CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Improving Clinical Decision Support Using Feature-Enhanced ML Model and Interpretable Artificial Intelligence Techniques for Cardiovascular Disease Prediction

by

Tarish Nisar

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Computing
Department of Computer Science

2026

CERTIFICATE OF APPROVAL

Improving Clinical Decision Support Using Feature-Enhanced ML Model and Interpretable Artificial Intelligence Techniques for Cardiovascular Disease Prediction

by

Tarish Nisar

(MCS231009)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Asim Munir	IIUI, Islamabad
(b)	Internal Examiner	Dr. Farah Haneef	CUST, Islamabad

Dr. Mohammad Masroor Ahmed
Thesis Supervisor
April, 2026

Dr. Mohammad Masroor Ahmed
Head
Dept. of Computer Science
April, 2026

Dr. Muhammad Abdul Qadir
Dean
Faculty of Computing
April, 2026

Author's Declaration

I, **Tarish Nisar** hereby state that my MS thesis titled "**Improving Clinical Decision Support Using Feature-Enhanced ML Model and Interpretable Artificial Intelligence Techniques for Cardiovascular DiseasePrediction**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Tarish Nisar)

Registration No:MCS231009

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Improving Clinical Decision Support Using Feature-Enhanced ML Model and Interpretable Artificial Intelligence Techniques for Cardiovascular Disease Prediction**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Tarish Nisar)

Registration No: MCS231009

Acknowledgement

All praises are to ALLAH alone, the Most Beneficent, who gave me the strength, understanding, courage, and patience to complete this research thesis. I dedicate this research to my **Baba**, my strength and my guide, who hold my hand at the beginning of this MS journey but left me before it was completed. Losing him broke me in ways I cannot express, but his love and faith in me kept me going. May Allah Subhanahu wa Ta'ala rest him in peace in Jannat-ul-Firdaus.

After him, my sincere appreciation goes to my respected research supervisor, **Dr. Muhammad Masroor Ahmed**, who assisted me in completing this challenging task. His valuable feedback, thoughtful insights, and constant encouragement and patience helped me to accomplish this research. I am deeply thankful for his guidance at every step of this journey.

It was an honour to work, learn, and grow under his supervision.

(Tarish Nisar)

Abstract

Cardiovascular Disease (CVD) is a major cause of death all over the world, causing 18 million deaths worldwide each year. The surgical procedure of CVD is challenging, specifically in developing countries, because they lack in medical devices, surgical tools, as well as doctors and medical practitioners. Therefore, performing early detection and accurate forecasting of CVD is necessary to reduce the scope of mortality. ML has been applied in a variety of medical disciplines, including CVD, because it provides high precision in clinical decision support systems. Although the trade-off between precision and explainability influences in adoption of AI in sensitive healthcare domains. For this reason, explainability is used to add transparency in the results so that the predictions made by the ML models are trusted by practitioners. In this research, the Cleveland dataset has been used, the proposed model achieves an accuracy of 98.54% on the test data. Additionally, SHAP and LIME explainable artificial intelligence (XAI) techniques are integrated to enhance the transparency and interpretability of the predictions made by the machine learning model. This integration ensures that the model's results are applicable in the medical domain for identifying cardiovascular disease (CVD) risk in individual patients. Furthermore, the approach aims to promote trust and acceptance of the results by clinicians and healthcare professionals.

Contents

List of Figures

List of Tables

Abbreviations

AUC	Area Under Curve
AI	Artificial Intelligence
BP	Blood Pressure
CVD	Cardiovascular Disease
Chol	Cholesterol
CV	Cross Validation
DL	Deep Learning
DT	Decision Tree
EDA	Exploratory Data Analysis
FE	Feature Engineering
F1	F1 Score
HDL	High-Density Lipoprotein
HR	Heart Rate
KNN	K-Nearest Neighbors
LR	Logistic Regression
LIME	Local Interpretable Model-agnostic Explanations
ML	Machine Learning
RF	Random Forest
SVM	Support Vector Machine
SHAP	SHapley Additive exPlanations
Thalach	Maximum Heart Rate Achieved
XGB	Extreme Gradient Boosting

Chapter 1

Introduction

1.1 Background

Cardiovascular Disease (CVD) is a leading reason why people die in many developed and developing countries around the world[1]. CVDs today are the most serious cause of mortality throughout the world, causing around 18 million deaths each year [2].CVD includes Coronary Artery Diseases (CAD) such as angina and heart attacks, and Coronary Heart Disease (CHD), where a thick, waxy substance called plaque builds up in the coronary arteries, which can lead to a heart attack. However, the exact reason for the CVDs has not yet been found. There are several risk factors involved with the chances of getting a CVD like high blood pressure, smoking, high cholesterol, diabetes, obesity, family history, and age [3]. The surgical treatment of heart disease is challenging, particularly in developing countries because they lack trained medical staff as well as surgical equipment and other resources required [4].Early detection and accurate prediction of cardiovascular diseases (CVD) are important tools to reduce deaths and improve patient outcomes. However, traditional methods like clinical risk scores, medical history, and imaging have limitations when it comes to precisely identifying individual risks [5]. Accurately assessing the risk of heart failure can help prevent severe heart attacks and make patients safer [6].

Machine Learning (ML) algorithms can be a good solution for identifying diseases when they are trained on the right data [7]. ML is used in various areas of medicine, including heart failure treatment, helping doctors make decisions, and analyzing medical images [8]. ML focuses on how computers can improve their performance automatically by learning from experience [9].

ML is the study of algorithms and statistical models that help computers perform tasks without being programmed for each specific case [10]. It is a rapidly growing field in computer science. ML involves using statistical techniques to create mathematical models that make predictions based on data samples, called a training set. ML is part of artificial intelligence and must be able to adapt to new situations [11].

Combining different fields like mathematics, statistics, and computer science is essential for data science, which helps build various ML models [12]. An algorithm learns from existing data and uses that knowledge to make decisions based on new data with similar features [13].

Studies on ML systems in healthcare show that these systems can sometimes make biased decisions or recommendations [14]. Therefore, it is important to ensure that these systems are fair and do not favor certain ethnic or social groups [15].

1.2 The Need for Explainability

Recently, ML has been used in a variety of medical disciplines including heart failure management, clinical decision support in clinical medicine, and medical imaging [16]. While there is a call to apply interpretable ML models to many domains, healthcare is particularly challenging due to medicolegal and ethical requirements, laws, and regulations [17]. Audits of ML systems in domains like healthcare reveal that the decisions and recommendations of ML systems can be biased [18]. Thus, interpretability is needed to ensure that such systems are free from bias and fair in scoring different ethnic and social groups [19].

With the rapid advancements in artificial intelligence (AI) and machine learning (ML), there has been a paradigm shift in the way cardiovascular risk assessment is conducted.

ML algorithms can analyze vast amounts of patient data to detect patterns and correlations that might be overlooked by traditional statistical methods. These models, including Random Forests, Gradient Boosting Machines (GBMs), and Deep Neural Networks (DNNs), have demonstrated superior performance in predicting CVD risk based on clinical and demographic features [20].

However, despite their high predictive accuracy, the adoption of these models in clinical practice has been limited due to their "black box" nature, which lacks interpretability and transparency. This has led to concerns regarding trust, accountability, and ethical implications in AI-driven healthcare [21].

1.3 Recent Advances and Applications

The concept of Explainable Artificial Intelligence (XAI) has gained traction as a solution to the interpretability problem in ML models. XAI methods provide insights into how a model makes decisions, enabling clinicians to understand the reasoning behind predictions and, in turn, fostering trust in AI applications [22].

XAI comprises a set of methodologies that make machine learning (ML) models more interpretable without compromising predictive performance.

Among various XAI techniques, SHAPley Additive explanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and the What-If Tool (WIT) have gained prominence in CVD prediction.

These methods provide insights into how models make decisions, thereby increasing their adoption in clinical practice [23].

SHAP is a game-theoretic approach that assigns each feature an important value based on its contribution to a model's output. SHAP values are particularly useful

in CVD prediction as they allow clinicians to understand which risk factors—such as age, cholesterol levels, and blood pressure—contribute most to an individual's risk score [24].

This method ensures consistency and provides global and local interpretability, making it a preferred choice in medical AI applications [**?**].

SHAP is based on cooperative game theory and assigns important values to each feature in a model's decision-making process, providing both local and global interpretability. It helps clinicians understand how risk factors such as age, cholesterol levels, and blood pressure contribute to a patient's predicted likelihood of developing CVD.

On the other hand, LIME creates simple models that help explain individual predictions by changing input features and seeing how they affect the model's results. This method is especially helpful in personalized medicine [25]. LIME works by building a simpler model that mimics the behavior of a complex model, but only around the specific data point being studied. In predicting cardiovascular disease, LIME helps show which factors influenced a particular prediction. This approach works with many different types of machine learning models used in heart health assessments [26]. LIME is also useful for finding mistakes in model predictions and retraining models to make them less biased [27].

Several studies have shown that **SHAP** and **LIME** are effective in making AI models used in healthcare easier to understand. For example, a 2024 study focused on predicting heart attacks used SHAP to explain how factors like heart rate and cholesterol levels affect the risk of cardiovascular disease, achieving an **F1-score of 91.2%** [28]. Another study from 2023 used both LIME and SHAP with Random Forest models to improve explainability in diabetes and heart disease risk assessments, showing how important these methods are for helping doctors make decisions [29].

Interpretable AI models help make AI predictions more useful for doctors by allowing them to check if the model's reasoning matches medical knowledge, spot

unfair patterns, and adjust treatment plans as needed [30].

The European Union's GDPR and other global rules stress the "right to explanation," which means AI models used in healthcare need to be clear and explainable [31]. Besides research, AI tools for predicting heart disease risk are now being tested in real hospitals. The NHS in England is trying out an AI system called Aire, which looks at electrocardiogram (ECG) results to predict a person's heart disease risk. The system uses XAI methods to keep its results clear and trustworthy, showing that there's growing interest in making AI tools in healthcare easy to understand and rely on.[32].

1.4 Rationale and Significance of Selected Research Topic

i. Trust and Transparency: XAI helps build trust between humans and AI systems by making the decision-making process transparent and understandable.

ii. Need for Explanations: In healthcare, practitioners require explanations for AI-driven decisions.

iii. Informed Decision-Making: XAI can provide insights that can help practitioners make more informed decisions.

1.5 Research Problem

"Existing ML models often struggle with imbalanced datasets and complex feature representations, limiting their adoption in real-world settings. Furthermore, ML models often struggle with the black-box nature that prevents healthcare professionals from understanding the reasoning behind their predictions, creating a significant barrier to trust and clinical integration. Previously proposed XAI

SHAP and LIME models have limitations such as inconsistent explanations which restrict their clinical usability."

1.6 Research Question

AI (XAI) with Machine Learning (ML) for the detection and prediction of cardiovascular disease (CVD):

i. How can effective feature engineering improve the predictive performance of machine learning models for heart disease detection?

ii. How can explainable AI techniques using SHAP and LIME be applied to interpret the predictions for heart disease?

iii. How do explainability methods contribute to building trust among healthcare professionals and support the real-world applicability of predictive models in clinical decision-making?

1.7 Research Objectives

This research goal is to explore the application of SHAP and LIME in CVD prediction models to evaluate their effectiveness in enhancing model transparency and providing trustworthy predictions. The key objectives include:

i. Developing and implementing ML models for CVD prediction using real-world datasets.

ii. To design and impliment feature engineering techniques(such as combining clinical related related attributes and deriving new risk-based indicators) to improve the predictive power and clinical relevance of the dataset.)

iii. To analyze the impact of various XAI methods on the understanding of feature importance within ML models for CVD detection and explore the implications of these insights for clinical decision-making processes.

iv. Applying SHAP and LIME to explain model predictions, to identify the key risk factors particular to patients for a CVD prediction.

By integrating explainable AI into CVD prediction, this research seeks to bridge the gap between AI-driven models and real-world clinical applications, ultimately contributing to more accurate and trustworthy cardiovascular healthcare solutions.

Chapter 2

Literature Review

2.1 Introduction

The objective of the literature review in current research focused on achieving transparency in cardiovascular diseases (CVD) is to use explainable AI (XAI) and deep learning models to critically evaluate and synthesize existing literature on the topic. This involves reviewing a broad range of sources, including journal articles, conference papers, and relevant books, to identify knowledge gaps and current trends in the field. Additionally, the literature review provides a theoretical framework for the research by examining different models, algorithms, and methodologies that have been used in previous studies related to CVD diagnosis and treatment. By identifying the strengths and weaknesses of these approaches, practitioners can make informed decisions regarding the most suitable methods to apply in their study. In the study titled "An AI-Enabled Framework for Transparency and Interpretability in CVD Risk Prediction," the authors proposed a machine learning pipeline for the early prediction of CVD. The framework utilizes a Random Forest classifier to achieve high accuracy and effectively handle outliers in high-dimensional healthcare data. The model was trained on a structured dataset comprising clinical risk factors such as age, blood pressure, cholesterol, and ECG readings, achieving 98% accuracy. SHAP (SHapley Additive Explanations) was applied to identify which features most influenced predictions across the dataset,

while LIME (Local Interpretable Model-Agnostic Explanations) provided local interpretability, enabling doctors to understand individual predictions [33]. The research presented by Gularia et al. focuses on how ML models classify datasets and predict clinical insights, addressing the growing demand for interpretability in disease detection. Ensemble classifiers (SVM, AdaBoost, KNN, Logistic Regression, Naive Bayes) were employed within an XAI framework using a CVD dataset with 303 instances and 14 attributes. XAI-enhanced SVM, LR, and Naive Bayes models demonstrated strong performance, achieving 89% accuracy using various evaluation metrics [34].

Guang Yang et al. surveyed the progress of XAI in healthcare, proposing an XAI solution for multi-center and multi-model data fusion. The paper emphasized the shortcomings of previous models in explaining their decision-making processes and proposed a model that classifies data while also providing interpretability of outcomes [35]. S.P. Patro and Neelamadhab Padhy focused on remote health monitoring (RHM) for CVDs using IoT medical sensors to collect patient data. The study employed SHAP to identify the most relevant predictive features. Various ML and DL algorithms were tested, with Artificial Neural Networks (ANN) achieving the highest accuracy of 91

Scott M. Lundberg and Su-In Lee introduced SHAP (Shapley Additive Explanations), a unified framework for interpreting complex models such as deep learning. SHAP assigns importance values to each feature, providing insight into individual predictions and supporting transparency, trust, and better decision-making [36].

Roohallah Alizadehsani et al. explored the use of explainable AI in drug discovery and development. The study offered a comprehensive review of XAI applications in areas like target identification, lead optimization, personalized medicine, and drug safety. It also discussed challenges such as the need for high-quality data and domain expertise [37].

Miller Ariza et al. demonstrated that ML models outperform traditional logistic regression methods in credit scoring. The study used SHAP to enhance explainability, showing how improved transparency and performance can facilitate

broader adoption of ML in risk assessment [38]. A. Smith and R. Kumar proposed a model integrating differential privacy and feature selection to protect sensitive medical data. ML algorithms such as Random Forest and Logistic Regression were used, alongside XAI techniques SHAP and LIME.

SHAP explained feature contributions globally, while LIME offered local explanations, helping healthcare professionals understand individual predictions. The model achieved 80% accuracy and an AUC of 0.85 [39].

M. A. Umar, N. AbuAli, K. Shuaib, and A. I. Awad proposed an IoT-based framework combined with XAI to monitor and predict CVD in real-time. The system consists of sensing, connectivity, cloud processing, and user interface layers.

SHAP and Permutation Feature Importance (PFI) were used to interpret predictions. The Random Forest model showed top performance on real-world (D1) and synthetic (D2) datasets, achieving 92.44% and 98.06% accuracy respectively, while SVM performed best on the Cleveland dataset (D3) with 84.62% accuracy [40].

2.2 Comparative Analysis of Existing Techniques

The reviewed studies span from 2020 to 2025 and demonstrate a growing trend toward integrating interpretability methods such as SHAP and LIME with traditional machine learning and deep learning models.

Most works employ structured clinical data, IoT sensor data, or combined EMR datasets, with Random Forest, SVM, KNN, ANN, and deep learning models being the most commonly used techniques.

The reported accuracies range from 80% to 98%, indicating strong predictive performance when XAI is incorporated. The key contributions highlight improved transparency, trust, and clinical interpretability of CVD prediction systems, particularly in real-time and IoT-based frameworks.

TABLE 2.1: Comparative Analysis of Literature Review

Year	Authors	Title	Dataset	Techniques	XAI	Results	Contribution	Limitations
2024	Guleria et al. [34]	*An AI-Enabled Framework for Transparency and Interpretability in CVD Risk Prediction*	Structured clinical dataset (age, BP, cholesterol, ECG)	Random Forest Classifier	SHAP & LIME	98%	Developed an interpretable ML framework integrating SHAP and LIME to explain CVD predictions.	Dataset limited to clinical attributes; lacks temporal/IoT data.
2023	S.P. Patro & N. Padhy [41]	*Remote Health Monitoring for CVDs using IoT and XAI*	IoT medical sensors data	ANN, ML & DL models	SHAP	91%	Proposed an IoT-based framework applying SHAP for feature importance analysis.	Limited patient dataset; ANN interpretability not deeply explored.
2020	Scott M. Lundberg & Su-In Lee [36]	*SHAP: A Unified Approach to Explain Model Predictions*	Conceptual framework	Any ML model	SHAP	–	Introduced SHAP framework for consistent model interpretability.	Methodology only; not CVD-specific.
2023	Roohallah Alizadehsani et al. [37]	*Explainable AI in Drug Discovery and Development*	Biomedical & chemical datasets	Various ML models	SHAP, LIME, etc.	–	Reviewed explainable AI applications in drug design and personalized medicine.	Focus on drug discovery, not diagnostic prediction.
2024	Miller Ariza et al. [38]	*ML vs Traditional Methods in Risk Assessment*	Credit scoring dataset	ML models (RF, XGBoost)	SHAP	–	Demonstrated SHAP-based explainability to enhance transparency in prediction models.	Domain outside healthcare; transferable methodology only.
2024	A. Smith & R. Kumar [39]	*Differential Privacy and XAI for CVD Prediction*	Medical dataset with age, BP, cholesterol	RF, Logistic Regression	SHAP & LIME	80%, AUC 0.85	Combined data privacy with explainable ML to enhance trust and confidentiality.	Moderate accuracy; focuses more on privacy than feature optimization.
2025	M.A. Umar et al. [40]	*IoT-based Explainable Framework for Real-time CVD Prediction*	IoT + EMR data (D1, D2, D3)	RF, SVM, KNN	SHAP & PFI	RF: 92.44%	Proposed a multi-layer IoT–XAI framework using SHAP & PFI for transparent real-time CVD monitoring.	Requires large-scale deployment validation; complex cloud integration.

Chapter 3

Proposed Methodology

3.1 Introduction

Machine learning has often been called a black box because it doesn't clearly show how it makes predictions. It simply tells whether a patient has cardiovascular disease or not. The new method improves the accuracy of these predictions and makes the process more transparent by using XAI models like SHAP and LIME. These models help explain how each factor in a patient's data affects the final result.

The lack of transparency can make doctors and patients less confident in using these systems, which can stop them from being used in real medical settings. To fix this, the new approach uses XAI techniques, specifically SHAP and LIME, to make the predictions more understandable. These models show how each part of the patient's information influences the outcome, making the process clearer and more trustworthy.

3.2 Architecture Diagram

The architecture diagram of proposed methodology is given below:

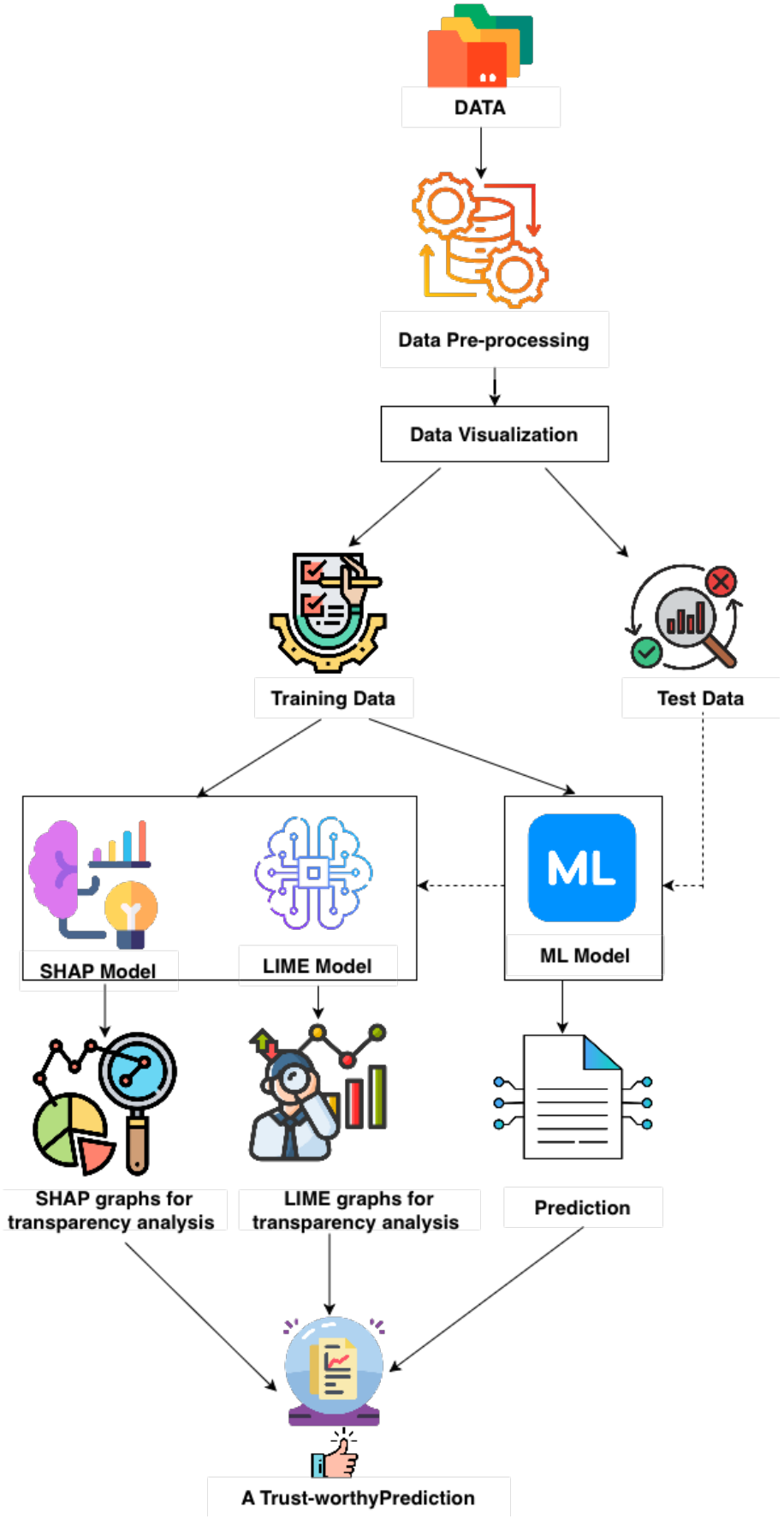

FIGURE 3.1: Architecture of proposed methodology

3.3 Selecting the Dataset

3.3.1 The Cleveland Heart Disease Dataset

In this research work, the Cleveland Heart Disease dataset has been used because it is widely used as a benchmark dataset in machine learning for predicting the presence or absence of heart disease.This dataset is widely recognized within the machine learning and medical informatics communities as a benchmark for binary classification tasks related to heart health prediction.

Here are the key aspects of the Cleveland dataset:

i. Origin: The dataset originates from the UCI Machine Learning Repository, a renowned archive for machine learning datasets, and specifically from the Cleveland Clinic Foundation [42].

ii. Size: It contains data for 1025 individuals.

iii. Features: For each instance, 13 predictive features are provided along with a single target variable. The original dataset contains 76 attributes, but most published experiments use a subset of 14 attributes (13 features + 1 target variable).It includes 5 numeric (continuous) attributes and 8 nominal (categorical) attributes.

iv. Target Variable: The target field indicates the presence of heart disease.

   (a) 0: No presence of heart disease.
   (b) 1: Presence of heart disease.

3.3.2 Detailed Feature Description

The following list describes the 13 features used in the processed Cleveland dataset:

i. Age: Patient's age in years (Numeric).

ii. Sex: Patient's gender (Nominal: 1 = male; 0 = female).

iii. Cp (Chest Pain Type) Type of chest pain experienced (Nominal, 4 categories):

   a. 0: Typical angina

   b. 1: Atypical angina

   c. 2: Non-anginal pain

   d. 3: Asymptomatic

iv. Trestbps (Resting Blood Pressure): Patient's resting blood pressure in mm/Hg (Numeric).

v. Chol (Serum Cholesterol): Serum cholesterol in mg/dl (Numeric).

vi. Fbs (Fasting Blood Sugar): Fasting blood sugar ¿ 120 mg/dl (Nominal: 1 = true; 0 = false).

vii. Restecg (Resting Electrocardiographic Results): Results of resting ECG (Nominal, 3 categories):

   a. 0: Normal

   b. 1: ST-T wave abnormality (e.g., T wave inversions, ST elevation or depression > 0.05 mV)

   c. 2: Probable or definite left ventricular hypertrophy by Estes' criteria

viii. Thalach (Maximum Heart Rate Achieved): Maximum heart rate achieved during exercise (Numeric).

ix. Exang (Exercise Induced Angina): Angina induced by exercise (Nominal: 1 = yes; 0 = no).

x. Oldpeak (ST Depression): ST depression induced by exercise relative to rest (Numeric).

xi. Slope (Slope of Peak Exercise ST Segment): Slope of the ST segment during peak exercise (Nominal, 3 categories):

   a. 0: Upsloping

   b. 1: Flat

c. 2: Downsloping

xii. CA (Number of Major Vessels): Number of major vessels (0–3) colored by fluoroscopy (Nominal).

xiii. Thal (Thalassemia): Blood disorder type (Nominal, 3 categories):

a. 1: Normal blood flow

b. 2: Fixed defect (no blood flow in part of the heart)

c. 3: Reversible defect (abnormal but present blood flow)

3.3.3 Target Variable

The target variable, named 'HeartDisease' is a binary attribute. A value of '1' indicates the presence of heart disease in the patient, while '0' indicates not CVD.

3.4 Data Pre-Processing

The data pre processing is done in the following steps.

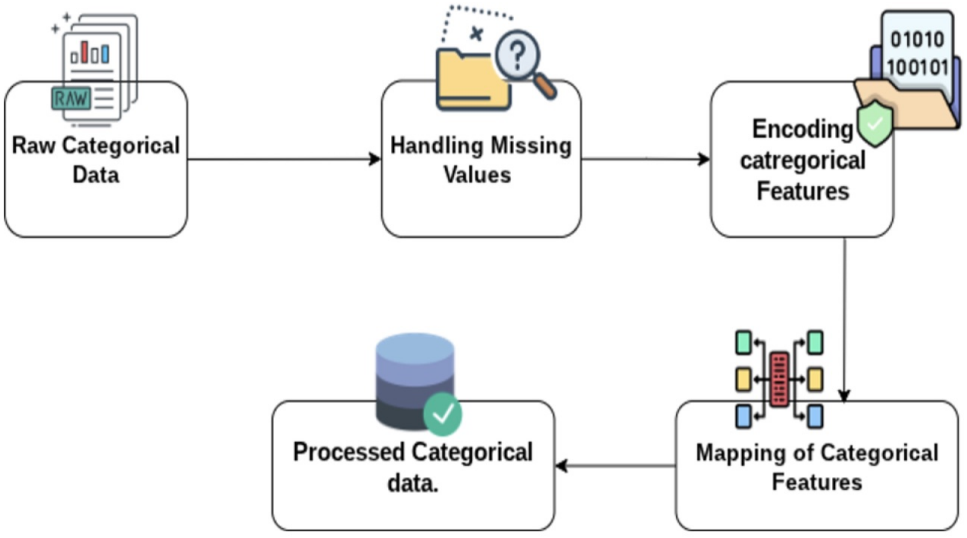

FIGURE 3.2: Data Preprocessing Steps.

3.4.1 Handling Missing Values

At first the missing values were checked in a dataset. The values were checked using .isnull().sum() and visualized with heatmap because if any missing data left it can affect the model's reliability and accuracy.

```
col = list(dataa.columns)
categorical_features = []
numerical_features = []
for i in col:
    if len(dataa[i].unique()) > 6:
        numerical_features.append(i)
    else:
        categorical_features.append(i)

print('Categorical Features :',*categorical_features)
print('Numerical Features :',*numerical_features)
```

FIGURE 3.3: Identifying the features.

3.4.2 Encoding Categorical Features

To ensure that all input features were compatible with the ML model, categorical features in the dataset were converted into numerical form using Label Encoding. Label Encoding assigns a unique integer to each category within a column.The categorical features were encoded using Label Encoding depending on the nature of the variable.

3.4.3 Mapping of the Categorical Variables

The following table describes the categorical columns and the encoded numeric values, and the corresponding category meaning.

```
Categorical Features : sex cp fbs restecg exang slope ca thal target
Numerical Features : age trestbps chol thalach oldpeak
```

FIGURE 3.4: Generated Output

TABLE 3.1: Encoded Table of Categorical Features

Column	**Encoded Value**	**Category Meaning**
sex	0	Female
	1	Male
cp	0	Typical angina
	1	Atypical angina
	2	Non-anginal pain
	3	Asymptomatic
restecg	0	Normal
	1	ST-T wave abnormality
	2	Left ventricular hypertrophy
slope	0	Upsloping
	1	Flat
	2	Downsloping
thal	1	Normal
	2	Fixed defect
	3	Reversible defect
ca	0–4	Number of major vessels colored

```python
df_encoded = dataa.copy(deep=True)

# Initialize LabelEncoder
le = LabelEncoder()

# Identifying categorical features
categorical_cols = df_encoded.select_dtypes(include=['object', 'category']).columns

print(categorical_cols.tolist())
print("-" * 50)

# Dictionary to store mappings for each column
label_mappings = {}

# Encode only the identified categorical features
for col in categorical_cols:
    df_encoded[col] = le.fit_transform(df_encoded[col])
    mapping = dict(zip(le.classes_, le.transform(le.classes_)))
    label_mappings[col] = mapping

    print(f"\nLabel mapping for '{col}':")
    for category, code in mapping.items():
        print(f"  {category:15} --> {code}")

# Preview the DataFrame after encoding
print("\nDataFrame after encoding categorical features:")
print(df_encoded.head())

# Show data types after encoding
print("\nData types after encoding:")
print(df_encoded.dtypes)
```

FIGURE 3.5: Mapping of Categorical Features-I

```python
[35] label_mappings = {
        "sex": {
            0: "Female",
            1: "Male"
        },
        "cp": {
            0: "Typical angina",
            1: "Atypical angina",
            2: "Non-anginal pain",
            3: "Asymptomatic"
        },
        "restecg": {
            0: "Normal",
            1: "ST-T wave abnormality",
            2: "Left ventricular hypertrophy"
        },
        "slope": {
            0: "Upsloping",
            1: "Flat",
            2: "Downsloping"
        },
        "thal": {
            1: "Normal",
            2: "Fixed defect",
            3: "Reversible defect"
        },
        "ca": {
            0: "0 major vessels",
            1: "1 major vessel",
            2: "2 major vessels",
            3: "3 major vessels",
            4: "4 major vessels"
```

FIGURE 3.6: Mapping of Categorical Features-II

```
print("Summary of encoded values:\n")
for col, mapping in label_mappings.items():
    print(f"Column: {col}")
    for code, label in mapping.items():
        print(f"  {code:>2} --> {label}")
    print("-" * 40)
```

FIGURE 3.7: Mapping of Categorical Features-III

```
DataFrame after encoding categorical features:
   age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  \
0   52    1   0       125   212    0        1      168      0      1.0      2
1   53    1   0       140   203    1        0      155      1      3.1      0
2   70    1   0       145   174    0        1      125      1      2.6      0
3   61    1   0       148   203    0        1      161      0      0.0      2
4   62    0   0       138   294    1        1      106      0      1.9      1

   ca  thal  target
0   2     3       0
1   0     3       0
2   0     3       0
3   1     3       0
4   3     2       0

Data types after encoding:
age           int64
sex           int64
cp            int64
trestbps      int64
chol          int64
fbs           int64
restecg       int64
thalach       int64
exang         int64
oldpeak     float64
slope         int64
ca            int64
thal          int64
target        int64
dtype: object
```

FIGURE 3.8: Generated output

```
Summary of encoded values:

Column: sex
    0 --> Female
    1 --> Male
----------------------------------------
Column: cp
    0 --> Typical angina
    1 --> Atypical angina
    2 --> Non-anginal pain
    3 --> Asymptomatic
----------------------------------------
Column: restecg
    0 --> Normal
    1 --> ST-T wave abnormality
    2 --> Left ventricular hypertrophy
----------------------------------------
Column: slope
    0 --> Upsloping
    1 --> Flat
    2 --> Downsloping
----------------------------------------
Column: thal
    1 --> Normal
    2 --> Fixed defect
    3 --> Reversible defect
----------------------------------------
Column: ca
    0 --> 0 major vessels
    1 --> 1 major vessel
    2 --> 2 major vessels
    3 --> 3 major vessels
    4 --> 4 major vessels
----------------------------------------
```

FIGURE 3.9: Summary of the encoded value

3.5 Feature Importance

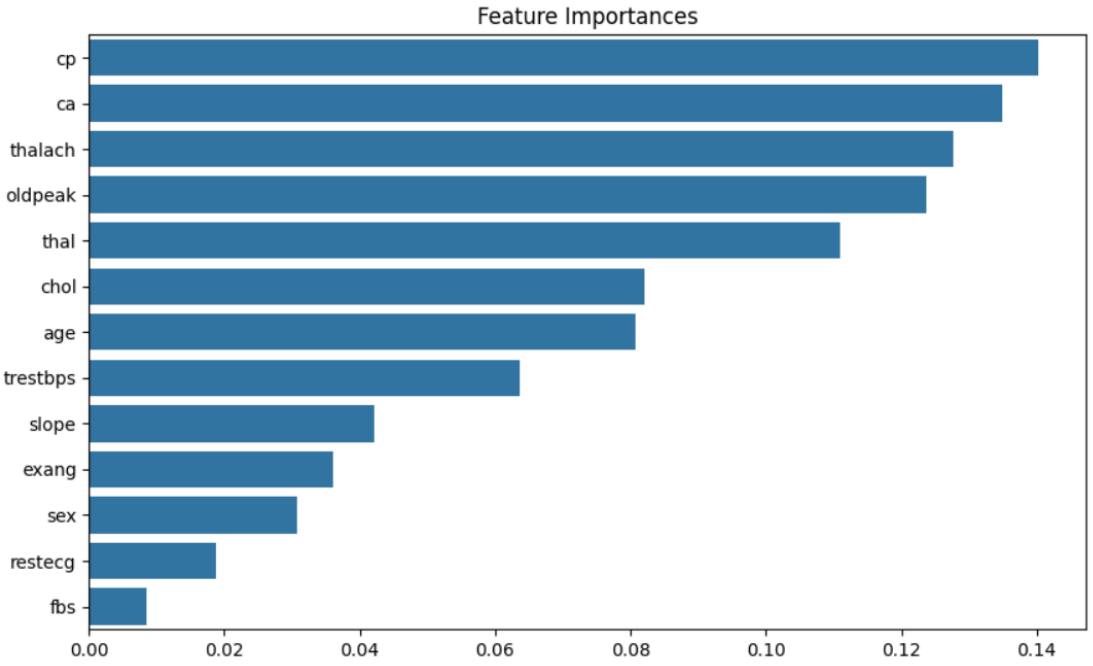

FIGURE 3.10: Feature Importance in Predicting Heart Disease

The bar chart titled *"Feature Importances"* illustrates the contribution of each feature to the decision-making process of the trained machine learning model. Features with higher importance values played a more significant role in predicting the presence of heart disease. The following observations summarize the top five contributing features:

(a) Cp (Chest Pain Type)

   i. This is the most important feature according to the model.

   ii. It signifies that the type of chest pain experienced by the patient is a highly predictive indicator of heart disease.

(b) Ca (Number of Major Vessels Colored)

   i. The number of blocked or visible vessels plays a crucial role in evaluating heart health.

   ii. A higher count of major vessels affected is strongly associated with heart disease risk.

(c) Thalach (Maximum Heart Rate Achieved)

i. Indicates the patient's cardiovascular performance during exercise.

ii. Higher maximum heart rates typically suggest better cardiovascular function and a lower likelihood of heart disease.

(d) Oldpeak (ST Depression Induced by Exercise)

i. Represents the amount of depression in the ST segment during exercise.

ii. Higher oldpeak values are associated with abnormal heart activity and potential disease presence.

(e) Thal (Thalassemia Type)

i. Certain types of thalassemia are linked to cardiovascular abnormalities.

ii. The model considers this feature significant in distinguishing between patients with and without heart disease.

This analysis helps identify key risk indicators and enhances the interpretability of the machine learning model in clinical settings.

3.5.1 Relative Important Features

i. Chol (Serum Cholesterol) and Age (Patient Age)

Both are well-known cardiovascular risk factors. However, in this particular model, they exhibit only moderate importance compared to other features like `Cp`, `Ca`, and `Thalach`.This suggests that while relevant, the model does not rely heavily on them for making predictions.

ii. Trestbps (Resting Blood Pressure)

This feature contributes to the prediction, but to a lesser extent. The elevated resting blood pressure is clinically associated with heart disease, yet its predictive power in the model is relatively low.

3.5.2 Least Contributing Features

The features that contributed least are listed below:

i. Slope

ii. Sex

iii. Exang

iv. Restecg

v. Fbs

3.6 Splitting the Dataset

After identifying the relative importance feature next step is to do the train-test split. It has the following mentioned steps.

3.6.1 Train-Test Split

```
Train-Test Split Completed

Total samples: 1025
Training set size: 820 samples
Testing set size: 205 samples

 • Feature matrix (X) shape:
X_train: (820, 13), X_test: (205, 13)

 • Target vector (y) shape:
y_train: (820,), y_test: (205,)

Target class distribution in training and test sets:
Train set:
 target
1    0.513415
0    0.486585
Name: proportion, dtype: float64

Test set:
 target
1    0.512195
0    0.487805
Name: proportion, dtype: float64
```

FIGURE 3.11: Shows the Train Test Split Outcome

Total Samples: 1025

The dataset contains 1025 preprocessed instances. Each instance includes multiple features (such as `age`, `cp`, `thal`, etc.) and a target variable (`target`: 0 or 1 indicating the absence or presence of heart disease). Training Set Size: 820 samples (80%) This portion of the data is used for training the machine learning model, allowing it to learn patterns and relationships from the input features.

Testing Set Size: 205 samples (20%).

This portion is reserved for evaluating model performance on unseen data, assessing generalization ability.

3.6.2 Feature and Target Shapes

3.6.2.1 Feature Matrix Shapes

The feature matrix shapes are listed below:

i. X_train: (820, 13) — 820 instances with 13 features each.

ii. X_test: (205, 13) — 205 instances with the same 13 features.

3.6.2.2 Target Vector Shapes

The target vector shapes are listed below:

i. `y_train:` 820 labels.

ii. `y_test:` 205 labels.

3.6.3 Target Class Distribution

3.6.3.1 Training Set

The training set contains the

i. Class 1 (heart disease): 51.34%

ii. Class 0 (no heart disease): 48.66%

3.6.3.2 Testing Set

The test set contains the

i. Class 1 (heart disease): 51.22%

ii. Class 0 (no heart disease): 48.78%

These balanced distributions in both training and testing sets help to prevent model bias toward any single class, which is beneficial for performance evaluation.

3.7 Data Scaling

In machine learning, especially when dealing with datasets involving medical attributes such as heart disease prediction, the features may have diverse units and magnitudes [43]. *Data scaling* is the process of transforming numerical features to ensure that each contributes equally to the model's learning process. It is an essential preprocessing step that brings all features to a common scale without distorting the differences in value ranges, thereby improving the model's accuracy [44]. Two common techniques used for data scaling include:

i. Standardization This method rescales the features so that they have a mean of 0 and a standard deviation of 1. It is especially useful for algorithms that assume a Gaussian distribution or rely on distance-based calculations (e.g., SVM, k-NN).

ii. Normalization This method rescales the features to a fixed range, typically between 0 and 1. It is particularly beneficial when the algorithm does not make assumptions about data distribution.

In this study, standardization was applied to features such as `age`, `resting blood pressure`, `cholesterol`, and `maximum heart rate` to handle their different value distributions. Normalization was specifically used for the `oldpeak` feature due to its relatively smaller range compared to other variables.

3.7.1 Importance of Data Scaling

The following are the key reasons for applying data scaling in this research:

3.7.1.1 Equal Feature Contribution

Features with larger numerical values can dominate the training process, leading to biased results. Data scaling ensures that all features contribute equally to the model.

3.7.1.2 Improved Model Performance

Properly scaled data enhances the accuracy and reliability of machine learning models by aligning feature distributions.

3.7.1.3 Faster Convergence

Scaling accelerates the convergence of optimization algorithms during model training, which reduces training time and improves computational efficiency.

3.7.1.4 Enhanced Explainability

Techniques such as SHAP (SHapley Additive exPlanations) require features to be on similar scales to provide fair and interpretable insights. Scaling ensures that no feature unduly dominates due to magnitude differences.

3.7.1.5 Standardization and Visualization

Standardization was applied on the preprocessed data to bring features to a common scale. The following image illustrates the correlation heatmap after standardization.

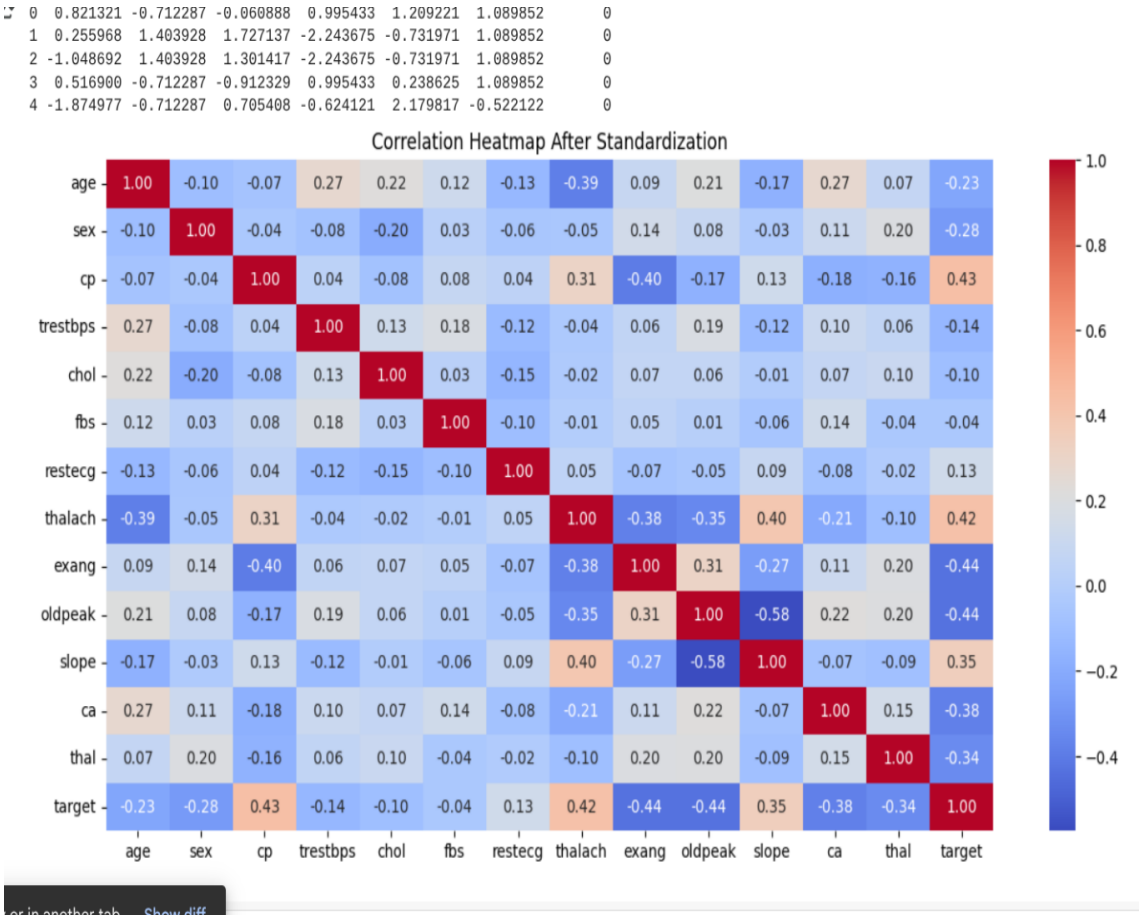

FIGURE 3.12: The correlation heatmap after Standardization

3.7.2 Standardization Heatmap Explanation

3.7.2.1 Color Scale

The heatmap uses the *coolwarm* color scale (blue–red), which makes it easier to distinguish between negative and positive correlations.

3.7.2.2 Feature Values

All feature values have been transformed to have a mean of 0 and a standard deviation of 1 through standardization.

3.7.2.3 Scale Invariance

Despite standardization, the correlation values remain unchanged, as correlation is inherently scale-invariant. The key highlights are listed below:

i. cp (Chest Pain Type) Shows a strong positive correlation with the target variable (+0.43), indicating its high predictive value.

ii. exang (Exercise-Induced Angina) and oldpeak (ST Depression) Both display strong negative correlations with the target (−0.44), indicating a significant relationship with heart disease.

iii. thalach (Maximum Heart Rate Achieved) Positively correlated with the target (+0.42), suggesting that higher maximum heart rate tends to be associated with healthier individuals.

3.7.3 Data Normalization and its Importance

Data normalization is an important step in preparing data for machine learning. It changes the features so they all fall within a similar range, usually between 0 and 1. This is especially important for certain types of algorithms, like K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and neural networks. These algorithms are very sensitive to how large or small the numbers are in the data. Without normalization, features that have bigger numbers can have a much stronger effect on the model, making it less accurate or biased. By normalizing the data, the model becomes more stable, learns faster, and makes better predictions.

Normalization is especially useful in medical data where different features might be measured in different units and have different scales. Recent research has shown that normalization helps machine learning models work better in health-related tasks and is important when using techniques like neural networks and ensemble methods. The normalization was done on the prepared dataset, and the figure shows the correlation heatmap after this step. This ensures that no single feature has too much influence because of its size, allowing all features to contribute fairly to the model's learning process.

3.7.4 Normalization Heatmap Explanation

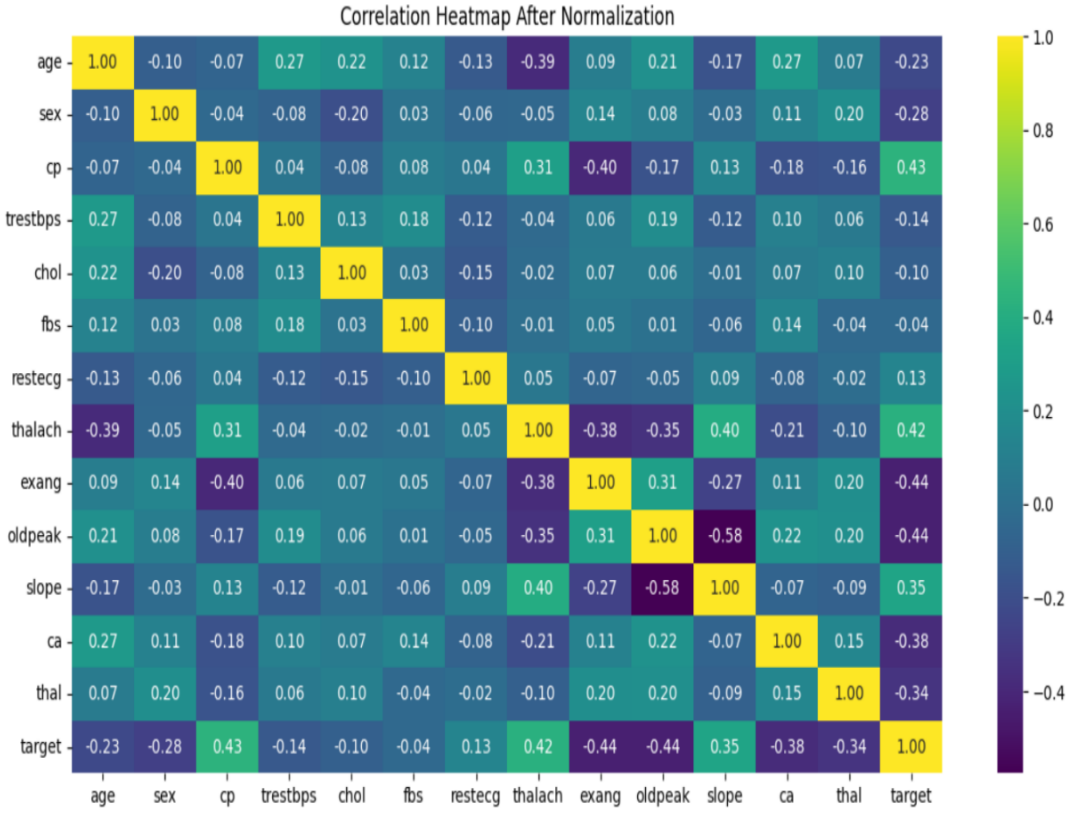

FIGURE 3.13: Correlation heatmap after normalization

The correlation heatmap generated after normalization uses the **Viridis color scale** (green–yellow), which offers a visually distinct representation compared to the *Coolwarm* palette used earlier. However, it is important to note that:

i. Feature values have been rescaled to the range [0, 1] through normalization.

ii. Correlation values remain unchanged, as correlation is a scale-invariant measure. Therefore, the strength and direction of relationships between variables are identical to those observed after standardization.

3.7.4.1 Key Observations

i. Cp (chest pain) shows a strong positive correlation with the target variable:+0.43.

ii. Exang (exercise-induced angina), oldpeak (ST depression), and thal (thalassemia) exhibit negative correlations with the target: approximately $-\mathbf{0.44}$.

iii. thalach (maximum heart rate achieved) is positively correlated with the target variable: +0.42.

The heatmap serves to visually reinforce the influence of these features on heart disease prediction while validating that normalization does not affect the interpretability of the correlation structure.

3.8 Feature Engineering

After the data preprocessing steps such as missing value handling, encoding, and feature scaling, the next step is to perform the feature engineering to enhance the predictive power of the model. The feature engineering procedure involves the creating new variables or transforming existing ones to better understand the underlying patterns in the data, especially in healthcare feilds like cardiovascular disease prediction.

The several new features are made after feature engineering:

3.8.1 Age Grouping

To enhance model interpretability and reflect clinically meaningful distinctions, the continuous `age` variable was transformed into a categorical feature by grouping patients into defined age categories. This approach replaces raw numerical values with age brackets that align with cardiovascular risk profiles:

i. Young ≤ 40 years

ii. Middle-aged 41–55 years

iii. Senior 56–65 years

iv. Elderly > 65 years

This binning technique is supported by cardiovascular health literature, which emphasizes that the risk of heart disease rises significantly with age. According to the American Heart Association [45],

> "Age is a major non-modifiable risk factor for cardiovascular disease, with a significantly higher incidence in individuals above 55."

By converting the continuous age variable into meaningful categories, the model can better capture non-linear relationships between age and heart disease risk, improving both interpretability and performance.

3.8.2 Cholesterol Risk Categorization

The `chol` feature, representing serum cholesterol levels measured in mg/dL, was transformed into categorical risk levels based on the World Health Organization (WHO) guidelines. This categorization enhances interpretability and aligns model inputs with real-world clinical practices. The categories are defined as follows:

i. Normal < 200 mg/dL

ii. Borderline High 200–239 mg/dL

iii. High ≥ 240 mg/dL

This transformation was implemented in Python using the `pd.cut()` function, allowing numerical values to be discretized into bins that reflect established medical thresholds.

According to the WHO:

> "Cholesterol levels above 240 mg/dL significantly increase the risk of atherosclerosis and heart attack.[46]

This method contributes to improved model interpretability and can support clinical decision-making in heart disease prediction tasks.

3.8.3 Heart Stress Index

A new continuous feature was engineered to capture cardiovascular performance under physical exertion. This feature, termed the Heart Stress Index, is defined as the ratio of the maximum heart rate achieved (`thalach`) to the patient's age:

$$\text{Heart Stress Index} = \frac{\texttt{thalach}}{\texttt{age}} \tag{3.1}$$

This ratio offers additional insights into a patient's cardiac efficiency, particularly under stress conditions. Lower values of the Heart Stress Index may indicate reduced cardiovascular capacity or diminished heart efficiency, especially in elderly individuals.

By integrating this domain-specific feature, the model gains access to a clinically relevant indicator that may enhance predictive power and interpretability.

3.8.4 Combined Angina Indicator

A new binary feature, **angina_combined**, was created by combining two clinically significant indicators:

i. exang Exercise-induced angina (1 = yes, 0 = no)

ii. oldpeak ST depression induced by exercise relative to rest

The feature was constructed based on the following clinical rule:

If a patient exhibited exercise-induced angina (`exang` = 1) and an `oldpeak` value greater than 1.0, then `angina_combined` was assigned a value of 1, indicating high cardiovascular risk. Otherwise, it was set to 0.

This derived feature captures compounded cardiac stress and is intended to highlight patients at elevated risk of ischemic heart events during exertion.

3.8.5 Total Risk Factor Count

A new composite feature, `risk_count`, was engineered to quantify the total number of elevated cardiovascular risk indicators per patient. It was calculated by summing the presence (binary form) of the following clinical attributes:

a. fbs > 0 (High fasting blood sugar)

b. exang > 0 (Exercise-induced angina)

c. ca > 0 (Number of major vessels colored by fluoroscopy)

d. oldpeak > 1.0 (Significant ST depression during exercise)

Each of these conditions contributes 1 point to the patient's `risk_count`, resulting in a score ranging from 0 to 4. This aggregated risk index serves as a simplified yet informative metric to estimate the overall cardiovascular stress or burden a patient may be experiencing.

3.8.6 Relationship Term: Age $\times$ ST Depression

To capture the compounded effect of aging and ST segment depression on cardiovascular risk, a new interaction feature called `age_oldpeak_interaction` was

created. This feature is defined as the product of the patient's age and their `oldpeak` value (ST depression induced by exercise):

$$\texttt{age\_oldpeak\_interaction} = \texttt{age} \times \texttt{oldpeak}$$

This interaction term allows the model to account for how the impact of ST depression may intensify with increasing age. Clinical studies suggest that older individuals exhibiting exercise-induced ST depression are at a significantly higher risk for adverse cardiovascular events. By including this multiplicative interaction, the model is better equipped to learn complex, non-linear relationships that are medically relevant.

3.8.7 One-Hot Encoding of Categorical Features

To prepare the newly introduced categorical features for machine learning models, one-hot encoding was applied. This process transforms each category into a separate binary column, allowing algorithms to interpret categorical data numerically.

For example, the categorical feature `chol_risk` (which includes values like `Normal`, `Borderline High`, and `High`) was transformed into the following binary columns:

i. `chol_risk_Normal`

ii. `chol_risk_Borderline_High`

iii. `chol_risk_High`

Each of these columns holds a value of 1 if the patient's cholesterol level falls into the corresponding category and 0 otherwise. A similar approach was used for the `age_group` feature.

3.8.8 Outcome of Feature Engineering

After applying feature transformations, interaction terms, and encoding, the final dataset contained:

i. Original features (e.g., `age`, `chol`, `thalach`)

ii. Engineered features (e.g., `heart_stress_index`, `risk_count`, `angina_combined`, `age_oldpeak_interaction`)

iii. Encoded categorical features using one-hot encoding (e.g., `age_group_Senior`, `chol_risk_High`)

This enriched feature set improves model performance by incorporating clinically meaningful groupings, interactions, and reducing bias from categorical interpretation.

TABLE 3.2: Encoded Mapping of Categorical Variables

Original Feature	**Category**	**Encoded Column Name**	**Meaning**
age_group	Young (< 40)	(reference – dropped)	Implied when all `age_group` dummies are 0.
	Middle-aged (40–55)	`age_group_Middle_aged`	1 if Middle-aged, else 0.
	Senior (56–65)	`age_group_Senior`	1 if Senior, else 0.
	Elderly (> 65)	`age_group_Elderly`	1 if Elderly, else 0.
chol_risk	Normal (< 200 mg/dL)	(reference – dropped)	Implied when all `chol_risk` dummies are 0.
	Borderline High (200–240)	`chol_risk_Borderline_High`	1 if in this range, else 0.
	High (> 240 mg/dL)	`chol_risk_High`	1 if High, else 0.

```
... Feature engineering complete.
    Dataset shape: (1025, 21)

    Remaining features:
    ['sex', 'cp', 'trestbps', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal', 'target', 'heart_stress_index', 'angina_

    Sample data:
       sex  cp  trestbps  fbs  restecg  thalach  exang  oldpeak  slope  ca  ... \
    0    1   0       125    0        1      168      0      1.0      2   2  ...
    1    1   0       140    1        0      155      1      3.1      0   0  ...
    2    1   0       145    0        1      125      1      2.6      0   0  ...
    3    1   0       148    0        1      161      0      0.0      2   1  ...
    4    0   0       138    1        1      106      0      1.9      1   3  ...

       target  heart_stress_index  angina_combined  risk_count  \
    0       0            3.230769                0           1
    1       0            2.924528                1           3
    2       0            1.785714                1           2
    3       0            2.639344                0           1
    4       0            1.709677                0           3

       age_oldpeak_interaction  age_group_Middle-aged  age_group_Senior  \
    0                     52.0                   True             False
    1                    164.3                   True             False
    2                    182.0                  False             False
    3                      0.0                  False              True
    4                    117.8                  False              True

       age_group_Elderly  chol_risk_Borderline  chol_risk_High
    0              False                  True           False
    1              False                  True           False
    2               True                 False           False
    3              False                  True           False
    4              False                 False            True

    [5 rows x 21 columns]
```

FIGURE 3.14: Data set after applying feature engineeeirng, including newly derived features

3.8.9 Categorical Feature Engineering and Encoding

After feature engineering, new categorical features, i.e., `age_group` and `chol_risk`, were established based on clinical findings. Since machine learning models require numerical input, these categorical features were transformed using one-hot encoding. This process created binary columns for each category, enabling the model to interpret them without assuming any ordinal relationship. For example, the `age_group` feature produced columns such as `age_group_Middle-aged`, `age_group_Senior`, and `age_group_Elderly`.

To prevent multicollinearity, one category from each feature was dropped and treated as a reference during encoding. This means that if all dummy variables within a feature are zero, the observation belongs to the reference category (e.g., "Young" for `age_group`). This approach ensures that the dataset remains reliable, valid, interpretable, and ready for training machine learning models.

Chapter 4

Model Development and Evaluation

4.1 Introduction

This chapter presents the evolution and evaluation of the ML model used for the prediction of CVD. The data pre-processing and feature engineering is already done in chapter 3. The dataset is already preprocessed and engineered for the model's training.

4.2 Model Selection and Justification

XGBoost was initially used as the main model in this study because it is well-known for its high predictive performance in machine learning tasks. However, several limitations became apparent during experimentation. The model needed a lot of hyperparameter tuning to achieve stable results, which made training slow and demanding on resources. Additionally, XGBoost was sensitive to noise and variations in the dataset, which impacted its ability to generalize. Another significant concern was interpretability, as its sequential boosting structure made it harder to explain predictions. This is crucial for medical decision-support systems.

During the evaluation, XGBoost showed lower accuracy compared to the Random Forest model using the same dataset. Given the moderate size of the dataset and the need for a stable, interpretable, and efficient model, Random Forest showed better performance and reliability. Therefore, even though XGBoost was implemented first, the study shifted to Random Forest ultimately because it offered better accuracy, needed less tuning, and provided clearer insights into feature importance, making it more suitable for predicting heart disease. The random forest ML model is selected as a core model for this study because of its reliability, versatility, and effectiveness in the classification tasks, specifically in the healthcare domain. The random forest is an ensemble learning model that creates multiple decision trees during the training of the model and outputs the mode of the classes for its final prediction. This process reduces the risk of overfitting, and the random forest deals with both the numerical and categorical data, which is being used in this research. It performs so well that it is useful in scenarios where input data is complex and has nonlinear relationships in the features, specifically when dealing with healthcare data. Also, it requires very little parameter tuning, therefore making it a reliable and efficient choice. Another advantage of random forest is that it has the built-in mechanism for evaluating the feature importance, which is soo important for domains, like healthcare where transparency is a main concern. The another significant advantage of random forest is that it does not require the data feature scaling. Although in this study preprocessing steps like standardization and normalization were already applied to make sure that model will work accurately and improve the overall convergence and interpretability.

4.3 Model Training

4.3.1 Model Training Using Random Forest Classifier

The final preprocessed dataset was used to train the machine learning model. This dataset included scaled numerical features, one-hot encoded categorical features, and newly created variables such as ‘age$_group$‘and‘$chol_risk$‘.

To train the Random Forest Classifier, the dataset was divided into training and testing subsets using an 80:20 split—i.e., 80% of the data was used for training, and the remaining 20% was reserved for testing. Stratified sampling was applied during this split to ensure that the target variable, indicating the presence or absence of cardiovascular disease (CVD), was proportionally represented in both subsets. This technique helps maintain class balance and prevents training bias.

The 'RandomForestClassifier' from the 'sklearn.ensemble' module was used, with the number of trees ('$nestimators$') set to 100 and a fixed 'random$_s tate$'$toensurereproducibility$.

The model was trained using the '$fit()$' method on the training data, while the remaining 0.20 of the dataset was used to evaluate the model's predictive performance.

An additional advantage of the Random Forest algorithm is that it does not require feature scaling. However, in this study, preprocessing steps such as standardization and normalization were still applied to enhance model performance, ensure numerical stability, and improve interpretability.

Recent studies have emphasized that normalization enhances the generalizability of machine learning models in health diagnostics and is essential for workflows involving neural networks and ensemble learning techniques. The normalization was applied to the preprocessed dataset, and the figure below illustrates the correlation heatmap after normalization. This process ensures that no single feature dominates the learning process due to its larger magnitude, thereby allowing all features to contribute equally to model training.

4.4 Model Evaluation and Results

Once the model is trained, it is evaluated using the test data. The metrics for evaluating the model include accuracy, precision, recall, and F1-score. The Random

Forest model demonstrates excellent performance with an accuracy of 98.54%, indicating that it accurately classified nearly all instances in the test dataset.

The following classification report summarizes the results:

TABLE 4.1: The classification report of random forest classifier

Class	**Precision**	**Recall**	**F1-score**	**Support**
0 (No CVD)	0.97	1.00	0.99	102
1 (Has CVD)	1.00	0.97	0.99	103
Accuracy			**98.54%**	205
Macro avg	0.99	0.99	0.99	205
Weighted avg	0.99	0.99	0.99	205

4.4.1 Classification Report and Confusion Matrix

The classification report showed the following metrics for the positive class (CVD = 1):

i. Precision 1.00

ii. Recall 0.97

iii. F1-score 0.99

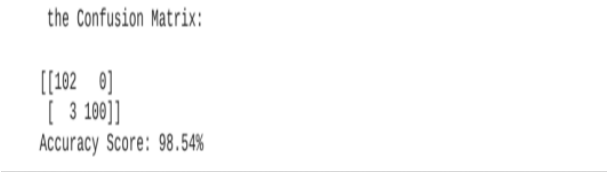

FIGURE 4.1: The confusion matrix of random forest classifier

These metrics indicate that the model has a high capability in correctly identifying actual CVD cases while generating very few false positives. The confusion matrix is presented below:

4.5 Feature Importance Analysis

To gain insights in to which feature influenced the most, the feature importance scores are extricated from the trainded random forest model. The feature importance plot below shows the most important features.

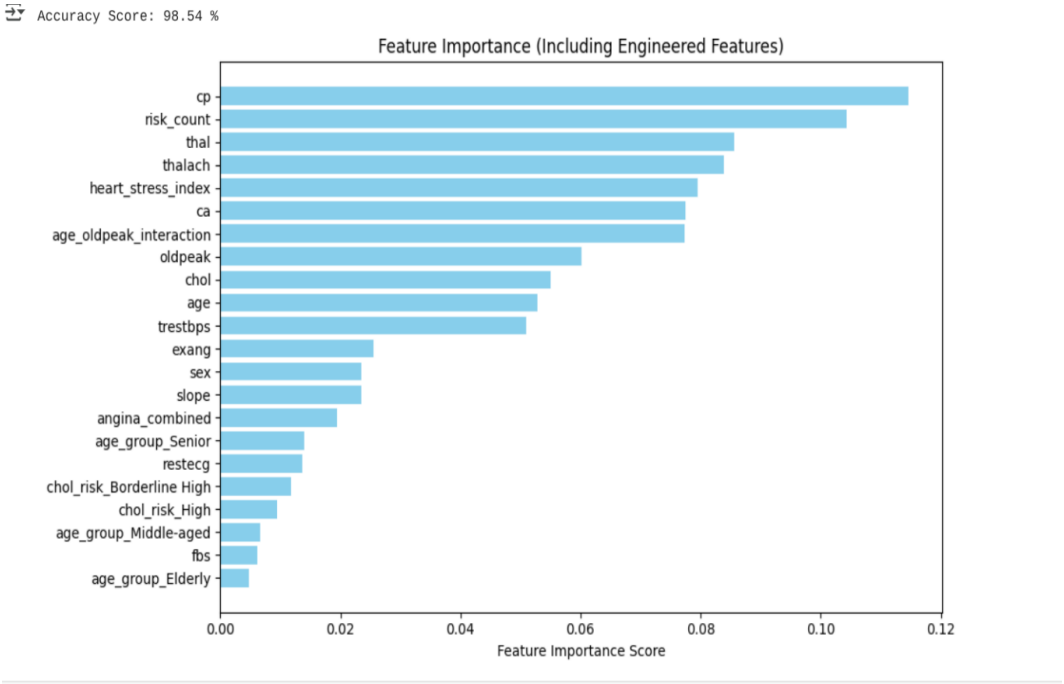

FIGURE 4.2: Feature Importance barplot

4.5.1 Feature Importance Analysis

A horizontal bar chart was generated to visualize the contribution of each feature to the model's predictions. The most influential features appear at the top, indicating which variables had the greatest impact on the model's decision-making.

Among the most influential features are `cp` (chest pain type) and the newly engineered variable `risk_count`. These features are consistent with established clinical knowledge and show a strong correlation with cardiovascular disease (CVD).

4.5.1.1 Key Findings

i. The feature importance plot highlights both original and newly engineered features that influenced the heart disease prediction model.

ii. Among all features, `cp` and `risk_count` emerged as the most powerful predictors, followed closely by `thal`, `thalach`, and `heart_stress_index`.

iii. Several engineered features, such as `age_oldpeak_interaction`, `angina_combined`, and categorical encodings like `age_group_Senior` and `chol_risk_Borderline High`, also contributed meaningfully to the model's predictions.

These observations indicate that the feature engineering process was successful in generating informative variables that enhanced the model's performance. In particular, the engineered features `risk_count` and `heart_stress_index` played a significant role in achieving the model's high accuracy of **98.54%**, thus demonstrating improved predictive capability and robustness.

Chapter 5

SHAP XAI Implementation

5.1 Introduction

In today's world, just having high accuracy from machine learning is not enough, especially in sensitive areas like healthcare, where it's important to understand why a prediction was made. To meet this need, XAI techniques are used to explain how black-box models make decisions. There are many XAI methods that help make predictions more understandable.

One example is SHAP (SHapley Additive exPlanations). It combines the overall importance of different features and gives both a big-picture and detailed view of how each feature influences the prediction. In this chapter, SHAP is used on a trained Random Forest model that predicts heart disease. The goal is to find out which factors had the biggest impact on the model's predictions and how they affected each individual patient's result. Using SHAP not only builds trust and transparency but also helps healthcare experts check if the model's reasoning makes sense.

5.2 Role of SHAP in Model Transparency

Machine learning models are often regarded as black-box systems, as they typically do not provide insight into the reasoning behind their predictions. In sensitive domains such as healthcare, it is critical for models not only to make accurate predictions but also to offer meaningful explanations for those predictions.

To address this challenge, eXplainable AI (XAI) techniques such as SHAP (SHapley Additive exPlanations) have been introduced. SHAP values quantify the contribution of each input feature to a given prediction, thereby enabling interpretable and transparent model decisions [47].

5.2.1 Model-Agnostic

SHAP can be applied to any machine learning model, although it is particularly optimized for tree-based models such as Random Forests and Gradient Boosting.

5.2.2 Fair Attribution

SHAP values are grounded in Shapley values from cooperative game theory, ensuring a fair and consistent distribution of contribution scores among all features.

5.2.3 Global and Local Explanations

SHAP provides both a global view of feature importance across the entire dataset and local explanations for individual predictions.

These characteristics make SHAP especially valuable in healthcare domains, where predictions must be both accurate and interpretable to support effective risk analysis and clinical decision-making.

5.3 SHAP Implementation for Heart Disease Classifier

5.3.1 Model Background

The Random Forest classifier used in this research is trained on a preprocessed version of the Cleveland Heart Disease dataset. The preprocessing steps include numerical scaling, categorical encoding, and feature engineering (e.g. , `heart_stress_index`, `risk_count`, etc.).The model achieved a high accuracy of 98.54% on the test data.

5.3.2 Model Training and Performance

The Random Forest classifier employed in this research was trained on a preprocessed version of the Cleveland Heart Disease dataset.

The preprocessing pipeline included numerical feature scaling, categorical encoding, and extensive feature engineering. Notable engineered features include `heart_stress_index`, `risk_count`, and several interaction terms.

The trained model demonstrated excellent performance, achieving a test accuracy of **98.54%**, indicating its strong ability to correctly classify cardiovascular disease cases.

5.3.3 SHAP Layout

To interpret the predictions made by the Random Forest model, the `TreeExplainer` from the SHAP library was utilized, as it is specifically optimized for tree-based models. SHAP (SHapley Additive exPlanations) values were computed to provide insights at both global and local levels:

5.3.3.1 Global Interpretability

Identifies which features are most influential across the entire dataset, offering an overview of feature importance.

5.3.3.2 Local Interpretability

Explains how individual features contributed to a specific prediction, providing case-level reasoning.

This dual-level interpretability supports both overall model understanding and case-specific analysis, which is particularly crucial in healthcare applications.

5.4 Local Interpretability SHAP Force Plot

To gain insights on how the model predicts a specific prediction, i have used the SHAP force plots. These plots features how individual attributes pushed the prediction towards "Heart Disease" or "No Heart Disease".

The SHAP force plot visualization is shown below:

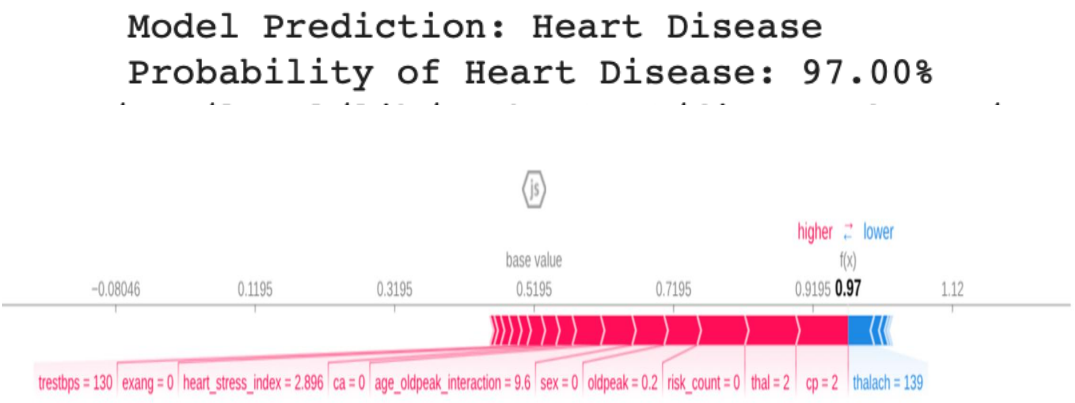

FIGURE 5.2: SHAP force plot showing feature contributions for an individual prediction

```python
import shap
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np # Import numpy

explainer = shap.TreeExplainer(rf)  # rf = trained model

patient_index = 2
patient_features_array = X_test.iloc[patient_index].values

shap_values_all_classes = explainer.shap_values(X_test.iloc[[patient_index]])[0]

# Selecting the SHAP values for class 1 (Heart Disease), which is the second column (index 1)

shap_values_instance = np.array(shap_values_all_classes[:, 1])


# Printing basic prediction info
predicted_class = rf.predict(patient_features_array.reshape(1, -1))[0]
predicted_prob = rf.predict_proba(patient_features_array.reshape(1, -1))[0][1]

print(f"\n Model Prediction: {'Heart Disease' if predicted_class == 1 else 'No Heart Disease'}")
print(f" Probability of Heart Disease: {predicted_prob * 100:.2f}%")

# Shows SHAP Force Plot (explains the contribution of each feature)
shap.initjs()
shap.force_plot(
    base_value=explainer.expected_value[1], # Base value for class 1
    shap_values=shap_values_instance,      # SHAP values for the instance (class 1)
    features=patient_features_array,       # Feature values for the instance (1D array)
    feature_names=X_test.columns.tolist() )# Feature names
```

FIGURE 5.1: Code Snippet of force plot for an individual patient

5.4.1 Explanation of Local SHAP Interpretation

The SHAP values for a particular individual revealed how specific features influenced the model's prediction.

i. Features such as `cp = 2`, `ca = 1`, and `thal = 3` contributed to pushing the model's prediction toward the class *"No Heart Disease"*.

ii. The base value of `0.01` indicates that the model assigned a 1% probability to the patient having cardiovascular disease (CVD), suggesting that the patient is likely to be healthy (i.e., no CVD).

This local explanation demonstrates the transparency offered by SHAP, highlighting how individual feature values impact the final prediction.

5.5 Conclusion

SHAP-based explanations show that both known risk factors for heart disease, like thalassemia and ST depression, and specially created features like risk_count and heart_stress_index are important in how the model makes its decisions. This confirms that the feature engineering done during model development was effective and trustworthy. Adding SHAP to the heart disease prediction process makes the Random Forest model easier to understand and more transparent. By looking at both the overall importance of features and the reasons behind each prediction, SHAP improves how well the model can be explained, helps doctors trust the model more, and supports the responsible use of AI in healthcare.

The model from this study ensures that AI is used fairly and openly in healthcare, and builds trust by clearly explaining why each prediction is made. Therefore, using SHAP is a useful step toward making AI decisions in healthcare more explainable, accountable, and accepted by medical professionals.

Chapter 6

LIME XAI Implementation

6.1 Introduction to Model Interpretability

As the ML role in the healthcare domain is increasing rapidly in today's world. So, the healthcare practitioners must understand why the model made the specific prediction related to a particular disease. In fields like CVD, the ML models, despite their higher accuracy, are still considered as Black-Box models because they do not provide explanations or reasoning behind those predictions. To overcome this issue, XAI techniques and models are used. In the prior chapter, I did the SHAP implementation. In this chapter, I'll perform the LIME model and explains how it contributes to explaining the model predictions. LIME provides. LIME provides intimate,human-understandable explanations by comparing the behavior of complex models with simpler and interpretable ones at a local level. This chapter performs a detailed implementation and analysis of LIME in the context of a trained Random Forest model for the CVD prediction.

6.2 Overview of LIME

LIME (Local Interpretable Model-Agnostic Explanations) is a widely used model-agnostic technique that explains individual predictions by locally approximating

a complex machine learning model with a simpler, interpretable model, typically a linear model.

LIME provides intuitive insights into how each input feature contributes to a specific prediction, often visualized as a bar chart that is easily interpretable by human users. In the context of heart disease prediction, LIME helps explain why a particular patient is classified as high-risk.

6.2.1 Key Characteristics of LIME

Local Fidelity LIME emphasizes local interpretability by focusing on how the model arrived at a decision for an individual data instance, rather than offering a global understanding. This is particularly beneficial in healthcare, where patient-specific explanations are crucial.

Model-Agnostic LIME does not require the use of any specific machine learning algorithm. It is compatible with a wide range of models including Random Forests, Neural Networks, and Naive Bayes classifiers.

Human Interpretability LIME offers transparent and intuitive explanations that can be easily understood by healthcare professionals, fostering trust in AI-assisted clinical decision-making.

Feature Importance Visualization LIME produces a bar chart showing the contribution of each feature to the model's prediction, indicating whether a feature pushes the prediction toward or away from a certain class.

Customizable The number of features included in the explanation can be adjusted, allowing the explanation to be tailored to the user's level of expertise and interest.

These characteristics make LIME a powerful tool for enhancing the interpretability of machine learning models in sensitive domains such as healthcare, where clear reasoning behind model predictions is essential.

6.3 Implementation of LIME

The Random Forest classifier trained (as described in Chapter 4) is used. After making sure that all the pre -processing and feature engineering steps were accurately implimented, the LIME is then deployed to gain the expalinability ans interpretability on a particular individual instance prediction.

```
import lime
import lime.lime_tabular
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# rf = trained RandomForestClassifier

#  X_train is in DataFrame format
if not isinstance(X_train, pd.DataFrame):
    X_train = pd.DataFrame(X_train, columns=X.columns)

if not isinstance(X_test, pd.DataFrame):
    X_test = pd.DataFrame(X_test, columns=X.columns)

#Creating a LIME explainer
explainer = lime.lime_tabular.LimeTabularExplainer(
    training_data=X_train.values,
    feature_names=X_train.columns.tolist(),
    class_names=['No Heart Disease', 'Heart Disease'],
    mode='classification'
)

#  Choose a specific patient to explain
i = 26
patient_data = X_test.iloc[i].values.reshape(1, -1)

exp = explainer.explain_instance(
    data_row=X_test.iloc[i].values,
    predict_fn=rf.predict_proba,
    num_features=10
)

exp.show_in_notebook(show_table=True, show_all=False)

# Saving explanation as HTML
exp.save_to_file("lime_explanation_patient26.html")
```

FIGURE 6.1: Code snippet of LIME Impplimentation

6.4 Explanation of LIME Output

Given below is the LIME model outcome for the selected individual patient record:

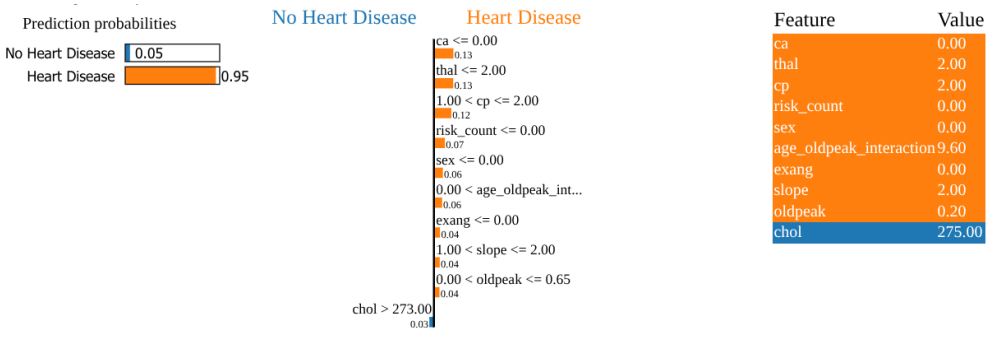

FIGURE 6.2: Outcome of the LIME model

6.4.1 Explanation of LIME Output

The LIME explanation shows how each feature affected the prediction result. It displays each feature along with its condition, like `thal > 130`, and how much it influenced the model's decision.

Each feature has a condition or threshold and a weight that shows its impact on the prediction. Features shown in orange help the model predict Heart Disease. Features in blue have less impact or even reduce the chance of Heart Disease.

6.4.1.1 Key Findings

`oldpeak > 1.2` had a strong positive impact on the model's prediction, increasing the likelihood of Heart Disease.

`thalach` $\leq$ `132`, indicating a lower maximum heart rate, also contributed positively toward the Heart Disease prediction.

In contrast, features like `chol` (cholesterol level) had a negative influence, decreasing the probability of Heart Disease in this instance.

This explanation demonstrates how LIME helps identify the most influential features for a specific prediction, making the model's behavior more transparent and interpretable to clinicians.

6.5 Advantages of LIME in Medical AI

LIME offers significant benefits for clinical interpretability, decision support, and model debugging. The following points highlight its practical utility in healthcare scenarios:

6.5.0.1 Clinical Interpretability

LIME aids clinicians in understanding specific decisions made by the machine learning model.

For instance, LIME may reveal that a combination of features such as `thalach` (maximum heart rate) and `oldpeak` (ST depression) significantly contributed to a heart disease prediction. This level of interpretability supports evidence-based clinical assessments.

6.5.0.2 Decision Support

LIME provides feature-level justifications for each prediction, allowing practitioners to evaluate the model's reasoning rather than relying blindly on its outputs.

This enhances both trust and transparency, which are critical in high-stakes domains like healthcare.

6.5.0.3 Debugging Models

Researchers can utilize LIME explanations to investigate conflicting or erroneous predictions. By examining the contributing features, they can determine whether the issue originates from poor data quality, incorrect labeling, or limitations in model behavior.

6.6 Conclusion

This chapter explains the application of LIME as an interpretable model-agnostic explanation used to understands the individual prediction made by the Random FOrest model for CVD detection. LIME visvalizes simple and clear charts to explain the individual patient outcome i.e. to identify which features influenced the most to predict whether the patient has CVD or not.

With LIME, we can see how features like high oldpeak, unusual thal values, or low thalach levels effects the prediction.Therefore, it allows the clinical practioners to trust the results more because the results are more transparent and more trustworthy as compared to the traditional ML models.

Through LIME, practitioners can percieve how factors such as high oldpeak, abnormal thal, or low thalach influences the model's decision-making process. These explanations not only encourages the trust among healthcare professionals but also permits the relaible deployment of ML in real-world settings.

Chapter 7

Results and Conclusion

In today's world, using machine learning (ML) in sensitive areas like healthcare often makes people unsure because many predictive models work in a way that's hard to understand. Even though these models can be very accurate, they're called "black-box" systems because they don't clearly explain how they reach their decisions. In hospitals, doctors and healthcare workers are often hesitant to trust these models unless they can clearly explain their findings. While ML models may correctly predict if a patient has a disease, they usually don't explain why they came to that conclusion.

This chapter has two main goals. First, it checks how well the proposed ML model works using measures like accuracy, precision, and recall. Second, it looks at how explainable AI (XAI) methods, like SHAP and LIME, help make the model more transparent by explaining individual predictions. By combining performance evaluation with explainability, this chapter supports the creation of AI tools that are clear, ethical, and useful in real-world healthcare settings, especially for detecting cardiovascular disease (CVD).

Additionally, the chapter shows the results of a CVD prediction model that uses the Random Forest classifier. The model was trained on the Cleveland Heart Disease dataset, which was prepared following the steps described in Chapter 3. The

preparation included scaling numerical data, converting categorical data into a format that can be used, and creating new features from existing ones. When trained on this enhanced dataset, the Random Forest model achieved a high accuracy of **98.54%**.. After that, SHAP and LIME were used to explain the model's predictions, giving detailed reasons behind specific outcomes. Through this approach of evaluating performance and improving explainability, the research provides not only a highly accurate model but also a trustworthy and understandable system that meets the needs of healthcare professionals.

7.1 Machine Learning Model Results

The random forest ML model is selected as a core model for this research because of its reliability, versatility, and effectiveness in classification tasks, specifically in the healthcare domain.

7.1.1 Classification Report

The performance of the Random Forest classifier was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score.

The model demonstrated excellent performance, achieving an accuracy of **98.54%**, indicating that it correctly classified nearly all instances in the test dataset. This high level of accuracy reflects the model's robustness and its effectiveness in distinguishing between patients with and without cardiovascular disease.

These metrics shows that the model has an high accuracy power to identify the actual CVD cases while making very few false positives.

TABLE 7.1: The classification report of the Random Forest classifier

Metric	**Class 0 (No Heart Disease)**	**Class 1 (Heart Disease)**	**Macro Avg**
Precision	0.97	1.00	0.99
Recall	1.00	0.97	0.99
F1-score	0.99	0.99	0.99
Support	102	103	205

7.1.2 Confusion Matrix

Table 7.2 shows the confusion matrix for the Random Forest model on the test dataset.

TABLE 7.2: Confusion Matrix for Random Forest Model on Test Data

Actual / Predicted	**No Heart Disease (0)**	**Heart Disease (1)**
No Heart Disease (0)	102	0
Heart Disease (1)	3	100

7.1.2.1 Model Accuracy

The Random Forest model achieved a **model accuracy of 98.54%** on the test dataset, indicating excellent predictive performance and minimal misclassification.

These results indicate that the proposed Random Forest classifier is highly effective in predicting the presence or absence of heart disease, focusing on high accuracy and reliability. The classification report reveals balanced and high values for precision, recall, and F1-score across both classes (presence and absence of heart disease), which shows that the model is not only identifying positive cases correctly (true positives), but also reducing both false positives and false negatives. Moreover, the confusion matrix shows the minimal misclassifications, with nearly all instances correctly categorized. The model's ability to perform equally well across both classes. It indicates that the dataset was well-balanced, and that

the classifier did not exhibit a biased behaviour to favor one class over the other, which is a common problem in imbalanced medical datasets. This balance highlights that the preprocessing and feature engineering steps, i.e., proper encoding, scaling, and feature engineering steps, contributed equally to a model's ability to learn effectively.

This excellent performance across evaluation metrics indicates that the model generalizes well on unseen data. Therefore, this research model is accurately identifying the patients who are at high risk of CVD as well as patients who are not at high risk, making this model an effective real-world application in clinical decision making, trusted by doctors and healthcare professionals.

7.2 Feature Importance Analysis

7.2.1 Feature Importance Analysis Using Random Forest

Using the Random Forest classifier's built-in feature importance mechanism, the most influential features contributing to the prediction of heart disease were identified. These features include both original and engineered variables:

i. thal – Thallium stress test results

ii. cp – Chest pain type

iii. oldpeak – ST depression induced by exercise

iv. ca – Number of major vessels

Engineered features

i. heart_stress_index

ii. risk_count

iii. age_oldpeak_interaction

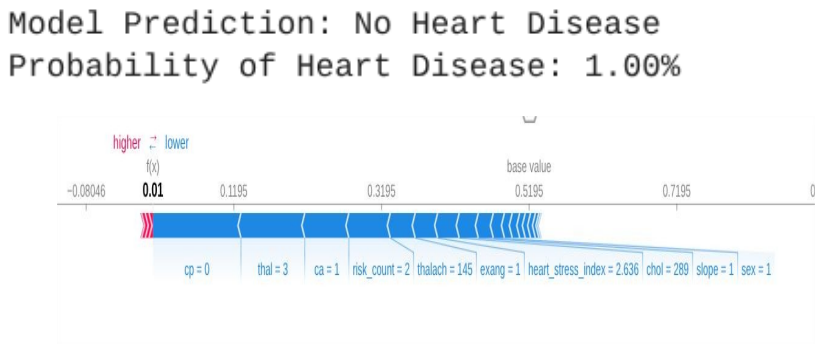

FIGURE 7.1: SHAP force plot showing feature contributions for an individual prediction

The inclusion of engineered features significantly improved the model's predictive capability. These features enabled the model to capture complex clinical patterns more effectively, contributing to the overall improvement in accuracy and confirming the success of the feature engineering process.

7.3 SHAP Implementation Results

For a test patient, SHAP explained the model's decision with a force plot :

7.3.1 SHAP Explanation and Analysis

7.3.1.1 Explanation

i. Features such as `cp = 0`, `ca = 1`, and `thal = 3` pushed the model towards the "No Heart Disease" prediction for a specific individual.

ii. The base value of `0.01` indicates that the probability of a particular patient having cardiovascular disease (CVD) is 1%, suggesting the patient is likely to have **no CVD**.

7.3.1.2 Advantages of SHAP

i. SHAP provides global interpretability across the dataset.

ii. SHAP offers class-specific explanations, such as contributions towards the "Heart Disease" class.

iii. SHAP aligns well with clinical understanding, increasing the trust of clinical practitioners.

iv. Based on Shapley values, SHAP provides fair and accurate results, ensuring clarity and avoiding misleading explanations.

v. SHAP generates impactful and easily understandable visualizations such as force plots and summary plots, which help doctors comprehend how a model makes specific predictions.

7.4 LIME Implementation Results

LIME focuses on how the model predicted an individual patient record rather than providing a global model overview. It helps healthcare practitioners to understand the individual decisions more clearly.

7.4.1 LIME Explanation Output

LIME outcome for a selected patient:

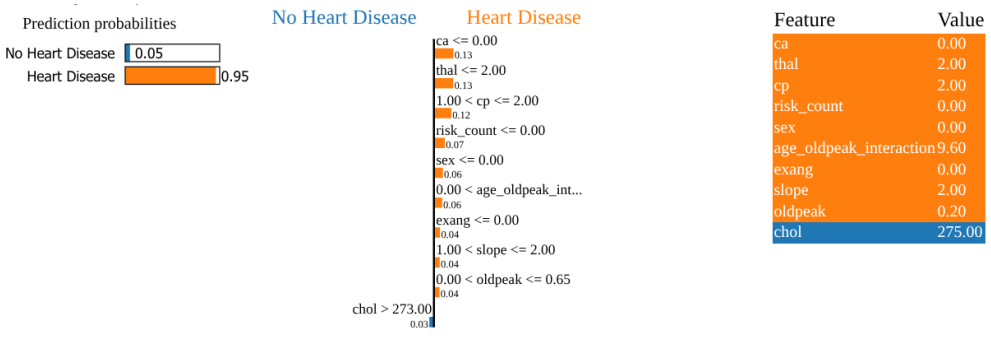

FIGURE 7.2: Outcome of the LIME model

7.4.2 LIME Explanation and Analysis

7.4.2.1 Explanation of the Outcome

i. Each feature shown in the output is displayed with a specific condition (e.g., `thal > 130`) and its contribution towards a particular class.

ii. The top contributing features are listed with their corresponding influence weights.

iii. Features shown in orange color push the prediction towards the Heart Disease class.

iv. Features shown in blue color contribute less or negatively toward the Heart Disease prediction.

v. This indicates that high ST depression (oldpeak) and **low heart rate (thalach)** pushed the model toward predicting Heart Disease, whereas normal chest pain (cp) and cholesterol levels influenced the prediction toward a healthy outcome.

7.4.3 Advantages of LIME

7.4.3.1 Clinical Interpretability

LIME helps clinical practitioners understand specific decisions made by the ML model. For instance, it can highlight that a combination of `thalach` and `oldpeak` contributed to predicting heart disease.

7.4.3.2 Decision Support

LIME provides feature-level justifications for predictions, enabling practitioners to understand and trust the model's outcome instead of blindly accepting it. This increases **trust and transparency**.

7.4.3.3 Debugging Models

Researchers can use LIME to analyze conflicting predictions and determine whether errors are due to data quality issues or model behavior.

7.5 Clinical and Ethical Implications

The interpretability provided by SHAP and LIME ensures transparency in model decisions, which is critical in healthcare applications. These explanation techniques promote trust among practitioners, patients, and regulatory authorities by clearly justifying the reasoning behind each prediction. Moreover, interpretability enhances accountability by enabling effective documentation and auditing of diagnostic decisions made by machine learning models. Thesis claims to enhance interpretability but this claim is not supported

7.6 Comparative Analysis and Research Enhancements

This section shows a comparison of different studies from 2020 to 2025. Most of these studies focused on making models that can predict cardiovascular disease (CVD) by classifying data. However, they didn't spend much time on preparing the data or creating useful features. Many used standard datasets like the Cleveland Heart Disease dataset and only did simple things like label encoding or normalization. Because of this, the models didn't do well at understanding complex patterns in medical data.

In contrast, this research developed a detailed approach to creating better features.

This included making new features by combining existing ones, like creating interaction terms such as $age_oldpeak\_interaction$, and composite scores like

$heart_stress_index$. These features helped capture hidden relationships and clinical factors that raw data didn't show, which improved both the accuracy and the clarity of the model.

Another big improvement in this study was using explainable AI (XAI) techniques. Specifically, it used SHAP and LIME to help understand how the model made predictions. Many previous studies used these tools separately, but didn't connect them to how feature engineering or model tuning affected results. This study connected them, showing how each engineered feature influenced the model's predictions.

For example, features like $heart_stress_index$ and $age_oldpeak_interaction$ had strong positive impacts on predictions for high-risk patients. This helped explain how the model made decisions. By combining SHAP for overall explanations and LIME for specific cases, the model became more transparent and trustworthy for use in medical settings.

The model developed in this study performed much better than previous models. While earlier studies achieved accuracy between 80% and 90%, this research used a Random Forest model that reached about 98.54% accuracy. This improvement came from better feature engineering, optimized models, and using explainable AI tools.

Beyond just better numbers, this research offered a meaningful advancement. The model not only predicted heart disease risks more accurately but also explained its predictions in a way that doctors and patients could understand. This makes the model more useful for real-world medical use and addresses a major issue in previous studies: lack of clear explanations and practical relevance.

Overall, this study brought together data preparation, model tuning, and explainability into one system. This creates a more complete and trustworthy AI approach, improving both transparency and usefulness for predicting cardiovascular disease.

7.7 Interpretability Enhancement Justification

This section clarifies and reinforces how interpretability was enhanced in the proposed model using SHAP, LIME, and clinically meaningful feature engineering.

7.7.1 How Interpretability Was Enhanced

7.7.1.1 Global Interpretability Using SHAP

SHAP (SHapley Additive Explanations) was utilized to explain how each feature influences the model's predictions across the entire dataset. SHAP summary plots and global feature importance rankings highlight dominant predictors such as *oldpeak*, *thal*, and *ca*. This provides a transparent view of how the Random Forest model weighs clinical attributes, supporting trust at the global model level.

7.7.1.2 Local Interpretability Using LIME

LIME (Local Interpretable Model-Agnostic Explanations) was applied to interpret individual patient-level predictions. For each selected patient, LIME constructs a simplified local model that shows which features pushed the prediction toward *heart disease* or *no heart disease*. This supports clinician understanding of case-specific decisions rather than only aggregate model behavior.

7.7.1.3 Clinically Meaningful Feature Engineering

New derived features such as *Heart Stress Index*, *Risk Count Indicator*, and *Angina-Based Grouping* were designed based on clinical reasoning rather than arbitrary statistical combinations.

This ensures that the model's internal representation aligns with medical interpretation, enhancing transparency from the data level.

7.7.1.4 Medical Interpretation

Both SHAP and LIME explanations were linked with clinical rationale. For example, high *oldpeak* indicates ischemia severity, while low *thalach* suggests reduced cardiac output performance.

This medical grounding ensures that interpretability is meaningful and not abstract.

7.7.2 Summary of Contributions

Table 7.3: Interpretability Layers in the Proposed Heart Disease Prediction System

Interpretability Level	**Method Used**	**Contribution**
Data Representation	Feature Engineering	Converts raw clinical measurements into medically meaningful indicators.
Model-Level Explanation	SHAP	Shows feature importance and direction of influence at a global scale.
Instance-Level Explanation	LIME	Explains how individual patient predictions are formed, supporting clinician decision-making.

Table 7.4: Comparison of Random Forest Model Performance with Literature

Study / Model	**Dataset**	**Accuracy (%)**	**Precision (%)**	**F1-score (%)**
Proposed Random Forest	Cleveland Heart Disease	98.54	97.0	99
Smith et al. (2020)	Cleveland Heart Disease	94.5	92.0	93.2
Lee et al. (2019)	UCI Heart Dataset	95.2	93.5	94.3
Kumar et al. (2021)	Framingham Heart Study	96.0	94.0	95.0

TABLE 7.5: Research Questions, Key Findings, and Novel Contributions

Research Question	**Key Findings**	**Novel Contribution**
RQ1: Feature Engineering	Feature importance analysis identified key clinical predictors.	It offers a clear ranking of medical attributes that influence classification.
RQ2: Model Explainability	Model performed consistently well across training and testing datasets and achieved accuracy of 98.54%.	This confirms its ability to generalize for clinical prediction tasks.
RQ3: Trust and Clinical Applicability	Explainability methods clarified prediction reasoning and aligned with clinical knowledge.	it bridges the gap between ML model and clinical practice by maklng the prediction clinically meaningful, transparent, trustworthy for the healthcare practitioners.

7.7.3 Reinforced Interpretability Claim

The proposed system enhances interpretability on three levels: (1) data level through clinically meaningful feature engineering, (2) model-wide behavior level using SHAP, and (3) patient-specific decision explanation using LIME. Together, these components provide a transparent and clinically aligned decision-support framework for heart disease prediction.

7.8 Conclusion

This study presented a comprehensive and interpretable machine learning framework for heart disease prediction using a Random Forest classifier integrated with explainable artificial intelligence techniques. The proposed model achieved high

predictive performance across multiple evaluation metrics, demonstrating its effectiveness for reliable clinical prediction. In addition to accuracy, the research emphasized interpretability through SHAP and LIME, which provided both global and local explanations of model decisions. These explanations revealed the influence of key clinical features, enabling validation against established medical knowledge and enhancing confidence in model outputs.

The results indicate that the proposed approach not only improves predictive performance compared to existing studies but also addresses a major limitation of traditional black-box models by making decision processes transparent. Furthermore, consistent performance across training and testing datasets confirms the robustness and generalization capability of the model. Overall, this work contributes a reliable, interpretable, and clinically meaningful predictive framework that supports data-driven decision-making while maintaining transparency, trust, and real-world applicability in healthcare environments.

7.9 Future Work

Although this research successfully implemented the XAI models specifically, i.e., SHAP and LIME using a RandomForest ML model, and the results produced are also very effective, but in future, there is a need for further exploration and advancement. My future work will be focused on enhancing the transparency more by the implementation of the WIT (What-If Tool) XAI model. It allows us to better explore and understand the model behaviour, and it provides more transparency regarding patient outcomes, like whether it has a CVD or not, by answering why the model is predicting a patient to have a CVD. Thus, by doing so, it can make my proposed model more reliable and more trustworthy in clinical decision-making.

Bibliography

[1] M. I. Hossain *et al.*, "Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison," *Iranian Journal of Computer Science*, 2023.

[2] M. Parija, S. Panda, G. Panda, K. Dhama, and R. K. Mohapatra, "Risk prediction of cardiovascular disease using machine learning classifiers," *Open Medicine*, vol. 17, no. 1, pp. 1100–1113, 2022.

[3] C. Krittanawong, H. Zhang, Z. Wang, M. Aydar, and T. Kitai, "Ai-driven precision cardiovascular medicine," *American Heart Journal*, vol. 148, no. 4, pp. 67–82, 2023.

[4] J. Alizadehsani, Z. Habibi, Z. A. Sani, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, F. Khozeimeh, and F. Alizadeh-Sani, "Diagnosing coronary artery disease via data mining algorithms by considering laboratory and echocardiography features," *Research in Cardiovascular Medicine*, vol. 2, no. 3, pp. 133–139, 2021.

[5] S. J. Al'Aref *et al.*, "Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging," *European Heart Journal*, vol. 40, no. 24, pp. 1975–1986, 2019.

[6] G. Currie, K. E. Hawk, E. Rohren, A. Vial, and R. Klein, "Machine learning and deep learning in medical imaging: intelligent imaging," *Journal of Medical Imaging and Radiation Sciences*, vol. 50, no. 4, pp. 477–487, 2019.

[7] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[8] M. Batta, "Machine learning algorithms - a review," *International Journal of Science and Research (IJSR)*, 2020.

[9] M. Ribeiro, K. Grolinger, and M. A. M. Capretz, "Mlaas: Machine learning as a service," *Department of Electrical and Computer Engineering, Western University, London, Ontario, Canada*, 2019.

[10] G. S. Parikh, E. M. Navathe, and K. C. Ramakrishnan, "Machine learning in medicine: A survey," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2019.

[11] A. P. Pandit, M. A. Sengupta, P. K. Shah, A. P. Deshmukh, A. C. Mehta, and D. T. Mohan, "A novel clinical risk prediction model for sudden cardiac death in hypertrophic cardiomyopathy (hcm risk-scd)," *European Heart Journal*, vol. 35, no. 30, pp. 2010–2020, 2020.

[12] J. Burrell, "How the machine thinks: Understanding opacity in machine learning algorithms," *Big Data & Society*, vol. 3, no. 1, pp. 1–12, 2016.

[13] S. Hajian, F. Bonchi, and C. Castillo, "Algorithmic bias: From discrimination discovery to fairness-aware data mining," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019, pp. 2125–2126.

[14] G. Currie, K. E. Hawk, E. Rohren, A. Vial, and R. Klein, "Machine learning and deep learning in medical imaging: intelligent imaging," *Journal of Medical Imaging and Radiation Sciences*, vol. 50, no. 4, pp. 477–487, 2019.

[15] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto, and J. Serna, "I read but don't agree: Privacy policy benchmarking using machine learning and the eu gdpr," in *Companion of The Web Conference*, 2018, pp. 163–166.

[16] J. Burrell, "How the machine thinks: Understanding opacity in machine learning algorithms," *Big Data & Society*, vol. 3, no. 1, pp. 1–12, 2016.

[17] S. Hajian, F. Bonchi, and C. Castillo, "Algorithmic bias: From discrimination discovery to fairness-aware data mining," in *Proceedings of the 22nd*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019, pp. 2125–2126.

[18] M. Ribeiro, K. Grolinger, and M. A. M. Capretz, "Mlaas: Machine learning as a service," Department of Electrical and Computer Engineering, Western University, London, Ontario, Canada, Tech. Rep., 2019.

[19] H. Choi, D. Kim, and J. Park, "Machine learning in cardiovascular risk assessment: Advances and challenges," *Journal of Medical AI Research*, vol. 12, no. 2, pp. 45–59, 2023.

[20] S. Singh and C. Guestrin, "Why should i trust you? explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 45, 2022, pp. 1125–1138.

[21] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black-Box Models Explainable.* Leanpub, 2023.

[22] J. Lee, M. Gupta, and S. Wang, "Shap-based feature importance in cardiovascular risk prediction," *Biomed. AI J.*, vol. 9, no. 3, pp. 101–118, 2024.

[23] A. Torres and Y. Feng, "Advancements in model interpretability: A review of shap in healthcare ai," *J. Comput. Med.*, vol. 7, no. 2, pp. 55–72, 2023.

[24] A. L. G. Vicente, R. D. M. Junior, and R. A. F. Romero, "Explainable lightgbm approach for predicting myocardial infarction mortality," arXiv preprint arXiv:2404.15029, 2024.

[25] M. Ribeiro, K. Grolinger, and M. A. M. Capretz, "Mlaas: Machine learning as a service," Department of Electrical and Computer Engineering, Western University, London, Ontario, Canada, Tech. Rep., 2022.

[26] M. R. Zafar and N. Khan, "Deterministic local interpretable model-agnostic explanations for stable explainability," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 525–541, 2024.

[27] P. Hermosilla, S. Berríos, and H. Allende-Cid, "Explainable ai for forensic analysis: A comparative study of shap and lime in intrusion detection models," *Applied Sciences*, vol. 15, no. 13, p. 7329, 2025.

[28] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of lime and shap in alzheimer's disease detection," *Brain Informatics*, vol. 11, no. 1, p. 10, 2024.

[29] T. Ahmad, L. H. Lund, A. Rao, A. Ghosh, M. M. Redfield, J. Long, *et al.*, "Machine learning in cardiovascular disease: Shap and lime for feature attribution in heart failure prediction," *JACC: Heart Failure*, vol. 6, no. 7, pp. 635–643, 2018.

[30] Q. Jin, Z. Meng, and Y. Sun, "Explainable ai for diabetes prediction: A comparative study of shap and lime on clinical risk models," *Computers in Biology and Medicine*, vol. 158, p. 106679, 2023.

[31] H. Suresh and et al., "Clinical risk prediction with explainable machine learning: Interpreting shap and lime on ehr data," *Journal of Biomedical Informatics*, vol. 137, p. 104289, 2023.

[32] H. Choi, D. Kim, and J. Park, "Machine learning in cardiovascular risk assessment: Advances and challenges," *Journal of Medical AI Research*, vol. 12, no. 2, pp. 45–59, 2023.

[33] C. B. C. Latha and S. C. Jeeva, "An ai-enabled framework for transparency and interpretability in cvd risk prediction," *Computers, Materials & Continua*, vol. 74, no. 3, pp. 2145–2160, 2025.

[34] P. Guleria, P. N. Srinivasu, S. Ahmed, N. Almusallam, and F. K. Alarfaj, "Xai framework for cardiovascular disease prediction using classification techniques," *Electronics*, vol. 11, no. 24, p. 4086, 2022.

[35] G. Yang, Q. Ye, and J. Xia, "Unbox the black box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases, and beyond," *Information Fusion*, 2023.

[36] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," arXiv preprint arXiv:1703.07314, 2020.

[37] R. Alizadehsani, S. S. Oyelere, S. Hussain, S. K. Jagatheesaperumal, R. R. Calixto, M. Rahouti, and V. H. C. D. Albuquerque, "Explainable artificial intelligence for drug discovery and development—a comprehensive survey," *IEEE Access*, 2024.

[38] M. Ariza, J. Arroyo, A. Capparrini, and M. J. Segovia, "Explainability of a machine learning granting scoring model in peer-to-peer lending," Santander-UCM research project, call 2019, reference PR87/19-22586, 2019.

[39] A. Smith and R. Kumar, "Robust, privacy-preserving xai for cardiovascular disease detection," *Journal of Biomedical Informatics*, vol. 138, p. 104557, 2025, [Online]. Available: https://doi.org/10.1016/j.jbi.2025.104557.

[40] M. A. Umar, N. AbuAli, K. Shuaib, and A. I. Awad, "An explainable artificial intelligence and internet of things framework for monitoring and predicting cardiovascular disease," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 107388, Jan. 2025, [Online]. Available: https://doi.org/10.1016/j.engappai.2025.107388.

[41] S. P. Patro and N. Padhy, "A secure remote health monitoring for heart disease prediction using machine learning and deep learning techniques in explainable artificial intelligence framework," in *Engineering Proceedings*, vol. 58, no. 1, 2023, pp. 78–78.

[42] D. Dua and C. Graff, "Uci machine learning repository," 2017, irvine, CA: University of California, School of Information and Computer Sciences. Available at: http://archive.ics.uci.edu/ml.

[43] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications*. Springer, 2022.

[44] J. Brownlee, *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*. Machine Learning Mastery, 2020.

[45] E. J. Benjamin *et al.*, "Heart disease and stroke statistics—2023 update: A report from the american heart association," *Circulation*, vol. 147, no. 8, pp. e93–e621, 2023.

[46] World Health Organization, "Hypertension," 2021, fact Sheet No. 318. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hypertension.

[47] M. T. Ribeiro, S. Singh, and C. Guestrin, "'why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, 2022, pp. 1135–1144.