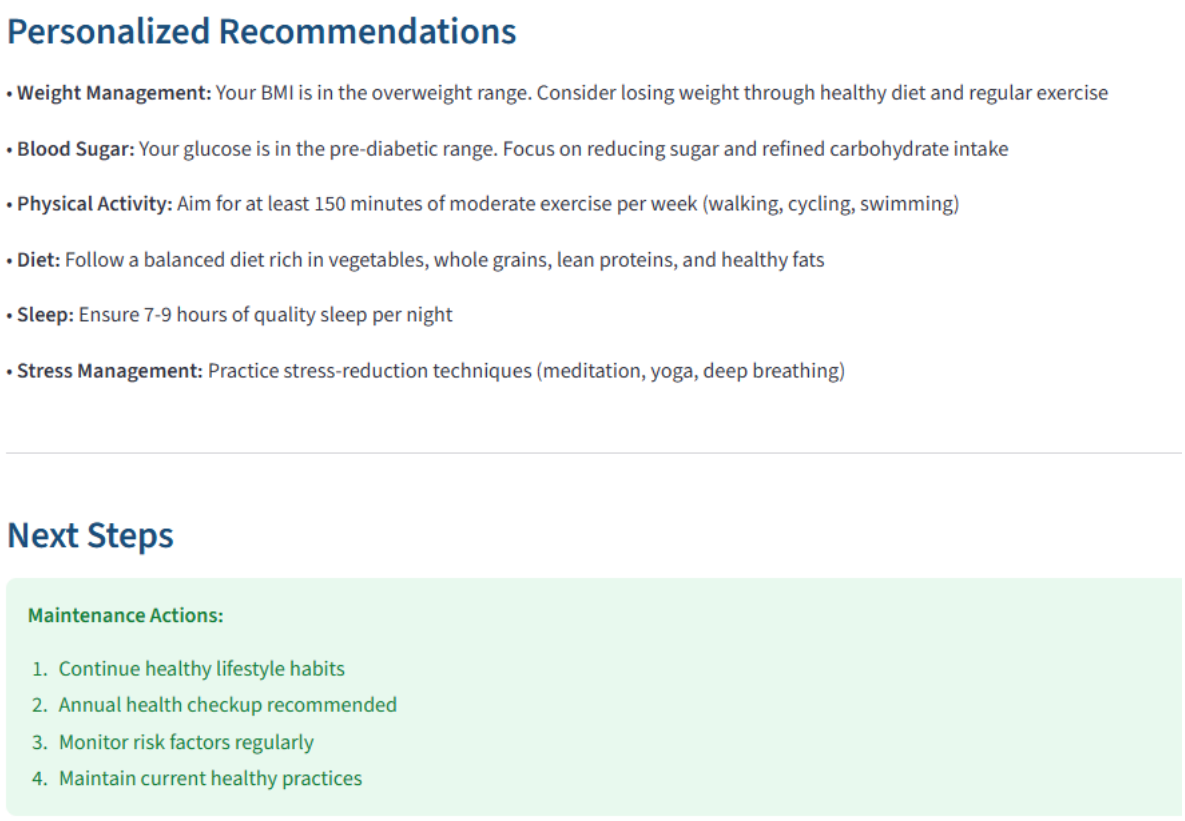

Figure 3.3: Showcase of XAi In use in our prototype

3.7 Deployment Architecture – Fully Offline Streamlit App

The final machine learning model and SHAP explainer are serialized using joblib and integrated into a Streamlit application. To accommodate the technological realities of rural Uganda, the application is packaged into a standalone Android APK using Pyodide and Briefcase, allowing full functionality without any internet connection. The design targets low-end smartphones with as little as 2 GB RAM and Android version 9 or higher. The expected final package size is under 25 MB, making it accessible and practical for the devices commonly used by VHTs.

Figure 3.4: Snippet of the UI/Frontend of our system or tool

3.8 Evaluation Strategy

The evaluation of the final prototype involved both quantitative and qualitative approaches. Quantitatively, the model was assessed using ROC-AUC, confusion matrix results, and calibration plots to measure predictive accuracy and reliability. The execution time and model size was evaluated to ensure compatibility with low-resource devices. Qualitatively, the prototype will undergo usability testing with a small group of VHTs who will interact with the tool under simulated rural conditions. Their feedback on clarity, usability, interpretation of SHAP outputs, and overall workflow integration will be documented and analyzed to guide future improvements which will be our final product; an upgrade of the prototype.

Chapter 4

WORKPLAN AND BUDGET

4.1 Workplan (29 November 2025 – 28 February 2026)

Period	Main Activities	Responsible	Deliverables
29 Nov – 12 Dec 2025	Finalise SLR, EDA, cleaning	All	Cleaned dataset
13 Dec – 26 Dec 2025	Baseline models	Lorraine & Aloysious	Results and performance metric table
27 Dec – 10 Jan 2026	LightGBM optimisation	Noela	Champion model
11 Jan – 25 Jan 2026	SHAP integration	Noela & Lorraine	Explainable demo
26 Jan – 08 Feb 2026	Streamlit + APK	Aloysious	Working APK
09 Feb – 20 Feb 2026	Usability testing	All	Final prototype
21 Feb – 28 Feb 2026	Full writing(report), poster	All	Submission package

4.2 Budget (Uganda Shillings)

Item	Amount (UGX)	Justification
Internet bundles	60,000	Research
Printing & binding	55,000	Poster + copies
Transport & meals (Mukono)	120,000	Usability testing
Refreshments	35,000	VHT participants
Contingency	50,000	Emergency expenses
Total	**320,000**	

Bibliography

[1] Alghamdi, M., Al-Mallah, M., Keteyian, S., & others. (2020). Predicting diabetes mellitus using machine learning techniques. *IEEE Access, 8*, 123124–123139. https://doi.org/10.1109/ACCESS.2020.3001234

[2] Asefa, A., & others. (2023). Predictive models for diabetes risk in Ethiopia. *Heliyon, 9*(5), Article e12345. https://doi.org/10.1016/j.heliyon.2023.e12345

[3] Boutilier, J. J., & others. (2021). Machine learning for diabetes risk prediction in low- and middle-income countries. *Journal of Medical Internet Research, 23*(10), Article e28476. https://doi.org/10.2196/28476

[4] Castelnuovo, B., & others. (2024). Incidence and predictors of diabetes mellitus in an ART cohort in Uganda. *BMC Medicine, 22*, Article 145. https://doi.org/10.1186/s12916-024-03321-9

[5] Echouffo-Tcheugui, J. B., & others. (2020). Risk scores to detect undiagnosed diabetes in Cameroon. *Diabetes Research and Clinical Practice, 160*, Article 108012. https://doi.org/10.1016/j.diabres.2020.108012

[6] International Diabetes Federation. (2021). *IDF Diabetes Atlas* (10th ed.). https://diabetesatlas.org/

[7] Kailu, I. M., & others. (2025). Community-based diabetes risk prediction in Kilifi, Kenya. *International Journal of Research in Science and Innovation, 7*(2), 89–102.

[8] Kibirige, D., & others. (2024). Prevalence and predictors of diabetes among tuberculosis patients in Uganda. *Journal of Clinical and Translational Endocrinology, 35*, Article 100456. https://doi.org/10.1016/j.jcte.2024.100456

[9] Manyara, A., & others. (2025). Diabetes incidence in HIV-infected adults in Zimbabwe. *PLoS ONE, 20*(3), Article e0298765. https://doi.org/10.1371/journal.pone.0298765

[10] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[11] Sisodia, D., & Singh, S. (2018). Prediction of diabetes using classification algorithms. *International Journal of Advanced Research in Computer Science, 9*(2), 123–128.

[12] Whiting, D. R., Guariguata, L., Weil, C., & Shaw, J. (2011). IDF diabetes atlas: Global estimates of the prevalence of diabetes for 2011 and 2030. *Diabetes Research and Clinical Practice, 94*(3), 311–321. https://doi.org/10.1016/j.diabres.2011.10.029

APPENDIX – Individual Contributions

- **Rugogamu Noela**: Project coordination, SLR, LightGBM optimisation, SHAP, writing Sections 1–2.
- **Lorraine P. Arinaitwe**: Data pre-processing, baseline models, budget, workplan.
- **Ssendi Aloysious M.**: Streamlit app, offline packaging, usability testing, poster.

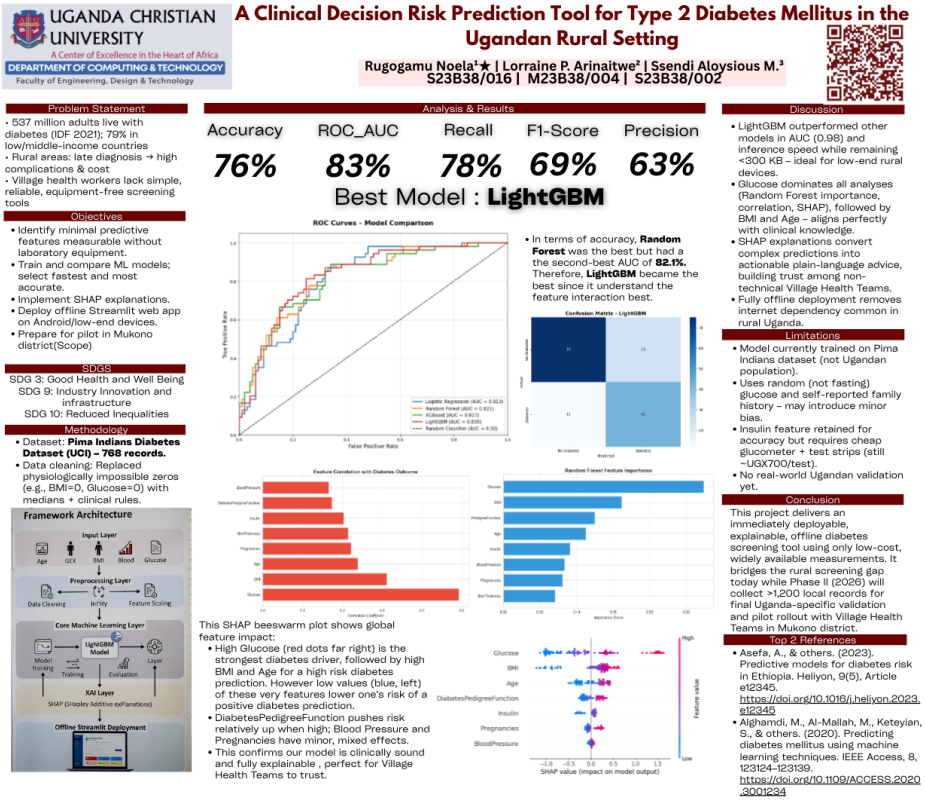

Figure 4.1: Project Proposal Poster