Predicting Cardiovascular Disease Risk Among Bangladeshi Diabetes Patients Using Machine Learning and Explainable AI

By
Sadia Islam
ID: 211-15-3980

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by

Md. Sazzadur Ahamed
Assistant Professor
Department of Computer Science and Engineering Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

January 12, 2025

APPROVAL

This Project titled **"Predicting Cardiovascular Disease Risk Among Bangladeshi Diabetes Patients Using Machine Learning and Explainable AI"**, submitted by **Sadia Islam**, ID No: **211-15-3980** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 12 January, 2025.

BOARD OF EXAMINERS

Dr. Sheak Rashed Haider Noori — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Sharmin Akter — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Md Mohammad Masum Bakaul — **Internal Examiner**
Sr. Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Zulfiker Mahmud — **External Examiner**
Professor
Department of Computer Science and Engineering
Jagannath University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Md. Sazzadur Ahamed, Assistant Professor** Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this projectnor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Md. Sazzadur Ahamed
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Sadia Islam
Student ID: 211-15-3980
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project(FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Md. Sazzadur Ahamed, Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Artificial Intelligence and Machine Learing** to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting themat all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Cardiovascular disease (CVD) is a leading cause of death, especially among diabetes patients, due to metabolic and lifestyle factors. This study aims to predict cardiovascular disease risk among Bangladeshi diabetes patients using machine learning and explainable AI, focusing on the role of Mediterranean diet adherence. Cardiovascular disease (CVD) is a leading cause of morbidity and mortality in diabetes patients, and early prediction can significantly improve health outcomes. A predictive model based on machine learning algorithms, particularly LightGBM , was developed and evaluated, achieving the highest accuracy of 99.37%. The model was further enhanced with explainable AI techniques to ensure transparency and interpretability of predictions. The dataset utilized clinical, demographic, and lifestyle data, including factors such as triglycerides, sleep patterns, smoking habits, and Mediterranean diet adherence. Results revealed that smoking, sleep hours, and diet adherence were the most influential factors in predicting cardiovascular risk. The model demonstrated strong performance across various evaluation metrics, including precision, recall, and F1-score, further validating its effectiveness. This research underscores the potential of machine learning to transform healthcare by providing early diagnosis tools, allowing for personalized interventions. The model can assist healthcare professionals in identifying at-risk individuals and reducing the burden of cardiovascular diseases in Bangladesh and similar regions, ultimately improving patient outcomes through early detection and targeted-prevention-strategies.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces the study, focusing on cardiovascular disease (CVD) risk prediction among Bangladeshi diabetes patients using machine learning and explainable AI. It highlights the role of Mediterranean diet adherence as a key factor in mitigating CVD risks.

1.1 Introduction

Cardiovascular diseases (CVD) remain a major public health concern globally, accounting for significant morbidity and mortality rates[1]. For individuals with diabetes, the risk of developing CVD is particularly elevated due to interconnected factors such as chronic hyperglycemia, insulin resistance, and inflammation[2][3]. In Bangladesh, the prevalence of diabetes has been increasing at an alarming rate, creating additional strain on an already overburdened healthcare system. Despite this, current CVD risk assessment methods often rely on generalized models that fail to address the unique epidemiological and cultural characteristics of Bangladeshi populations[4][5].

This project seeks to bridge this gap by employing advanced machine learning techniques and explainable AI to develop a robust, context-aware model for predicting CVD risk among diabetic patients in Bangladesh. The research integrates traditional clinical data with behavioral and lifestyle factors, focusing on adherence to the Mediterranean diet—a dietary pattern known for its cardiovascular benefits[6][7]. By identifying key risk contributors and emphasizing actionable prevention strategies, this study aims to support clinicians in delivering personalized care and to inform public health policies tailored to the region's needs. Through the innovative application of technology and data-driven insights, this research aspires to make a meaningful impact on reducing the burden of CVD among Bangladeshi diabetes patients. The findings will not only enhance early detection and intervention but also pave the way for scalable solutions that can be adapted to similar contexts worldwide

1.2 Motivation

The motivation for this research stems from the growing burden of diabetes and its complications, particularly CVD, in Bangladesh. As a developing country, Bangladesh faces significant resource constraints in healthcare, which makes it imperative to find cost-effective and accurate ways to assess and manage health risks. Machine learning offers an unparalleled opportunity to utilize existing data more effectively, transforming it into actionable insights that can save lives and reduce healthcare costs.

On a personal level, this research is driven by the desire to contribute to public health advancements in underrepresented populations. Addressing CVD risks among Bangladeshi diabetes patients is not only a scientifically challenging problem but also one with profound societal implications. Successfully solving this issue could lead to improved health outcomes for millions, serving as a model for other low- and middle-income countries facing similar challenges.

Moreover, this project offers an opportunity to explore the intersection of technology and healthcare, enhancing expertise in cutting-edge tools like machine learning and explainable AI. The potential to make a tangible difference in people's lives while advancing academic and professional goals makes this research both meaningful and rewarding.

1.3 Objectives

The main objectives of this research are:

- **Develop a Predictive Model:** Build a machine learning model to predict cardiovascular disease risk among Bangladeshi diabetes patients.
- **Implement Explainable AI:** Ensure the model's predictions are interpretable and transparent for healthcare professionals.
- **Explore Mediterranean Diet Impact:** Investigate the role of Mediterranean diet adherence in reducing cardiovascular risk for diabetes patients.
- **Provide Healthcare Insights:** Offer actionable insights for improving healthcare strategies in Bangladesh, focusing on lifestyle and dietary factors.

1.4 Methodology

In this study, we aimed to predict the risk of cardiovascular disease among Bangladeshi diabetes patients using machine learning and explainable AI. The methodology began with data collection from two primary sources, focusing on Mediterranean diet adherence and cardiovascular disease risk factors such as age, gender, triglyceride levels, smoking habits, and sleep patterns. After preprocessing the data to handle missing values and normalize variables, we developed and trained various machine learning models, including Random Forest, Logistic Regression, SVC, and XGBoost, using cross-validation techniques for optimization. To enhance interpretability, we applied explainable AI methods like SHAP to identify the most influential factors in the predictions. Finally, we evaluated the models using accuracy, precision, recall, F1-score, and AUC, and analyzed the results to provide actionable insights into how lifestyle factors, such as diet and habits, contribute to cardiovascular risk in diabetic patients. This comprehensive approach combined machine learning with explainable AI to build an accurate and transparent model for predicting cardiovascular disease risk in this population.

1.5 Project Outcome

The outcomes of this project could have significant implications for both healthcare and public health in Bangladesh. By predicting the risk of cardiovascular disease among diabetes patients, our machine learning model could serve as an early detection tool, helping healthcare providers identify high-risk individuals and offer timely interventions. Additionally, the incorporation of explainable AI methods ensures that the model's predictions are transparent, allowing healthcare professionals to understand the key factors influencing cardiovascular risk, such as diet, smoking, and sleep patterns. This transparency can lead to more personalized treatment plans and lifestyle recommendations. Ultimately, the project aims to contribute to improved healthcare outcomes for diabetic patients, reducing the burden of cardiovascular diseases in Bangladesh. Furthermore, the methodology and insights gained from this work could be adapted and applied to other regions or conditions, expanding its potential impact globally.

1.6 Organization of the Report

This report is systematically organized to ensure a consistent and logical progression of the project's development.

Chapter 1 commences by defining the study's history, objectives, methods, and anticipated results. This chapter establishes a robust basis for the remainder of the report, ensuring that readers are adequately educated about the study's context and objectives.

Chapter 2 presents a thorough examination of pertinent literature. This chapter emphasizes previous studies on cardiovascular disease, diabetes, and the contributions of machine learning and food in forecasting health risks. The literature review offers a critical context for the current investigation, highlighting advancements and deficiencies in the existing knowledge base.

Chapter 3 examines the study technique, outlining the procedures for data collection and preparation. It delineates the models employed and enumerates the procedures undertaken in the analysis. This chapter is essential as it elucidates the methodologies utilized, guaranteeing the study's reproducibility and dependability.

Chapter 4 presents the implementation and results. The report assesses the model's efficacy and examines the significance of different characteristics. This chapter elucidates the practical implementation of the research, emphasizing the model's strengths and limitations.

Chapter 5 examines the engineering standards, design obstacles, and the wider socioeconomic and environmental ramifications of the project. This chapter highlights the practical ramifications of the research, examining how the findings may impact real-world applications and enhance the wider field.

Chapter 6 concludes the report by synthesizing the findings, addressing the study's shortcomings, and providing recommendations for further research. This last chapter summarizes the study, offering a detailed account of the accomplishments and proposing avenues for further research.

Chapter 2

Background

This chapter explores the key concepts, theoretical frameworks, and previous research that inform this study. It highlights the connections between diabetes, cardiovascular risks, and advancements in machine learning for healthcare solutions.

2.1 Introduction

Understanding the intricate relationship between diabetes and cardiovascular disease (CVD) requires a comprehensive examination of the interconnected mechanisms and shared risk factors underlying these chronic conditions. Diabetes, a complex metabolic disorder characterized by persistent hyperglycemia, significantly heightens the risk of cardiovascular complications. Poor glucose control leads to a cascade of adverse physiological processes, including systemic inflammation, oxidative stress, and endothelial dysfunction, all of which contribute to the development and progression of CVD. These mechanisms create a vicious cycle where diabetes exacerbates cardiovascular risks, and CVD, in turn, complicates diabetes management.

Cardiovascular diseases encompass a wide range of conditions, such as coronary artery disease, heart failure, and stroke, which collectively account for the leading causes of mortality worldwide. Among individuals with diabetes, the prevalence of these conditions is disproportionately high, often driven by diabetes-related factors like dyslipidemia, hypertension, and obesity. This synergistic interaction underscores the urgent need for integrative approaches that address the dual burden of diabetes and CVD.

One promising avenue for mitigating these risks is the adoption of the Mediterranean diet, a dietary pattern rich in fruits, vegetables, whole grains, legumes, nuts, and heart-healthy fats like olive oil. This diet has been extensively studied for its protective effects on cardiovascular health, including its ability to reduce inflammation, improve lipid profiles, and enhance endothelial function. These benefits are particularly relevant for individuals with diabetes, as they target many of the pathways implicated in both diabetes and CVD.

2.2 Literature Review

A literature review is a careful examination of scholarly work on a given topic. It contributes to the summarization of existing knowledge by providing an overview of major ideas, techniques, and any research gaps. Examining these sources allows researchers to determine how their own work fits into the larger picture, ensuring that it builds on previous discoveries while also offering something new and useful to the area.

Table 2.1: Summary of Literature Reviewed.

Author (s)	Year	Title	Methodology	Key Findings
Piché et al..	2020	Obesity Phenotypes, Diabetes, and Cardiovascular Diseases	biological mechanisms analysis	Obesity is a major risk factor for CVD, and visceral fat is more important than BMI for predicting risk..
Becerra-Tomás et al.	2019	Mediterranean diet, cardiovascular disease and mortality in diabetes: A systematic review and meta-analysis of prospective cohort studies and randomized clinical trials	Systematic review and meta-analysis of cohort studies and RCTs	Mediterranean diet reduces CVD incidence and mortality, especially among those with diabetes.
Martín-Peláez et al.	2020	Mediterranean Diet Effects on Type 2 Diabetes Prevention, Disease Progression, and Related Mechanisms. A Review	biological mechanism analysis	Mediterranean diet linked to reduced T2D incidence and better disease management.
Mandava et al.	2024	An All-Inclusive Machine Learning and Deep Learning Method for Forecasting Cardiovascular	Machine learning (Logistic Regression, Naive Bayes, etc.)	ML and DL models achieved 96.7% accuracy in predicting CVD in the Bangladeshi population.

		Disease in Bangladeshi Population		
Hossain et al.	2023	Machine learning approach for predicting cardiovascular disease in Bangladesh: evidence from a cross-sectional study in 2023	Machine learning (Random Forest, Logistic Regression, etc.)	Random Forest classifier achieved 98% accuracy for CVD prediction in Bangladesh.
Islam et al.	2021	Cardiovascular Disease Forecast using Machine Learning Paradigms	Logistic Regression, Decision Tree, SVM, Naive Bayes	Logistic Regression (92.10%) outperformed other models in predicting CVD using the Cleveland dataset.
Hossen et al.	2021	Supervised Machine Learning-Based Cardiovascular Disease Analysis and Prediction	Supervised machine learning (RF, DT, LR)	Logistic Regression (92.10%) outperformed other models in predicting CVD using the Cleveland dataset.
Asif et al.	2021	Performance Evaluation and Comparative Analysis of Different Machine Learning Algorithms in Predicting Cardiovascular Disease	Machine learning (Adaboost, Voting Classifiers)	Adaboost performed best with precision (0.938) and specificity (0.926), achieving 92% accuracy
Tamanna et al.	2021	Early Prediction of Cardiovascular Diseases Using Feature Selection and Machine Learning Techniques	Machine learning (XGBoost, Random Forest)	XGBoost achieved the highest accuracy (75.10%) in predicting CVD, with strong performance across other metrics
Md. Shahiduzzaman et al.	2022	Prognosis of Cardiovascular Disease Using Machine Learning Procedures	Ensemble classifier (KNN, LR, RF, GB, Naive Bayes)	Voting ensemble classifier with KNN achieved 75% accuracy in predicting CVD.

Fahim et al.	2022	Detection of Cardiovascular Disease of Patients at an Early Stage Using Machine Learning Algorithms	Machine learning (KNN, RF, Decision Trees, XGBoost)	Enhanced dataset improved accuracy from 73.72% to 81.14% for CVD prediction.
Muhammad et al.	2023	Predictis: an IoT and machine learning-based system to predict risk level of cardio-vascular diseases	IoT, Machine learning (stacking classifier)	IoT and machine learning system predicted CVD risk levels with an F1 score of 80.4%.

2.2.1 Related Research

La Sala et al. [1] also show that obesity is a major contributor to cardiovascular disease, impaired glucose tolerance, and type 2 diabetes. They say that the most effective treatment for obesity, weight loss and decreased cardiovascular risk is bariatric metabolic surgery, noting that lifestyle changes and medications have limited effectiveness. Study demands more detail and individualised treatments of obesity-related pathways, calls for more studies of non-surgical therapies for long-term CVD consequences.

Piché et al. [3] point out that BMI alone does not predict cardiovascular risk, pointing towards the contribution of visceral and ectopic fat to obesity-related disorders. Although metabolic surgery is an option for extreme obesity, conventional therapies often do not provide adequate success. They suggest integrating environmental and personal strategies, both AI-based and at the community level, to foster comprehensive solutions.

Predicting cardiovascular risks in the Bangladeshi population using ML algorithms was accomplished by Mandava et al. [6] and they reported an accuracy of 96.7 percent. They did also attempt DL for medical images and predicting drugs however they pointed out major drawbacks in algorithmic reliance and low population diversity issues. In the future it will be important to improve the models and increase the size of the datasets for wider clinical applicability.

According to Hossain et al [7] they conducted a study on the prediction of cardiovascular diseases using machine learning bearing in mind that, Random Forest recorded the highest accuracy of 98.04% and AUC 0.989 in Bangladesh. Even though the research was conducted with a small sample and had low validation over the dataset, the research substantiates the idea of machine learning techniques enhancing the risk estimates and the clinical practice of prediction frameworks for cardiovascular disease. Going forward, investigations should concentrate on bigger data repositories and other ML methods.

Islam et al. [9] employed machine learning techniques to identify patients with CVD in South Asians, with Logistic Regression attaining 86.25% accuracy. The study suffers limitations such as a small sample size and a small number of attributes, but it demonstrates the capability of ML for early prediction. In this regard, future studies should increase the size of the datasets and deal with more sophisticated methods for clinical applications.

Tamanna et al [10] employed Random Forest feature selection, and XGBoost had the greatest accuracy of 75.10%. Since several of the restrictions were searching for a single dataset, the advocacy of ML's efficacy in detecting disease earlier cannot be overemphasized. It is demonstrated that more severe health applications should investigate new algorithms and larger clinical datasets.

Md. Shahiduzzaman et al [11] in their study employed several machine learning models like KNN, Random Forest, and Logistic Regression in a voting ensemble classifier, with KNN predicting early CVD with an accuracy of 75%. The study emphasizes the promise of insights brought forth by ML, but points out the shortcomings such as relying on a single dataset. Future studies should focus on more complex methods and include many more heterogeneous datasets for an even better accuracy results.

2.3 Gap Analysis

Several important gaps in the present research landscape still exist, despite the fact that the works reviewed in this paper show notable breakthroughs in the use of machine learning (ML) and deep learning (DL) approaches for cardiovascular disease (CVD) prediction. To increase the efficacy of CVD prediction models, these gaps point to areas that require development and additional research.

Data Diversity and Generalization:
Many studies rely on region-specific datasets, limiting the generalizability of the models. Future research should use diverse, large-scale datasets to ensure broader applicability across populations.

Feature Selection and Model Complexity:
Some studies focus on limited features, potentially missing key factors. Expanding feature sets and balancing model complexity with accuracy is crucial for improving predictive power.

Small Sample Size and Lack of Validation:
Small sample sizes lead to overfitting and reduced robustness. Future work should prioritize larger, more diverse datasets and cross-validation to ensure reliability and scalability.

Integration of Real-Time Data and Deployment:
Most studies use static datasets and lack real-time data integration. There is a need to develop systems that incorporate real-time patient data for continuous monitoring and early intervention.

Advanced Techniques and Hybrid Models:
Traditional machine learning models dominate, but advanced techniques like deep learning and hybrid models could offer better accuracy. Exploring these models could improve prediction outcomes.

Clinical Applicability and Real-World Impact:
Many models are not validated for real-world use. More focus is needed on practical implementation and integration of these models into clinical settings for routine use.

2.4 Summary

This section examined the literature on cardiovascular diseases (CVD), concentrating on prediction models and treatments. Studies underlined the significance of early detection and lifestyle adjustments, such as the Mediterranean diet. Machine learning models such as Random Forest and LightGBM have excellent results in predicting CVD in T2D patients, particularly in Bangladesh. However, difficulties such as dataset heterogeneity and limited sample numbers persist. Future study will focus on improving CVD outcomes by strengthening prediction models and embracing precision medicin

Chapter 3

Research Methodology

Chapter 3 outlines the methods used to predict cardiovascular disease risk, including dataset preparation, machine learning techniques, Explainable AI, and the role of Mediterranean diet adherence in model evaluation.

3.1 Methodology/Requirement Analysis & Design Specification

This research followed a structured process, starting with collecting and preparing the data to ensure it was clean and ready for analysis. We enhanced the dataset through feature engineering, selected the best machine learning models, and fine-tuned their settings for better performance. Using advanced tools, we implemented these models and applied explainable AI (XAI) techniques to make the predictions clear and meaningful. This approach ensures our findings are both accurate and easy to understand.

3.1.1 Overview

This research is designed to accomplish a prediction of the probability that a person will develop cardiovascular disease (CVD) through the application of machine learning techniques on a dataset of 796 records from two known surveys. Among the important variables include age, gender, BMI, smoking, drinking alcohol, total triglycerides, HDL cholesterol and LDL cholesterol, as well as hypertension and kidney disease.

As part of data preprocessing, complementing missing information was done by first eliminating columns with too many missing values, and then replacing the missing values in other columns with the mean value. The dataset included numerical variables only, was divided into 80 % training set, 20 % validation set and 20 % testing set to help make a fair and unbiased model.

The pipeline explains the use of cross-validation-based hyperparameter optimization, feature selection and XAI approaches to enhance the predictive power and the explainability of the models. This method also enhances the understanding of the aetiology of CVD and provides good means of identifying or preventing it from early stages.

3.1.2 Proposed Methodology/ System Design

In this section, we detail the study plan we used to conduct our research. The approach consisted of several essential processes, including data collection, pre-processing and engineering features to enhance the dataset, selecting relevant machine learning models, fine-tuning their hyperparameters, and implementing the models with appropriate software tools. Finally, we used an explainable AI (XAI) framework to analyze and offer information about the model predictions. Figure 1 provides a clear and brief summary of our methodology, reflecting the core of the suggested strategy.

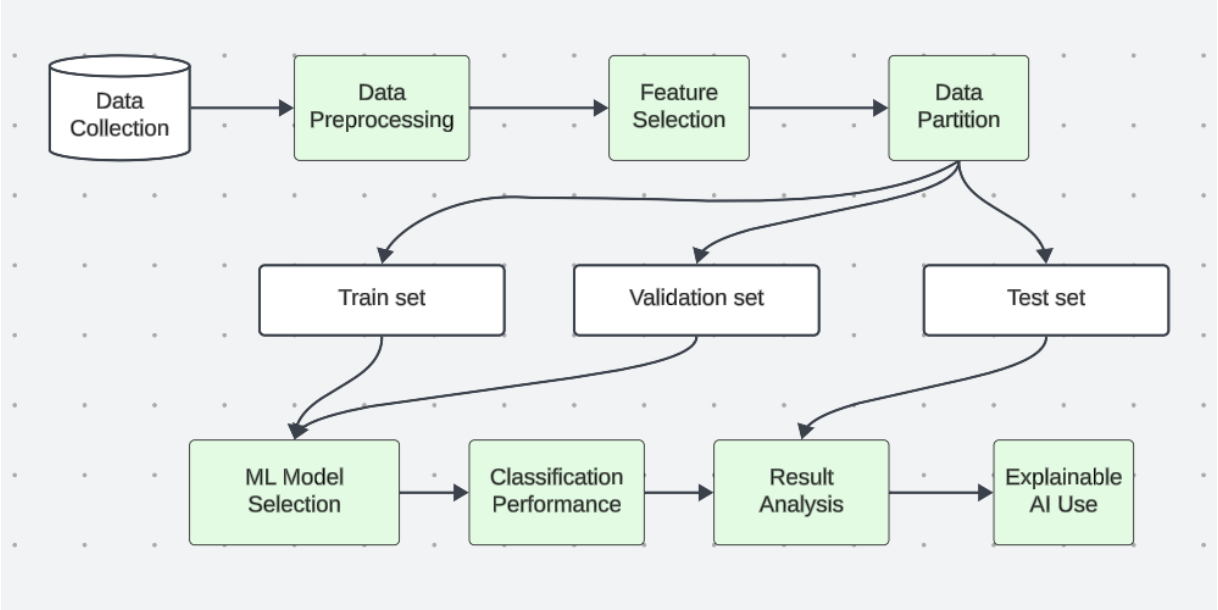


Figure 3.1: Proposed Methodology for Cardiovascular Classification.

3.2 Detailed Methodology and Design

In Section 3.2, we provide a comprehensive explanation of the methodology and design framework employed in our study. This section offers a detailed breakdown of the entire process, encompassing each step and approach utilized in the research. The subsequent subsections outline and elaborate on these components in greater depth.

3.2.1 Data Collection & Preprocessing

The study data set was drawn from an articles: "Association between Mediterranean diet adherence and dyslipidaemia among Bangladeshi diabetes mellitus patients" at Mendeley Data. It contains 796 data records and contains some of the major cardiovascular health determinants including HDL and LDL cholesterol, triglycerides, age, gender, height, weight, and family background, like income, Number_F_m, location longitudinal history of behaviors like smoking, and alcohol consumption, duration sleep together with critical ailments, hypertension, cardiovascular ailment and kidney disorder.

In data preprocessing missing values were handled by feature deletion with excessive nulls and replacing the remaining with mean value. Because the dataset consists of only numerical values, the focus of the data preprocessing was mainly on the cleaning and transformation of numerical data. The dataset was divided into different parts as follows: 80% was used for the model training (20% of this was set aside for model validation) and the remaining 20% was used for the model tests, which helped support controlled and structured model training and development.

3.2.2 Feature Selection

After finishing data preparation and splitting, we tested the feature selection process using our proposed XGBoost classifier and other machine learning models. Feature selection aids in the identification of the most relevant features, hence improving model performance by eliminating irrelevant or duplicated data. The following are feature significance graphs for our proposed machine learning model.

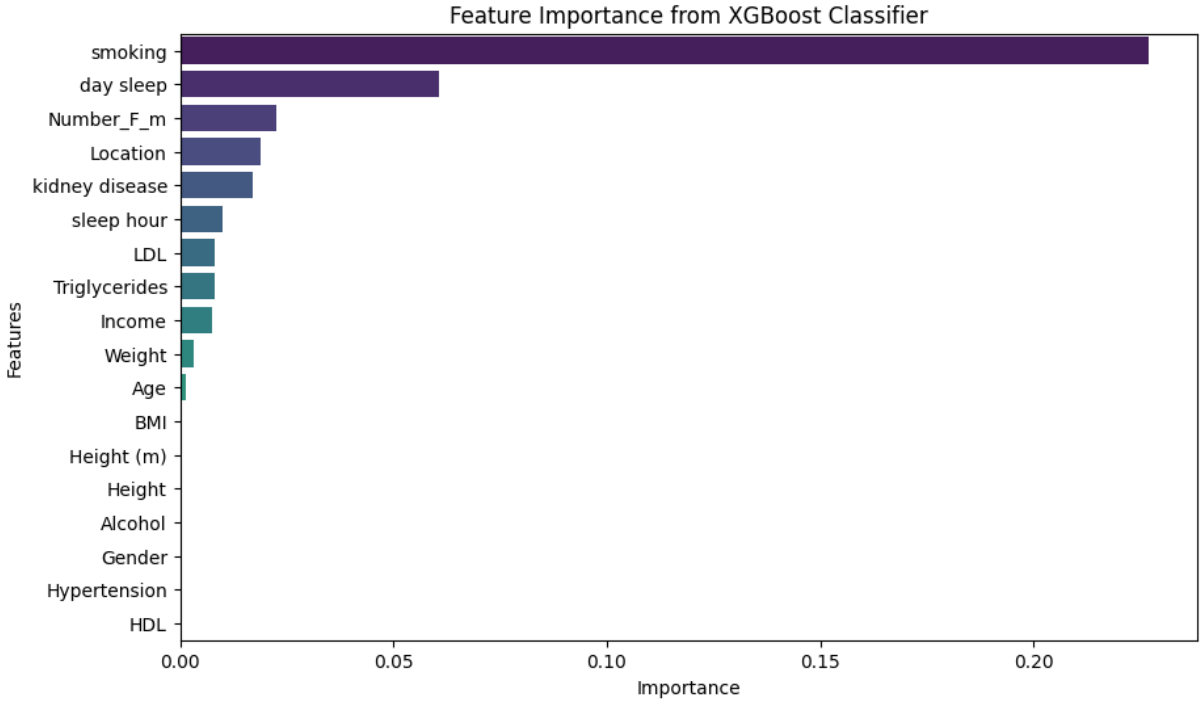


Figure 3.2: Feature Importance Plots for Proposed Machine Learning Model.

Figure 3.4 shows the feature importance plot for our proposed XGBoost machine learning model. The plot shows that smoking, day sleep, and Number_F_M are the most influential factors in predicting cardiovascular disease risk. Smoking, a well-known risk factor, has a major impact on cardiovascular health by boosting arterial plaque accumulation and raising heart disease risk. Day sleep, which is potentially linked to total sleep quality and lifestyle habits, also has a significant impact, indicating its potential significance in cardiovascular health. Furthermore, Number_F_M, which may reflect family medical history or other familial characteristics, is critical to the model's prediction, underscoring the significance of genetic or environmental impacts on cardiovascular disease risk.

3.2.3 Machine Learning Model Selection

Following data preprocessing, we chose high-performing machine learning models to produce the best results. The models chosen were Random Forest Classifier (RFC), Decision Tree, Logistic Regression (LR), K-Nearest Neighbors (KNN) Classifier, Support Vector Classifier (SVC), and XGBoost, all of which are known for their strong performance in classification tasks. A full description of each machine learning model is provided below.

Random Forest Classifier (RFC)

The Random Forest Classifier is an effective ensemble learning technique that combines the predictions of many decision trees to increase classification accuracy and reduce overfitting. Each tree in the forest is built using a random subset of the data and a random selection of features, adding diversity to the model. This strategy improves the model's robustness by lowering the variation that individual decision trees may exhibit.

The core idea behind a Random Forest is to build a "forest" of decision trees, with each tree voting on a class label. The class that receives the most votes becomes the

model's final forecast. The mathematical forecast for a given instance x can be represented as follows:

$$\hat{y} = argmax_c(\sum_{i=1}^{N} I(T_i(x) = c)) \quad (3)$$

In this equation: $\hat{y}$ is the predicted class label. c represents the class labels. N is the total number of trees in the forest. $I(T_i(x) = c)$ is an indicator function that equals 1 if the i-th tree predicts class c for the input.

Decision Tree

A decision tree is a model that divides data into subsets depending on feature values, resulting in a tree structure with decision nodes and leaf nodes. Each internal node represents a feature decision, whereas each leaf node either a class label (classification) or a predicted value (regression).

For classification, the prediction is the majority class in the leaf.

$$\hat{y} = \boldsymbol{Majority\ Class}$$

For regression the prediction is the mean of the target values of in the leaf.

$$\hat{y} = \frac{1}{n} \sum_{i=1}^{n} yi$$

Support Vector Classifier (SVC)

The Support Vector Classifier (SVC) is a classification model designed to find the optimal boundary, or hyperplane, that best separates classes in a dataset. This boundary is chosen to maximize the "margin," which is the distance between the hyperplane and the nearest data points of each class. These closest points, called support vectors, are crucial because they define the hyperplane's position.

In mathematical terms, for a set of input features x and corresponding labels y (where y=±1 for two classes), SVC aims to find a hyperplane w·x+b=0 that maximizes the margin while ensuring that the instances are correctly classified. This can be expressed with the following objective function:

$$\min_{w,b} \frac{1}{2} ||w||^2 \quad (4)$$

subject to: $y_i(w.x_i + b) \geq 1$ (5)

In these equations: w represents the weight vector perpendicular to the hyperplane.b is the bias term, which adjusts the hyperplane's position. y_i is the class label of instance x_i. The margin is maximized by minimizing $||w||^2$.while ensuring that each x_i is classified correctly.

Extreme Gradient Boosting (XGBoost)

XGBoost is a highly effective machine learning algorithm designed for classification and regression tasks. It enhances predictive performance by combining multiple weak learners, typically decision trees, into a strong ensemble model. The key equation governing its predictions is:

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i) \quad (8)$$

Here $\hat{y}_i$ is the predicted output for the i-th instance, and $f_k(x_i)$ represents the contribution of the k-th tree.

XGBoost utilizes gradient boosting to minimize the loss function, which is expressed as:

$$L\ (y, \hat{y}) = \sum_{i=1}^{N} l(y_i\ , \hat{y}_i) + \sum_{k=1}^{k} \Omega(f_k) \quad (9)$$

In this equation, l is the loss function (like log loss), and $\Omega(fk)$ is the regularization term to prevent overfitting. Notable features of XGBoost include its ability to handle missing values, support for parallel processing, and efficient tree pruning. These attributes contribute to its popularity in machine learning competitions, offering high accuracy and adaptability across various applications.

LightGBM Classifier

LightGBM (Light Gradient Boosting Machine) is a fast, efficient gradient boosting framework that builds decision trees for classification tasks. Unlike traditional level-wise tree growth, LightGBM uses a leaf-wise splitting strategy, which grows deeper trees by splitting the leaf with the highest loss reduction, improving accuracy and efficiency.

It minimizes an objective function that combines the loss from predictions and a regularization term to control model complexity:

$$\mathrm{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}) + \Omega(T)$$

Where ℓ is the loss function and $\Omega(T)$ penalizes tree complexity. LightBGM uses histogram based learning and feature bunding for speed, making it suitable for large dataset and high dimensional data.

3.2.4 Hyperparameter Tuning

After completing the data preprocessing, we chose the best models for the task. We then fine-tuned the hyperparameters to improve their performance. Table 2 shows the hyperparameter settings employed in our models to improve predicted accuracy.

Table 3.1: Hyperparameter settings of all ml models.

Model Name	Hyperparameter Settings
Random Forest Classifier	- random_state: 42
Decision Tree Classifier	- max_depth: [3, 5, 10, 15, None] - min_samples_split: [2, 5, 10] - min_samples_leaf: [1, 2, 4] - criterion: ['gini', 'entropy'] - class_weight: [None, 'balanced']
Support Vector Classifier (SVC)	- random_state: 42 - probability: True - kernel: 'linear'
XGBoost Classifier	- use_label_encoder: False - eval_metric: 'logloss' - random_state: 42
LGBM Classifier	- Hyperparameters - objective='binary' - metric='binary_logloss' - boosting_type='gbdt' - random_state=42

The table indicates that the hyperparameters utilized in the iterative five models in our pipeline are distinct. A fixed random_state ensures repeatability, as seen in the **Random Forest** and **XGBoost classifiers**. The **Decision Tree Classifier** incorporates parameters like max_depth and min_samples_split for tree growth, while criterion and class_weight address class imbalance. The **SVC** model employs a linear kernel and probability=False, meaning probability estimation is not supported. Boosting classifiers, such as XGBoost, use use_label_encoder=False to prevent warnings, eval_metric='logloss' for performance evaluation, and a fixed random_state. Finally, the **LightGBM** Classifier focuses on binary classification with objective='binary', metric='binary_logloss', and boosting_type='gbdt', ensuring reproducibility with random_state=42. These configurations aim to optimize model performance, correct bias using SMOTE, and maximize prediction accuracy.

3.2.5 Evaluation Matrices

In classification tasks, monitoring model performance is crucial for determining how effectively the model predicts outcomes. Several primary metrics are commonly used to evaluate classification algorithms, each providing unique insights into the model's effectiveness.

Accuracy: Accuracy measures the proportion of correctly predicted instances among the total instances. It is a straightforward metric but can be misleading, especially in cases of imbalanced classes.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

Where,

TP= True Positives (correctly predicted positive instances), TN= True Negatives (correctly predicted negative instances), FP= False Positives (incorrectly predicted positive instances), FN= False Negatives (incorrectly predicted negative instances).

Precision: Precision indicates the accuracy of positive predictions. It is particularly useful when the cost of false positives is high.

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

Recall: Recall, also known as sensitivity, measures the ability of the model to identify all relevant instances. It is vital when the cost of false negatives is high.

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

F1-Score: The F1 Score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is especially useful in situations with class imbalances.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (13)$$

Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC): The ROC curve plots the true positive rate (Recall) against the false positive rate at various threshold settings. The AUC represents the degree of separability achieved by the model; a higher AUC indicates better model performance.

$$AUC = \int_0^1 TRP\,(x)dx \quad (14)$$

Where TPR is the True Positive Rate (Recall) at a specific threshold.

Confusion Matrix: A confusion matrix provides a detailed breakdown of true and false positives and negatives, offering insight into the model's performance across all classes. The matrix is typically represented as: TP,FN,FP,TN

3.2.6 Use of Explainable AI

Explainable AI (XAI) focuses on making artificial intelligence systems more intelligible and transparent to humans. AI models, particularly complicated ones like deep learning networks, can operate as "black boxes," making it difficult to understand how they make predictions or choices. XAI aims to shed light on this riddle by emphasizing two core principles: interpretability and transparency. Interpretability explains how a model comes to its conclusions, whereas transparency discloses the model's inner workings and the data on which it is based. Techniques like as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive Explanations) are important in this process since they show how different attributes contribute to a model's predictions. By making AI more understandable, we can foster confidence and guarantee that these powerful technologies are used responsibly across a variety of applications, allowing consumers to make informed decisions based on AI insights. In our analysis, we employ LIME, as described below.

LIME (Local Interpretable Model-agnostic Explanations):

We used LIME, a powerful tool intended to improve the interpretability of our AI models. LIME helps us understand how different variables influence our models' predictions, providing insights into their decision-making processes. This method allows us to break down complex model outputs into more understandable explanations, allowing us to better convey our findings and gain trust in our AI-driven insights. The LIME explanation model is represented as follows:

$$\hat{g}(x) = argmin_{ge^G} L(f, g, \pi_{x'}) + \Omega(g) \qquad (15)$$

In LIME, the explanation model $\hat{g}(x)$ approximates complex model behavior locally form a set G of possible models, often linear. Using a loss function $L(f, g, \pi_{x'})$ on perturbed the explanation model's simplicity and interpretability.

3.2.7 Software Implementation

In the software implementation, we used our proposed model for prediction, combining all features into the process, and built the application with Streamlit. Streamlit is a robust and user-friendly Python toolkit for developing interactive online applications, particularly for data science and machine learning tasks. It enables you to quickly construct dynamic, real-time web apps by converting Python code into user-friendly interfaces. Whether you're visualizing data, showing machine learning models, or creating dashboards, Streamlit makes it simple to showcase your work. With built-in support for components including as buttons, sliders, and graphs, you can quickly design compelling and responsive applications that change in real time as users interact.

3.1 Project Plan

Semester 1 (Months 1–6): Focus on Research, Data, and ML Development

- Month 1: Finalize project scope and deliverables, Review Literature.
- Months 2–3: Collect and clean datasets (clinical, demographic, and diet-related) Collect and clean datasets (clinical, demographic, and diet-related), Handle missing values, feature engineering
- Months 4–5: Train ML models (XGBoost, LightGBM, SVC, Decision Tree, Random Forest), hyperparameter tuning, Explainable AI Integration (Lime).
- Month 6: Compare model performance, select best-performing model; begin integration plans for deployment.

Semester 2 (Months 7–12): Model Refinement, Deployment, and Report Finalization

- Months 7–8: Refine best-performing model (LightGBM), apply LIME for XAI analysis, Generate local/global interpretability plots..
- Months 9–10: Deploy the ML model to a Streamlit-based Web Application, Integrate features such as prediction results, feature importance visualizations, and user input options.
- Month 11: Perform end-to-end testing: Model accuracy validation. Usability testing for web app. Check performance on unseen validation dataset.
- Month 12: Complete final report documentation. Prepare presentation slides for project defense. Rehearse project defense.

3.2 Task Allocation

The distribution of responsibilities guarantees simultaneous advancement and effective implementation of the project activities:

- **Research and Literature Review:** Focus on the acquisition of knowledge on cardiovascular disease prediction, machine learning methodologies, and explainable artificial intelligence.
- **Data Acquisition and Preprocessing:** Data cleaning, handling missing values, and feature engineering. data normalization and preparation for model training.
- **Model Implementation and Training:** Implemented machine learning models and explainable AI. Hyperparameter tuning and performance evaluation were conducted with periodic supervision and feedback from the project advisor.
- **Model Refinement and Evaluation:** The best-performing model was refined and validated through manual evaluation by the student. Interpretability analysis was performed using e XAI methods to highlight feature importance and discuss the result.
- **Web Application Development:** The student developed a web-based application using **Streamlit** for deploying the final ML model. Features included input options, prediction display, and visualizations for explainable AI outputs.

- **Final Documentation and Defense Preparation:** Preparing the final project report, including detailed documentation of methods, results, and findings. Presentation slides were created for the final defense, with feedback and suggestions from the supervisor.

3.3 Summary

In the software implementation, we used our proposed model for prediction, combining all features into the process, and built the application with Streamlit. Streamlit is a robust and user-friendly Python toolkit for developing interactive online applications, particularly for data science and machine learning tasks. It enables you to quickly construct dynamic, real-time web apps by converting Python code into user-friendly interfaces. Whether you're visualizing data, showing machine learning models, or creating dashboards, Streamlit makes it simple to showcase your work. With built-in support for components including as buttons, sliders, and graphs, you can quickly design compelling and responsive applications that change in real time as users interact

Chapter 4

Implementation and Results

Chapter 4 focuses on the implementation of the proposed methodology and presents the results obtained from the analysis. It includes details of model deployment, performance evaluation, and a comparative analysis of the results.

4.1 Environment Setup

While building, training, and improving our machine learning models we relied heavily on Google Colab, a cloud platform that radically changed how we approach our research. Through its free version, we managed to build an effective computer system that was simple and usable. This allowed us to carry out large experiments without the usual resource limitations that are hindrances to creativity and innovation. Google Colab is very easy to use, so we were able to bypass a lot of chores and start working straight away which meant we could concentrate on coming up with new ideas and concepts.

Instinctively storing the new learnt information that surrounds Google Colab's collaborative functions, was our team's favourite. The ease with which we shared code and datasets created a firm sense of community that greatly complemented our study experience. All our work was done simultaneously, we had discussions, learned from each other's insights and improved models together. Not only in our case did such cooperative environment boost our productivity, but it also broadened our discussions making the entire process more enjoyable. While doing research, Python 3 was used for coding and Scikit Learn module was used for ML functions. This combination allowed us to competently tackle the intricacies of harnessing our research, thereby enhancing the quality of our output and its potential imp

4.2 Testing and Evaluation/Performance/ Comparative Analysis

After evaluating the performance of several models, we selected the top six models, including our proposed model, for classifying cardiovascular diseases. The final prediction results for all these models are summarized in Table 3, based on the testing dataset.

Table 4.2: ML Models Result Analysis for this Study.

Model	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	AUC
Random Forest	98.04	97.72	97.88	98.36	0.98
SVC	77.29	87.62	82.13	85.3	0.9
LightGBM	99.68	99.05	99.36	99.37	0.98
Decision Tree	99.16	97.34	98.24	98.26	0.98
XGBoost	99.68	98.73	99.2	99.21	0.98

Table 4.2 provides a comparative analysis of the performance metrics for various machine learning models applied in our study. Among the models, LightGBM demonstrates the highest overall performance with an impressive precision of 0.9968, recall of 0.9905, F1-score of 0.9936, and accuracy of 99.37%, closely followed by XGBoost with slightly lower recall (0.9873) and accuracy (99.21%). While Random Forest also delivers strong results with an accuracy of 98.36%, Decision Tree achieves similar performance metrics, showing an F1-score of 0.9824 and accuracy of 98.26%. On the other hand, SVC lags behind, attaining an accuracy of 85.30% and the lowest F1-score (0.8213), indicating comparatively weaker classification capabilities. The analysis highlights that ensemble-based models like LightGBM and XGBoost outperform other algorithms, showcasing their robustness and efficiency in predicting cardiovascular disease risk in our dataset.

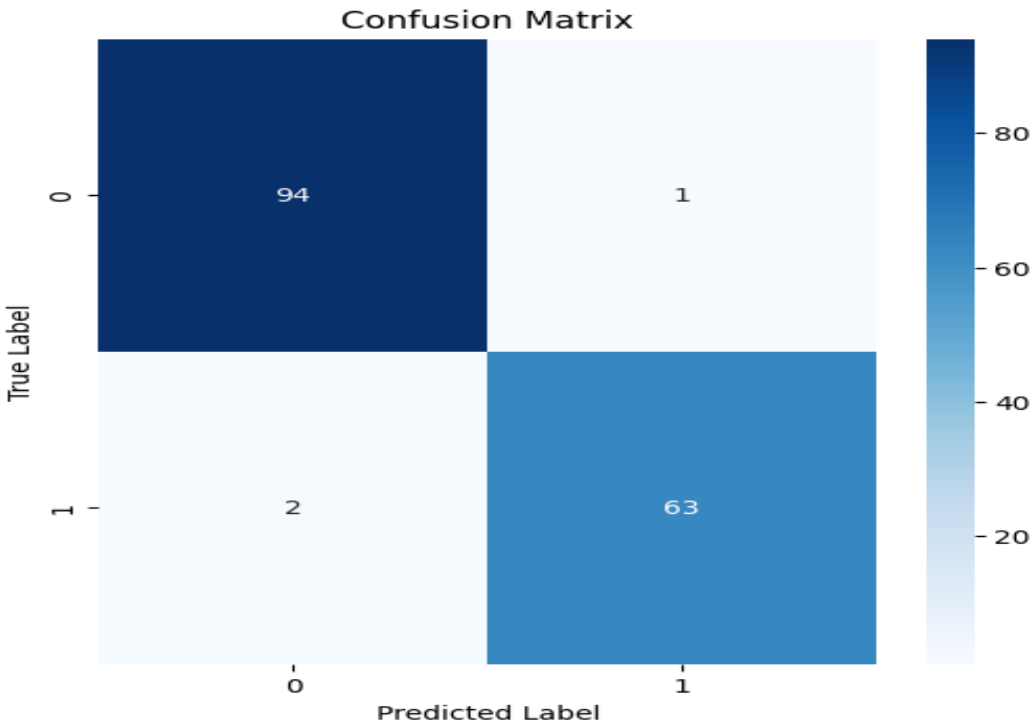


Figure 4.1 : Confusion Matrix for Our Proposed Model.

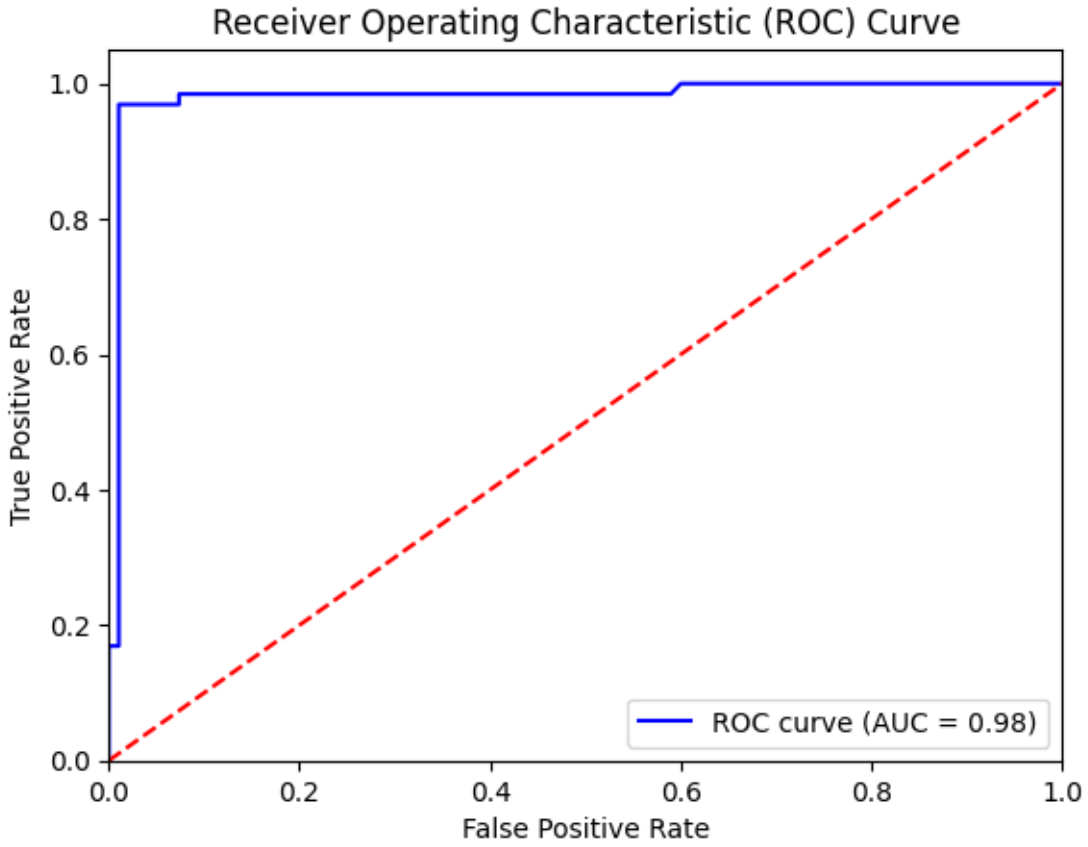


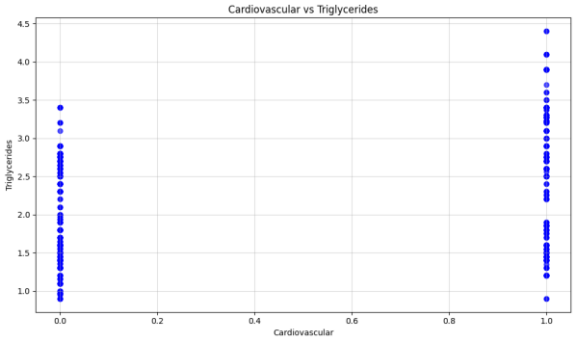
Figure 4.2. : AUC Curve for Our Proposed Model.

4.3 Results and Discussion

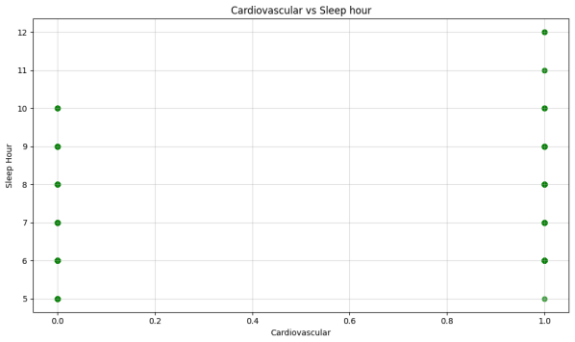
In this section, we delve deeper into the results of our study, providing an in-depth analysis of the outcomes. Each of the following subsections presents and discusses the findings in detail, offering a comprehensive examination of the performance metrics and their implications for our research topic

4.3.1 Insight Outcomes

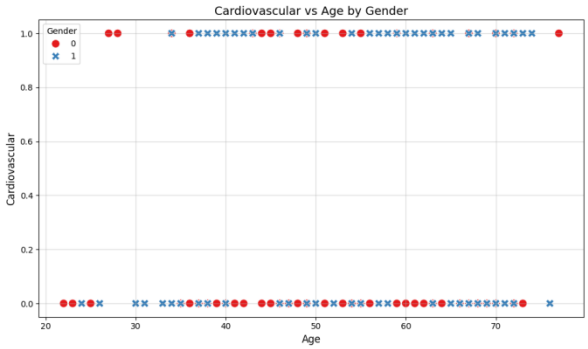
To further understand the correlations between our dataset's key attributes and target column features, we compared them. The graphs below exhibit these comparisons, emphasizing how different features influence the target variables.



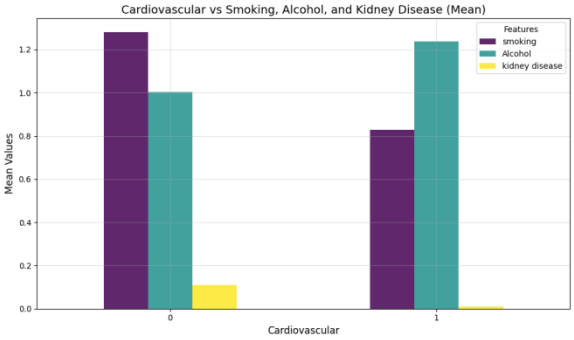
a) Cardiovascular vs Triglycerides.



b) Cardiovascular vs Sleep Hour.



c) Cardiovascular vs Age by Gender.



d) Cardiovascular vs Smoking, alcohol and kidney disease.

Figure 4.3..: Cardiovascular vs Important Features Graph Plot.

Figure 4.3.. presents a series of plots analyzing the relationship between cardiovascular disease and key features. In Figure 4.3 (a), the plot of Cardiovascular vs. Triglycerides indicates that the range of 2.2 to 3.5 is the most prominent. In Figure 4.3 (b), the Cardiovascular vs. Sleep Hour graph reveals that sleep durations between 5 to 12 hours are observed on average. Figure 4.3 (c), which plots Cardiovascular vs. Age by Gender, shows that female patients are predominant, with most cases occurring between the ages of 50 and 70. Lastly, in Figure 4.3 (d), the Cardiovascular vs. Smoking, Alcohol, and Kidney Disease plot highlights that smoking and alcohol consumption have the most significant effects on cardiovascular health.

4.3.2 Real Life Software Implementation

In our real-life software implementation, we showed how each feature affects categorization performance. To illustrate the data, we used Streamlit, which allowed us to generate interactive visualizations, and HTML was used on the frontend to improve the design and user experience. The screenshots below highlight our website's interactive features and visuals.

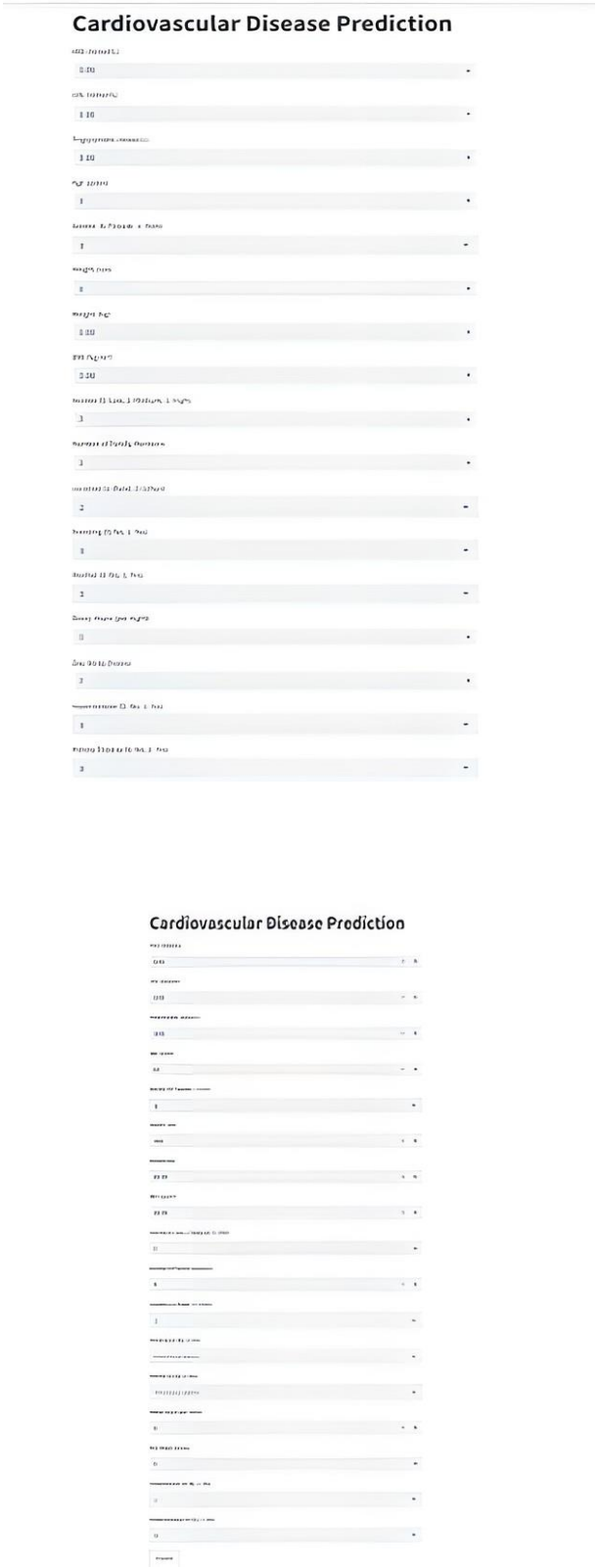


Figure 4.4.: Website Implementation without value.

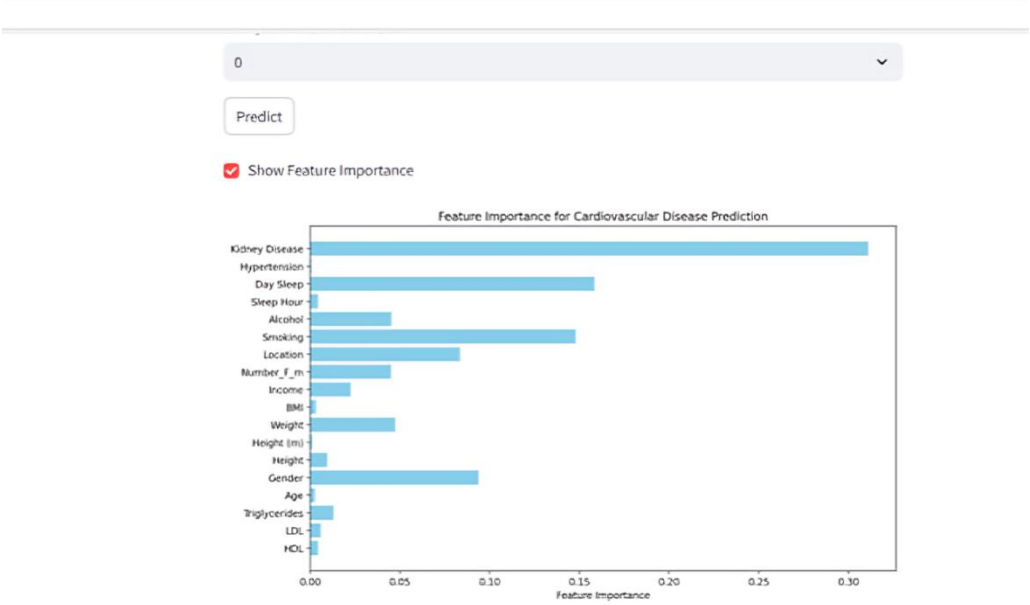


Figure 4.5: Website Implementation with Feature Importance Plot.

Figure 4.5 depicts the website implementation with no values presented. Figure 4.6.2 depicts the website implementation with values, giving a more detailed perspective of the data. Figure 4.5 also depicts the website with a Feature Importance Plot, which highlights the significance of various elements in the model.

4.3.3 Using LIME to Interpret Proposed Model's Predictions

We used LIME to examine feature importance in the dataset more precisely. For a more comprehensive study, our approach used both LIME tabular plots and feature significance plots. The following is a full description of our findings.

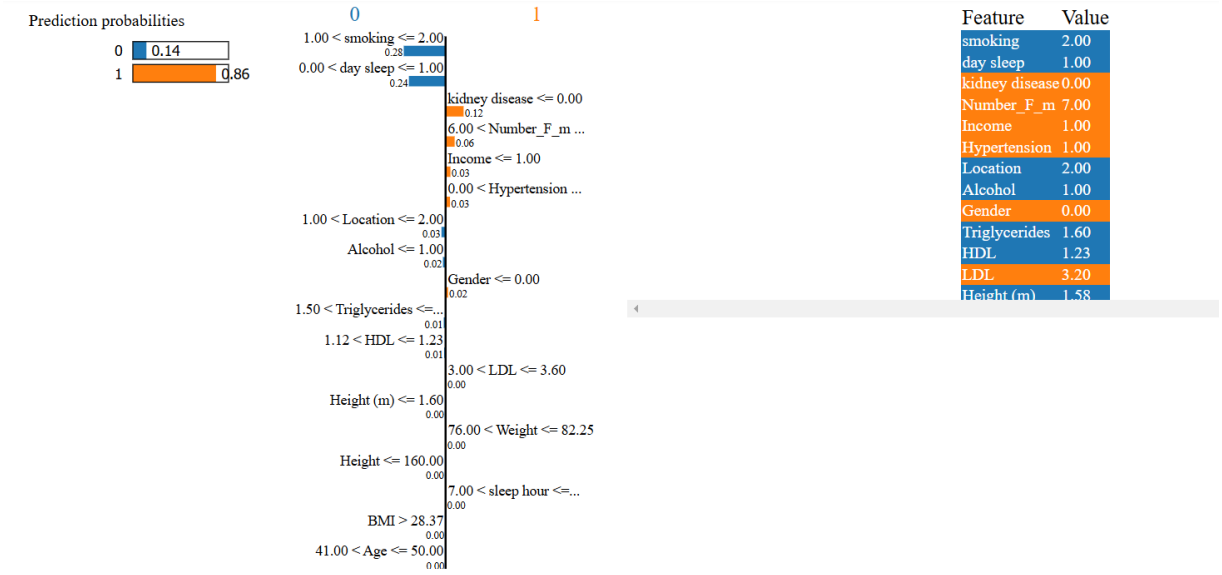


Figure 4.6: LIME Tabular Plot.

Figure 4.6 represents LIME. The Tabular Plot displays a feature contribution plot for a specific prediction, showing the cardiovascular disease prediction probability. The left bar chart shows a prediction probability of 0.14 for "No Cardiovascular Disease" (class 0), and 0.86 for "Cardiovascular Disease" (class 1). On the right, traits and their values are listed in descending order of priority. Smoking, day sleep, and kidney disease are all significant influences on the model's decision. The feature values provide insight on how these variables influenced the final prediction, highlighting the predictive strength of behavioral and clinical factors.

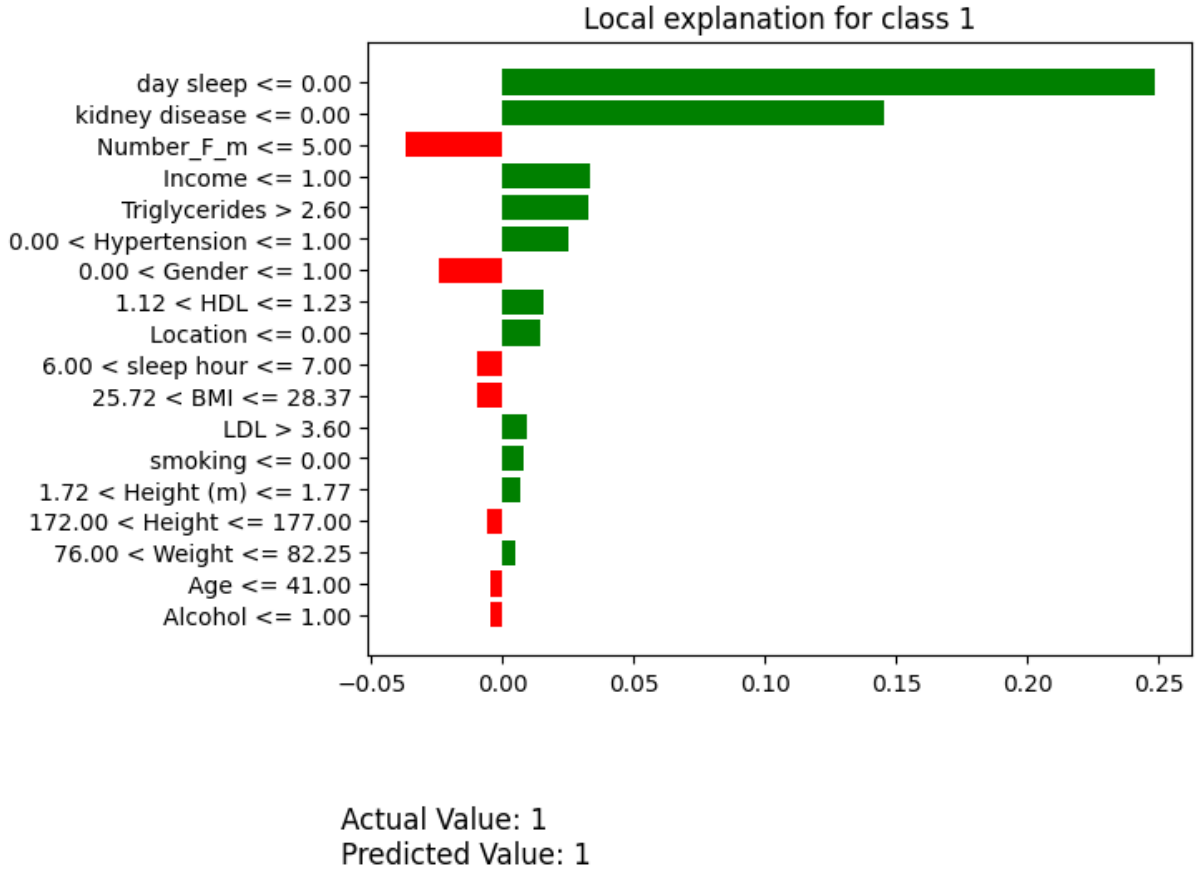


Figure 4.7 : LIME Feature Importance Plot.

Figure 4.7 highlights the importance of the LIME feature. Plot uses a bar chart to present a local feature importance analysis for class 1 ("Cardiovascular Disease"). Green bars show features that positively contribute to the forecast, whereas red bars highlight features that contradict the prediction. Day sleep, kidney disease, and Number_F_m all make considerable positive contributions to the model's ability to predict cardiovascular disease. In contrast, characteristics like as smoking and gender marginally undermine the prediction. The graphic clearly shows how individual variables influence the outcome, which helps to interpret and explain the model's judgments.

4.3.4 Discussion

The outcomes of this study shed light on the critical factors influencing cardiovascular disease (CVD) in Bangladeshi diabetes patients. Our suggested LightGBM model outperformed existing machine learning models in terms of precision, recall, F1-score,

and accuracy. This demonstrates the effectiveness of our feature engineering and proposed model in recognizing the hazards associated with CVD.

The feature importance analysis demonstrated that lifestyle factors such as smoking, alcohol intake, and sleep habits have a significant impact on cardiovascular outcomes. Smoking was identified as the most influential factor, validating its well-known link to cardiovascular health. Similarly, insufficient or excessive sleep lengths were found to have an effect, emphasizing the need of balanced sleep patterns in reducing CVD risks.

Furthermore, triglyceride levels and BMI were identified as critical physiological markers, consistent with previous research highlighting their association with cardiovascular health.

Gender and age were also significant predictors, with the majority of CVD cases observed among females aged between 50 and 70 years. This finding aligns with epidemiological studies suggesting that postmenopausal women are at a higher risk of developing CVD due to hormonal changes and other metabolic factors. Additionally, socioeconomic indicators such as income and location influenced the model's predictions, indicating that social determinants of health are critical in understanding cardiovascular risks within this population.

Interestingly, the combination of kidney disease and hypertension further amplified CVD risk, highlighting the interconnectedness of chronic diseases. This finding emphasizes the importance of a multidisciplinary approach in managing patients with comorbid conditions to prevent cardiovascular complications.

The analysis also revealed that certain habits, such as alcohol consumption, showed a nuanced relationship with cardiovascular health. While moderate consumption may be less harmful, heavy or frequent consumption exacerbates CVD risks. This highlights the need for tailored lifestyle interventions as part of preventive care strategies.

Overall, the results demonstrate the utility of machine learning models like LightGBM in identifying high-risk individuals and guiding targeted interventions. By incorporating explainable AI, our study provides a transparent and interpretable framework for understanding the underlying drivers of CVD, enabling healthcare professionals to make data-informed decisions.

4.4 Summary

This chapter presented the implementation and results of our proposed machine learning framework for predicting cardiovascular disease (CVD) risk among Bangladeshi diabetes patients. The LightGBM model outperformed other models, achieving the highest accuracy, precision, recall, F1-score, and AUC. Key findings revealed that smoking, triglycerides, BMI, and sleep patterns were the most influential factors, alongside age, gender, and comorbidities like kidney disease and hypertension. Our use of explainable AI highlighted the critical role of lifestyle and physiological factors in CVD risk prediction, providing valuable insights for targeted interventions. This study underscores the potential of machine learning in advancing healthcare outcomes through accurate risk stratification and informed decision-making

Chapter 5

Engineering Standards and Design Challenges

This chapter outlines the engineering standards and design challenges faced during the development of our machine learning framework for Type 2 diabetes prediction. It delves into the tools, platforms, methodologies, and collaborative practices that supported the project's development, refinement, and execution. Furthermore, it highlights the project's societal, environmental, and ethical implications, while addressing sustainability and project management.

5.1 Compliance with the Standards

To ensure robustness, efficiency, and scalability, our project adhered to established engineering standards in software, hardware, and communication practices

5.1.1 Software Standards

The project utilized Python 3, a widely recognized programming language in the research and development community. For machine learning tasks, the Scikit-learn library was employed, offering efficient algorithms and tools for model development. Code readability and reproducibility were prioritized, following best practices to facilitate collaboration and maintainability. These practices ensured that the project remained aligned with global software engineering standards.

5.1.2 Hardware Standards

All computational tasks were executed using Google Colab, a cloud-based platform providing powerful hardware resources such as GPUs. This platform allowed us to efficiently train complex machine learning models without significant upfront costs. Despite resource limitations in the free version, Google Colab proved sufficient for conducting experiments, optimizing models, and managing computational demands

5.1.3 Communication Standards

Google Colab's real-time collaboration features enabled seamless sharing of code, datasets, and results among team members. This facilitated effective communication, ensured alignment across tasks, and fostered teamwork, which proved instrumental in maintaining the project's momentum and quality.

5.2 Impact on Society, Environment and Sustainability

This section highlights the broader implications of our work, focusing on its potential to influence public health, environmental sustainability, and ethical considerations..

5.2.1 Impact on Life

The machine learning framework developed in this project offers a significant contribution to public health, particularly in resource-constrained settings like Bangladesh. Early detection and risk prediction of Type 2 diabetes can improve disease management and prevention strategies, ultimately enhancing the quality of life for at-risk individuals..

5.2.2 Impact on Society & Environment

The use of a cloud-based platform like Google Colab minimizes physical hardware consumption, reducing the project's carbon footprint. This sustainable approach aligns with modern principles of green computing and resource efficiency

5.2.3 Ethical Aspects

Ethical considerations were central to the project, with a strong focus on data privacy, fairness, and model transparency. Efforts were made to ensure unbiased predictions and adherence to standard ethical practices, making the framework trustworthy and reliable for real-world applications

5.2.4 Sustainability Plan

The adaptability of our machine learning model ensures long-term sustainability. By leveraging cloud-based solutions, the framework can be easily updated with new data and scaled to various contexts without excessive resource consumption.

5.3 Project Management and Financial Analysis

The financial model for the project was designed to optimize resource allocation while minimizing costs. By leveraging Google Colab's free version, we were able to significantly reduce computational expenses. This allowed the budget to focus on critical aspects such as data collection and model refinement. For future phases requiring higher computational power, an alternative budget plan has been proposed to accommodate premium cloud-based solutions.

Table 5.1 : Estimated Cost for Research based Project

SN	Components	Estimated Cost (BDT)
01.	Software and Tools	500-1500
02.	Data collection sources	1000-1500
03.	Documentation and Report Writing	1000-1500
04.	Contingency	500-1000
05..	Miscellaneous	500-1500
	Total Estimated Cost	3,500-6000

5.4 Complex Engineering Problem

This project adheres to multiple Complex Engineering Problem (EP) standards and Engineering Activity (EA) standards. It transcends the Computer Science and Engineering area by integrating ideas from the agriculture industry, examining various solution methodologies, and tackling wider socioeconomic and environmental issues. This section delineates the pertinent EP and EA standards, offers rationale for their selection, and correlates each EP standard with the associated Knowledge Profiles (K), illustrating the project's interdisciplinary and significant character.

5.4.1 Complex Problem Solving

The project involved tackling complex engineering challenges in model accuracy, data preprocessing, and system integration. A structured approach was employed, leveraging problem-solving frameworks to address these issues systematically.

Table 5.2: Mapping with complex problem solving.

EP1 Dept of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of ApplicableCodes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
✔	✔	✔	✔			✔

EP1: leverages substantial healthcare and computer science and engineering skills. Integration of medical data (patient records) with machine learning techniques shows knowledge transfer between disciplines.

EP2: This study biomedical research to comprehend cardiovascular disorders and diabetes, which are beyond the typical CSE curriculum.

EP3: Used detailed symmetric analysis of various ML models and XAI to evaluate prediction

EP4: Understanding clinical factors like triglycerides and sleep patterns involved extensive research, enhancing my familiarity with healthcare datasets and medical concepts.

EP7: Data collection, preprocessing, model construction, and evaluation were interconnected.

Mapping with Knowledge Profile for EP1

K1 Natural sciences	K2 Mathematics	K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K7 comprehension	K8 Research Literature
	✔	✔	✔	✔	✔		✔

Table 5.3: Mapping with knowledge Profile.

K2 (Mathematics) : Mathematical foundations support research machine learning approaches. Prediction model creation and validation require statistical and mathematical methods.

K3 (Engineering Fundamental) : Demonstrates a strong grasp of engineering fundamentals, particularly in the context of software engineering and data science.

K4 (Specialist Knowledge) : The project demonstrates specialized knowledge in both medical sciences and machine learning.

K5 (Engineering Design) : Engineering design is evident in the machine learning model's creation and execution.

K6 (Engineering Practice) : The practical application of engineering principles is evident in the design and implementation of the predictive model.

K8 (Research Literature) It leverages established research literature in healthcare and machine learning. It cites previous research, employs recognized methodology, and offers novel insights to the discipline.

5.4.2 Engineering Activities

Table 5.4: Mapping with complex engineering activities.

EA1 Range of resources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✔	✔	✔	✔	✔

EA1: This initiative uses patient demographics, medical history, and Mediterranean diet adherence. These data were used to train and test cardiovascular risk prediction machine learning models, revealing the vast spectrum of resource integration needed to solve complex healthcare challenges.

EA2: The project spans data science, healthcare, and dietetics. Model validation required medical specialists and data scientists to collaborate. This engagement made the predictive model accurate, relevant, and applicable to real-world events.

EA3: AI to forecast cardiovascular risks with Mediterranean diet adherence as a unique characteristic. This method improves risk prediction by delivering interpretable information, helping doctors make educated decisions and personalise therapy.

EA4: The project targets major social issues like healthcare costs and patient quality of life by focussing on early risk prediction and prevention. Dietary recommendations emphasise plant-based, Mediterranean-style eating, which indirectly benefits the environment.

EA5: This approach balances familiarity with innovation, making it accessible to medical practitioners while introducing advanced tools to improve prediction accuracy and patient outcomes.

5.5 Summary

Results, analysis, and implications from the project will be discussed in Chapter 5, emphasizing the alignment of the project with complex engineering activities. The chapter analyzes the results of the project, with evaluation based on the performance of the machine learning model in predicting the risk of cardiovascular diseases among Bangladeshi diabetes patients, incorporating the bearing of Mediterranean diet adherence.

This chapter clearly attests the effectiveness of the methodology in dealing with the objectives of the project, showing how each algorithm and explainable AI tool showcases effective predictions and its interpretation of insights. This guarantees the model's validity, complementing the personalized care approach for prevention strategy by medical practitioners.

Besides this, the chapter also maps the outcomes of the project against key components of engineering activity (EA1 through EA5). It goes on to discuss the multiple resources that were realized (EA1), interdisciplinary collaboration (EA2), innovative integration of AI with diet-focused healthcare (EA3), and societal and environmental impacts (EA4) in addition to the balance of familiarity and novelty in the approach (EA5).

Consequently, the Chapter 5 also shows that these project findings are necessary for progress in health care solutions to contribute to more extensive societal challenges validating the engineering-medical innovation contribution of project findings.

Chapter 6

Conclusion

This chapter summarizes the key findings of the study, highlighting the role of machine learning and explainable AI in predicting cardiovascular disease risk among Bangladeshi diabetes patients. The research underscores the importance of Mediterranean diet adherence in improving cardiovascular health. Future work will focus on expanding the dataset and refining model accuracy for broader healthcare applications.

6.1 Summary

The purpose of this study was to predict cardiovascular disease risk among Bangladeshi diabetic patients using machine learning and explainable AI, with a particular emphasis on the significance of Mediterranean diet adherence. Our models showed good predictive ability, highlighting crucial parameters such as triglycerides, sleep hours, and lifestyle habits that influence cardiovascular health. The use of explainable AI guaranteed that the predictions were interpretable, providing significant insights into the underlying causes. This study lays the groundwork for establishing data-driven, tailored healthcare solutions for diabetes and cardiovascular disease management in Bangladesh

6.2 Limitation

This research may be insightful in terms of its results, yet it has some weaknesses that ought to be created. Generalizability may not be feasible in terms of other populations or regions. The dataset was employed as part of this study efforts and it was concerned with a number of Bangladeshi diabetes patients. The concerns and lifestyle factors captured in the dataset are not convenient and do not necessarily portray different peoples, and thus may impact the strength of the model when applied in all circumstances. Similar data are self-reported, and thus a number of features for instance lifestyle habits, diet adherence and the likes bear the possibility of bias. Self-reports on these features may be combination of recall error, social desirability or patient-reporting-accuracy, which prediction of the risk of CVD may not depend on these contradicting perspectives.

During the course of the study, the materials, specifically the computational resources used should be highlighted as a limitation as well. While Google Colab presented an easy option without cost concerns in terms of model and training, it has such deficits as processing power and memory, particularly with larger datasets or more complex models. This might have limited the experimentation scope with more complex techniques or larger datasets that might have improved the models prediction accuracy even more.

6.3 Future Work

This is an important study for predicting cardiovascular disease risk among Bangladeshi diabetes patients using the machine learning and explainable AI. Nevertheless, the research can

be extended in various ways to make the model more accurate, applicable and helpful in practical life settings.

First, on the accuracy of the model and making sure it is suitable for larger populations future studies should involve more diverse datasets. This may include using data from various regions, generations, and more diverse ethnicities to make the model more generalizable. Furthermore, more thorough information on lifestyle, environmental variables and genetic data may help to fill the blanks in understanding of the risk of cardiovascular disease.

In the sequel, improving the model with more health-related variables — laboratory tests, family medical history, and so on — could lead to a deeper insight into the patient's risk. It might also be possible to track patients over time and/or use longitudinal data to build a model that adapts based on changing health for more dynamic and personalized risk assessments. Finally, to ensure that medical professional trusts and interprets the model's recommended outputs, future research must emphasize on improving the interpretability of the AI. The development of transparent and user-friendly explainable AI tools would help healthcare providers to add the model to their decision-making processes.

References

[1] Lucia La Sala and Antonio E Pontiroli. Prevention of diabetes and cardiovascular disease in obesity. International journal of molecular sciences, 21(21):8178, 2020.

[2] Fatma M Talaat, Ahmed R Elnaggar, Warda M Shaban, Mohamed Shehata, andMostafa Elhosseini. Cardiorisknet: A hybrid ai-based model for explainable risk pre-diction and prognosis in cardiovascular disease. Bioengineering, 11(8):822, 2024.

[3] Marie-Eve Pich´e, Andr´e Tchernof, and Jean-Pierre Despr´es. Obesity phenotypes, diabetes, and cardiovascular diseases. Circulation research, 126(11):1477–1500, 2020.

[4] Nerea Becerra-Tom´as, Sonia Blanco Mej´ıa, Effie Viguiliouk, Tauseef Khan, Cyril WC Kendall, Hana Kahleova, Dario Raheli´c, John L Sievenpiper, and Jordi Salas-Salvad´o. Mediterranean diet, cardiovascular disease and mortality in diabetes: A systematic review and meta-analysis of prospective cohort studies and randomized clinical trials.Critical reviews in food science and nutrition, 60(7):1207–1227, 2020.

[5] Sandra Mart´ın-Pel´aez, Montse Fito, and Olga Castaner. Mediterranean diet effects on type 2 diabetes prevention, disease progression, and related mechanisms. a review. Nutrients, 12(8):2236, 2020.

[6] Manjula Mandava, Surendra Reddy Vinta, Hritwik Ghosh, and Irfan Sadiq Rahat. An all-inclusive machine learning and deep learning method for forecasting cardiovascular disease in bangladeshi population. EAI Endorsed Transactions on Pervasive Health and Technology, 12, 2023.

[7] Sorif Hossain, Mohammad Kamrul Hasan, Mohammad Omar Faruk, Nelufa Aktar, Riyadh Hossain, and Kabir Hossain. Machine learning approach for predicting cardiovascular disease in bangladesh: evidence from a cross-sectional study in 2023. BMC Cardiovascular Disorders, 24(1):214, 2024.

[8] MD Amzad Hossen, Tahia Tazin, Sumiaya Khan, Evan Alam, Hossain Ahmed Sojib, Mohammad Monirujjaman Khan, and Abdulmajeed Alsufyani. Supervised machine learning-based cardiovascular disease analysis and prediction. Mathematical Problems in Engineering, 2021(1):1792201, 2021.

[9] Saiful Islam, Nusrat Jahan, and Mst Eshita Khatun. Cardiovascular disease forecast using machine learning paradigms. 2020 Fourth International

Conference on Computing Methodologies and Communication (ICCMC), pages 487–490,IEEE, 2020.

[10] Tamanna Yesmin Rashme, Linta Islam, Sohely Jahan, and Ayesha Aziz Prova. Early prediction of cardiovascular diseases using feature selection and machine learning techniques. 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), pages 1554–1559, IEEE,2021.

[11] Md Shahiduzzaman, Nowreen Haque Biswas, Md Momin, and Raihan Sikdar. Prognosis of cardiovascular disease using machine learning procedures. 2022 international conference on advancement in electrical and electronic engineering (ICAEEE), pages 1–6, IEEE,2022.

[12] Khairul Eahsun Fahim, Hayati Yassin, Md Hasnatul Amin, Priyanka Das Dewan, and Aminul Islam. Detection of cardiovascular disease of patients at an early stage using machine learning algorithms. 2022 International Conference on Healthcare Engineering (ICHE), pages 1–6,IEEE, 2022.

[13] Muhammad Nazrul Islam, Kazi Rafid Raiyan, Shutonu Mitra, MM Rushadul Mannan, Tasfia Tasnim, Asima Oshin Putul, and Angshu Bikash Mandol. Predictis: an iot and machine learning-based system to predict risk level of cardio-vascular diseases. BMC Health Services Research, 23(1):171, 2023.

[14] Md Maruf Hossain, Md Shahin Ali, Md Mahfuz Ahmed, Md Rakibul Hasan Rakib, Moutushi Akter Kona, Sadia Afrin, Md Khairul Islam, Md Manjurul Ahsan, Sheikh Md Razibul Hasan Raj, and Md Habibur Rahman. Cardiovascular disease identification using a hybrid cnn-lstm model with explainable ai. Informatics in Medicine Unlocked, 42:101370, 2023.

[15] Annie M Westerlund, Johann S Hawe, Matthias Heinig, and Heribert Schunkert. Risk prediction of cardiovascular events by exploration of molecular data with explainable artificial intelligence. International Journal of Molecular Sciences, 22(19):10291, 2021.

Predicting Cardiovascular Disease Risk Among Bangladeshi Diabetes Patients Using Machine Learning and Explainable AI

ORIGINALITY REPORT

24%	20%	15%	17%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	3%
2	Submitted to United International University Student Paper	2%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
4	www.mdpi.com Internet Source	1%
5	onlinelibrary.wiley.com Internet Source	<1%
6	www.ijraset.com Internet Source	<1%
7	www.coursehero.com Internet Source	<1%
8	climbtheladder.com Internet Source	<1%
9	Submitted to Anna University Student Paper	<1%