Predicting Cardiovascular Diseases using Machine Learning

Fathima Zaineb Ismath

BSc (Hons.) Computer Science
Honours Dissertation

Supervised by Dr. Cristina Turcanu

Heriot-Watt University
School of Mathematical and Computer Sciences

November 2024

DECLARATION

I, Fathima Zaineb Ismath, confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. A list of the references employed is included.

Signed: Fathima Zaineb Ismath

Date: 1 November 2024

ABSTRACT

Cardiovascular diseases affect a huge number of individuals worldwide. Early detection and accurate risk prediction can reduce its impact. Traditional risk factors drive the urgency of developing predictive models that can effectively identify individuals at high risk.

This project explores multiple machine learning techniques, including logistic regression, random forests, and deep learning algorithms, to build a robust predictive model for cardiovascular disease. A key innovation in this work is the integration of risk stratification and Explainable AI (XAI) to provide transparency and interpretability in predictions, enabling healthcare professionals to understand the rationale behind model decisions. This is critical for gaining clinical trust and promoting the adoption of AI-driven diagnostic tools in healthcare settings.

In addition to model accuracy, this study emphasizes the identification of significant predictors of CVD, optimizing the model for both performance and interpretability. Hence, this project seeks to contribute meaningfully to the field of cardiovascular health, offering insights that could support personalized treatment plans. Through these efforts, this study aims to bridge the gap between complex machine learning models and their practical use in early CVD detection and prevention.

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my supervisor, Dr. Cristina, for her guidance, encouragement, and expertise throughout this research. Her unwavering support has been instrumental in shaping this study, and I am immensely grateful for her mentorship.

I would also like to thank my parents for their constant love, care, patience, and encouragement.

Finally, I would like to extend my appreciation to the School of Mathematics and Computer Science for providing the resources and environment needed to pursue this work.

Thank you all for your invaluable contributions to this journey.

Table of Contents

List of Figures

List of Tables

GLOSSARY

AI Artificial Intelligence.

ANN Artificial Neural Networks.

AUC Area Under the Curve.

AUC-ROC Area Under the Receiver Operating Characteristic Curve.

CHD Coronary Heart Disease.

CNN Convolutional Neural Networks.

CVD Cardiovascular Disease.

DL Deep Learning.

EDA Exploratory Data Analysis.

FN False Negative.

FP False Positive.

FR Functional Requirements.

IELBT Improved Explainable Learning-Based Technique.

KNN K-Nearest Neighbors.

LIME Local Interpretable Model-agnostic Explanations.

MAE Mean Absolute Error.

ML Machine Learning.

NFR Non-Functional Requirements.

NN Neural Network.

PCA Principal Component Analysis.

PLES Professional, Legal, Ethical and Social Issues.

PSO Particle Swarm Optimization.

SHAP SHapley Additive exPlanations.

SVM Support Vector Machines.

TN True Negative.

TP True Positive.

UCI University of California, Irvine.

XAI Explainable Artificial Intelligence.

1 INTRODUCTION

This research explores Cardiovascular Disease (CVD) prediction and patient risk assessment using Machine Learning (ML) techniques.

1.1 Motivation

The heart is considered as one of the most vital organs in the human body. It is responsible for circulating blood throughout the body. According to the World Health Organization, cardiovascular diseases are one of the main causes of death worldwide. It results in about 17.9 million deaths every year [World Health Organization 2019]. In the United States alone, an individual experiences a heart attack every 40 seconds [Centers for Disease Control and Prevention 2024]. Predictive models for early Cardiovascular Disease identification have become crucial because of the rising growth of risk factors such as diabetes, obesity and sedentary activities.

This project has been driven by the potential of harnessing machine learning techniques to improve the early diagnosis of cardiovascular diseases. Predictive models would help healthcare professionals identify people at a high risk of heart disease and provide personalised treatment plans. This strategy can greatly minimize the financial burden on healthcare systems and improve patient outcomes. Additionally, this project incorporates Explainable Artificial Intelligence (XAI) to make the model's predictions more understandable, building physicians' confidence in applying machine learning tools in clinical settings.

1.2 Aim and Objectives

Aim

The aim of this project is to enhance existing efforts in cardiovascular disease prediction by developing an explainable machine learning model. This project seeks to provide healthcare professionals with a better understanding of the predictions by incorporating Explainable AI and risk stratification. It also aims to evaluate the performance of traditional machine learning and deep learning algorithms. Hence, this project aims to make a significant contribution to the domain of cardiovascular health.

Objectives

The key objectives are listed below.

O1: Conduct a detailed study on cardiovascular diseases - Research and understand the basic concepts of CVD and its risk factors.

O2: Perform Exploratory Data Analysis (EDA)- Apply appropriate data preprocessing techniques to ensure the dataset used in this study is of high quality.

O3: Evaluate Established Machine Learning Models- Investigate and evaluate the suitability of various models, such as Logistic Regression, Random Forests, and Deep learning techniques, for predicting cardiovascular disease.
O4: Optimal Feature Selection- Identify significant predictive features whose presence improves the accuracy of CVD prediction.
O5: Machine Learning Model Development- Develop a model for predicting cardiovascular diseases with various ML algorithms, deep learning techniques, and risk stratification.
O6: Model Performance Evaluation- Evaluate the developed model to check its predictive accuracy and interpretability by employing relevant evaluation metrics.

1.3 Contributions

This project contributes to the following fields in disease detection in healthcare.

(1) Model development for cardiovascular disease prediction.
(2) Integration of Explainable AI techniques.
(3) Risk stratification for patients.

1.4 Organisation

This report is structured to progressively illustrate the flow of the research.
Introduction: The introduction gives an overview of the project with its aims and objectives.

Background: This section presents the fundamental concepts of the project in sub-sections, followed by a detailed literature review of prior work in the field. It ends with a critical analysis of the existing literature.
Requirements: The Requirements Analysis chapter lists the study's objectives through functional requirements and outlines the non-technical aspects via non-functional requirements. This includes a traceability matrix that links these requirements to the project's objectives.
Design: The Research Methodology section offers an overview of how the research will be executed in the upcoming semester. The **Evaluation Strategy** section details the metrics and measures that will be employed to assess the model's performance.
References: A list of references used in this study is included at the end.
Appendices: The appendices include the Professional, Legal, Ethical and Social Issues (PLES) and Project Management sections. Appendix A presents the steps taken to address PLES and Appendix B outlines the risks and the project plan for implementation. Finally, Appendix C comprises project journals from weeks 4,7 and 10 which document incremental progress made during the study.

2 BACKGROUND

This chapter explores the primary concepts and offers an overview of existing literature in this domain. Following this, it presents a critical analysis of the studies discussed, aiming to establish a solid foundation for the research.

2.1 Background

2.1.1 Cardiovascular diseases.

Cardiovascular disease is a broad term used to describe a variety of disorders affecting the heart and blood arteries. There are four different types of CVD such as coronary heart disease, stroke, peripheral arterial disease and aortic disease [Yassine et al. 2014]. It induces many illnesses, disabilities, and deaths[Special report: 2001 – 2003, New Mexico [n. d.]]. Disease diagnosis is a significant part of healthcare. A heart attack occurs when a coronary artery becomes suddenly blocked, typically due to a blood clot. Symptoms of a heart attack include chest pain, shortness of breath, and fatigue.

2.1.2 Machine Learning.

Machine learning is a subset within the broader field of Artificial Intelligence (AI). It focuses on developing systems capable of learning from experience and making predictions accordingly [Shah et al. 2020].

Machine Learning models are broadly divided into two types: supervised and unsupervised models. Supervised learning involves training a model with labelled data to make predictions or classifications. It is suitable for classification scenarios where a set of inputs can result in an output. Popular supervised machine learning algorithms include Logistic Regression, Support Vector Machines (SVM), and Random Forests [Lee et al. 2023]. For example, Random Forest has shown high accuracy in CVD predictions due to its ability to handle large feature spaces and reduce overfitting through ensemble learning.[Breiman 2001].

In contrast, unsupervised learning models aim to identify hidden patterns or relationships within the data without any predefined guidance [Careervira 2023]. Patients can be clustered into different groups based on similarities in their health data, potentially revealing new subgroups of risk factors or disease phenotypes. K-Means clustering and Principal Component Analysis (PCA) are common unsupervised algorithms that have been employed.

2.1.3 Deep Learning.

Deep Learning (DL) is a subset of machine learning that uses multilayered Neural Network (NN) to imitate the decision-making power of the human brain. The main difference between deep learning and machine learning is the framework of the neural network architecture. Traditional ML methods rely on simple neural networks with one or two computational layers. Deep learning models train with three or more layers, typically hundreds or thousands of layers [Holdsworth and Scapicchio 2024]. They can identify patterns in patient data which might not be captured by traditional machine learning models. A conventional deep learning model consists of three main types of layers. They are the input layer- the layer which receives the input from the data, the hidden layers – intermediate layers that process the data to learn and the output layer which generates the prediction. The information learnt is propagated throughout the network using the hidden layers. Commonly used deep learning frameworks include Artificial Neural Networks (ANN), fully connected neural networks and Convolutional Neural Networks (CNN). Deep learning models are increasingly being used due to their multiple advantages such as increasing scalability with large datasets and the ability to learn and process complex data.

2.1.4 Risk Factors associated with cardiovascular diseases and Risk Stratification.

Several factors contribute to an individual's susceptibility to cardiovascular disease. Main influences include age, gender, cholesterol levels, blood sugar levels, and heart rate, among others [García-Ordás et al. 2023]. The identification of heart disease is particularly challenging due to a multitude of risk factors, such as diabetes, hypertension, elevated cholesterol levels, abnormal pulse rates, and various other elements that can significantly impact cardiovascular health. This complexity highlights the necessity for comprehensive evaluations and tailored interventions to effectively address these risks [Mohan et al. 2019].

Risk stratification is a process that categorises input into different risk levels based on the likelihood that it experiences a certain outcome. It plays a pivotal role in healthcare by identifying patients with high-risk scores and providing immediate treatment. For instance, in cardiovascular disease prediction, patients can be stratified into risk categories such as low, medium and high. Hence, risk stratification aids in providing patients with personalized treatment based on their risk profiles.

2.1.5 Explainable AI (XAI).

Explainable artificial intelligence is a set of techniques used to help human users understand and trust the results predicted by machine learning models [IBM 2023]. It advocates for the development of ML approaches that maintain strong performance while being more interpretable, enabling humans to interpret and trust the next generation of AI systems more effectively [Barredo Arrieta et al. 2020]. Its goal is to promote greater transparency in the realm of artificial intelligence. Artificial intelligence is becoming more accessible. When AI suggests a decision, the decision-makers must understand the underlying cause. However, when making significant life-changing decisions, such as a medical diagnosis, it is crucial to understand the reasoning behind such a critical decision. This highlights the importance of explainability in AI predictions [Adadi and Berrada 2018].

Common explainable AI methods used are SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). These two insightful methods help to explain the behaviour of the trained ML model. LIME provides localized insights for the user whereas SHAP gives a broader overview which is critical for complex models. SHAP values are derived from game theory and help quantify how each feature contributes to a prediction. A high SHAP value indicates that the feature has a major impact on the prediction. The formula for SHAP value for a feature in prediction is illustrated below.[Molnar 2024]

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left(f(S \cup \{i\}) - f(S)\right)$$

where:

ϕ_i is the Shapley value for feature i, representing its contribution to the model prediction. S is any subset of features excluding i, N is the total set of features in the model, and $f(S)$ represents the model prediction using only the features in subset S. $f(S \cup \{i\})$ computes the model prediction using the features in S along with feature i. $|S|$ is the number of features in subset S, and $|N|!$ is the factorial of the number of total features, which represents all possible combinations.

2.2 Relevant Work

This section explores the research landscape of CVD prediction using ML techniques, drawing insights from existing research papers on supervised learning models, ensemble techniques, deep learning, and explainable AI.

Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms (2019) In their 2019 study, Divya Krishnani et al. performed a comparative analysis of the performance of various ML algorithms in predicting coronary heart disease risk. Their main research objective was to improve the accuracy of predictions by using algorithms such as Random Forests, Decision Trees, and K-Nearest Neighbours [Krishnani et al. 2019]. The study shows the performance of a model using Random Forests in handling tabular and noisy data. The usage of the Framingham Heart Study dataset played a vital role in training the model with 4,240 records and 16 different features. The authors optimized the model by applying random oversampling to address null values and class imbalance. The use of random oversampling helps to mitigate the issue of class imbalance in datasets. It ensured both classes had equal representation. The study found that the Random Forest algorithm outperformed the other algorithms examined, with a remarkable accuracy of 96.8%. It outperforms Decisions Trees and K-Nearest Neighbour algorithms. Additionally, Random Forests showed resilience to data inconsistency.

However, this study concentrated solely on three specific algorithms, ignoring a broader review of various techniques that could enhance cardiovascular disease predictions. It missed a broader exploration of alternative methods such as built-in explainability features. Evaluating the model with more metrics and using a larger dataset could enhance its application in clinical settings. Future work could include developing a hybrid model that combines the interpretability of simple models with the effective predictive power of complex models. Adding methods like

SHAP or LIME could strengthen interpretability and provide a clearer explanation of which features influence predictions.

Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques (2019) The authors investigated the effectiveness of ensemble classification techniques for predicting heart disease risk. Their research aimed to improve the predictive accuracy of various weak classifiers by employing methods such as bagging, boosting, and stacking on a heart disease dataset [Latha and Jeeva 2019]. They noted that SVMs resulted in the highest accuracy (92.1%), followed by neural networks (91%). Decision trees were slightly lower at around 89.6%. They found that ensemble methods led to a significant increase in prediction accuracy, achieving up to a 7% improvement for weaker classifiers. This study highlights the potential of ensemble techniques to enhance predictive models in the medical field. Ensemble methods are effective in addressing the limitations of single weak classifiers, such as high variance or bias. Even though ensemble techniques improve accuracy, they could increase computational complexity, posing challenges for real-time deployment. It primarily emphasizes ensemble approaches without a thorough comparison to other classification methods, potentially overlooking alternative strategies that could yield similarly high accuracy. Bagging is a method that reduces model variance by training models. Boosting builds sequential models that rectify errors during each iteration. For example, the accuracy of the Naïve Bayes classifier increased by 0.99% with boosting. But, it may amplify noise in the data which could result in overfitting. [Chen and Guestrin 2016].
The authors evaluated multiple classifiers including Naïve Bayes, Random Forest, C4.5, Multilayer Perceptron, and PART, both individually and in ensemble models. Random Forest showed a good performance due to its ability to handle non-linear data [Breiman 2001]. However, it lacks interpretability which is essential in healthcare models[Rudin 2019]. The authors did not investigate advanced feature selection methods like genetic algorithms which could refine feature sets. [Guyon et al. 2002]. They mainly focused on accuracy as a performance metric. Even though accuracy is an important metric, it may not fully reflect the model's reliability in situations where false positives and false negatives are used. Implementing precision and recall would provide further insights into the model's effectiveness in prediction while minimizing unnecessary false positives [Davis and Goadrich 2006]. Future research could focus on balancing accuracy with interpretability by combining interpretable models like decision trees with ensemble methods.

Deep Learning for Cardiovascular Risk Stratification (2020) In 2020, Schlesinger and Stultz investigated the potential of deep learning in the realm of cardiovascular risk categorization [Schlesinger and Stultz 2020]. Schlesinger and Stultz argue that a predictive model must offer insights and be interpretable by doctors. This is a significant reasoning as understanding model interpretation is crucial. Several methods like Shapley values and Gradient-weighted

Class Activation Mapping (Grad-CAM) are used to improve the model's interpretability. This approach adds meaning in clinical usage. They identify the most influential features in the model's predictions. They focused their study on the need for better accuracy in risk assessment tools in the healthcare sector. The main purpose was to analyse the effectiveness of the deep learning models when compared to standard ones, especially in identifying high-risk patients who might benefit from quick intervention. The authors mentioned several deep-learning techniques for medical image analysis and structured electronic health record data. The study presents numerous benefits of DL models. They can analyse raw data while minimising the need for preprocessing and feature engineering. Despite these benefits, the need for large and high-quality datasets, and the possibility of model overfitting poses some challenges. The study does not address how the model might predict different demographics and socio-economic backgrounds depicting the importance of choosing diverse datasets for model training. They also suggest identifying failure modes which are the situations where the model may perform poorly.

Although they are preferred for their interpretability, conventional risk models like logistic regression and Cox proportional hazards models, fail to capture complex, non-linear relationships between patient characteristics and results. To ensure the model's fit for real-world scenarios, they also highlighted how important it is for the predictions to be transparent and comprehensible. The study also concentrates on CNN for image-based applications. Cardiovascular disease prediction may rely on other data sources such as structured electronic health records (EHR). The paper finishes by advocating that deep learning models must be integrated which would enhance the predictive accuracy in clinical diagnosis. Future work could include developing a hybrid model that integrates both traditional and deep learning techniques. This integration would bridge the gap between accuracy and interpretability, ensuring models are clinically useful and trustworthy.

Heart Disease Prediction using Hybrid Machine Learning Model (2021) This research paper proposed a novel hybrid ML model to enhance heart disease prediction [Kavitha et al. 2021]. The authors implemented a hybrid model as a novel technique for better-optimized results by merging Decision Tree and Random Forest algorithms. They aimed to combine the strengths of both algorithms to enhance prediction accuracy. In the hybrid approach, the probabilities generated from one model (Random Forest) are used as input for the other model (Decision Tree), thereby improving robustness and reducing the likelihood of overfitting. By integrating the probabilistic outputs of both machine learning techniques, it provides a comprehensive approach to disease prediction.

The study sought to leverage the Cleveland heart disease dataset and split the data into 70% training and 30% testing subsets. It used data mining techniques, including regression and classification, to effectively analyse patient health data. The proposed work used a TkInter Python-designed application with sklearn libraries, pandas and matplotlib. Evaluation metrics such as mean square error (MSE), mean absolute error (MAE), R-Squared parameter, root mean square error (RMSE) and accuracy were implemented to evaluate the effectiveness of the proposed model. The authors evaluated model performance based on accuracy. Experimental results demonstrated that the hybrid model achieved 88% accuracy, outperforming the Decision Tree (79%) and Random Forest (81%) when used individually. The research highlights the potential of machine learning techniques in improving early detection and intervention for heart diseases. Additionally, the paper mentions optimization techniques like Genetic Algorithms and Particle Swarm Optimization (PSO) for feature selection. However, it does not implement these techniques, leaving potential improvements unexplored. Future research could focus on expanding the dataset and applying deep learning techniques for higher predictive accuracy. Researchers could also explore multi-class classification approaches and deep learning techniques to examine the severity of heart disease.

XAI Framework for Cardiovascular Disease Prediction Using Classification Techniques (2022) In their 2022 research, Guleria et al. presented an explainable artificial intelligence framework aimed at enhancing cardiovascular disease prediction using advanced classification methods [Guleria et al. 2022]. They investigated the benefits of ensemble classifiers on the XAI framework [Baghdadi et al. 2023]. Their main objective was to improve the interpretability and reliability of machine learning models, by employing ensemble classifiers such as SVM,K-Nearest Neighbors (KNN), and AdaBoost. By combining results from multiple classifiers, ensemble techniques such as bagging and boosting were used to increase prediction robustness by improving the overall predictive stability. These methods enable to create a balanced predictive framework by addressing the limitations of individual models.

This study correlates with [Bizimana et al. 2024], which showed the role of SHAP and LIME in enhancing model interpretability. The study uses correlation analysis, neighbourhood component analysis, SHAP values, and LIME to identify key features in a predictive model, focusing on age and sex as critical predictors. SVM is highlighted for its capability to handle multi-dimensional data. The dataset used in this study was sourced from the UCI ML repository comprising 303 instances and 14 attributes. The study illustrated the need for explainable models in clinical settings, where understanding the reasoning behind predictions is crucial for effective decision-making. The effectiveness of the XAI-based cardiovascular disease prediction models was evaluated using a variety of evaluation criteria, such as Area Under the Curve (AUC), accuracy, sensitivity, and F1-score, to assess model performance.

According to their findings, the ensemble classifiers performed better in distinguishing between patients who were at risk of heart disease from those who were not. They attained an accuracy of 89%. Outcomes showed that the XAI-driven ensemble classifiers outperformed traditional classification models in terms of efficiency, with the SVM algorithm obtaining the maximum accuracy of 82.5%.

This research underscores how crucial it is to incorporate XAI concepts into healthcare applications to improve diagnostic procedures and foster greater clinician trust in predictive models. The significance of explainable artificial intelligence was emphasised by the researchers, who cited Gunning D. (2019) as saying that "Explainable AI will develop a set of machine learning algorithms which will allow individual users to comprehend, adequately trust, and manage the next era of artificially intelligent companions". The reliance on a small-sized dataset limits the model's generalizability. To improve the robustness of heart disease classification algorithms, the study recommends that future research concentrate on real-time data collection and further statistical analysis using a variety of datasets, including the UK Biobank dataset and the Statlog heart disease dataset.

Advanced machine learning techniques for cardiovascular disease early detection and diagnosis (2023) In their 2023 study, Baghdadi et al. contributed to the growing body of evidence by emphasising the importance of machine learning in the early detection and diagnosis of cardiovascular diseases [Baghdadi et al. 2023]. The study aimed to leverage health data from hospital databases to create effective machine learning algorithms that enhance the predictive accuracy of CVD risk assessments. The researchers introduced a Gradient Booster model, which achieved an F1-score of 92.3% and an overall accuracy of 90.94%. This highlights the critical need for early diagnosis and treatment to improve patient outcomes and reduce healthcare costs. It also tried to determine the feature that contributes the most to the prediction.

The dataset used in this study was synthesized from the UCI ML Repository, combining five heart datasets to create the largest heart disease dataset available for research, featuring 11 common attributes. The authors used Google Colab as the framework for implementing machine learning models. Their approach comprised a thorough preprocessing of the data, feature selection using Shapley values along with rigorous model evaluation based on various performance metrics. To ensure reliable estimates of the model's generalization capability, Baghdadi et al. implemented K-Fold cross-validation. The top 20 predictors of heart disease were shown in order of relevance using a summary plot of Shapley values. The authors use CatBoost with feature engineering, which handles both structured and unstructured data effectively.

The findings of this study revealed that machine learning techniques not only facilitate early diagnosis but also significantly lessen the financial strain on healthcare systems. Since the research relied on secondary data, there were some missing values present. Additionally, Gradient Boosting methods require high computational power, which can limit the model's applicability in resource-constrained settings. This research underlines the potential of machine learning in enhancing cardiovascular care through evidence-based decision-making. Hence, it shows the importance of integrating advanced analytics in clinical practice. The authors suggest that future research should concentrate on using datasets that cover a wider range of risk variables, both modifiable and non-modifiable.

Automated heart disease prediction using improved explainable learning-based technique (2024) In their 2024 study, Bizimana et al. proposed an Improved Explainable Learning-Based Technique (IELBT) focusing on improving heart disease prediction [Bizimana et al. 2024]. The authors refined the dataset to select only the most predictive features using techniques like Random Forest, Recursive Feature Elimination, and Support Vector Machine. The use of the SHAP method enhances the interpretability of the model. However, the use of SHAP doesn't guarantee computational demands especially when scaling the model to larger datasets. To build a reliable predictive model, the authors integrated different feature selection strategies, data normalisation strategies, and machine learning algorithms. They obtained an excellent accuracy of 96.00% using SVM, greatly surpassing previous models in the literature, utilising the Alizadeh Sani heart disease dataset. A standout feature of this study is the use of explainable AI methods, specifically SHAP and LIME. It emphasises model interpretability. However, the study's dependence on a single dataset raises questions on how broadly applicable the findings might be. Testing the model on diverse datasets would strengthen the model's application in clinical settings. Further studies may build on this work by evaluating the IELBT on various datasets to determine its effectiveness in various clinical settings. It could explore the use of deep learning models, such as Convolutional Neural Networks and Long Short-Term Memory networks. These models can handle large healthcare datasets.

2.3 Critical Analysis of Relevant Works

Author(s)	**Year**	**Article Name**	**Description**	**Limitations**
Divya Krishnani, Anjali Kumari, Akash Dewangan, Aditya Singh, Nenavath Srinivas Naik	2019	Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms	This study aims to predict the risk of Coronary Heart Disease (CHD) by using various machine learning algorithms, including Random Forest, Decision Trees, and K-Nearest Neighbours. In order to increase model accuracy, the authors use the Framingham Heart Study dataset and concentrate on thorough preprocessing approaches; using Random Forest, they were able to achieve an accuracy of 96.8%.	The comparative analysis lacks qualitative explanations for why some models perform better and is mainly quantitative. It misses opportunities to delve deeper into the behaviour of the models. Despite its high accuracy, Random Forest lacks interpretability, which restricts its clinical use.

C. Beulah Christalin Latha, S. Carolin Jeeva	2019	Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques	This study explores the use of ensemble classification techniques to improve the accuracy of heart disease risk prediction. The authors conducted experiments using various ensemble methods, including bagging, boosting, and stacking, on a heart disease dataset. They reported an accuracy increase of up to 7% for weak classifiers through ensemble techniques, demonstrating the effectiveness of these methods in enhancing predictive performance.	The study focuses primarily on the use of ensemble methods without a comprehensive comparison to other individual classification techniques that may also provide high accuracy. The study focused on accuracy and overlooked other critical metrics like precision and recall which are crucial in healthcare applications.
Daphne E. Schlesinger, Collin M. Stultz	2020	Deep Learning for Cardiovascular Risk Stratification	Using health data from hospital databases, this study suggests innovative machine learning methods for the early diagnosis of cardiovascular illnesses. The authors provide a Catboost model with an accuracy of 90.94% and an F1-score of 92.3%, emphasising feature selection.	While the study demonstrates remarkable accuracy, it primarily focuses on one machine learning model without conducting a comparative analysis with several other cutting-edge methods. Even though deep learning models like CatBoost are beneficial, it could be difficult to comprehend them in clinical settings.

Dr. M. Kavitha, G. Gnaneswar, R. Dinesh, Y. Rohith Sai, R. Sai Suraj	2021	Heart Disease Prediction using Hybrid Machine Learning Model	This research proposes a hybrid machine learning model combining Random Forest and Decision Trees. The hybrid technique achieves an accuracy of 88.7%. The authors highlight the advantages of using a hybrid model to improve prediction accuracy and use the Cleveland Heart Disease dataset.	Although the hybrid model shows improved accuracy, the study lacks a comprehensive evaluation of the distinct contributions made by each of the hybrid model's algorithms. The dataset is limited to Cleveland which may limit the model's generalizability to broader populations.
Pratiyush Guleria, Parvathaneni Naga Srinivasu, Shakeel Ahmed, Naif Almusallam, Fawaz Khaled Alarfaj	2022	XAI Framework for Cardiovascular Disease Prediction Using Classification Techniques	This review discusses that in comparison to conventional techniques, deep learning has the potential to increase prediction accuracy when creating risk stratification models for cardiovascular illnesses. It stresses the importance of comprehending failure mechanisms and evaluating models in clinical situations.	The study offers an excellent overview, but it does not offer new insights. While ensemble methods enhance accuracy, they increase computational complexity which could affect deployment in resource-limited situations.
Nadiah A. Baghdadi, Sally Mohammed Farghaly Abdelaliem, Amer Malki, Ibrahim Gad, Ashraf Ewis, Elsayed Atlam	2023	Advanced machine learning techniques for cardiovascular disease early detection and diagnosis	This research focuses on using cutting-edge machine learning methods, such as Catboost, to identify and diagnose cardiovascular illnesses. The authors emphasize the role of feature selection in improving model performance, obtaining an F1-score of 92.3%.	While the paper highlights the importance of feature selection, it does not go into detail about the particular features which constitute the model, impacting its utility in real-world clinical settings.

Pierre Claver Bizimana, Zuping Zhang, Alphonse Houssou Hounye, Muhammad Asim, Mohamed Hammad, Ahmed A. Abd El-Latif	2024	Automated heart disease prediction using improved explainable learning-based technique	IELBT for predicting cardiac disease is presented. The authors combine feature selection techniques, data normalisation, and ML algorithms. Using a support vector machine, they tested their model on the Alizadeh Sani heart disease dataset and obtained an accuracy of 96.00%. The study highlights the model's interpretability and offers insights using SHAP and LIME methods.	While the IELBT achieved great accuracy, the study's concentration on a single dataset may restrict the generalisability of results to larger populations.

Table 1. Critical Analysis of Cardiovascular Disease Prediction Studies

2.4 Conclusion

The reviewed literature provides a substantial contribution to the field of machine learning in cardiovascular disease prediction. It examines a variety of models using different methodologies and datasets but there are notable limitations.

Many of these studies focus on providing binary outcomes without categorizing individuals into distinct risk levels, which is a significant drawback. While significant strides have been achieved in CVD prediction, there exists a tradeoff between accuracy and interpretability. Existing studies, including those by Krishnani et al. and Schlesinger and Stultz, demonstrate that even highly accurate models are limited by their lack of explainability. While Bizimana et al. mainly focused on refining a single dataset with SHAP, Baghdadi et al. leveraged multiple datasets, making their results more generalizable. The studies emphasize feature selection and ensemble methods, showing that they enhance prediction accuracy. They highlight the need for a balance between accuracy and explainability. Understanding the risk level is crucial, especially for high-risk patients, as it allows medical professionals to prioritize treatments effectively rather than relying solely on basic predictions.

As a result, this project aims to address these limitations by developing a model to predict heart disease and stratify patients into risk levels. It will evaluate the performance of various machine-learning algorithms.

3 REQUIREMENT ANALYSIS

The upcoming section addresses the essential steps needed for the successful implementation of the project. The main functionalities of the proposed system are stated as requirements, which are categorized into Functional and Non-Functional Requirements (NFR). Functional Requirements (FR) define what the system should do, while non-functional requirements describe how the system should do it [GeeksforGeeks 2024]. Functional Requirements are further prioritized using the MoSCoW prioritization method and colour-coded for clarity. This analysis delineates the system's primary functions and operational attributes, thereby ensuring a comprehensive understanding of the project's scope.

- **M (Must Have):** Necessary features that are critical for the successful completion of the project.
- **S (Should Have):** Important features that should be included if feasible.
- **C (Could Have):** Desirable attributes that are not essential but can be incorporated as enhancements.
- **W (Won't Have):** Initiatives that are not prioritized within the scope of the project.

3.1 Functional Requirements

The functional requirements obtained from the objectives of this study are mentioned below.

FR	Description	Priority
FR-1	**Feature Selection:** Perform feature engineering to identify the most significant factors that influence cardiovascular disease outcomes.	**M**
FR-2	**Machine Learning Model Development:** Develop a predictive model using machine learning algorithms, such as Logistic Regression and Random Forests, specifically for cardiovascular disease prediction.	**M**
FR-3	**Risk Stratification Implementation:** Integrate risk stratification techniques within the model to categorize patients into high-risk, medium-risk, or low-risk groups.	**M**
FR-4	**Model Performance Evaluation:** Evaluate the model's performance using various metrics such as accuracy, precision, recall, F1 score,AUC, and confusion matrix.	**M**

FR-5	**Explainable AI Integration:** Incorporate explainable AI techniques, such as SHAP or LIME to assess the model's interpretability.	**M**
FR-6	**Risk Assessment Visualization Generation:** Create visualizations, such as graphs, to illustrate how each feature contributes to the risk predictions for individual patients.	**S**
FR-7	**Prototype Development:** Develop a graphical user interface that enables healthcare professionals to input patient data and get risk predictions in a user-friendly manner.	**C**
FR-8	**User Interface Functionality:** The system could enable healthcare professionals to navigate through results using a graphical user interface.	**C**

Table 2. Functional Requirements

3.2 Non-Functional Requirements

Non-Functional Requirements list the quality attributes of the system.

FR	**Description**	**Priority**
NFR-1	**GDPR-Compliant Security:** The dataset used in this research is open source and complies with GDPR regulations governing data protection and ethical standards.	**M**
NFR-2	**Code Documentation:** The code written will be documented to allow easy readability and maintainability.	**M**
NFR-3	**Performance Efficiency:** The model will be optimized for performance to ensure efficient data processing.	**S**
NFR-4	**Response Time:** The system shall process user inputs and deliver predictions within 10 seconds to provide an effective user experience.	**S**

Table 3. Non-Functional Requirements

3.3 Traceability Matrix

A traceability matrix provides a framework to map each research objective with the corresponding functional requirements. This approach ensures a clear connection between the objectives of the study and the specific system functionalities.

Requirement/Objectives	**O1**	**O2**	**O3**	**O4**	**O5**	**O6**
FR-1	✓			✓		
FR-2			✓		✓	
FR-3					✓	
FR-4						✓
FR-5				✓	✓	✓
FR-6					✓	
FR-7					✓	
FR-8						✓
NFR-1	✓	✓				
NFR-2			✓		✓	
NFR-3						✓
NFR-4						✓

Table 4. Traceability Matrix

3.4 Research Questions

The main research questions discussed in this report are listed below.

(1) What are the key risk factors associated with cardiovascular diseases, and how do they influence disease prediction outcomes?
(2) How do various machine learning models and deep learning models compare in their effectiveness for predicting cardiovascular disease?
(3) Which predictive features significantly contribute to the accuracy of cardiovascular disease predictions, and how can these features be effectively identified and validated?
(4) How can the incorporation of explainable AI techniques, such as SHAP or LIME, enhance the interpretability of the model's predictions?

4 DESIGN: RESEARCH METHODOLOGY SKETCH

The steps below outline the research methodology that will be adopted for the implementation of the proposed model. Figure 1 illustrates the architecture of the project work flow.

4.1 Research Design and Intent

The main research questions have been stated in section 3.6. The research design has been formulated by conducting a critical analysis of previous studies and identifying potential research gaps. It was evident that future research could focus on explainable AI and risk stratification on machine learning models. A two-way approach will be adopted for model development. The initial focus will be on traditional machine learning models, followed by integrating deep learning techniques for analysing tabular data in healthcare.

Traditional Machine Learning Models such as Decision trees, random forests and other popular ML techniques will be used.

- **Random Forest**: A Random Forest is a learning technique that creates multiple decision trees and combines their results. It is used to boost prediction accuracy and reduce overfitting.
- **Decision Tree**: A Decision Tree is used to classify CVD risk by using a tree-like structure where each node is based on certain patient criteria.

Deep Learning Model

- **Fully Connected Neural Network**: It is a type of neural network where every neuron in the current layer is connected to every neuron in the previous and next layer. It enhances prediction accuracy by capturing non-linear relationships in the data. [GeeksforGeeks 2021]

4.2 Research Methodology

4.2.1 Data Acquisition and Preprocessing. The primary datasets considered include the Heart Disease dataset which was obtained from University of California, Irvine (UCI) Machine Learning Repository [Janosi et al. 1989], Framingham dataset [Kaggle 2023] and Statlog heart disease dataset [UCI Machine Learning Repository 2023]. The UCI Heart Disease dataset, ethically cleared and sourced from reputable sources, includes 14 attributes and 303 rows, including patient heart disease factors like age, cholesterol, blood pressure, and diabetes, to enhance the model's risk assessment. This stage involves cleaning the dataset to address null values and inconsistencies within the data. Identifying missing and duplicate values is essential to prevent bias during model training. The dataset will be divided into training and testing subsets which will be used to train and test the model's performance respectively.

4.2.2 Feature Selection. In this stage, significant features are identified and selected. This step is crucial as it enhances the model's accuracy. The core features are presented using explainable AI techniques in the form of graphs to facilitate better understanding. This increases transparency and user confidence in the model by helping users understand which features are highly influencing its predictions.

4.2.3 Model Development. Next, the model will be developed using a wide range of machine learning techniques, including Random Forests and Decision Trees. A fully connected neural network will be implemented using deep learning techniques. This phase will use XAI techniques to examine how different algorithms handle attributes and prediction processes, ensuring that the model's outputs remain comprehensible even as complexity grows. SHapley Additive exPlanations (SHAP) is a powerful explainable AI technique that offers clarity to the opaque predictions of ML models [Yasunobu et al. 2022]. Patients will be classified using risk categories such as low, medium, and high enabling the model to stratify patients by their risk of developing cardiovascular disease.

4.2.4 Model Testing and Evaluation. The model will be tested by using a subset of the dataset that was reserved earlier for testing purposes. The predictive performance of the developed model will be assessed through standard quantitative evaluation metrics, such as accuracy, precision, and recall. Chapter 5 details on further analysis of the evaluation strategy used.

4.3 Development Tools

Python will be used as the primary programming language for this project, due to the extensive range of libraries it offers for machine learning and data analysis, including tools like scikit-learn, tensorflow/keras, pandas, and numpy. Github will be used as a code repository.

4.4 Significance: Aiming for Clinical Utility

The study aims to develop a practical model for early heart disease risk prediction, integrating SHAP values and risk stratification. Hence, it enables clinicians to assess patient risk more accurately and deliver personalized treatment plans in cardiovascular care, bridging the gap between predictive performance and interpretability.

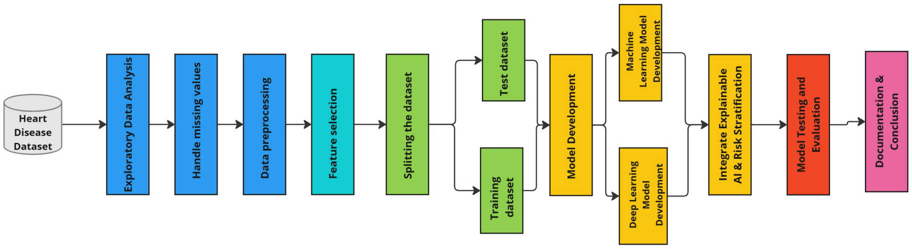

Heart Disease Dataset
Exploratory Data Analysis
Handle missing values
Data preprocessing
Feature selection
Splitting the dataset
Test dataset
Training dataset
Model Development
Machine Learning Model Development
Deep Learning Model Development
Integrate Explainable AI & Risk Stratification
Model Testing and Evaluation
Documentation & Conclusion

Fig. 1. Model Development Architecture

5 EVALUATION STRATEGY

This chapter details the evaluation metrics used to test the performance of the developed CVD prediction model. The metrics used are mentioned below along with their purpose and formula.

5.1 Performance Metrics

True Positive (TP)- The cases that are correctly classified as positive results.
False Positive (FP)- The cases that are predicted as positive but were negative.
True Negative (TN)- The cases that are correctly predicted as negative.
False Negative (FN)- The cases that were actually positive but predicted as negative.

5.1.1 Accuracy.

The ratio of correctly classified predictions by the total number of cases examined.
Purpose: This metric highlights the percentage of outcomes that are correctly classified. [Devi et al. 2021]
Formula:

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

5.1.2 Precision.

The ratio of true positive predictions divided by the total predicted positives.
Purpose: It demonstrates the relevance of the results.
Formula:

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

5.1.3 Recall (Sensitivity or True Positive Rate).

The ratio of true positive predictions predicted by the model to the total actual positives.
Purpose: It presents the model's ability to find relevant results. A high recall indicates that at-risk patients are not disregarded.
Formula:

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

5.1.4 F1 Score.

The harmonic mean of precision and recall, offers a balance between the two metrics.
Purpose: It is useful when the class distribution is imbalanced.
Formula:

$$F1\text{ Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

5.1.5 Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

AUC measures the model's capacity to differentiate between positive and negative classes.
Purpose: It visualizes how well the model performs, especially in imbalanced datasets. The equation for AUC is given below where TPR denotes the True Positive Rate and FPR denotes the False Positive Rate.

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR}^{-1}(x))\,dx$$

5.1.6 Confusion Matrix.

The confusion matrix is a table that plots the performance of a model by showing the true positives, true negatives, false positives, and false negatives.
Purpose: It helps to focus on the types of errors made by the model and refine performance.

	Predicted True	**Predicted False**
Actual True	True Positive (TP)	False Negative (FN)
Actual False	False Positive (FP)	True Negative (TN)

Table 5. Confusion Matrix

5.2 Additional Evaluation Strategies

5.2.1 Cross-Validation.

Purpose: K-fold Cross-validation is used to assess a model's performance more reliably by testing it across multiple data subsets. This ensures that the model is trained well and prevents overfitting.
In k-fold cross-validation, the dataset is partitioned into k equal-sized segments. [Katabathina 2024]. The model is then trained on k-1 folds and tested on the remaining reserved fold. During the k repetitions of the process, each fold serves as the test set exactly once. The final metric is the average of the metrics from each fold, giving an overall average of the model's performance.

5.2.2 Mean Absolute Error (MAE). MAE is known as the average (mean) of the absolute differences between the actual values and predicted values. Purpose: MAE provides a simple interpretation of prediction errors.

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}\left|y_i - y_i^*\right|$$

where: y_i is the actual value, y_i^* is the predicted value and m is the total number of predictions.

5.3 Interpretability and Explainability Assessment

Tools like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) provide insights into how the model makes its predictions, which can improve trust and understanding in clinical settings. By visualising contributions to individual predictions, it helps to identify the most influential factors.

REFERENCES

Amina Adadi and Mohamed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

N.A. Baghdadi et al. 2023. Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. *Journal of Big Data* 10, 1 (2023). https://doi.org/10.1186/s40537-023-00817-1

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Andres Barbado, Sergio Garcia, Sonia Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

P.C. Bizimana et al. 2024. Automated heart disease prediction using improved explainable learning-based technique. *Neural Computing and Applications* 36, 26 (2024), 16289–16318.

Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

Careervira. 2023. *Machine learning for beginners: A comprehensive 2023 guide.* Available at: https://medium.com/@careervira.community/machine-learning-for-beginners-a-comprehensive-2023-guide4ec02d5caab7 (Accessed: 22 September 2024).

Centers for Disease Control and Prevention. 2024. Heart Disease Facts. Available at: https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).* 785–794. https://doi.org/10.1145/2939672.2939785

Jesse Davis and Mark Goadrich. 2006. The Relationship between Precision-Recall and ROC Curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML).* 233–240. https://doi.org/10.1145/1143844.1143874

D. Devi, S. Sophia, and S. R. Boselin Prabhu. 2021. Deep learning-based cognitive state prediction analysis using Brain Wave Signal. In *Cognitive Computing for Human-Robot Interaction.* 69–84. https://doi.org/10.1016/b978-0-323-85769-7.00017-3

María Teresa García-Ordás et al. 2023. Heart disease risk prediction using Deep Learning techniques with feature augmentation. *Multimedia Tools and Applications* 82, 20 (2023), 31759–31773. https://doi.org/10.1007/s11042-023-14817-z

GeeksforGeeks. 2021. What is Fully Connected Layer in Deep Learning? https://www.geeksforgeeks.org/what-is-fully-connected-layer-in-deep-learning/. [Accessed November 7, 2024].

GeeksforGeeks. 2024. Functional vs Non-Functional Requirements. https://www.geeksforgeeks.org/functional-vs-non-functional-requirements/ Accessed: 2024-11-09.

Deepak Guleria et al. 2022. A Comprehensive Review of Machine Learning Techniques for Predicting Diseases. *Journal of Biomedical Informatics* 133 (2022), 104166. https://doi.org/10.1016/j.jbi.2022.104166

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Feature Selection for Machine Learning with Support Vector Machines. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS).* 1–7.

J. Holdsworth and Mark Scapicchio. 2024. *What is deep learning?* Available at: https://www.ibm.com/topics/deep-learning (Accessed: 22 September 2024).

IBM. 2023. *What is explainable AI?* Available at: https://www.ibm.com/topics/explainableai (Accessed: 11 October 2024).

Andras Janosi, W Steinbrunn, M Pfisterer, and R Detrano. 1989. Heart Disease Dataset. https://archive.ics.uci.edu/ml/datasets/heart+disease. Accessed: 2023-10-31.

Kaggle. 2023. *Framingham Heart Study Dataset.* Available at: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset?select=framingham.csv

(Accessed: 11 October 2024).

Sai Krishna Katabathina. 2024. Understanding Cross-Validation in Machine Learning. https://medium.com/@katabathina44313/understanding-cross-validation-in-machine-learning-a172b3c8ce84 Accessed: 2024-11-09.

M. Kavitha et al. 2021. Heart disease prediction using Hybrid Machine Learning Model. In *2021 6th International Conference on Inventive Computation Technologies (ICICT).* https://doi.org/10.1109/icict50816.2021.9358597 [Preprint].

Divya Krishnani et al. 2019. Prediction of coronary heart disease using supervised machine learning algorithms. In *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON).* https://doi.org/10.1109/tencon.2019.8929434 [Preprint].

C.B. Latha and S.C. Jeeva. 2019. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked* 16 (2019), 100203. https://doi.org/10.1016/j.imu.2019.100203

Young-Gon Lee, Ji-Yong Oh, Do-Kwan Kim, et al. 2023. SHAP Value-Based Feature Importance Analysis for Short-Term Load Forecasting. *Journal of Electrical Engineering & Technology* 18 (2023), 579–588. https://doi.org/10.1007/s42835-022-01161-9

S. Mohan, C. Thirumalai, and G. Srivastava. 2019. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access* 7 (2019), 81542–81554.

Christoph Molnar. 2024. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* Available at: https://christophm.github.io/interpretable-ml-book/ (Accessed: 31 July 2024).

Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1 (2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x

D.E. Schlesinger and C.M. Stultz. 2020. Deep Learning for Cardiovascular Risk Stratification. *Current Treatment Options in Cardiovascular Medicine* 22, 8 (2020). https://doi.org/10.1007/s11936-020-00814-0

D. Shah, S. Patel, and S.K. Bharti. 2020. Heart disease prediction using Machine Learning Techniques. *SN Computer Science* 1, 6 (2020). https://doi.org/10.1007/s42979-020-00365-y

Special report: 2001 – 2003, New Mexico. [n. d.]. Hospitalization for Heart Attack, Stroke, or Congestive Heart Failure among Persons with Diabetes.

UCI Machine Learning Repository. 2023. Statlog (Heart) Dataset. Available at: https://archive.ics.uci.edu/dataset/145/statlog+heart (Accessed: 11 October 2024).

World Health Organization. 2019. Cardiovascular diseases. Available at: https://www.who.int/health-topics/cardiovascular-diseasestab=tab$_1(Accessed : 11October2024)$.

Ihab Yassine, Doaa Mousa, and Nagy Zayed. 2014. Automatic Cardiac MRI Localization Method. In *Proceedings of the 7th Cairo International Biomedical Engineering Conference (CIBEC).* https://doi.org/10.1109/CIBEC.2014.7020942

N. Yasunobu, M. Koutarou, S. Hidehisa, and N. Naoki. 2022. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Computer Methods and Programs in Biomedicine* 214 (2022), 106584. https://doi.org/10.1016/j.cmpb.2021.106584 (Accessed: 11 October, 2024).

H.A. Zaidi and P. Jain. 2024. A Review of Machine Learning Models for Predicting Agile Methodology. In *2024 2nd International Conference on Disruptive Technologies (ICDT)* (Greater Noida, India). 971–974. https://doi.org/10.1109/ICDT61202.2024.10489437

A APPENDIX: PROFESSIONAL, LEGAL, ETHICAL AND SOCIAL ISSUES (PLES)

The use of machine learning for cardiovascular disease prediction raises many professional, legal, ethical, and social considerations. Taking these aspects into account is crucial to ensure that the research adheres to the industry standards, legal practices and ethical concerns. This project will adhere to data protection regulations, uphold ethical standards in model development, and consider the broader societal impact to promote responsible AI in cardiovascular disease prediction.

A.1 Professional

The project adheres to the British Computer Society's code of conduct, ensuring integrity and professionalism. It provides code readability and proper maintenance of code. This study aligns with best practices in software development and data science, adhering to high standards in programming and research methodology. All code written will be documented. The data used is sourced from reputable and open-source databases. No personal or sensitive information will be collected during the course of this project. Open-source Python libraries like Scikit-learn, Matplotlib, and TensorFlow will be used. The project has a proper structure, adequate time for tasks, and regular meetings with the supervisor. Project risks have been considered, and appropriate mitigation strategies have been developed. Secondary data is licensed for research purposes, and ideas and information are cited and referenced. The use of any content from both online and offline sources will be cited and referenced. The project's scope is confined to developing a Machine Learning model for academic and exploratory purposes, eliminating client or customer involvement and additional professional issues.

A.2 Legal

All information used in this study including the dataset, source code, ideas and any work that is not my own has been given due credit by citing and referencing. All research papers or articles are either open-source or have been granted permission for usage. These papers have been referenced in the bibliography. The project will use open-source tools and platforms for its development.

A.3 Ethical

This project aims to develop a machine-learning model that predicts the presence of cardiovascular disease. The dataset contains human information, but the dataset is open-source and protected by open-source license rules. It has been obtained from reputable sources and has been cited in previous academic articles. Human information is anonymized and unlinked to ensure the privacy and security of people. An ethics form stating the research aims and methods has been approved by Heriot-Watt University. This ensures that the research aligns with the Data Protection Act (DPA) and ethical guidelines. As the study focuses on system performance metrics and doesn't involve human subjects, ethical data collection tools are not required.

A.4 Social

This system is used for Cardiovascular disease prediction. Hence, its social impact is large. This project is developed primarily for research and academic purposes and is not intended for immediate public use. If the project is used, all results must be consulted with a healthcare professional and must not be the sole decision-making factor. Source code access will be limited to the supervisor and the research evaluators at Heriot-Watt University who will grade the project. Hence, concerns related to misuse of the application or authorized use of source code are not applicable.

B APPENDIX: PROJECT MANAGEMENT

This chapter presents a detailed examination of the project management approaches used for this research. It begins with an overview of the project management plan, followed by a Gantt chart illustrating the project timeline and concludes with a risk analysis. This section provides an overview of the project management techniques and planning strategies adopted to ensure the successful completion of this research project, scheduled from September 2024 to May 2025.

Effective project management is necessary to structure various phases of the research including model development and documentation within a limited timeframe. The plan is carefully designed to accommodate potential challenges faced during the course of the project. Hence, it enables flexibility throughout the project.

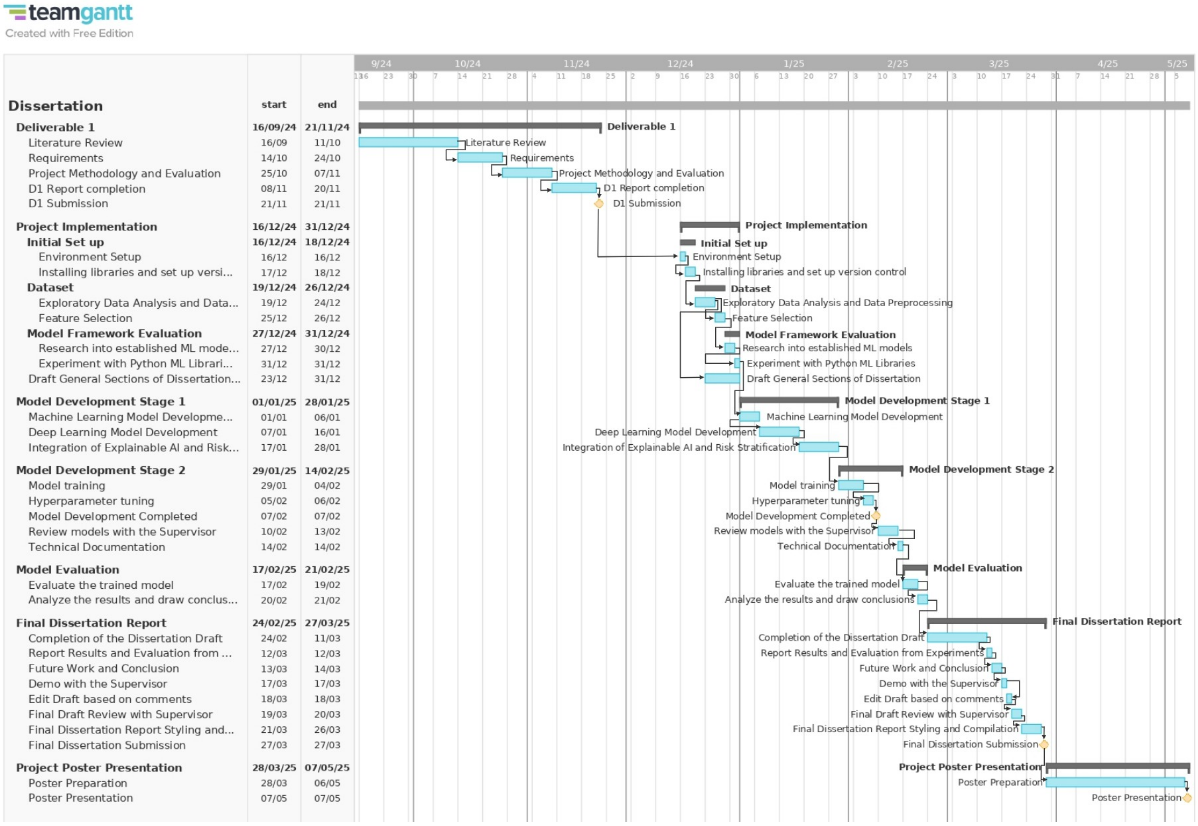
B.1 Overview of Project Management Approach

This research project uses an individual Agile approach, focusing on flexibility, iterative development, and continuous improvement. Agile and Scrum methodologies are used for team-based projects and individual research projects.

The project is divided into sprints, each with a review phase to evaluate progress and assess challenges. This structured yet flexible management style supports continuous improvement and responsiveness to new findings. Agile methodologies are particularly beneficial in healthcare research, where data complexity and critical outcomes demand a flexible approach. Agile methodologies are being widely used across various industries such as financial services, energy, healthcare, pharmaceuticals, and education [Zaidi and Jain 2024].

B.2 Project Plan

A Gantt Chart is a tool suggested by Agile project management. It is used for tracking the project's progress during its timeline. Figure 2 below shows the Gantt chart demonstrating the stages of development of the project. Big tasks have been broken down into smaller sub-tasks for easier implementation. Dependencies between tasks have been added. A buffer period has been allocated to each task to ensure its timely completion.



Dissertation	start	end
Deliverable 1	**16/09/24**	**21/11/24**
Literature Review	16/09	11/10
Requirements	14/10	24/10
Project Methodology and Evaluation	25/10	07/11
D1 Report completion	08/11	20/11
D1 Submission	21/11	21/11
Project Implementation	**16/12/24**	**31/12/24**
Initial Set up	**16/12/24**	**18/12/24**
Environment Setup	16/12	16/12
Installing libraries and set up versi...	17/12	18/12
Dataset	**19/12/24**	**26/12/24**
Exploratory Data Analysis and Data...	19/12	24/12
Feature Selection	25/12	26/12
Model Framework Evaluation	**27/12/24**	**31/12/24**
Research into established ML mode...	27/12	30/12
Experiment with Python ML Librari...	31/12	31/12
Draft General Sections of Dissertation...	23/12	31/12
Model Development Stage 1	**01/01/25**	**28/01/25**
Machine Learning Model Developme...	01/01	06/01
Deep Learning Model Development	07/01	16/01
Integration of Explainable AI and Risk...	17/01	28/01
Model Development Stage 2	**29/01/25**	**14/02/25**
Model training	29/01	04/02
Hyperparameter tuning	05/02	06/02
Model Development Completed	07/02	07/02
Review models with the Supervisor	10/02	13/02
Technical Documentation	14/02	14/02
Model Evaluation	**17/02/25**	**21/02/25**
Evaluate the trained model	17/02	19/02
Analyze the results and draw conclus...	20/02	21/02
Final Dissertation Report	**24/02/25**	**27/03/25**
Completion of the Dissertation Draft	24/02	11/03
Report Results and Evaluation from ...	12/03	12/03
Future Work and Conclusion	13/03	14/03
Demo with the Supervisor	17/03	17/03
Edit Draft based on comments	18/03	18/03
Final Draft Review with Supervisor	19/03	20/03
Final Dissertation Report Styling and...	21/03	26/03
Final Dissertation Submission	27/03	27/03
Project Poster Presentation	**28/03/25**	**07/05/25**
Poster Preparation	28/03	06/05
Poster Presentation	07/05	07/05

teamgantt
Created with Free Edition
9/24
10/24
11/24
12/24
1/25
2/25
3/25
4/25
5/25
Deliverable 1
Project Implementation
Initial Set up
Dataset
Model Framework Evaluation
Model Development Stage 1
Model Development Stage 2
Model Evaluation
Final Dissertation Report
Project Poster Presentation

Fig. 2. Gantt Chart showing the project timeline

B.3 Risk Management

Risk analysis is an important component of project management. It involves identifying potential risks, assessing their impact, and strategically developing techniques to mitigate or eliminate the risks. Risk management is essential for ensuring project success, especially in sensitive fields like disease prediction in the healthcare sector.

The following table presents the risk analysis for this project. Each identified risk is assigned a unique risk ID and classified by type, depending on whether it pertains to people, tools, or project requirements. Risks are further categorized based on their likelihood of occurrence and impact on the project, with high, medium, and low priority levels. Additionally, an action type and an action plan have been planned for managing any issues that may arise during the development process. The priority of each risk is colour-coded for the reader's clarity:

- **High Priority risks**: Risks that need immediate action.
- **Medium Priority risks**: Risks that need to be addressed but are less critical than high-priority risks.
- **Low Priority risks**: Risks that have a lower likelihood of occurrence or impact but should still be monitored.

Risk ID	Risk	Type	Likelihood	Impact	Action	Action Plan
R1	Data Accessibility	Requirements	Low	Medium	Mitigate	Use alternate datasets proposed in other research papers.
R2	Model Complexity	Requirements	Medium	High	Eliminate	Research necessary skills and topics before feature implementation and begin promptly.
R3	Poor Model Performance	Requirements	Medium	Low	Recognize	Investigations into suboptimal performance will be conducted, with findings documented for field advancement. The technique will be refined to improve effectiveness.
R4	Development Tools and Limitations in Computational Resources	Tools	High	Medium	Eliminate	Secure more powerful computational resources beforehand or resort to services/platform like Google Collab. Use alternate tools.
R5	Project Scheduling	Time	Medium	High	Eliminate	Allow sufficient time and allocate resources for tasks expected to be more time-consuming in the planned schedule.
R6	Supervisor Availability	People	Medium	Medium	Recognize	Regular check-ins with the supervisor, with contingency plans if supervisor availability is limited.

Risk ID	Risk	Type	Likelihood	Impact	Action	Action Plan
R7	Student Health	People	Low	Medium	Recognize	Prioritize personal health and ensure work-life balance to maintain productivity. Plan for breaks if needed.
R8	SHAP Implementation Complexity	Requirements	Medium	High	Eliminate	Plan time for SHAP testing, use documentation and refer to similar case studies to address implementation challenges early. Allocate extra time for SHAP setup.

Table 6. Risk Analysis

A risk matrix plotting each risk ID against its likelihood and impact is presented below.

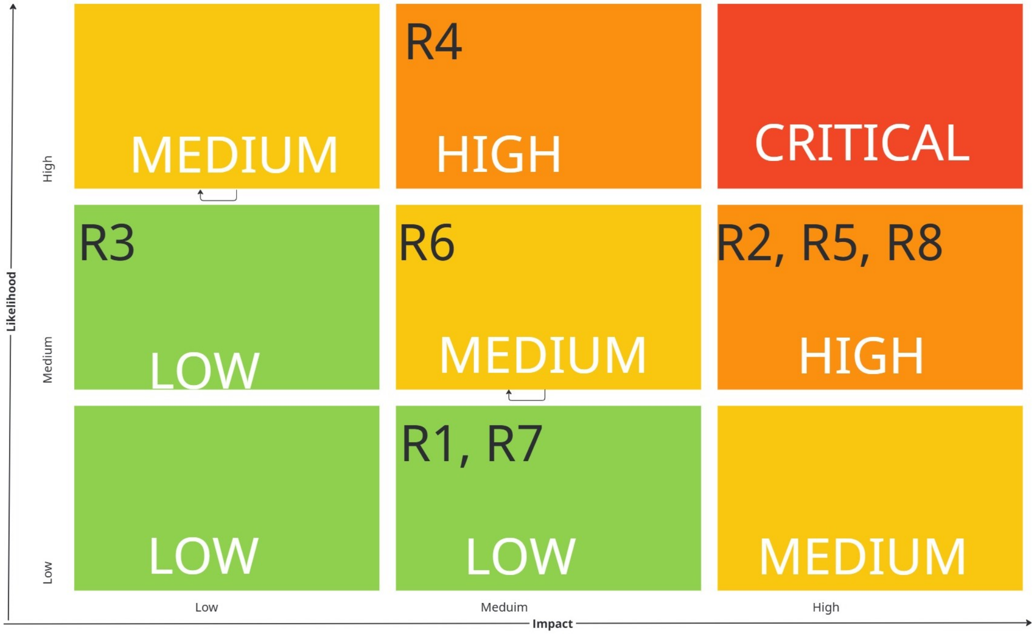


Fig. 3. Risk Matrix

C APPENDIX: PROJECT JOURNALS

C.1 Project Journal 1

Section	**Details**
Title of Dissertation and Brief Description	Title: Predicting Cardiovascular Diseases using Machine Learning Description: This project aims to advance the prediction of heart disease through the application of machine learning algorithms, with an emphasis on integrating explainable artificial intelligence (XAI) techniques and risk stratification models. By incorporating XAI, the project seeks to enhance the interpretability of predictive models, enabling both healthcare professionals and non-specialists to better understand the underlying factors contributing to the predictions and support in accurate clinical decision-making and treatment.
Communicating with Supervisor	I messaged Professor Cristina expressing interest in her supervision for the dissertation and shared the research proposal. The supervisor agreed, said she found the topic interesting and updated the allocation in the project system. We agreed on a preferred time and location for our first meeting. First meeting with Professor Cristina to get started with Dissertation (In person) – 20/09/2024. We discussed the D1 approaches, and the professor explained the details about the ethics form, and project handbook & asked me to find a dataset used in previous research papers. She suggested I start writing from the background chapter of D1 and submit a draft to her by the end of Week 4 or the start of Week 5. I presented my research done during the summer on the proposed research topic. The supervisor advised to include 5-7 research papers due to word limit constraints.

References consulted	I have researched on some papers found online through Google Scholar and Discovery. • Baghdadi, N.A. et al. (2023) ‘Advanced machine learning techniques for cardiovascular disease early detection and diagnosis’, Journal of Big Data, 10(1). doi:10.1186/s40537-023-00817-1. • Bizimana, P.C. et al. (2024) ‘Automated heart disease prediction using improved explainable learning-based technique’, Neural Computing and Applications, 36(26), pp. 16289–16318. doi:10.1007/s00521-024-09967-6. • Careervira (2023) Machine learning for beginners: A comprehensive 2023 guide, Medium. Available at: https://medium.com/@careervira.community/machine-learning-for-beginners-a-comprehensive-2023-guide-4ec02d5caab7 (Accessed: 22 September 2024). • García-Ordás, M.T. et al. (2023) ‘Heart disease risk prediction using Deep Learning techniques with feature augmentation’, Multimedia Tools and Applications, 82(20), pp. 31759–31773. doi:10.1007/s11042-023-14817-z. • Holdsworth, J. and Scapicchio, & Mark (2024) What is deep learning? IBM. Available at: https://www.ibm.com/topics/deep-learning (Accessed: 22 September 2024). • "Hospitalization for Heart Attack, Stroke, or Congestive Heart Failure among Persons with Diabetes", Special report: 2001 – 2003, New Mexico. • Kavitha, M. et al. (2021) ‘Heart disease prediction using Hybrid Machine Learning Model’, 2021 6th International Conference on Inventive Computation Technologies (ICICT) [Preprint]. doi:10.1109/icict50816.2021.9358597. • Krishnani, D. et al. (2019) ‘Prediction of coronary heart disease using supervised machine learning algorithms’, TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON) [Preprint]. doi:10.1109/tencon.2019.8929434. • Latha, C.B. and Jeeva, S.C. (2019) ‘Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques’, Informatics in Medicine Unlocked, 16, p. 100203. doi:10.1016/j.imu.2019.100203.

References used	• Mohan, S., Thirumalai, C. and Srivastava, G., 2019. Effective heart disease prediction using hybrid machine learning techniques. IEEE access, 7, pp.81542-81554. • Schlesinger, D.E. and Stultz, C.M. (2020) ‘Deep Learning for Cardiovascular Risk Stratification’, Current Treatment Options in Cardiovascular Medicine, 22(8). doi:10.1007/s11936-020-00814-0. • Shah, D., Patel, S. and Bharti, S.K. (2020) ‘Heart disease prediction using Machine Learning Techniques’, SN Computer Science, 1(6). doi:10.1007/s42979-020-00365-y. • Sudhakar, K. and Manimekalai, D.M., 2014. Study of heart disease prediction using data mining. International journal of advanced research in computer science and software engineering, 4(1), pp.1157-1160. • UCI: Heart Failure Prediction Dataset. UCI Machine Learning Repository. (Accessed: October 2024). https://archive.ics.uci.edu/dataset/45/heart+disease • Shankar et al. (2020) Heart Disease Prediction Using CNN Algorithm.
Tools explored/used	• Github • Jupyter Notebook for practising data cleaning and feature selection. • Overleaf LaTeX editor • Anaconda Python environment. • Libraries such as numpy, scikit-learn, pandas and other data mining libraries were explored. • Experimented by creating a basic model that can denote 0 and 1 for predicting heart disease. • Used TensorFlow library, Gradio, Shapley • Experimented with sklearn.metrics and evaluation strategies such as accuracy, precision, etc. • Logistic Regression and random forests

Other work carried out	• Structure of Chapters 1 and 2 is planned out. • Analysis of datasets from Kaggle and other resources. • Completed writing project journal 1. • Made an Excel spreadsheet on relevant prior works on the topic. • Explored various Python libraries available online and referred to similar implementations on GitHub. • Narrowed down the literature review to 7 research papers. • Wrote a rough draft for Chapter 2 Background and cited various sources using Harvard citation style.
Plan for the next 2 to 3 weeks	• Finalize the datasets, methodologies, research questions, and objectives of the project. • Analyze existing literature and choose 5-7 research papers for critical analysis. • Research machine learning algorithms used in prior work and identify gaps in existing literature. • Inform the supervisor about the dataset, methodologies, and objectives of the project. • Complete chapter 2 of D1 and submit the draft to the supervisor. • Modify and refine the draft using feedback. • Familiarize with deep learning and machine learning techniques. • Perform data preprocessing and understand the finalized dataset. • Submit the ethics approval form in the project system.

Overall Reflection	The dissertation work is progressing well, with significant progress made in the initial stages of the project. Communication with my supervisor has been very productive and has helped me clarify the direction of the research. The project title, objectives, and a basic research proposal have been decided, and the supervisor provided valuable feedback during our first meeting, particularly in relation to structuring the dissertation and the ethical considerations involved. I have explored various tools, datasets, and machine learning libraries, successfully building a preliminary model for heart disease prediction. The exploration of Python libraries such as NumPy, scikit-learn, and TensorFlow has helped me gain hands-on experience in data preprocessing and model development. However, selecting the most appropriate dataset for the final project was a challenge due to the diversity of available datasets. After careful consideration to ensure the chosen dataset aligns with the project's goals, I have narrowed it down to 3 datasets. The literature review is coming together, with 12 research papers identified, and I have begun drafting Chapter 2 on the background, narrowing down the review to a smaller set of key papers to meet the word limit constraint. Time management has been a challenge, particularly balancing background research with hands-on technical work and other courses. However, breaking down tasks into smaller milestones, such as completing a rough draft of Chapter 2 by the end of Week 4, has helped in staying on track. Looking ahead, my primary focus will be on finalizing the dataset, methodologies, and research questions, as well as critically analyzing 5-7 research papers for the literature review. Ensuring that my supervisor is kept updated on the project's progress and incorporating feedback promptly will also be key to refining the work. The project will be fruitful through continued research and structured planning.

Table 7. Project Journal 1 - Week 4

C.2 Project Journal 2

Section	**Details**
Title of Dissertation and Brief Description	Title: Predicting Cardiovascular Diseases using Machine Learning Description: This project aims to advance the prediction of heart disease through the application of machine learning algorithms, with an emphasis on integrating explainable artificial intelligence (XAI) techniques and risk stratification models. By incorporating XAI, the project seeks to enhance the interpretability of predictive models, enabling both healthcare professionals and non-specialists to better understand the underlying factors contributing to the predictions and support in accurate clinical decision-making and treatment. Description (Updated) This project aims to advance the prediction of heart disease through the application of machine learning algorithms, and deep learning techniques with an emphasis on integrating explainable artificial intelligence (XAI) techniques and risk stratification models. By incorporating XAI, the project seeks to enhance the interpretability of predictive models, enabling both healthcare professionals and non-specialists to better understand the underlying factors contributing to the predictions and support in accurate clinical decision-making and treatment.
Communicating with Supervisor	I messaged Professor Cristina about the Ethics form submission and the Supervisor agreed to an online meeting where she guided me with the submissions. After I submitted the Ethics form, the supervisor approved it. Expressed interest in exploring both machine learning techniques and deep learning. The supervisor encouraged me to do it provided it doesn't complicate the dissertation, owing to other courses and integration of explainable AI. I informed the supervisor that I'll be submitting a draft of some chapters towards the end of October.

References consulted	New references were consulted between Week 4 to Week 7. • Tsao, C.W. et al. (2023) 'Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association,' Circulation, 147(8). https://doi.org/10.1161/cir.0000000000001123. (Tsao et al., 2023) • Heart Disease Facts (2024). https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html. • World Health Organization: WHO (2019) Cardiovascular diseases. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (Accessed: 11 October, 2024)
Tools explored/used	• Github • Jupyter Notebook for practicing data cleaning and feature selection. • Overleaf LaTeX editor • Anaconda Python environment. • Libraries such as numpy, scikit-learn, pandas and other data mining libraries were explored. • Experimented by creating a basic model that can denote 0 and 1 for predicting heart disease. • Used TensorFlow library, Gradio, Shapley • Experimented with sklearn.metrics and evaluation strategies such as accuracy, precision, etc. • Logistic Regression and random forests

Other work carried out	• Submitted the Ethics form for approval • Combined datasets from UCI and integrated them into a single Excel file • Expressed interest in exploring deep learning models. • Made an Excel spreadsheet on relevant prior works on the topic. • Completed writing project journal 2. • Explored various python libraries available online and referred to similar implementations on GitHub. • Refined Chapters 1 and 2. • Drafted Chapters 3,4 and 5. • Completed risk analysis and the Gantt Chart for project management which are included in the Appendix. • Formatted the report in the latex editor
Plan for the next 2 to 3 weeks	• Complete chapters 3,4 and 5 of D1 and submit the draft to the supervisor for reviewing by the end of week 8. • Complete the draft of the Requirements Analysis and Methodology chapter. • Refine Functional and Non-Functional Requirements • Connect them with the objectives using a traceability matrix. • Formulate research questions for the project. • Modify and refine the draft using feedback provided by the supervisor. • Familiarize with deep learning and machine learning techniques. • Perform data preprocessing and understand the finalized dataset. • Finally, draft other sections of D1. • Draft the PLES section. • Add references. • Feedback Iterations: After getting feedback from the supervisor for the draft, add her insights and refine the report accordingly.

Overall Reflection	The deliverable 1 is progressing well. I have completed chapters 1 and 2 and refined them. Chapters 3,4 and 5 are drafted and will be refined for clarity. I converted the report from Word format to latex for better presentation. The appendix section has been drafted. It has been a little challenging with all other courses having submissions within the same week, but with a better time management plan, the D1 will successfully be completed.

Table 8. Project Journal 2 - Week 7

C.3 Project Journal 3

Section	**Details**
Title of Dissertation and Brief Description	Title: Predicting Cardiovascular Diseases using Machine Learning Description: This project aims to advance the prediction of heart disease through the application of machine learning algorithms, and deep learning techniques with an emphasis on integrating explainable artificial intelligence (XAI) techniques and risk stratification models. By incorporating XAI, the project seeks to enhance the interpretability of predictive models, enabling both healthcare professionals and non-specialists to better understand the underlying factors contributing to the predictions and support in accurate clinical decision-making and treatment.
Communicating with Supervisor	I submitted drafts of Chapters 1,2,3 and 5 along with appendices Dr Cristina. Feedback from the Supervisor: She appreciated my works and asked for some minor modifications such as adding the word Chapter in front of Chapter headings. She also asked me to modify some functional requirements and remove the validation section. I have completed those modifications. I will be submitting the final draft for review on the 12th or 13th of November 2024. My ethics form got approved and I plan to submit the D1 on Canvas a few days in advance to allow time for any last-minute adjustments.

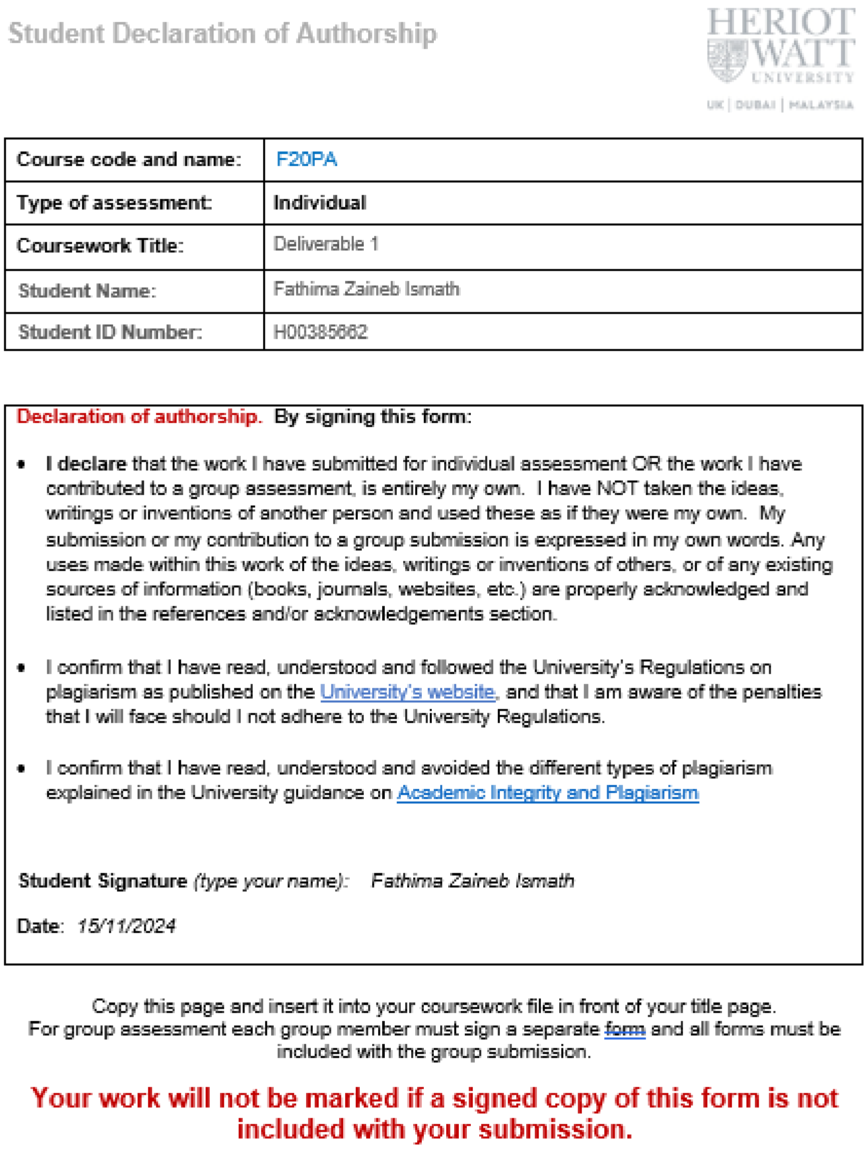
References consulted	New references were consulted from Week 7 onwards. • Multi Disease Prediction System using Random Forest Algorithm in Healthcare System. • A Systematic Review of Artificial Intelligence Models for Time-to-Event Outcome Applied in Cardiovascular Disease Risk Prediction. • Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: a systematic review. • Predicting Heart Attack through Explainable Artificial Intelligence. • Automated heart disease prediction using improved explainable learning-based technique. • Optimized Ensemble Learning Approach with Explainable AI for Improved Heart Disease Prediction. • https://link.springer.com/article/10.1007/s00521-024-09967-6 • https://www.iraj.in/journal/journal_file/journal_pdf/6-71-140490825388-92.pdf • https://ieeexplore.ieee.org/abstract/document/8971374 • https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6164626 • https://link.springer.com/article/10.1007/s11042-022-14305-w • https://link.springer.com/article/10.1007/s00500-022-07788-0?fromPaywallRec=true

Tools explored/used	• Data visualization tools • Python libraries such as numpy, scikit-learn, pandas • Experimented by creating a basic model that can denote 0 and 1 for predicting heart disease. • Pytorch • Implementation of Decision trees, random forests, • Deep learning models like perceptron and fully connected neural networks Other work carried out • Submitted the draft of Chapters 1,2,3 and 5 to the supervisor. • Incorporated supervisor's feedback to the report. • Explored various python libraries available online and referred to similar implementations on GitHub. • Finalized Chapters 1 and 2. • Completed Chapters 3,4 and 5- the draft of the Requirements Analysis and Methodology chapter. • Completed risk analysis and the Gantt Chart for project management which are included in the Appendix • Completed the PLES section • Completed grammar, and spelling checks. • Completed plagiarism check and ensured all sources consulted are referenced, • Formatted the report and enhanced its structure. • Ensured the main body doesn't exceed 20 pages. • Research on evaluation strategies for assessing model performance. • Read through other D1 provided on Canvas for inspiration.

Other work carried out	• Submit the final draft with all chapters for review on the 12th or 13th of November 2024 ensuring adequate time is left for refinements. • Feedback Iterations: After getting feedback from the supervisor for the draft, incorporate her insights and refine the report accordingly. • Create a folder that contains D1 report, Standard Declaration of Authorship form and project journals. • Once the supervisor's feedback is added and finalized, submit the final version of D1 on Canvas. • Familiarize with deep learning and machine learning techniques. • Perform data preprocessing and understand the finalized dataset.
Plan for the next 2 to 3 weeks	• Complete chapters 3,4 and 5 of D1 and submit the draft to the supervisor for reviewing by the end of week 8. • Complete the draft of the Requirements Analysis and Methodology chapter. • Refine Functional and Non-Functional Requirements • Connect them with the objectives using a traceability matrix. • Formulate research questions for the project. • Modify and refine the draft using feedback provided by the supervisor. • Familiarize with deep learning and machine learning techniques. • Perform data preprocessing and understand the finalized dataset. • Finally, draft other sections of D1. • Draft the PLES section. • Add references. • Feedback Iterations: After getting feedback from the supervisor for the draft, add her insights and refine the report accordingly.

Overall Reflection	Overall Reflection Deliverable 1 is almost completed. I have completed the main body of D1- Chapters 1,2,3,4 and 5 and I'm refining them further. The appendix section is complete with the PLES section, project management and project journals. It has been a little challenging with all other courses having submissions within the same week. But I'm on track. I will be having another detailed review of the report to ensure it adheres to the guidelines given in the project handbook. I have simultaneously gained knowledge of the technical aspects of the project. I have familiarized myself with the data preprocessing techniques and various machine learning models. I have also researched the evaluation strategies to assess how the model is performing. I will refine my report and submit it to my supervisor on 12th or 13th November for her feedback. I feel so happy. Completion of the main body of D1 marks an important milestone in the Dissertation. I'm grateful for Dr Cristina and her valuable feedback throughout this journey.

Table 9. Project Journal 3 - Week 10



Student Declaration of Authorship

HERIOT WATT UNIVERSITY
UK | DUBAI | MALAYSIA

Course code and name:	F20PA
Type of assessment:	**Individual**
Coursework Title:	Deliverable 1
Student Name:	Fathima Zaineb Ismath
Student ID Number:	H00385662

Declaration of authorship. **By signing this form:**

- **I declare** that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.
- I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.
- I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism

Student Signature *(type your name):* *Fathima Zaineb Ismath*

Date: *15/11/2024*

Copy this page and insert it into your coursework file in front of your title page.
For group assessment each group member must sign a separate form and all forms must be included with the group submission.

Your work will not be marked if a signed copy of this form is not included with your submission.

Fig. 4. Standard Declaration of Authorship