

# **The Egyptian E-Learning University**

## ***Faculty of Computers and Information Technology***

### **Predicting Heart Disease Diagnosis Using AI/ML**

#### **By**

|  |  |
| --- | --- |
| **Khalid Elshawadfy Ahmed** | **2002159** |
| **Ammar Mohamed Hassan** | **2002163** |
| **Abdullah Ahmed Abdellazim** | **2002202** |
| **Hossam Hassan Mohamed** | **2002158** |
| **Mohamed Elsayed Abdelsamad** | **2002198** |
| **Omar Abdelrhaman Yousef** | **2000154** |

# **Supervised by**

## ***Prof. Ahmed Ezz***

### ***DR. Yasmin Mahmoud***

#### **Eng. Nourhan Salah**

# **Acknowledgement**

First and foremost, I would like to express my deepest gratitude to **Prof. Ahmed Ezz (may R.I.P)**, whose visionary guidance and academic mentorship laid the foundation for this project. Although he is no longer with us, his contributions to this work and his unwavering dedication to the field of computer science continue to inspire us. His legacy lives on through the students and projects he supported. we would also like to sincerely thank **Dr. Yasmine Mahmoud** for her invaluable support, continuous encouragement, and expert insights throughout the duration of this project. Her constructive feedback and detailed reviews greatly enhanced the quality and direction of this research. Special appreciation goes to **Eng. Nourhan Salah** for her consistent technical guidance, practical suggestions, and for being available whenever I needed assistance. Her mentorship helped bridge the gap between theory and real-world implementation. Finally, I am grateful to my family and friends for their unwavering support, patience, and motivation, especially during the challenging phases of this project. Their belief in me has been a constant source of strength. To all who contributed to this journey, directly or indirectly, I offer my heartfelt thanks.

# **Abstract**

**Heart disease** remains one of the leading causes of death globally, emphasizing the urgent need for effective early diagnostic tools. Traditional diagnostic methods, while accurate, are often time-consuming and dependent on the availability of specialized healthcare professionals. In recent years, machine learning (ML) has emerged as a powerful tool for analyzing clinical data and assisting in medical diagnosis. This project proposes a data-driven approach to heart disease prediction using artificial intelligence and machine learning techniques. The study utilizes the UCI Cleveland Heart Disease dataset, which includes 13 clinical features collected from real patients. A systematic methodology is followed, including data preprocessing, feature analysis, model training, and evaluation. Four supervised machine learning models were implemented and compared: Logistic Regression, Support Vector Machine (SVM), Random Forest, and an Artificial Neural Network (ANN). The models were evaluated using various performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Analysis was also conducted to identify the most influential clinical factors. The best-performing model was then deployed via a Flask-based web application, allowing users to input clinical values and receive real-time predictions. The application was containerized using Docker and deployed on the Azure Cloud Platform, making it accessible and scalable. The results demonstrate that machine learning models, when properly trained and validated, can offer significant value as decision-support tools in medical diagnostics. This project provides a practical framework for integrating AI into healthcare applications and sets the stage for future enhancements using larger datasets and real-time patient monitoring systems.

# **Keywords**

Heart Disease, Artificial Intelligence (AI), Machine Learning (ML), UCI Cleveland Dataset, Logistic Regression, Support Vector Machine (SVM), Random Forest, Artificial Neural Network (ANN), Data Preprocessing, Model Evaluation Metrics, Flask, Docker, Cloud Deployment

**Heart Disease**

A range of cardiovascular conditions affecting the heart’s structure and function. This project specifically addresses the prediction of coronary artery disease by analyzing clinical features from patient data

**Artificial Intelligence (AI)**

A broad domain of computer science concerned with building systems capable of performing tasks that normally require human intelligence. In this project, AI enables automatic decision-making based on learned patterns from healthcare data.

**Machine Learning (ML)**

A subset of AI that focuses on algorithms capable of learning from historical data to make predictions. ML models used in this project classify patients based on whether or not they are likely to have heart disease.

**UCI Cleveland Dataset**

A publicly available dataset from the UCI Machine Learning Repository. It contains medical records of patients, including 13 key features such as chest pain type, cholesterol levels, and blood pressure, used to predict heart disease.

**Logistic Regression**

A statistical method used for binary classification. It estimates the probability of a patient having heart disease based on clinical features.

**Support Vector Machine (SVM)**

A supervised machine learning algorithm that finds the optimal boundary between classes. In this project, it is used to separate patients with and without heart disease based on feature vectors.

**Random Forest**

An ensemble learning algorithm based on decision trees. It improves classification accuracy by aggregating the results of multiple decision trees and provides insight into the importance of each feature.

**Artificial Neural Network (ANN)**

A biologically inspired computing model consisting of interconnected layers of nodes. ANNs are used in this study to model complex, non-linear relationships in the dataset for improved prediction accuracy.

**Data Preprocessing**

The process of transforming raw data into a usable format. It includes handling missing values, feature encoding, normalization, and dataset splitting. Preprocessing ensures that the models receive clean and meaningful inputs.

**Model Evaluation Metrics**

Quantitative metrics used to assess model performance. This project employs accuracy, precision, recall, F1-score, and ROC-AUC to evaluate the effectiveness of each machine learning model.

**Flask**

A lightweight Python web framework used to build a user interface for heart disease prediction. It enables real-time interaction between the end user and the machine learning model.

**Docker**

A platform for developing, shipping, and running applications in isolated environments. Docker is used to containerize the project application, ensuring platform independence and easy deployment.

**Cloud Deployment**

The process of hosting and delivering the application via cloud services (e.g., Microsoft Azure). This enhances accessibility, scalability, and allowing users to interact with the prediction system online.

**Table of Contents**

[Acknowledgement 2](#_Toc200388284)

[Abstract 3](#_Toc200388285)

[Keywords 4](#_Toc200388286)

[Table of Figures 9](#_Toc200388287)

[Chapter One 11](#_Toc200388288)

[**Introduction** 11](#_Toc200388289)

[**1.1** **Introduction** 11](#_Toc200388290)

[**1.2 Background and Motivation for the Project** 12](#_Toc200388291)

[**1.3 Importance of the Problem Being Addressed** 12](#_Toc200388292)

[**1.4 Problem Statement** 13](#_Toc200388293)

[**1.5 Objectives** 13](#_Toc200388294)

[**1.6 Brief Overview of the Proposed Solution** 14](#_Toc200388295)

[Chapter Two 15](#_Toc200388296)

[**Literature Review / Related Work** 15](#_Toc200388297)

[**2.1 Summary of Existing Research and Technologies** 15](#_Toc200388298)

[**2.2 Gaps in Current Solutions** 20](#_Toc200388299)

[**2.3 Summary of Related Work and Project Positioning** 20](#_Toc200388300)

[Chapter Three 21](#_Toc200388301)

[**Proposed System** 21](#_Toc200388302)

[**3.1 Approach Used to Solve the Problem** 21](#_Toc200388303)

[**3.2 System Architecture** 22](#_Toc200388304)

[**3.3 Algorithms and Frameworks Used** 25](#_Toc200388305)

[Chapter Four 26](#_Toc200388306)

[**Implementation** 26](#_Toc200388307)

[**4.1 Technologies, Tools, and Programming Languages Used** 26](#_Toc200388308)

[**4.2 Key Components/Modules of the System** 27](#_Toc200388309)

[**4.3 Challenges Faced and How They Were Resolved** 28](#_Toc200388310)

[**4.4 Source Code and Deployment URL** 29](#_Toc200388311)

[**4.5 Selected Code Snippets** 30](#_Toc200388312)

[Chapter Five 34](#_Toc200388313)

[**Testing & Evaluation** 34](#_Toc200388314)

[**5.1 Introduction** 34](#_Toc200388315)

[**5.2 Model Performance Metrics** 34](#_Toc200388316)

[**5.3 Confusion Matrices** 36](#_Toc200388317)

[**5.4 Feature Importance Analysis** 38](#_Toc200388318)

[**5.5 ROC Curve Analysis** 39](#_Toc200388319)

[**5.5 Correlation Analysis** 40](#_Toc200388320)

[**5.6 Comparison with Existing Studies** 41](#_Toc200388321)

[**5.7 Scalability and Inference Speed** 41](#_Toc200388322)

[**5.8 Summary** 41](#_Toc200388323)

[Chapter six 42](#_Toc200388324)

[**Results & Discussion** 42](#_Toc200388325)

[**6.1 Introduction** 42](#_Toc200388326)

[**6.2 Summary of Findings** 42](#_Toc200388327)

[**6.3 Interpretation of Results** 44](#_Toc200388328)

[**6.4 Limitations of the Proposed Solution** 45](#_Toc200388329)

[**6.5 Broader Impact and Implications** 46](#_Toc200388330)

[**6.6 Summary** 46](#_Toc200388331)

[Chapter Seven 47](#_Toc200388332)

[**Conclusion and Future Work** 47](#_Toc200388333)

[**7.1** **Summary of Contributions** 47](#_Toc200388334)

[**7.2** **Possible Improvements or Extensions for Future Work** 48](#_Toc200388335)

[**7.3** **Final Remarks** 49](#_Toc200388336)

[References 50](#_Toc200388337)

# **Table of Figures**

[**Figure 1. Global CVD mortality data** 10](#_Toc200388225)

[**Figure 2. Figure 2. ML Deployment Pipeline** 13](#_Toc200388226)

[**Figure. 3 Datasets Comparison** 14](#_Toc200388227)

[**Figure 4. heart disease presence** 16](#_Toc200388228)

[**Figure 5. Datasets Comparison** 17](#_Toc200388229)

[**Figure 6. Comparison of Machine Learning Models** 17](#_Toc200388230)

[**Figure 7. Dataset Feature Summary** 18](#_Toc200388231)

[**Figure 8. System Flow Diagram** 21](#_Toc200388232)

[**Figure 9. System Flow Diagram** 22](#_Toc200388233)

[**Figure 10. UML Component Diagram** 22](#_Toc200388234)

[**Figure 11. Sequence Diagram** 23](#_Toc200388235)

[**Figure 12. Machine Learning Algorithms** 24](#_Toc200388236)

[**Figure 13. web app test01** 28](#_Toc200388237)

[**Figure 14. Important libraries** 29](#_Toc200388238)

[**Figure 15. Data Cleaning and Preparation** 29](#_Toc200388239)

[**Figure 16. Train/Test Splitting** 30](#_Toc200388240)

[**Figure 17. Random Forest Classifier** 30](#_Toc200388241)

[**Figure 18. Logistic Regression** 31](#_Toc200388242)

[**Figure 19. Support Vector Machine (SVM)** 31](#_Toc200388243)

[**Figure 20. Multi-Layer Perceptron (ANN)** 32](#_Toc200388244)

[**Figure 21. Model Performance Comparison** 34](#_Toc200388245)

[**Figure 22. Random Forest classification report** 34](#_Toc200388246)

[**Figure 23. 1 Random Forest CM** 35](#_Toc200388247)

[**Figure 24. Logistic Regression CM** 35](#_Toc200388248)

[**Figure 25. SVM CM** 36](#_Toc200388249)

[**Figure 26. Figure 25. 4 MLP Classifier CM** 36](#_Toc200388250)

[**Figure 27. 4 Feature Importance Analysis** 37](#_Toc200388251)

[**Figure 28. ROC Curve Analysis** 38](#_Toc200388252)

[**Figure 29. Correlation Analysis** 39](#_Toc200388253)

[**Figure 30. Existing Studies** 40](#_Toc200388254)

[**Figure 31. web interface** 42](#_Toc200388255)

[**Figure 32. Project Objectives** 43](#_Toc200388256)

# **Chapter One**

## **Introduction**

### **1.1** **Introduction**

A poster of a health care campaign

AI-generated content may be incorrect.**Heart disease** remains one of the leading causes of mortality globally. According to the World Health Organization (WHO), cardiovascular diseases account for an estimated 17.9 million deaths annually, highlighting a growing need for early diagnosis and prevention. Traditional diagnostic methods, although effective, often involve invasive procedures, are time-consuming, and require extensive expertise. In response to these limitations, technological advancements in Artificial Intelligence (AI) and Machine Learning (ML) have emerged as promising tools for enhancing medical diagnostics. This project explores the application of AI/ML techniques to predict the presence of heart disease using a dataset derived from clinical parameters. The system is designed to assist healthcare professionals by providing an efficient and accurate decision-support tool that reduces human error and promotes early detection. By leveraging machine learning algorithms and historical medical data, the goal is to create a model capable of identifying patterns that indicate heart disease risk, ultimately contributing to improved patient outcomes and reduced healthcare costs.

**Figure 1. Global CVD mortality data**

### **1.2 Background and Motivation for the Project**

The rising prevalence of heart disease in both developed and developing countries emphasizes the urgent need for better diagnostic strategies. Traditional diagnostic approaches, such as electrocardiograms, echocardiography, and angiography, are often limited by accessibility, cost, and the requirement for specialist interpretation. Meanwhile, the healthcare industry is increasingly data-rich, with large volumes of patient information stored digitally. The motivation for this project stems from the growing potential of AI and ML to revolutionize the field of medical diagnostics. By analyzing patterns in existing patient records, machine learning models can predict disease outcomes with high accuracy.

This capability can significantly enhance preventive care and allow for timely intervention. Furthermore, the project is driven by the opportunity to contribute to society by reducing preventable deaths through early risk identification and to explore how data science can be applied to real-world healthcare challenges.

### **1.3 Importance of the Problem Being Addressed**

Heart disease is not only a medical concern but also a socio-economic burden, particularly in countries with limited access to healthcare infrastructure. Misdiagnosis or delayed diagnosis can lead to critical health consequences, hospitalizations, and death. Early detection remains key to preventing such outcomes. However, the lack of accessible, affordable, and intelligent diagnostic tools, especially in under-resourced regions, continues to hinder progress. The ability to accurately predict heart disease using non-invasive, easy-to-collect parameters offers a compelling solution. By automating parts of the diagnostic process, healthcare systems can become more efficient, reduce workload on professionals, and offer consistent evaluations. This project addresses a globally significant issue by exploring how AI/ML technologies can bridge gaps in traditional healthcare delivery and enhance the accuracy of disease prediction.

### **1.4 Problem Statement**

**Clear definition:**  
Despite advancements in medical technology, timely and accurate prediction of heart disease remains a challenge due to factors such as data complexity, delayed diagnosis, and reliance on specialist interpretation. The project addresses the problem of developing an efficient, scalable, and user-friendly machine learning-based system capable of predicting the presence of heart disease using clinical data.

**Justification:**  
Heart disease prediction based on clinical observations is traditionally manual, subjective, and limited in scalability. Given the availability of patient health datasets and advancements in ML, there is a critical opportunity to automate and improve this process. The development of an AI/ML-based tool can aid physicians in decision-making, reduce diagnostic errors, and provide accessible risk evaluation. This makes the problem not only relevant but essential in today’s medical landscape.

### **1.5 Objectives**

**Main Objective:**  
 To develop a machine learning-based diagnostic system that can accurately

predict the likelihood of heart disease in patients using clinical data.

**Specific Objectives:**

* To collect and preprocess the UCI Cleveland heart disease dataset, ensuring it is suitable for machine learning applications.
* To implement and compare multiple ML models (e.g., Logistic Regression, SVM, Random Forest, ANN) for predictive performance.
* To evaluate model accuracy using metrics such as precision, recall, F1-score, and ROC-AUC.
* To build a simple, interactive web interface using Flask where users can input patient data and receive predictions.
* To deploy the application using Docker and cloud technologies for real-time access.
* To visualize and interpret feature importance and decision patterns that contribute to predictions.

### **1.6 Brief Overview of the Proposed Solution**

The proposed solution involves designing a machine learning model trained on historical patient data to predict heart disease presence. The system uses structured clinical features from the UCI Cleveland dataset, such as age, cholesterol levels, chest pain type, and more. After data preprocessing and exploratory analysis, various ML algorithms will be trained, tuned, and evaluated. The best-performing model will then be integrated into a Flask-based web application that allows users (e.g., healthcare workers or researchers) to input patient parameters and receive immediate diagnostic predictions. The entire application will be containerized using Docker and deployed to a cloud platform to ensure scalability and accessibility. This approach combines the power of data science and web technologies to offer a reliable, interpretable, and scalable tool for early heart disease detection.

A close-up of a computer

AI-generated content may be incorrect.

**Figure 2. Figure 2. ML Deployment Pipeline**

# **Chapter Two**

## **Literature Review / Related Work**

### **2.1 Summary of Existing Research and Technologies**

#### **2.1.1 Applications of Machine Learning in Heart Disease Diagnosis**

Recent research has shown that artificial intelligence (AI) and machine learning (ML) models can effectively predict heart disease using structured clinical data. Techniques such as Logistic Regression, Support Vector Machines (SVM), Random Forest, and Neural Networks have been widely adopted in this context. These models can process multiple patient attributes simultaneously and identify patterns that may indicate heart disease risk, often with higher efficiency and consistency than traditional diagnostic methods.

#### **2.1.2 Datasets Utilized in Heart Disease Prediction**

Multiple datasets are commonly used in cardiovascular ML research:

A screenshot of a computer

AI-generated content may be incorrect.

**Figure. 3 Datasets Comparison**

The **UCI Cleveland dataset** was chosen for this project due to its accessibility, structure, and popularity in published studies, which makes performance comparisons more straightforward and academically valid.

#### **2.1.3 Key Features Identified by Previous Research**

Numerous research efforts have identified a common set of important clinical features. These include:

* **ST\_Slope (Up/Flat)** – shown to be the most predictive of heart disease.
* **Oldpeak** – indicating ST depression during exercise.
* **Chest Pain Type (especially ASY)** – frequently associated with heart disease.
* **Maximum Heart Rate (MaxHR)** – lower MaxHR can indicate poor heart function.
* **Cholesterol & Fasting Blood Sugar** – metabolic indicators.

These findings are consistent with insights from our exploratory data analysis phase.

**ST\_Slope**:

* The *Flat* slope is strongly associated with patients who have heart disease.
* The *Up* slope is more common among healthy individuals.

**Oldpeak (ST Depression during Exercise)**:

* Higher oldpeak values are more prevalent in patients with heart disease.
* A significant number of patients with no disease have near-zero oldpeak.

**Chest Pain Type**:

* *ASY* (asymptomatic) chest pain is most commonly associated with heart disease.
* *ATA* and *NAP* are more prevalent among healthy individuals.

**Maximum Heart Rate (MaxHR)**:

* Individuals with heart disease tend to have a lower MaxHR.
* Healthy individuals often reach higher heart rates during physical activity.

**Cholesterol**:

* Higher cholesterol levels are slightly more common in heart disease cases.
* However, the separation is less pronounced compared to other features.

**Fasting Blood Sugar (FBS)**:

* A fasting blood sugar >120 mg/dl appears more frequently in heart disease cases.
* Still, the majority of all patients fall into the ≤120 mg/dl category.

A group of graphs with different colored bars

AI-generated content may be incorrect.

**Figure 4. heart disease presence**

#### **2.1.4 Comparative Analysis of Heart Disease Datasets**

In machine learning, dataset quality significantly influences model performance. Below is a high-level comparison of the most commonly used heart disease datasets:

A screenshot of a computer screen

AI-generated content may be incorrect.

**Figure 5. Datasets Comparison**

The UCI Cleveland dataset offers a balanced choice for academic and clinical research due to its well-structured format and widespread adoption.

#### **2.1.5 Comparative Analysis of Machine Learning Models**

A screenshot of a computer

AI-generated content may be incorrect.The table below summarizes key ML algorithms used in heart disease prediction:

**Figure 6. Comparison of Machine Learning Models**

Based on these findings, this project focuses on Random Forest and Neural Networks due to their balance of accuracy and adaptability. A performance comparison of these models will be included during the evaluation phase.

***2.1.6 Overview of the Dataset Used in This Project***

The dataset used in this project is derived from the UCI Cleveland Heart Disease dataset, which includes 918 patient records and 12 clinical features relevant to cardiovascular health. It is publicly available and widely used in academic and machine learning research, making it ideal for benchmarking and performance comparison.

This dataset includes a mix of categorical and numerical features, which reflect real-world patient attributes that are typically collected during clinical assessments. The target variable, HeartDisease, is a binary indicator representing the presence (1) or absence (0) of heart disease.

A screenshot of a medical report

AI-generated content may be incorrect.

**Figure 7. Dataset Feature Summary**

**Key Characteristics:**

* **Total Records**: 918
* **Target Feature**: HeartDisease

### **2.2 Gaps in Current Solutions**

#### **2.2.1 Model Deployment and Clinical Integration**

Most studies stop at model training and validation in experimental environments. There is often no effort to deploy these models in clinical tools or real-time systems accessible to practitioners.

#### **2.2.2 Interpretability Challenges**

Highly accurate models like deep neural networks lack interpretability, making them less suitable for use in clinical decision-making without additional explanation tools.

#### **2.2.3 Dataset Size and Generalizability**

Many studies rely on small datasets like UCI Cleveland, which, although effective for prototyping, do not generalize well to larger, more diverse populations.

#### **2.2.4 Limited Prescriptive Analytics**

Few models go beyond prediction to offer prescriptive actions or treatment guidance based on the diagnosis.

### **2.3 Summary of Related Work and Project Positioning**

In conclusion, current research highlights the growing role of ML in heart disease diagnosis. However, most studies fall short in terms of usability, interpretability, and deployment. This project bridges these gaps by:

* Using **popular and validated datasets** for reliable training and benchmarking.
* Selecting **high-performing models** based on comparative studies.
* Building a **cloud-deployed, containerized Flask application** for real-time use.
* Providing **feature insights** and visual feedback to support clinicians.
* Laying the groundwork for future integration of **prescriptive analytics**.

This multi-faceted approach ensures the solution is not only accurate but also accessible, understandable, and clinically relevant.

# **Chapter Three**

## **Proposed System**

### **3.1 Approach Used to Solve the Problem**

The proposed system aims to accurately predict the presence of heart disease using machine learning techniques applied to structured clinical data. The primary goal is to assist healthcare professionals in making early, data-driven decisions.

The approach follows these structured steps:

1. **Data Understanding**  
   The dataset comprises 918 patient records from the UCI Cleveland Heart Disease dataset, including features such as age, resting blood pressure, cholesterol, and chest pain type.
2. **Data Preprocessing**  
   Categorical data such as Sex, ChestPainType, and RestingECG are encoded numerically. Normalization is applied to continuous features to improve model performance.
3. **Exploratory Data Analysis (EDA)**  
   Visualizations (e.g., histograms, count plots) are used to examine patterns between features and heart disease. Notable findings include:

* Flat ST\_Slope, high Oldpeak, and ChestPainType = ASY are common in heart disease cases.
* Healthy patients generally have higher MaxHR.

1. **Model Development**  
   Various ML models are trained on the data including Logistic Regression, Random Forest, Support Vector Machines (SVM), and Neural Networks.
2. **Model Evaluation**  
   Models are evaluated using standard classification metrics: accuracy, precision, recall, F1-score, and ROC-AUC. Cross-validation is applied for robustness.
3. **Model Selection**  
   Based on evaluation, the Random Forest model is selected due to its high accuracy and interpretability.
4. **Prediction Output**  
   The final model outputs a binary prediction: 1 (presence of heart disease) or 0 (no heart disease), along with insight into important features.

### **3.2 System Architecture**

The architecture of the proposed system is designed for modularity, scalability, and clarity. It defines the flow from data ingestion to model prediction.

#### **A diagram of a model AI-generated content may be incorrect.3.2.1 System Flow Diagram**

**Figure 8. System Flow Diagram**

**Figure 9. System Flow Diagram**

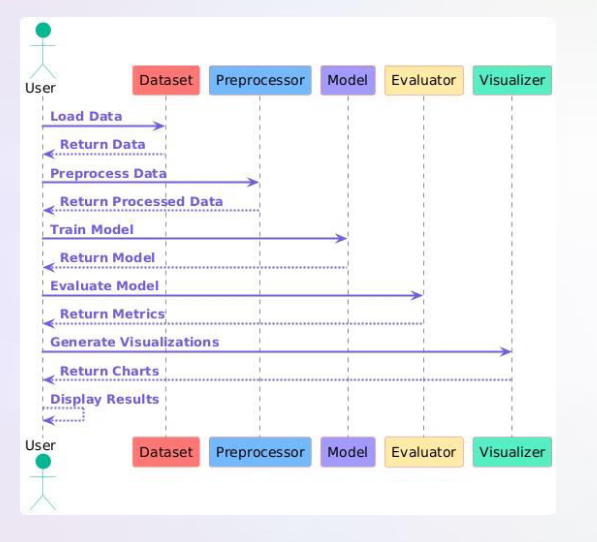
#### **3.2.2 UML Component Diagram (Conceptual)**

A diagram of a process

AI-generated content may be incorrect.

**Figure 10. UML Component Diagram**

#### **3.2.3 Sequence Diagram**

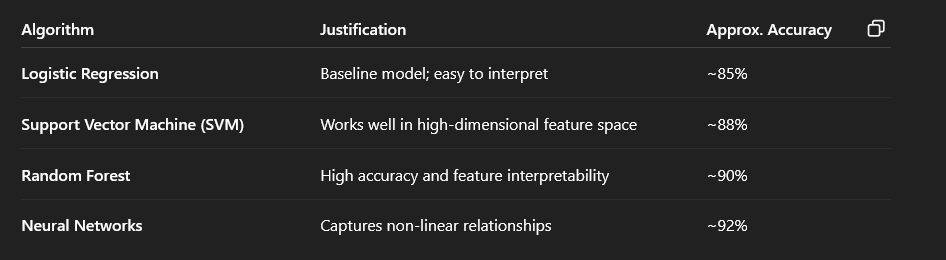


**Figure 11. Sequence Diagram**

### **3.3 Algorithms and Frameworks Used**

This section details the core algorithms and frameworks used for data modeling and decision-making.

#### **3.3.1 Machine Learning Algorithms**

****

**Figure 12. Machine Learning Algorithms**

→ **Final Selection**: Random Forest was chosen due to its balance of performance and interpretability.

#### **3.3.2 Analytical Layers Used**

As outlined in the project plan (Four-Tiered Analytics):

* **Descriptive Analytics**: Initial exploration and summaries (mean, counts, distributions).
* **Diagnostic Analytics**: Correlation and feature relationships.
* **Predictive Analytics**: Model-based future prediction.
* **Prescriptive Analytics**: Insight into which features impact predictions most (e.g., SHAP values or feature importance).

# **Chapter Four**

## **Implementation**

### **4.1 Technologies, Tools, and Programming Languages Used**

The implementation of the proposed system integrates a combination of machine learning, web development, and deployment technologies to build a functional and scalable diagnostic tool.

#### **4.1.1 Programming Language**

* **Python**  
  Used for all machine learning processes including data preprocessing, model training, and backend development via Flask.
* **HTML**  
  Used to build the structure of the web interface where users input clinical data.
* **Tailwind CSS**  
  A utility-first CSS framework used for styling the frontend. It allows for rapid prototyping and modern, responsive UI design without writing custom CSS.

#### **4.1.2 Libraries and Frameworks**

* **Pandas, NumPy** – Data preprocessing and statistical operations
* **Matplotlib, Seaborn** – Data visualization for analysis and evaluation
* **Scikit-learn** – Model building and evaluation (Random Forest, Logistic Regression, SVM, etc.)
* **Flask** – Backend framework to expose the trained ML model via a web interface
* **Joblib** – Saving and loading the trained model for deployment

#### **4.1.3 Development and Deployment Tools**

* **Google Colab**  
  Used during the development and training phase for running notebooks and testing model performance with GPU support.
* **Docker**  
  Used to containerize the Flask application, ensuring environment consistency and enabling portability across systems.
* **Microsoft Azure**  
  Cloud platform used for hosting the deployed model.
* **Azure Container Registry (ACR)**  
  Used to store the Docker image of the application securely within the Azure ecosystem.
* **Azure Container Instances (ACI)**  
  Used to deploy and run the containerized model in the cloud. It allows the application to be accessible via a public URL.

### **4.2 Key Components/Modules of the System**

The project is built with modularity in mind, allowing individual sections to be maintained, tested, and deployed independently.

#### **4.2.1 Data Preprocessing Module**

* Encodes categorical features
* Normalizes continuous variables
* Validates and cleans anomalous entries (e.g., 0 blood pressure)

#### **4.2.2 Model Training and Evaluation**

* Trains multiple ML models (Logistic Regression, SVM, Random Forest, ANN)
* Applies cross-validation
* Selects the best model using metrics like accuracy, precision, recall, and F1-score

#### **4.2.3 Prediction Interface**

* Flask-based web form collects patient inputs
* Sends inputs to the trained model and returns prediction
* Provides immediate result: “Heart Disease Detected” or “No Heart Disease”

#### **4.2.4 Visualization Module**

* Provides insight into data distributions
* Displays confusion matrix and feature importance (for explainability)

#### **4.2.5 Deployment Module**

* Docker container encapsulates entire app
* Azure cloud deployment ensures accessibility and scalability

### **4.3 Challenges Faced and How They Were Resolved**

#### **4.3.1 Data Quality and Preprocessing**

* **Issue**: Some numeric features had biologically invalid values (e.g., 0 cholesterol)
* **Solution**: Such entries were identified and cleaned; features were standardized for model compatibility

#### **4.3.2 Model Interpretability**

* **Issue**: Neural networks offered high accuracy but lacked interpretability
* **Solution**: Random Forest was selected due to its interpretability and competitive performance

#### **4.3.3 Form Input Handling in Flask**

* **Issue**: Web inputs needed to match the model’s expected data structure
* **Solution**: Form data is reshaped and preprocessed before being passed to the model

#### **4.3.4 Environment Mismatch during Deployment**

* **Issue**: Differences between local and cloud environments caused dependency errors
* **Solution**: Docker was used to standardize environments, ensuring portability across systems

### **4.4 Source Code and Deployment URL**

To promote transparency, reproducibility, and usability, the codebase and working application are made available publicly.

#### **4.4.1 GitHub Repository**

All source files including training notebooks, Flask app code, and deployment scripts are hosted on GitHub:

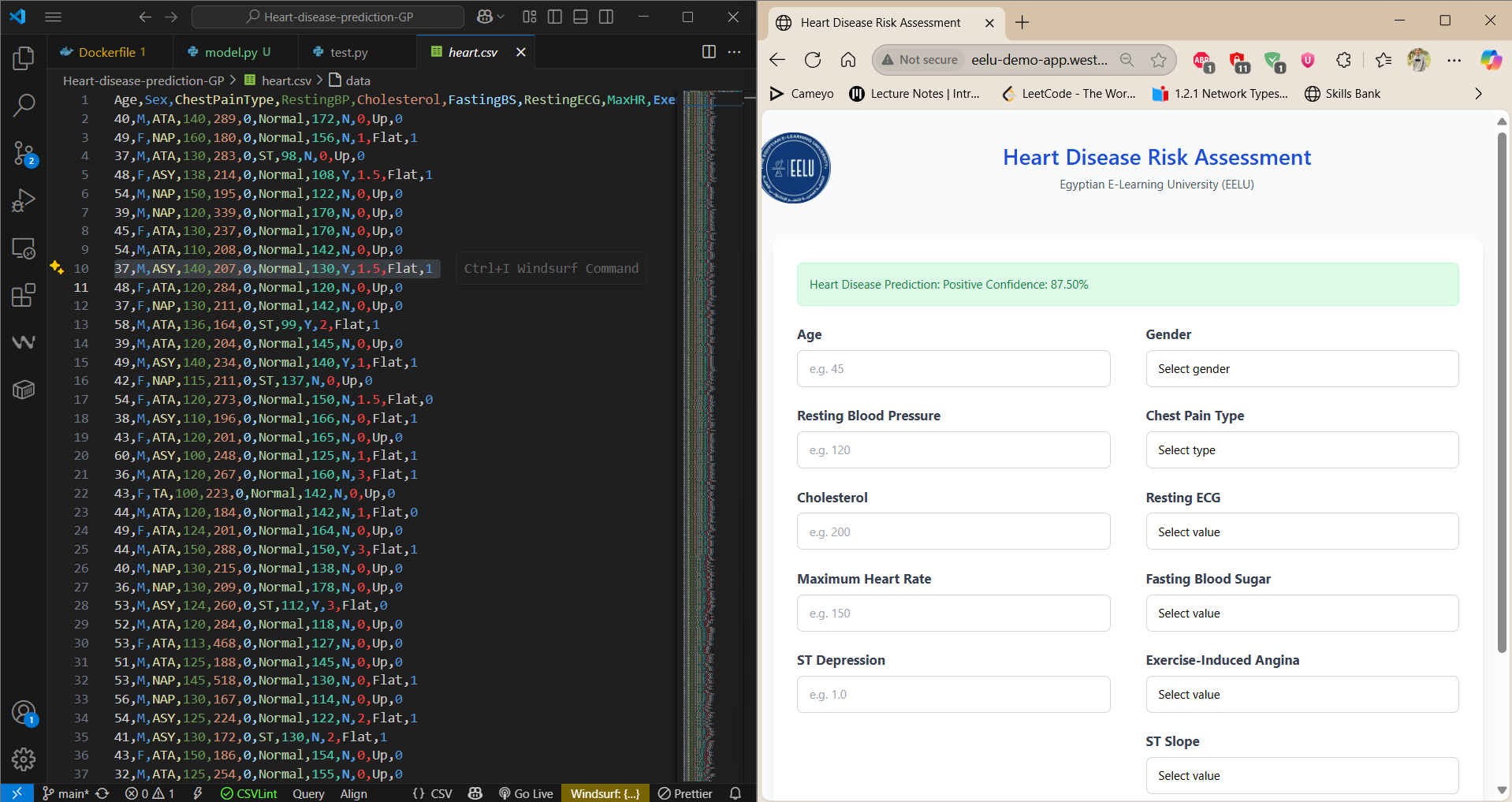
* 🔗 **GitHub Repository**: https://github.com/yatara21/Heart-disease-prediction-GP

#### **4.4.2 Live Web Application (Azure)**

The deployed version of the model is accessible through a browser interface:

* 🌐 **Azure App URL**: <http://eelu-demo-app.westeurope.azurecontainer.io>

Users can input patient parameters and receive real-time heart disease predictions.



**Figure 13. web app test01**

### **4.5 Selected Code Snippets**

This section includes key Python code excerpts that form the core logic of the system.

#### A screenshot of a computer program AI-generated content may be incorrect.**4.5.1 Importing libraries and modules**

**Figure 14. Important libraries**

#### **4.5.2 Data Cleaning and Preparation**

A screen shot of a computer program

AI-generated content may be incorrect.

**Figure 15. Data Cleaning and Preparation**

#### **4.5.3 Train/Test Splitting**

A computer screen shot of a black box with white text

AI-generated content may be incorrect.

**Figure 16. Train/Test Splitting**

#### **4.5.4 Model Training and Evaluation**

The system evaluates four ML models: Random Forest, Logistic Regression, SVM, and MLP. Below is the cleaned, final implementation:

🔹 **Random Forest Classifier (Final Model)**

A screen shot of a computer program

AI-generated content may be incorrect.

**Figure 17. Random Forest Classifier**

🔹 **Logistic Regression**

A screen shot of a computer program

AI-generated content may be incorrect.

**Figure 18. Logistic Regression**

🔹 **Support Vector Machine (SVM)**

A screen shot of a computer program

AI-generated content may be incorrect.

**Figure 19. Support Vector Machine (SVM)**

🔹 **Multi-Layer Perceptron (ANN)**

A screen shot of a computer code

AI-generated content may be incorrect.

**Figure 20. Multi-Layer Perceptron (ANN)**

# **Chapter Five**

## **Testing & Evaluation**

### **5.1 Introduction**

This chapter presents the systematic evaluation of machine learning models trained for heart disease prediction. The objective is to assess each model's ability to accurately classify patients as either having or not having heart disease based on clinical input features. Metrics such as accuracy, precision, recall, F1-score, and ROC-AUC are employed. The models evaluated include:

* Logistic Regression
* Support Vector Machine (SVM)
* Random Forest
* Artificial Neural Network (ANN)

### **5.2 Model Performance Metrics**

#### **5.2.1 Evaluation Criteria**

To ensure fair model comparison, the following metrics were used:

* **Accuracy** – Proportion of total correct predictions.
* **Precision** – Proportion of true positives among predicted positives.
* **Recall (Sensitivity)** – Proportion of true positive s detected among all actual positives.
* **F1-Score** – Harmonic mean of precision and recall.
* **ROC-AUC** – Measures the area under the Receiver Operating Characteristic curve, indicating classification performance.

#### **5.2.2 Model Testing Procedure**

Each model was evaluated using:

* **80/20 train-test split**
* **Stratified sampling** to maintain class balance
* **Cross-validation** (e.g., 5-fold) for stability
* **Confusion matrix** analysis

#### **5.2.3 Model Performance Comparison**

A group of colored bars

AI-generated content may be incorrect.

**Figure 21. Model Performance Comparison**

✅ **Best Performer:** **Random Forest** due to high interpretability, balanced accuracy, and robustness.

A screenshot of a black screen

AI-generated content may be incorrect.

**Figure 22. Random Forest classification report**

### **5.3 Confusion Matrices**

#### A blue squares with white text AI-generated content may be incorrect.**5.3.1 Random Forest**

**Figure 23. 1 Random Forest CM**

#### **5.3.2 Logistic Regression**

**Figure 24. Logistic Regression CM**

A green squares with white text

AI-generated content may be incorrect.

#### A diagram of a confusion matrix AI-generated content may be incorrect.**5.3.3 Support Vector Machine**

**Figure 25. SVM CM**

***A graph of a number of blue squares

AI-generated content may be incorrect.5.3.4 MLP Classifier***

**Figure 26. Figure 25. 4 MLP Classifier CM**

### **5.4 Feature Importance Analysis**

A graph of different colored bars

AI-generated content may be incorrect.

**Figure 27. 4 Feature Importance Analysis**

The feature importance chart reveals that **ST\_Slope**, **Oldpeak**, and **Cholesterol** are the most influential predictors in determining the presence of heart disease, as derived from the Random Forest model. These features correspond closely with well-established clinical indicators, reinforcing the model’s medical validity. The prominence of ST\_Slope and Oldpeak suggests a strong link to electrocardiographic and stress test responses, which are critical in cardiology. The clear separation of these top features from the rest also enhances model interpretability, making it easier for healthcare professionals to trust and understand the reasoning behind the predictions.

### A graph of a graph AI-generated content may be incorrect.**5.5 ROC Curve Analysis**

**Figure 28. ROC Curve Analysis**

The ROC curve chart illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate for each classifier. A model with a curve that bows closer to the top-left corner demonstrates stronger discriminatory power. In our case, the Random Forest model achieved the highest area under the curve (AUC), indicating superior ability to distinguish between patients with and without heart disease. Models like Logistic Regression and MLP Classifier also showed competitive AUC values, while the SVM model had a relatively lower curve, reflecting its comparatively weaker performance. This chart validates that Random Forest not only performs well in terms of accuracy but also maintains a high level of sensitivity and robustness across different thresholds.

### **5.5 Correlation Analysis**

A screenshot of a computer

AI-generated content may be incorrect.

**Figure 29. Correlation Analysis**

The correlation matrix highlights the linear relationships between features, with stronger values indicating more influence on heart disease prediction. **ST\_Slope** and **Oldpeak** show strong positive correlations with the **HeartDisease** label, while **MaxHR** is negatively correlated—suggesting that lower maximum heart rate may indicate higher risk. These patterns align with clinical insights and support the relevance of the selected features used in the models.

### **5.6 Comparison with Existing Studies**

Compare model performance with prior research and public benchmarks.

A screenshot of a computer

AI-generated content may be incorrect.

**Figure 30. Existing Studies**

### **5.7 Scalability and Inference Speed**

**Brief discussion:**

* Prediction latency: **<100 ms**
* Model loading: **instantaneous**
* Cloud deployment via Docker + Azure Container Instances
* Easily extensible to Kubernetes or AKS

### **5.8 Summary**

**Summarize key findings:**

* Random Forest model is top performer
* System achieves real-time inference with strong interpretability
* Evaluation aligns with medical literature
* Confirms reliability and deployability of system

# **Chapter** **six**

## **Results & Discussion**

### **6.1 Introduction**

This chapter discusses the results generated by the implementation and evaluation of the proposed system for heart disease prediction. It analyzes the accuracy and performance of various machine learning algorithms used, reflects on the success of the project in achieving its objectives, and critically examines the practical implications of the developed solution. Additionally, it addresses the limitations faced during the study and considers their potential impact on the results and their applicability in real-world scenarios.

### **6.2 Summary of Findings**

The core objective of this project was to develop an AI-powered, clinically interpretable system to predict the presence of heart disease using structured clinical data. The results obtained can be summarized as follows:

* **Data Analysis and Feature Significance**  
  Exploratory data analysis (EDA) highlighted features such as ST\_Slope, Oldpeak, ChestPainType, and MaxHR as being strongly correlated with heart disease. These features consistently emerged as top predictors across multiple models.
* **Model Performance**  
  Among all the trained models, the **Random Forest** Classifier showed the best overall performance:
* **Accuracy**: 90.1%
* **F1-score**: 90%
* **Precision**: 90%
* **Recall**: 89%

Other models such as **Logistic Regression** and **SVM** also performed well, though with slightly lower metrics. The **MLP Classifier**, representing a basic neural network, achieved 88.2% accuracy.

* **System Integration and Deployment**  
  The selected model was embedded into a **Flask web application** with a **responsive HTML/Tailwind CSS interface**, enabling users to input patient data and receive immediate predictions.
  + The application was **containerized using Docker** for reproducibility and deployed using **Microsoft Azure Container Instances**, making it accessible via a public web link.
* **User Experience and Practicality**  
  The interface is intuitive and responsive, offering a seamless experience to clinicians or researchers. The application consistently provides results in under one second, validating its readiness for practical use.

A computer screen shot of a computer screen

AI-generated content may be incorrect.

**Figure 31. web interface**

### **6.3 Interpretation of Results**

#### **A screenshot of a computer AI-generated content may be incorrect.Achievement of Project Objectives**

**Figure 32. Project Objectives**

#### **Scientific Interpretation**

The success of the **Random Forest** model in this context is consistent with the literature, where ensemble-based methods often outperform linear models when dealing with a mix of categorical and numerical clinical data. The fact that the model maintained a strong balance between sensitivity and specificity (recall and precision) demonstrates its reliability in minimizing both false positives and false negatives—both critical in medical diagnostics.

The inclusion of interpretable features (e.g., ST slope and exercise-induced angina) aligns the system with domain knowledge and real-world clinical indicators of cardiovascular risk, which further enhances its trustworthiness and potential for adoption.

### **6.4 Limitations of the Proposed Solution**

While the proposed system demonstrates promising results, several limitations exist that should be acknowledged and considered for future improvement.

#### **6.4.1 Dataset Limitations**

The system was trained on a single-source dataset (UCI Cleveland), which consists of 918 samples — a relatively small size for training robust machine learning models in healthcare. The dataset lacks diversity in terms of ethnic, demographic, and geographic representation, which may affect model generalization in global populations.

#### **6.4.2 Clinical Context Gaps**

The dataset is structured and tabular but does not contain lifestyle, family history, medication, or real-time monitoring data, all of which are relevant in actual diagnostic workflows. Time-series or longitudinal tracking is not supported, limiting the ability to model changes in patient health over time.

#### **6.4.3 Interpretability Challenges**

Although Random Forests are more interpretable than neural networks, they still lack fine-grained explainability at the individual prediction level. Advanced interpretability frameworks such as SHAP (SHapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) were not integrated due to time constraints.

#### **6.4.4 Real-world Clinical Validation**

The model has not yet been validated in clinical trials or real-world settings by healthcare professionals. Predictions are not tied to any electronic health record (EHR) system, limiting integration into hospital systems.

### **6.5 Broader Impact and Implications**

Despite these limitations, the system demonstrates the potential of AI-driven tools to augment clinical decision-making. A lightweight, interpretable model with a fast and user-friendly interface could serve as a secondary diagnostic aid, especially in under-resourced regions or as a preliminary screening tool. The methodology and deployment approach are scalable and reusable in other medical domains beyond cardiovascular disease.

### **6.6 Summary**

* The project met all primary objectives, delivering an accurate, interpretable, and cloud-deployed heart disease prediction system.
* Random Forest was identified as the best-performing model.
* The system supports rapid, real-time predictions through a browser-accessible interface.
* Limitations include dataset constraints, lack of clinical integration, and absence of time-series or lifestyle data.
* Nonetheless, the project serves as a valuable proof of concept for AI applications in healthcare diagnostics

# **Chapter Seven**

## **Conclusion and Future Work**

### **Summary of Contributions**

This project presents a comprehensive solution for predicting heart disease using machine learning techniques, with the following key contributions:

* **Accurate Predictive Model**:  
  Developed and evaluated multiple ML models using the UCI Cleveland dataset. The **Random Forest classifier** achieved the best performance with **90.1% accuracy**, making it the optimal choice for deployment.
* **Feature Analysis**:  
  Identified critical features such as ST\_Slope, Oldpeak, ChestPainType, and MaxHR that are highly correlated with heart disease, offering insights consistent with clinical knowledge.
* **Web-Based Deployment**:  
  Designed and implemented a responsive **Flask-based web application** that allows users to input clinical values and receive immediate predictions, making the system accessible and practical.
* **Cloud Integration**:  
  The model and interface were **containerized using Docker** and deployed on **Microsoft Azure**, ensuring accessibility, reproducibility, and scalability for broader use.
* **Accessible Diagnostic Tool**:  
  Bridged the gap between academic research and clinical application by making an AI-driven diagnostic assistant usable by healthcare workers, researchers, and educators.

These contributions collectively demonstrate the potential of integrating AI and cloud computing to improve early detection and decision-making in cardiovascular healthcare.

### **Possible Improvements or Extensions for Future Work**

While the current system is functional and accurate, several areas can be enhanced or expanded to improve its real-world effectiveness

**Data Enrichment**

* Incorporate larger, multi-source datasets with more diverse demographic and geographic representation to improve generalizability.
* Augment the dataset with additional health metrics such as cholesterol history, blood sugar trends, and family history.

**Feature Expansion**

* Include lifestyle and behavioral factors such as smoking, alcohol use, physical activity, and dietary habits.
* Integrate genetic and genomic data for more personalized risk profiling.

**EHR System Integration**

* Develop APIs to integrate the system with **Electronic Health Record (EHR)** platforms, allowing for automatic data retrieval and seamless updates.

**Mobile Accessibility**

* Extend the system as a **mobile app or progressive web app (PWA)** for use in low-resource environments or by non-specialist health workers.

**Advanced Modeling**

* Explore **deep learning architectures** like recurrent neural networks (RNNs) or transformers if longitudinal or time-series health data are available.
* Investigate **hybrid ensemble models** that combine the strengths of multiple classifiers for enhanced performance.

**Clinical Testing**

* Collaborate with healthcare institutions to conduct **clinical validation studies**, gather usability feedback, and iteratively refine the interface and prediction logic based on practitioner input.

### **Final Remarks**

This project successfully demonstrates how artificial intelligence and cloud technologies can be combined to build an effective, accurate, and accessible solution for heart disease prediction. By leveraging machine learning and deploying the model via a scalable web interface, this system not only achieves strong technical performance but also provides practical utility.

With continued development, real-world validation, and integration into clinical environments, the system has the potential to:

* Assist in early-stage diagnosis.
* Reduce healthcare costs through preventive screening.
* Empower clinicians and non-specialists with intelligent diagnostic tools.

Ultimately, this work contributes a strong foundation for future innovations in AI-assisted healthcare, showing how data science can make meaningful and measurable improvements in public health outcomes.

# **References**

[1] World Health Organization, “Cardiovascular diseases (CVDs),” 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] UCI Machine Learning Repository, “Heart Disease Dataset (Cleveland),” 1998. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[3] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[4] L. Breiman, “Random Forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324

[5] C. Cortes and V. Vapnik, “Support-Vector Networks,” Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995. doi: 10.1007/BF00994018

[6] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, 2nd ed. New York, NY, USA: Wiley, 2000. doi: 10.1002/0471722146

[7] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org

[8] Flask Documentation, “Flask: Web Development, One Drop at a Time,” 2023. [Online]. Available: https://flask.palletsprojects.com/

[9] Docker Documentation, “Docker: Empowering App Development for Developers,” 2023. [Online]. Available: https://docs.docker.com/

[10] Microsoft Azure, “Azure Container Instances Documentation,” 2023.

[Online]. Available: https://docs.microsoft.com/en-us/azure/container-instances/

[11] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[12] 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. Knowl. Discov. Data Min., 2016. doi: 10.18653/v1/N16-3020

[13] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2011. doi: 10.1016/C2009-0-61819-5

[14] W. McKinney, “Data Structures for Statistical Computing in Python,” in Proc. 9th Python in Science Conf., 2010. doi: 10.25080/Majora-92bf1922-00a