

KARATINA UNIVERSITY SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATICS

COURSE: BACHELOR OF SCIENCE IN COMPUTER SCIENCE

ACADEMIC YEAR 2024/2025

FINAL YEAR PROJECT PROPOSAL

NAME: NICODEMUS MUIMI MALOMBE

REG NO: P101/1749G/21

PROJECT TITTLE: **AI-POWERED DIABETES PREDICTION SYSTEM**.

# **LIST OF ABBREVIATIONS**

* AI: Artificial Intelligence
* BMI: Body Mass Index
* ML: Machine Learning
* WHO: World Health Organization
* API: Application Programming Interface

# ABSTRACT

Diabetes mellitus poses serious complications if not properly managed due to elevated blood glucose levels.

Early diagnosis of diabetes mellitus requires effective management and prevention of associated problems.

This project aims to develop a predictive model predicting diabetes probability based on physiological and medical profiles.

Supervised machine learning uses publicly accessible diabetes data including factors like pregnancy and glucose levels.

Selected algorithms optimize accuracy and computational efficiency enabling reliable predictions.

A web-based interface allows users to enter health information and receive instant forecasts with intuitive design.

System provides information on diabetes risk and encourages consultation with healthcare professionals when necessary.

Review emphasizes need for accessible and accurate tools by addressing gaps in existing diabetes prediction models.

Project contributes to population health through early detection and intervention using data preprocessing and feature engineering.

Machine learning demonstrates importance of technology in improving healthcare access and outcomes with further developments.

Model could become an important tool in fighting diabetes with integration into health care systems.

Table of Contents

[**LIST OF ABBREVIATIONS** ii](#_Toc188648478)

[ABSTRACT iii](#_Toc188648479)

[CHAPTER ONE vii](#_Toc188648480)

[1.1 INTRODUCTION vii](#_Toc188648481)

[1.2 BACKGROUND OF THE STUDY viii](#_Toc188648482)

[1.3 Problem Statement ix](#_Toc188648483)

[1.4 Objectives ix](#_Toc188648484)

[1.4.1 General Objective ix](#_Toc188648485)

[1.4.2 Specific Objectives ix](#_Toc188648486)

[1.5 SCOPE x](#_Toc188648487)

[1.6 Limitations of the study x](#_Toc188648488)

[Chapter Two xi](#_Toc188648489)

[Literature Review xi](#_Toc188648490)

[CHAPTER THREE xiii](#_Toc188648491)

[3.1 METHODOLOGY xiii](#_Toc188648492)

[3.2 Input Design xiii](#_Toc188648493)

[3.3 Output Design xiii](#_Toc188648494)

[**3.4 Tools and Technologies** xiv](#_Toc188648495)

[3.5 Deployment Tools xiv](#_Toc188648496)

[3.6 Workflow xv](#_Toc188648497)

[3.7 BUDGET xv](#_Toc188648498)

[**3.8 Gantt Chart** xv](#_Toc188648499)

# CHAPTER ONE

## 1.1 INTRODUCTION

Diabetes mellitus, is a chronic disease that has a significant impact on the health and well-being of people around the world. Characterized by high levels of blood glucose from the pancreas without sufficient insulin levels or the body’s inability to properly utilize insulin production Diabetes has become more common in recent decades, and has been one of the leading causes of death and disability worldwide. Several factors are responsible for this increase, including population aging, urbanization, and social change.

With the increasing prevalence of diabetes, controlling and avoiding its consequences now increasingly depends on early detection and early intervention. Early detection before serious consequences such as heart disease, kidney failure, and neurological disorders enables people to make necessary lifestyle changes and seek treatment but can be prevented due to how ways standardized diagnostic methods are resource intensive, often requiring a laboratory, skilled personnel, and more time especially in areas of limited resources

The availability of machine learning (ML) and artificial intelligence (AI) is providing new opportunities to develop predictive algorithms that can accurately identify those at risk for developing diabetes. Machine learning algorithms use existing patient data such as age, blood pressure, glucose levels and BMI. This technology makes it productive, affordable and accessible substitute for conventional diagnostic techniques.

This study presents a diabetes prediction system that uses machine learning to identify people at risk of developing the disease. This system is able to make accurate predictions based on patient data and is easy to use. It can also be accessed through a web-based interface. This project aims to support early detection efforts, close access gaps, and integrate AI and healthcare to improve diabetes management worldwide.

.

## 1.2 BACKGROUND OF THE STUDY

Diabetes manifests with elevated blood glucose levels and can lead to complications such as heart disease kidney failure eye damage.

Traditional diagnostic methods often require laboratory tests medical advice that may not be available to everyone.

Technological advances over years have paved way for new approaches to prediction and management of diabetes.

Historically diagnosis of diabetes relied heavily on clinical methods like fasting glucose levels oral glucose levels.

Although effective these methods are resource intensive may not be feasible in resource-limited situations.

Digital health tools emerged in early 2000s with basic decision support systems designed to help clinicians diagnose manage diabetes.

Machine learning and artificial intelligence have revolutionized medical research providing original applications with powerful predictive models.

Today web-based platforms using machine learning models have increased accessibility enabling patient healthcare professionals to interact with predictive tools in real time.

This work builds on advances to develop a web-based system to better predict diabetes risk.

Proposed system incorporating learning methods aims to provide prediction accuracy and highlight utility accessibility for wider audience including individuals in underserved regions.

## 1.3 Problem Statement

Diabetes affects millions worldwide and grows at an unprecedented rate. Chronic disease leads to serious pathologies like cardiovascular disease neuropathy and kidney failure highlighting necessity of early detection and management. Diagnostic modalities like fasting glucose determination and HbA1c measure are resource demanding needing clinical settings and specific devices. These methods provide challenges particularly in low-resource regions where access to medical facilities and trained personnel is larger. Many people are undiagnosed as early-stage diabetes is asymptomatic meaning treatment is delayed and risk for complications is increased. Absence of easily accessible and intuitive tools to assess diabetes risk aggravates problem. Recent developments in ML algorithms and AI techniques show promise but most existing solutions end up being either too complex for non-specialists or inaccessible to general population. Creating a web-based machine learning-based diabetes prediction system fills these gaps by providing easy to use and intuitive platform. System seeks to close access gap by precisely determining risk of diabetes through cutting-edge machine learning techniques. Suggested approach aims to equip people and healthcare professionals with resources that facilitate early detection and prompt action leads to better health consequences and lessens burden of disease worldwide.

.

## **1.4 Objectives**

1.4.1 General Objective**:**

This project aims to build a web-based diabetes prediction system through machine learning techniques that can serve as an accessible, effective, and accurate tool for detecting diabetes risk at an early stage.

* + 1. Specific Objectives**:**
* Develop a model that predicts the diabetes risk using the patient data.
* To assess the efficiency of predictive model considering suitable metrics.
* To make the system user friendly and easily accessible.
* To Evaluate and Review the Existing Diabetes Prediction Systems

## 1.5 SCOPE

The scope of this project includes the development and implementation of a web-based diabetes prediction system using machine learning. The system is designed to cater to individuals and healthcare professionals by providing early detection of diabetes risk based on user-provided information such as age, sex, lifestyle, medical history and health on the issues

## Limitations of the study

* The accuracy of the system depends on the quality of the training data.
* It does not reflect genetic traits or lifestyle factors beyond the investment range.
* The model is based on the prediction of the risk of developing diabetes and not on other relevant conditions.

# Chapter Two

## Literature Review

Diabetes, is a major global health issue, with millions of individuals affected worldwide. Early detection and accurate prediction of diabetes are critical for preventing complications such as cardiovascular diseases, kidney failure, and neuropathy. Traditional diagnostic methods, while effective, are often limited by the complexity and variability of patient data. To address these challenges, artificial intelligence (AI) and machine learning (ML) techniques have emerged as powerful tools for predicting diabetes based on a variety of patient data.

AI and ML algorithms have shown considerable promise in healthcare, particularly in predictive modeling. By analyzing large datasets, these models can identify patterns and correlations that may not be immediately apparent to healthcare professionals. In the context of diabetes prediction, machine learning models such as logistic regression, decision trees, support vector machines, and neural networks have been widely used. These models analyze clinical features like glucose levels, age, BMI, blood pressure, and family history to classify individuals as either at risk for diabetes or not.

Various studies have employed a range of ML techniques to dev diabetes prediction models. Logistic regression, although simple, is commonly used for binary classification tasks, estimating the probability of a patient developing diabetes. Decision trees and random forests are popular for their ability to handle both numerical and categorical data, providing clear decision rules that make them easy to interpret. Support vector machines (SVM), known for their robustness in high-dimensional spaces, have also been applied to diabetes prediction. Additionally, neural networks, particularly deep learning models, have been explored for their ability to learn complex, non-linear relationships in large datasets.

Datasets such as the PIMA Indians Diabetes Database and the Diabetes 130-US hospitals dataset are commonly used for training and evaluating these models. These datasets include features like glucose concentration, BMI, blood pressure, and family history, which are known to be significant factors in predicting diabetes. Feature selection and engineering are important steps in preparing data for machine learning models, as identifying the most relevant features can improve model performance.

The evaluation of machine learning models in diabetes prediction typically involves metrics like accuracy, precision, recall, F1-score, and ROC-AUC, which measure the model’s ability to correctly classify individuals as diabetic or non-diabetic. Cross-validation techniques are used to assess the model’s generalization ability and prevent overfitting, ensuring the model performs well on unseen data.

Despite the promising results of AI in diabetes prediction, several challenges remain. The quality of data can vary significantly, with missing values, biases, and imbalances in the dataset potentially affecting model performance. Moreover, the complexity of deep learning models, while providing high accuracy, can lead to a lack of interpretability, making it difficult for healthcare professionals to understand the reasoning behind predictions. Ethical concerns surrounding patient privacy and data security also need to be addressed, particularly with the increasing use of electronic health records and wearable devices.

Looking ahead, the integration of AI-powered diabetes prediction systems with real-time data from wearable devices, such as continuous glucose monitors, presents exciting opportunities. These devices can provide continuous data streams, enabling early and personalized interventions. The future of diabetes prediction may also benefit from the use of genetic data, which could further improve the accuracy of risk assessments and facilitate more personalized treatment plans.

In conclusion, AI-powered diabetes prediction systems have the potential to significantly improve early detection, diagnosis, and prevention of diabetes. While progress has been made, there are still challenges related to data quality, model interpretability, and ethical concerns. However, ongoing advancements in AI, data collection, and healthcare technologies hold the promise of more accurate, efficient, and personalized diabetes care in the future.

# CHAPTER THREE

## 3.1 METHODOLOGY

For my project (AI-powered diabetes prediction system) a **Supervised Machine Learning** methodology is appropriate. This approach is best suited for classification tasks, such as predicting whether a patient is at risk of developing diabetes based on their clinical data. Here's why this methodology is chosen:

1. **Supervised Learning**: The project will use labeled data, where each patient record includes both input features (e.g., glucose level, BMI, age, etc.) and an associated label indicating whether the person has diabetes or not (binary classification: diabetic or non-diabetic).
2. **Feature Selection and Engineering**: By selecting the most relevant features from the dataset (e.g., glucose levels, age, BMI, blood pressure), the model can be trained to recognize patterns that indicate diabetes risk.
3. **Model Training and Evaluation**: The project will involve training multiple machine learning models (e.g., logistic regression, decision trees, random forests, SVM, neural networks) and evaluating their performance using standard metrics like accuracy, precision, recall, and F1-score.
4. **Validation**: Cross-validation will be used to ensure that the model generalizes well to unseen data, preventing overfitting.

## 3.2 Input Design

The input for the system will be clinical data from patients. Each record will contain the following features, which are typically used for diabetes prediction:

* **Glucose level**: The blood sugar level of the patient (a key indicator of diabetes risk).
* **BMI (Body Mass Index)**: The body mass index of the patient, which is a significant risk factor.
* **Age**: The age of the patient, as diabetes risk increases with age.
* **Blood pressure**: Blood pressure levels, which correlate with diabetes and heart disease.
* **Family history**: Information about whether the patient has a family history of diabetes
* **Insulin level**: Measures the insulin levels, which can indicate insulin resistance, a precursor to diabetes

3.5 **Output**

* Displays prediction results as either “Diabetes” or “No Diabetes.”
* **Probability**: The model can also output the probability of the patient being diabetic, which provides more granularity on the prediction and allows healthcare professionals to assess the risk level.

## **3.4 Tools and Technologies**

The following tools and technologies will be used to develop and deploy the AI-powered diabetes prediction system:

1. **Programming Languages**:
   * **Python**: The primary language for implementing machine learning models due to its rich ecosystem for data analysis and AI development.
2. **Libraries and Frameworks**:
   * **Scikit-learn**: A powerful library for machine learning in Python, offering implementations for a wide range of algorithms like decision trees, random forests, SVM, logistic regression, and more.
   * **Pandas**: For data manipulation and preprocessing (e.g., handling missing data, scaling features).
   * **NumPy**: For numerical operations and handling arrays.
   * **Matplotlib/Seaborn**: For data visualization and model performance evaluation.
3. **Tools for Model Evaluation**:
   * **Cross-validation**: To prevent overfitting and ensure the model generalizes well.
   * **Evaluation Metrics**: Accuracy, Precision, Recall, F1-score, ROC-AUC.

## 3.5 Deployment Tools

**Flask**: For building a web API to serve the prediction model

**Development environment**: **VS Code**: For coding and experimenting with machine learning models

**Deployment:** Flask server with static HTML, CSS, and JavaScript

## 3.6 Workflow

* **Data Preprocessing**: Import the dataset, clean the data (handling missing values, encoding categorical features), and normalize the numerical features.
* **Model Selection and Training**: Train models using algorithm and evaluate its performance.
* **Model Evaluation**: Assess the models using cross-validation and performance metrics.
* **Deployment**: Deploy the best-performing model as a web API for real-time predictions, if necessary.

## 3.7 BUDGET

|  |  |
| --- | --- |
| ITEM | COST(Ksh) |
| Internet | 5000 |
| Development tools | free |
| Data acquisition | 1500 |
| Model training | 2000 |
| Miscellaneous | 1000 |
| Total | 9500 |

## **3.8 Gantt Chart**

|  |  |
| --- | --- |
| TASK | DURATION |
| Literature Review | 2 WEEKS |
| Data Collection | 3 WEEKS |
| Model Development | 2 WEEKS |
| System Design and Integration | 3 WEEKS |
| Final Documentation | 2 WEEKS |