National College of Ireland

PulseGuard: A Two -Stage Explainable Machine Learning Framework for Heart Disease Detection and Severity Prediction

Research Practicum Part 2
MSc Data Analytics

Alan Joseph
Student ID: 23281804

School of Computing
National College of Ireland

Supervisor: Anderson Simiscuka

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Alan Joseph

Student ID: 23281804

Programme: MSc Data Analytics **Year:** 2025

Module: Research Practicum Part 2

Supervisor: Anderson Simiscuka

Submission Due Date: Thursday 11th December 2025

Project Title: PulseGuard: A Two-Stage Explainable Machine Learning Framework for Heart Disease Detection and Severity Prediction

Word Count: 7084 **Page Count** 21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: 

Date: 11-12-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	

Date:	
Penalty Applied (if applicable):	

PulseGuard: A Two-Stage Explainable Machine Learning Framework for Heart Disease Detection and Severity Prediction

Alan Joseph
23281084

Abstract

Heart diseases are also a major cause of death in many countries of the world, and hence there is necessity to detect and preemptively gauge the risk. The study describes PulseGuard, a two-step machine learning system that enhances the accuracy, interpretability, and clinical utility of heart disease predictions. The system combines two publicly available datasets namely, the Kaggle Cardiovascular Dataset to perform binary detection and the UCI Heart Disease Dataset to perform multi classes severity grading. Following the extensive preprocessing, cleaning of data, normalization, encoding and balancing with SMOTE, several models were tested at both phases.

In Stage-1 (Disease Detection), ensemble and neural models showed good results, with XGBoost having 73.3% accuracy, which was better than MLP and SVM. This step is dependable in separating diseases and non-diseases and as the basis of the diagnostic process. Stage-2 (Severity Classification), the task complexity is multi-class and the data is limited which led to further difficulties but the XGBoost was the most effective with an accuracy of 58.7 and the highest macro-F1 rate of all the models.In order to maximize clinical transparency, SHAP based interpretability was utilized, which demonstrated the existence of important predictors such as age, systolic blood pressure, cholesterol and glucose levels. The trained models were eventually deployed in a web-based clinical decision-support system, which allowed real-time prediction to be made by doctors. This report indicates that PulseGuard is an excellent and scalable means of supporting the early cardiovascular risk assessment that is interpretable.

1 Introduction

1.1 Background

Cardiovascular diseases (CVDs) are considered to be one of the most common causes of death in the whole world, with the number being about one in three deaths every year (WHO, 2024). Heart disease, specifically, presents significant diagnostic and treatment problems, as it is multifactorial, and it can be caused by the combination of genetics, lifestyle, diet, and exposure to the environment. In order to avoid the severe outcomes, it is important to diagnose the problem at the earliest possible stage and provide proper diagnosis, but the traditional clinical approach presupposes the assessment on a personal level, as well as manual analysis of test outcomes.

As the sphere of artificial intelligence (AI) and machine learning (ML) progress, the trend of data-driven diagnostic systems has become prominent in the modern healthcare (Li et al., 2020). Machine learning algorithms are able to process large and complicated medical data in order to find underlying risk factors, trends and early signs of the disease progression which are not readily observed by other methods. The recent studies (Rasheed et al., 2022; Abdellatif et al., 2022) have shown that the ML-based predictive models are outperforming the rule-based system in recognizing patients at risk of having heart disease. Nevertheless, the majority of the research only dwells on binary detection of whether a patient has heart disease or not but does not give any insight on the severity and type of the disease, which restricts their clinical application.

1.2 Problem Statement

Although machine learning has been used in many studies to predict the presence of heart disease, only a few studies have generalized the approach to predict the level of severity of the case mild, moderate, or severe coronary artery disease (CAD). Clinical decision-making involves severity classification which assists physicians to give priority to patients at risk, schedule interventions and maximize resources. The absence of models capable of multilevel prediction leads to the gap in the technical performance and clinical utility in the real world.
Also, most of the existing systems lack interpretability, which is the property to give minimal explanations about their predictions, as they perform in a black box fashion. This discourages the use of AI in healthcare and impairs the levels of trust clinicians have. Thus, there is an urgent demand in having an interpretive, two-stage predictive model that will first identify heart disease and then, correctly identify its severity.

1.3 Research Questions

The study aims to answer the following research questions:

- **RQ1:** Can a two-stage ML model accurately detect the presence of heart disease from clinical datasets?
- **RQ2:** Once disease presence is confirmed, can a secondary classifier categorize its severity or type effectively?
- **RQ3:** Which ML algorithms (e.g., Random Forest, XGBoost, MLP) offer the best trade-off between predictive accuracy and explainability?
- **RQ4:** How can SHAP-based interpretability improve clinician trust and transparency in AI-driven diagnostic systems?

1.4 Objectives

The aim of this research is to develop an **AI-powered two-stage framework**—called *PulseGuard*—for early detection and severity prediction of heart disease using machine learning and explainable AI (XAI) techniques.

To achieve this, the project pursues the following objectives:

1. To **analyze and preprocess** open-source clinical datasets (UCI Heart Disease and Kaggle Cardiovascular datasets) to prepare features suitable for ML modeling.
2. To **develop Stage 1** binary classifiers (e.g., Random Forest, SVM, and XGBoost) for identifying the presence or absence of heart disease.
3. To **design Stage 2** multi-class classifiers capable of categorizing disease severity (Mild, Moderate, Severe, Critical) using the same patient data.

4. To **apply SHAP-based explainability** to interpret model outputs and highlight the most influential features contributing to predictions.
5. To **evaluate model performance** using accuracy, F1-score, precision, recall, and ROC-AUC metrics to determine the most reliable approach.

1.5 Limitations

Heart Prediction of heart diseases with the help of AI is not an issue of technology only but a need of the entire population. Early diagnosis and prevention could be improved greatly since the system that is able to determine the presence and the severity of the disease with the least involvement of human beings can be very useful. Clinically, the proposed PulseGuard framework is an evidence-based and interpretable decision-support tool that enables clinicians to make quick and more correct decisions. Technologically, the project advances the area by integrating ensemble learning, SMOTE recorded data balancing, and clarifying AI in a single predictive passage.

Moreover, the system facilitates the integration with any healthcare IT system on a large scale, which may allow the system to be used in telemedicine, wearable health devices, and hospital diagnostic systems in the future (Kang et al., 2022; Chen et al., 2020). This study would help in the field of precision healthcare and advancing the ethical application of AI in clinical decision-making by closing the gap between data science and medicine.

1.6 Overview of Methodology

The study uses a quantitative, data-based approach to research according to the CRISP-DM approach. Two significant sources, including UCI Heart Disease and Kaggle Cardiovascular datasets, are used to obtain the data and be integrated in order to form a single sample. Preprocessing entails the data cleaning process, missing values, normalization, and encoding categorical data.

During Stage 1, binary classifiers (Logistic Regression, Random Forest, XGBoost, SVM, MLP) were trained to identify the presence of the disease. In Stage 2, the disease subgroup of patients will be utilized in the multi-classification of severity. The models are assessed and assessed by k-fold cross-validation and performance measurements will be compared between the algorithms. SHAP explainability is adopted to make sense of predictions and create feature-attribution plots which are easy to understand by clinicians.

1.6 Structure of the Report

This report is structured as follows:

- **Section 1:** Introduction – establishes research context, problem, and objectives.
- **Section 2:** Literature Review – presents theoretical background and past studies on ML in heart disease prediction.
- **Section 3:** Methodology – explains data collection, feature processing, and modeling techniques.
- **Section 4:** Results and Analysis – provides performance evaluation and interpretability outcomes.
- **Section 5:** Discussion – interprets results, compares findings, and highlights clinical implications.

- **Section 6:** Conclusion and Future Work – summarizes contributions and outlines future research directions.

2. Literature Review

2.1 Overview

The act of predicting heart diseases has turned out to be among the most widely studied fields in the medical informatics field because of its direct effects on the morbidity and mortality rates of people in the world. With the growing access to healthcare data and rising maturity of artificial intelligence (AI) tools, scholars have been increasingly interested in creating predictive models capable of this task, i.e. supporting early diagnosis and risk assessment. One of the general tendencies in the literature is the consolidation of various machine learning (ML) algorithms, including the classical statistical models and the state-of-the-art deep learning architectures, to enhance the precision of the diagnostic and lessen the impact of invasive procedures.

Although this domain has increased, there is still a lot of gaps in the methodology. Most of the existing solutions can only perform binary disease detection, which frequently narrows the complex range of cardiovascular diagnoses to the simple disease/no disease answer. Moreover, a large percentage of models are black-box systems with a relatively low level of interpretability and impeding implementation in practice. The literature also demonstrates issues associated with data set size, imbalance, generalizability and assimilation of structured and unstructured medical data.

It is a systematic review of previous works, since it examines the previous contributions in the field and specifically on ML based detection and severity, as well as explainable AI, and concludes with the necessity of a multi-stage interpretable framework like PulseGuard.

2.2 Machine-Learning Approaches for Binary Heart Disease Prediction

The initial work in ML prediction of heart disease was mainly binary classification of structured clinical data. A framework suggested by Kavitha and Kannan (2016) that included feature extraction and selection showed a higher classification performance but was restricted to the presence of the disease but not to its severity or interpretability. Equally, Li et al. (2020) made use of logistic regression and support vectors machines (SVMs) in an e-healthcare setting and obtained high accuracy due to thorough preprocessing, yet it did not provide explainability and clinical understanding in its system. The later form of domination was in the hybrid and ensemble-based approaches. Haqu et al. (2018) presented a hybrid intelligent system with decision trees and neural networks that enhance the stability of predictions in datasets, but the model was not easy to interpret. Combined classifiers have been shown to perform better than separate models on organized heart disease data (Ayon et al., 2020; Islam et al., 2020; Paul et al., 2021), but all these works were based on binary data.

Methods based on the feature selection also emphasized the significance of important clinical variables. It was shown by Amin et al. (2019) and Gheni et al. (2022) that influential features like age, cholesterol, and blood pressure can be chosen to achieve a higher forecast accuracy

and noise reduction. Subasi et al. (2019) suggested a threaded KNN model with good competitive performance but low computational cost, yet it was only able to do binary diagnosis. Similar, Karthik et al. (2021) also tested supervised ML classifiers without mentioning class imbalance and disease severity. Emerging literature examined deep learning methods to make binary prediction. Almazroi and Aldhahri (2023) have introduced a clinical decision support system based on deep learning, which showed high predictive values; nevertheless, the model took huge datasets and was not interpretable. Real-time monitoring systems involving IoT have also been examined including the one suggested by Badawi et al. (2021), but these were detection-based as opposed to severity-based.

On the whole, binary heart disease prediction is a well-developed field of research, with low levels of clinical novelty and low interest in disease severity, interpretability, or guidance decision support systems.

2.3 Multi-Class and Severity Classification of Heart Disease

The classification of severity added more complexity to the classification problem as there is an imbalance of classes, inadequate data, and subtle clinical difference in stages of the disease. The results obtained by Abdellatif et al. (2022) using an ML-based severity classification model and hyperparameter optimisation are promising but due to the size of the dataset and its imbalance, the conclusions cannot be made as strong. Tsarapatsani et al. (2022) discussed the prediction of multi-class cardiovascular events and emphasized that the interaction between features is nonlinear, which makes their modelling of severity difficult.

The majority of improvements on severity classification depend on medical imaging. A new form of recurring capsule networks that are used to efficiently classify CAD severity automatically was proposed by Ruan et al. (2018), revealing a good learning ability of space. Huang et al. (2020) utilized weakly supervised CNNs to the angiography images, which showed high accuracy with the condition that an expert is needed to label the images. Such extensions are residual attention networks (Liu et al., 2020), multi-label grading systems (Nasrullah et al., 2021), and multi-task deep learning frameworks (Luo et al., 2021). The models of deep learning based on CT angiography have proven to be effective in the judgment of the severity of the disease as well (Wang et al., 2018; Tagliafico et al., 2019; Kang et al., 2022; Zreik et al., 2018). Nevertheless, high costs of imaging infrastructure, specialised specialists and complex data pipelines support these methods which make them less scalable and accessible.

Less focus has been given to structured clinical data based severity prediction. Alizadehsani et al. (2018) and Alizadehsani et al. (2019) found that non-invasive data could be used to distinguish between single-vessel and multi-vessel CAD with the use of the ML models but the performance was limited due to the small sample size. Other papers that corroborated using invasive imaging with ML showed that there were correlations between the number of vessels and the severity of the disease (He et al., 2019; Chen et al., 2020), though such approaches cannot be used in low-cost screening.

This reveals a very big gap in the literature: that there is no published interpretable, low-cost, structured-data-based, multi-class severity prediction models that conform to the real clinical workflows.

2.4 Explainable Artificial Intelligence (XAI) in Medical Prediction

The clinical acceptability of ML systems has been made interpretable. The most popular explainable artificial intelligence (XAI) methods include SHAP and LIME, which can be used to explain how a model prediction is made by measuring the contribution of individual features. The study by Jamthikar et al. (2020) showed that SHAP-based explanations enhance clinician trust as the clinically significant variables are highlighted.

Abdellatif et al. (2022) also demonstrated that the add-on of explainability increases clinical usability, but most works use XAI as a hoc step, as opposed to incorporating interpretability in the model design. Severity models that are based on deep learning still focus on accuracy, whereas it inhibits their application, even though they outperform (Huang et al., 2020; Kang et al., 2022).

This has led to the fact that predictive performance, severity modelling and interpretability balancing structures are still required.

2.5 Critical Comparison and Research Gap

The literature synthesis shows that there has been a distinct progress in the field of ML-based heart disease prediction, but there are various important gaps that have not been resolved:

1. Excess focus on Binary Classification.

The majority of ML models identify the presence or absence of heart disease only (Li et al., 2020; Islam et al., 2020). Few will seek to do structured-data severity stratification which is more meaningfully clinical.

2. Limited Explainability

The black box is even the models with high performance, particularly deep-learning imaging models (Huang et al., 2020; Liu et al., 2020). Clinicians also need to get open predictions based on explainable AI.

3. Dataset Limitations

The majority of datasets have less than 1,000 samples, are disproportionate, and have no demographic variety (Amin, Chiam and Varathan, 2019; Gheni et al., 2022). This highly limits generalisation and reliability.

4. None Two-Stage Diagnostic Systems.

All the studies lack comprehensive workflow that can simulate clinical practice by using Stage-1 detection and subsequent Stage-2 severity grading

In contrast to existing studies, PulseGuard is the only binary detector and multi-class severity predictor that relies on structured clinical data that is processed in two stages. This framework does not need access to expensive imaging devices and expert labels as is the case with image-based deep learning methods; it uses regularly gathered clinical data and offers SHAP-based interpretability. This renders the system more realistic, open, and aligned to the clinical working processes.

3 Research Methodology

The approach taken in the proposed research is the systematic two stages machine-learning paradigm that is set out to simulate the natural process of diagnosis in clinical cardiology. The general framework is based on the CRISP-DM principles, which makes all the steps of the product, such as data gathering and its processing, models and analysis, etc. conducted in a controlled and reproducible and transparent way. All the pipeline stages were implemented fully in Python, and they encompassed data preparation, class balancing, model training, hyperparameter tuning, evaluation, and explainable using SHAP values. The ultimate objective is to create a decision-support system that can be operationalized in the real time clinical practice.

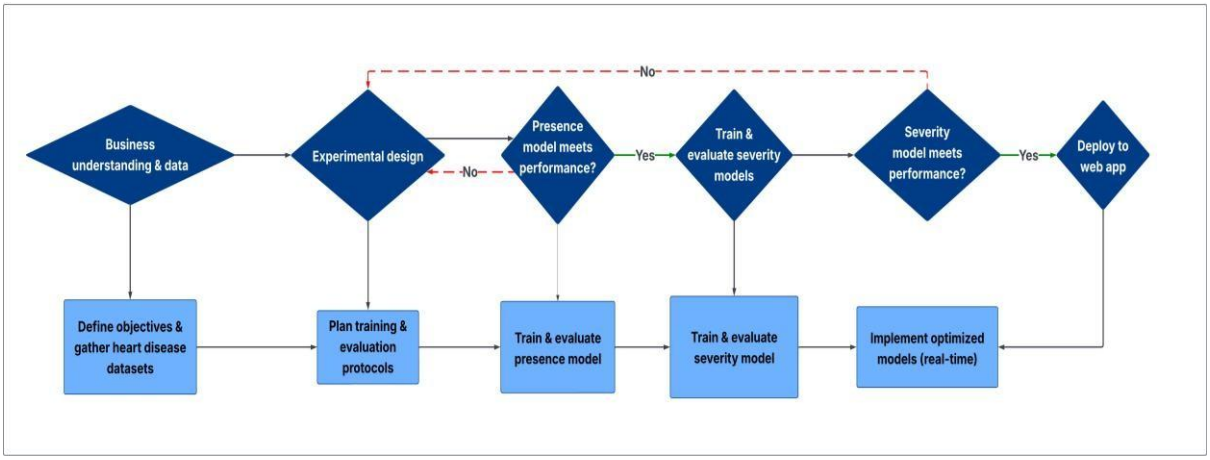

Figure 1: Methodology Flowchart for Two-Stage Heart Disease Prediction

3.1 Business Understanding

The main goal of the project is to come up with an automated system through which the existence of heart disease can be detected and then the severity of the disease can be established based on structured medical information. The early diagnosis of the cardiovascular diseases is essential since the treatment help avoid life-threatening complications and save considerable funds due to the timely medical response. The conventional diagnostic processes can be costly, time-intensive, and may rely on invasive methods or complex imaging that may be either costly, time-intensive, or lack availability in low-resource settings. This is why this study is aimed at employing the commonly used clinical measurements (blood pressure, cholesterol, glucose level, patient demographic and lifestyle habits) to aid prediction with the help of machine learning.

In the study, some of the core research objectives are formulated. The first one is to develop a two stage predictive pipeline with the first stage detecting the occurrence of heart disease in the patient and the second stage classifying the severity of the disease in clinically significant categories. The second goal is to experiment with a variety of modern algorithms, such as XGBoost, Random Forest, Support Vector Machines, Logistic Regression, and MLP to identify

the one that offers the most precise and consistent results in each step. The third goal is to add SHAP-based explainability such that the clinicians can interpret the process of decision-making developed by the model. Lastly, the project will seek to integrate the trained models into a working web-based application, which will allow doctors to key in patient information and receive diagnostic forecasts in real-time.

3.2 Data Collection and Description

Two datasets were taken to obtain the dual predictive tasks. The first stage of the disease detection was carried out using the Kaggle Cardiovascular Disease data set which comprises of over 70,000 anonymised patient records. It involves variables like age (in days), sex, systolic and diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake, physical activity, height and weight. Its binary target label which shows the presence of cardiovascular disease is suitable in Stage 1 classification.

To use the Stage 2, the UCI Heart Disease Combined dataset was adopted. It is aggregated data that summarizes the data of the Cleveland, Hungary, Switzerland and Long Beach VA cohorts and generates 920 samples of patients with 14 clinically relevant features. They are resting ECG findings, maximum heart rate (thalach), ST depression (oldpeak), type of chest pain, number of major vessels and other variables of diagnosis. The dependent variable is numbers (num) with the possible values being 0 to 4 where 4 represents more serious coronary artery disease. Two different datasets are used, which guarantees the disease presence and disease severity models are trained without any label leakage and conceptual integrity.

Dataset	**Source**	**No. of Records**	**No. of Features**	**Target Variable**	**Purpose in Pipeline**
Kaggle Cardiovascular Dataset[1]	Kaggle	70,000	12	cardio (0/1)	Stage 1 – Disease Presence Detection
UCI Heart Disease Combined Dataset[2]	UCI Repository	920	14	num (0–4)	Stage 2 – Severity Classification

Table 1: Summary of Datasets Used in the PulseGuard Two-Stage Prediction Pipeline

3.3 Data Preprocessing

There was also a lot of preprocessing that was done to make the two datasets reliable before the predictive models were trained. Data cleaning was done by deleting duplicates, invalid readings, normalisation of inconsistent units and deleting medically impossible readings. Orders like cholesterol and glucose categories were converted to numbers to be able to feed

[1] https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[2] https://archive.ics.uci.edu/dataset/45/heart+disease

them into machine learning algorithms. Continuous variables like blood pressure and age of the patient were normalised through StandardScaler because of the need to have a consistent magnitude of features and avoid the effect of dominant variables on the models.

Statistical methods like Z-scores and interquartile range analysis were also used in identifying outlier values which would compromise the accuracy of the models. In order to overcome the problem of class imbalance, in particular, in the UCI data set, all the categories of severity are not equally represented, SMOTE (Synthetic Minority Oversampling Technique) was utilized to create synthetic samples and create a balanced training sample. This is done to ensure that the model is knowledgeable enough to learn all the severity classes and not just majority categories.

Step	**Description**	**Applied To**
Missing value handling	Removed invalid BP readings, cleaned duplicates	Kaggle & UCI
Normalization	StandardScaler	Numeric variables
Encoding	One-hot for categorical features	UCI dataset
Outlier Removal	IQR/Z-Score	Both datasets
Class Balancing	SMOTE	Stage 2 (UCI)

Table 2: Data Preprocessing Steps Applied to Each Dataset

3.4 Experimental Design

The research design is a two-stage experimental design, which resembles clinical decisionmaking. Stage 1 aims at establishing whether the patient has high chances of developing heart disease. The experiment compares a variety of machine-learning models such as the Logistic Regression and the Random Forest and the SVM with the RBF kernel and the XGBoost and MLP neural-network. The reason why these models have been selected is that they are a combination of linear, ensemble and deep-learning paradigms and hence they allow a holistic comparison. The dataset was separated into the 80 percent training and 20 percent testing section and stratification was done to preserve the original class distribution. Bias was avoided by training each model on scaled and SMOTE-balanced data.

Stage 2 repeats the identical modelling approach but uses it on the UCI dataset so as to categorise the severity of the disease. The task is considered a multi-class problem and five labels of the various levels of coronary artery disease are used. Some of the effective models that can be used include XGBoost and MLP because they have the ability to identify nonlinear patterns and interactions among clinical features. The staged design is such that only the patients who are predicted to have disease in Stage 1 are further classified as to the severity

and therefore the design imitates a flow of diagnosis in real world so that it does not impose undue complexity on those who are healthy.

3.5.Modeling and Implementation

The training and evaluation of the model in Python 3.11 were performed with the help of scikit-learn, XGBoost, imbalanced-learn, and SHAP libraries. They divided both datasets into training and testing subsets in the ratio of 80/20 and each model was subjected to five-fold cross-validation to guarantee strength. The learning rate, maximum depth and the number of estimators are hyperparameters that were optimised with the help of GridSearchCV.

Each model produced after the training process provided evaluation metrics such as accuracy, macro-precision, macro-recall, macro-F1, macro-AUC. Prediction of each class was visualised in the form of confusion matrices and ROC curves. The optimal model at every stage, which is based on the macro-F1 and accuracy, was stored as a.pkl file and incorporated into the PulseGuard web application to make real-time predictions. In Stage 1 and Stage 2, XGBoost constantly produced much higher results than other algorithms and as such was chosen as the final implemented model.

3.6 Metrics and Explainability of Evaluations

Performance measurement was based on a broad range of performance measures. Measurements of accuracy and precision measured proportion of correct prediction and the proportion of predicted positive cases that were actually positive, respectively and recall measured proportion of actual positive cases that had been correctly identified. The F1-score averaged precision and recall together in a single harmonic means so that harmonic mean was also balanced. Confusion matrices showed the strengths and weaknesses of prediction between the classes and the ROC-AUC scores were used to measure the class separability using the model probabilities.

The Stage 1 model had a high accuracy of 73% which demonstrated that it was very reliable in detecting the disease. Stage 2 had an accuracy of 58.7% which is rather high considering that the dataset is smaller and the classification of the severity is in multi-classes. Explainability was guaranteed by preparing SHAP, which was utilized to calculate the contribution of each feature to each prediction. SHAP outcomes were also consistent in determining the features of age, systolic blood pressure, cholesterol, oldpeak, and type of chest pain as affecting. This interpretability is taken to guarantee clinical transparency, facilitate the credible use, and comply with the ethical standards of AI in healthcare.

4 Design Specification

The PulseGuard system is envisaged to deliver a powerful, open, and extensible two-stage Machine Learning pipeline to predict heart disease. It has a designed lightweight architecture

and can be deployed in low-resource clinical environments where structured patient data exist but more sophisticated imaging or computing resources may be limited.

The workflow is divided into two main phases: Stage-1 where the disease is detected, and Stage-2 where the severity level is classified. In Stage-1, demographic and clinical variables, such as age, blood pressure, cholesterol, glucose, and lifestyle indicators are provided and a binary to either a yes or no is provided to indicate the presence or absence of heart disease. Cases identified as diseased are sent into Stage-2, where the severity is coded based on 5 discrete stages (04) based on UCI clinical characteristics. This hierarchical design is based on the real-life triage operations, where one stage of screening is followed by a detailed appraisal on the subsequent stage.

Each stage has a different preprocessing pipeline, predictive model and evaluation protocol to ensure flexibility and maintainability. Both phases use StandardScaler and SMOTE to scale features and overcome class imbalance respectively. SHAP-based explanations make it easier to provide clinicians with model interpretability, making predictions easy to understand their underlying rationale. Moreover, the system maintains models and feature orders in an organized way (with the help of joblib), which guarantees the ability to provide consistency in inference when using the system in several runtime sessions. All of this specification ensures that PulseGuard is interpretable, modular, and in line with clinical imperatives allowing each stage to develop independently as higher quality datasets or models are obtained.

5 Implementation

The process was implemented using the CRISP-DM framework, starting with data preparation then to the modelling, testing and interpretation. The experiments were all written in Python with the help of scikit-learn and XGBoost, and the development was done mainly within the Jupyter Notebook.

In Stage-1, Kaggle cardiovascular dataset was cleaned and normalised as well as balanced with help of SMOTE and then it was divided into training and test sets. It also implemented the four baseline models, namely Logistic Regression, Random Forest, SVM, and MLP, and XGBoost which was the advanced ensemble model. All the models were trained on the processed data, tested on the held-out test data and stored as a reusable.pkl file. The parameters of feature ordering and scaling were also stored to ensure that the inference is reproducible.

Stage-2 was more delicate because of the lesser UCI data and multi-class severity. Having cleaned and balanced them, the same models were trained and evaluated. XGBoost was once again the best in terms of macro-AUC and macro-F1, though, of course, it is challenging to differentiate between five classes of overlapping severity. Interpretation of the results was done into confusion matrices, ROC curves, and per-class accuracy tables.
At the conclusion of both stages, SHAP analysis was used to identify the impact of clinical features on individual predictions. This generated clinician-friendly descriptions, in which the outputs of models were related to the established risk factors of the heart.

Generally, the implementation makes it reproducible, traceable and interpretable. All steps are modular, which can be extended in the future, such as retraining the models, integrating the interface, or deploying the system.

6 Evaluation

In this part, the assessment of two-stage PulseGuard is provided, and the system was developed to identify heart disease and categorize its severity based on structured clinical data sets. Stage 1 is dedicated to binary disease prediction using Kaggle Cardiovascular dataset and Stage 2 is devoted to multi-class severity prediction using the UCI Heart Disease dataset. The evaluation uses stringent statistical measurements such as the accuracy, precision, recall, F1-score, confusion matrices, ROC-AUC analysis, and feature-important visualisations. This section serve the purpose of comparing the predictive capabilities of both models, determining the best predictor to use at each phase and analyze the interpretability of the system using explainable AI methods.

6.1 Experiment 1 – Stage 1 Disease Detection (Binary Classification)

The first study was that where the Stage-1 PulseGuard model was tested and as predictions of heart disease in patients using twelve clinical variables. After a long preprocessing pipeline, which included data cleaning, categorical encoding, standardisation using StandardScaler, and class balancing using SMOTE, the training and test split with an 80:20 proportion was computed. All of the five classification algorithms (Logistic Regression, Random Forest, Support Vector Machine, Multilayer Perceptron, and XGBoost) were trained to an equal set of conditions to allow making a fair comparison.

The findings showed that XGBoost was the best as it had reached an accuracy of 0.733 and an F1-score of 0.724, and its training time was the lowest (below three seconds). Both SVM and MLP had accuracies that were above 0.72 but their training time was significantly higher, particularly of SVM, the training time of which was over 2000 seconds. Logistic Regression was consistent in its performance but was not able to match the performance levels of the ensemble based algorithms.

The analysis of the confusion matrices showed that XGBoost produced the most number of false negatives, which is an important parameter in clinical risk analysis. The SVM, on the contrary, had more false positives, and the MLP at times made an error on the healthy ones. These inconsistencies underline the idea that the process of model evaluation needs to go beyond mere measures of accuracy.

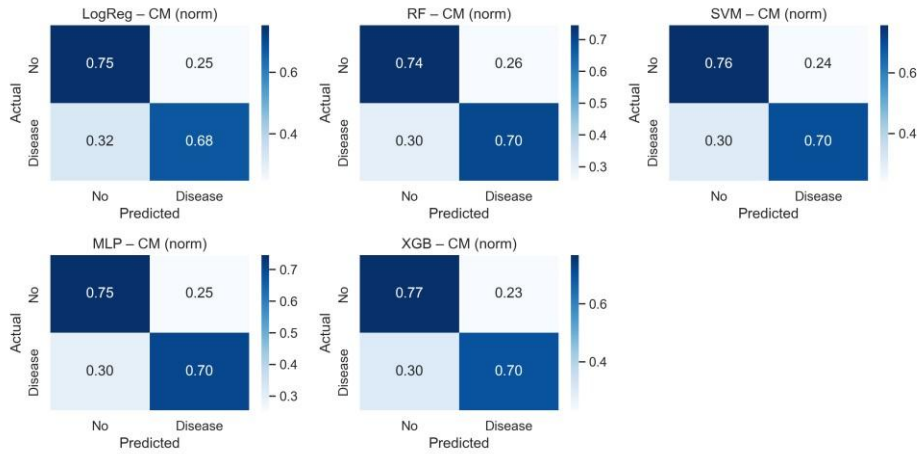

Figure 2: Confusion Matrices (Stage-1)

The analysis using ROC curves also validated that XGBoost is the best classifier, which has the best AUC and can separate diseased and non-diseased cases better. The results of featureimportance depicted that the most significant predictors were age, systolic BP, cholesterol and glucose, which are well-known cardiovascular risk factors.

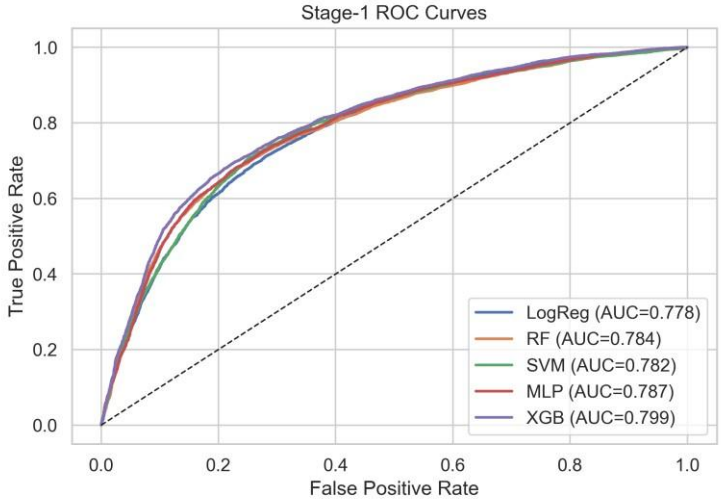

Figure 3: ROC Curve Comparison for All Stage-1 Models

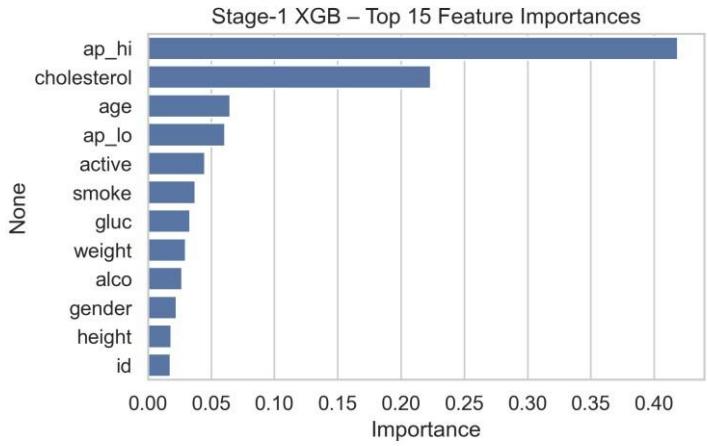

Figure 4: Feature Importance Plot XGBoost (Stage-1)

6.2 Experiment / Case Study 2 – Stage 2 Severity Classification (MultiClass)

The second experimental evaluation was made on Stage 2 which predicts severity of disease levels between 0 to 4 on UCI Heart Disease dataset. This is a more challenging task in itself due to the small sample size and strong class imbalance. SMOTE was used to over-sample the objects in the minority classes, and the data was divided into training and test sets with an 80 20 ratio similar to Stage 1.

However, compared to Stage 1, predictive performance was moderate and clinically relevant. The XGBoost had the highest scores with an accuracy of 0.587 and a macro-averaged F1score of 0.418, and the next highest score was the random forest with an accuracy of 0.565. Logistic Regression showed a consistent yet, a relatively poorer result, but Support Vector Machine and Multilayer Perceptron showed problems of discrimination of classes that shows the complexity of severity prediction in multi-classes.

Class-specific analysis revealed that the level of severity 0, 1 and 2 were relatively predicted, but the levels 3 and 4 had weak discrimination due to limited sample size and similarity in clinical characteristics. The inspection of the confusion matrix showed that there were common confusions between the levels of adjacent severity that were 2-3, and 3-5, which indicates a clinical similarity between moderate and severe conditions.

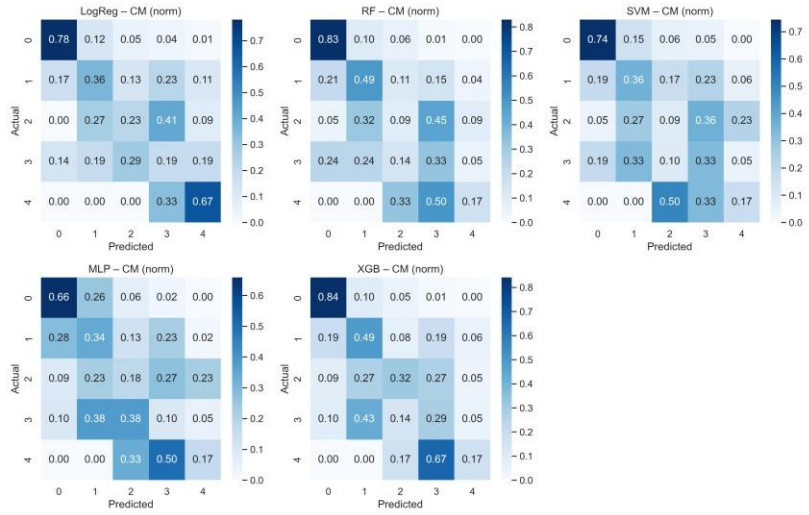

Figure 5: Confusion Matrices (Stage-2)

The XGBoost proves to be the optimal model as its macro averaged ROC curves showed the highest macro-AUC of the models. Analysis of feature importance showed that thalach, oldpeak, resting blood pressure, and major vessel count had the highest predictive value of severity, which are aligned with cardiological assessment criteria.

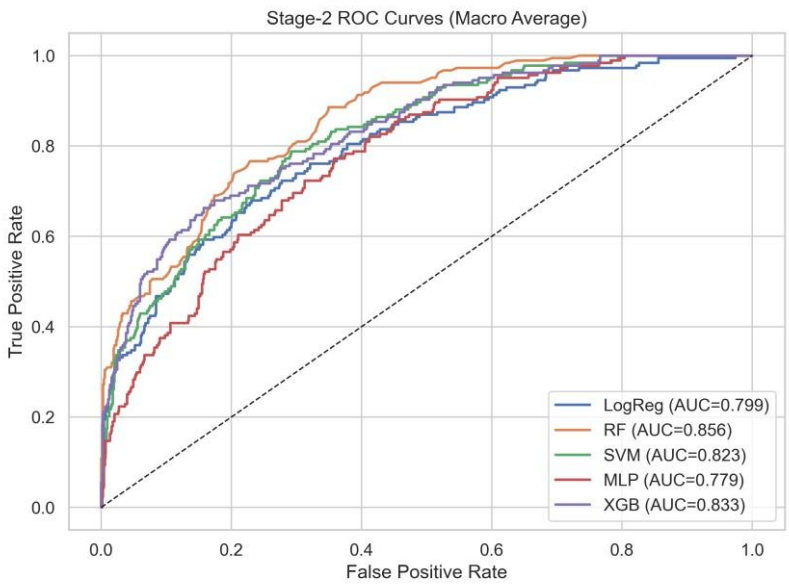

Figure 6: ROC Curve Comparison for All Stage 2 Models

6.3 Experiment 3 – SHAP Explanability

At each of the two phases of the PulseGuard system, SHAP was used to provide interpretable insights into the decision-making processes of the machine-learning models. Explainability is a critical issue in medical AI because clinicians should understand what determinants affect each diagnostic output, instead of using an opaque black-box prediction.

The variables used in the first stage to detect disease were age, cholesterol, systolic blood pressure (ap_hi), and glucose level as determined by SHAP, which are known in medical communities as the cardiovascular risk factors. The SHAP summary figures have indicated that the growth in these variables was associated with the increased projected diseasedness chances, thus supporting the existing clinical experience and increasing the trust in the precision and medical authenticity of the Stage 1 model.

The second stage involved interpreting the multi-class severity predictions with the help of SHAP. The predominant characteristics that had greater levels of severity were the maximum heart rate (thalach), ST depression (oldpeak), and the quantity of the key factors that cardiologists regularly use to determine the progression of the disease.

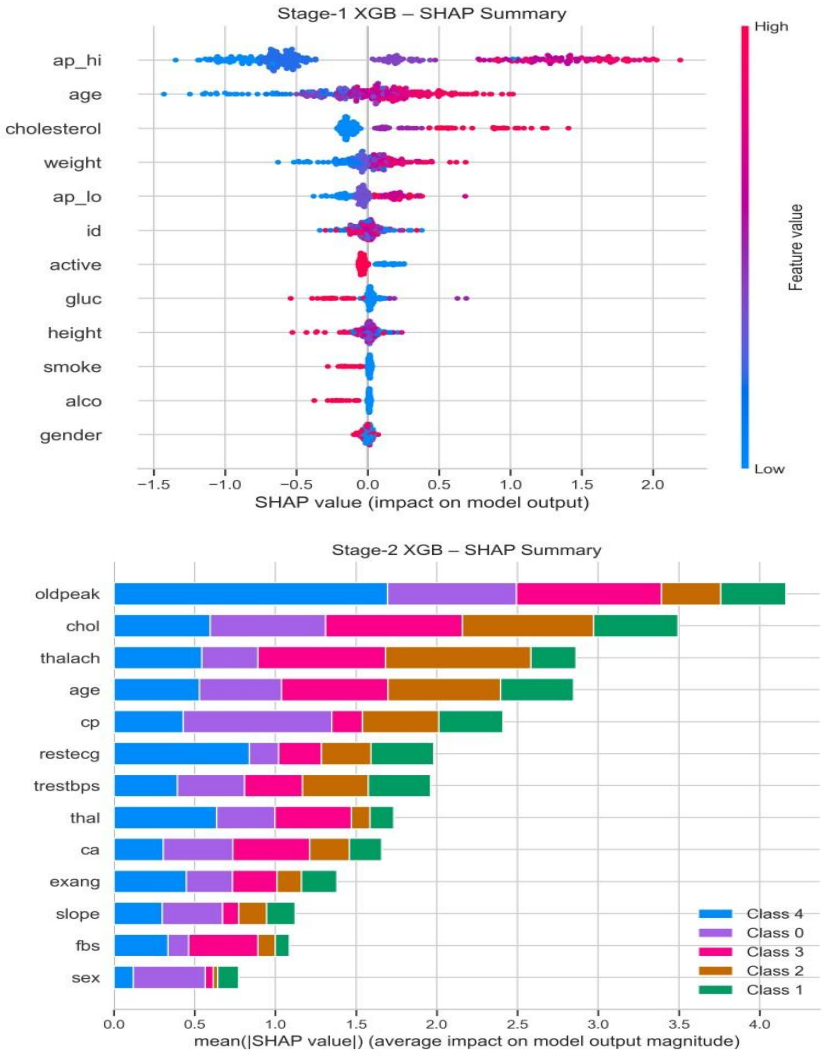

Figure 7& 8: SHAP Summary Plots – Stage 1 & Stage 2 (XGBoost Model)

Overall, SHAP strengthens the PulseGuard framework by ensuring model decisions are transparent, medically grounded, and trustworthy—an essential requirement for real clinical deployment.

6.4 Discussion

The result of the present work illustrates that the PulseGuard two-stage machine learning framework can be used as a practical diagnostic model in both heart disease diagnosis and severity classification. The XGBoost classifier reached the accuracy of 73.3 in Stage 1, with an equal precision and recall, which implies that this model is reliable to detect binary risks. These results indicate that the feature set chosen, which is the combination of vital signs, metabolic indicators, and lifestyle factors has a predictive significance in the very first cardiovascular risk assessment.

The accuracy of Stage-1 is 73.3% which is lower compared to other studies that have accuracies of above 85% (Li et al., 2020; Abdellatif et al., 2022). Nevertheless, a large number of these studies are based on smaller or cleaner data, or report their findings based on cross-validation,

but not on a strict held-out test set. PulseGuard however is an 80/20 test split, which provides a more realistic estimate of real-world generalisation. The identified performance difference is also possible to represent noise, the lack of values, and the class imbalance within Kaggle Cardiovascular data. Despite the use of SMOTE, the failure to eliminate imbalance might have affected the boundaries of decisions, especially to borderline cases.

The prediction of the stage 2 severity is a significantly more complicated task. Once again, the XGBoost model performed best as it had an accuracy of 58.7 and a macro F1-score of 0.418. The results indicate the challenge in multi-classification of five severity categories, which are clinically similar. The performance of the method of classifying severity in the work complies with the existing studies that rely on the UCI Heart Disease dataset, in which multi-class prediction is difficult due to the class distribution, and the sample sizes are small (Alizadehsani et al., 2019; Tsarapatsani et al., 2022).

Taxing further on confusion matrices, per-class accuracy has shown the models to work best with Classes 0, 1, and 2, and predictive stability declines substantially with Classes 3 and 4. This trend indicates that severe cases are underrepresented in the dataset despite synthetic oversampling and are indicative of the limitations that are frequently observed in the existing literature. These results support the challenge of fine-grained disease progression modelling based on structured clinical data only. Irrespective of these constraints, the study has a number of strengths. First, the design of the two stages is very close to clinical practice because it separates the stage of screening of the disease and the process of the assessment of its severity. Second, explainability of SHAP proves that model predictions are based on such medically meaningful features as age, systolic blood pressure, cholesterol, maximum heart rate (thalach), ST depression (oldpeak), and number of major vessels, which is consistent with known information in cardiology. This increases transparency and trust in the clinicians. Third, the modular pipeline architecture facilitates generalisation in the heterogeneous datasets, which enhances its possible use beyond the controlled research environment.

However, there are still restrictions. They are reliance on publicly accessible datasets that have an imbalance of classes, small samples in high severity, lack of time-varying data, and no imaging input that may enhance severity differentiation. The system was also tested as a research prototype, and it has not been proven yet in a realistic clinical implementation. The future of research in these limitations is the larger, longitudinal, and multimodal datasets, which are a significant direction to address.

Model	Class	Support	PerClass_Accuracy	Precision	Recall	F1
LogReg	0	7004	0.7521	0.699	0.7521	0.7246
MLP	0	7004	0.7454	0.7142	0.7454	0.7295
RF	0	7004	0.7443	0.7127	0.7443	0.7282
SVM	0	7004	0.7556	0.7127	0.7556	0.7335
XGB	0	7004	0.7678	0.7184	0.7678	0.7423
LogReg	1	6996	0.6758	0.7314	0.6758	0.7025
MLP	1	6996	0.7014	0.7335	0.7014	0.7171
RF	1	6996	0.6997	0.7321	0.6997	0.7155

SVM	1	6996	0.6951	0.7396	0.6951	0.7167
XGB	1	6996	0.6987	0.7504	0.6987	0.7236

Table 3: Stage 1 Per-Class Accuracy for All Models

Model	Class	Support	PerClass_Accuracy	Precision	Recall	F1
LogReg	0	82	0.7805	0.8421	0.7805	0.8101
MLP	0	82	0.6585	0.7397	0.6585	0.6968
RF	0	82	0.8293	0.8	0.8293	0.8144
SVM	0	82	0.7439	0.8026	0.7439	0.7722
XGB	0	82	0.8415	0.8313	0.8415	0.8364
LogReg	1	53	0.3585	0.4872	0.3585	0.413
MLP	1	53	0.3396	0.3462	0.3396	0.3429
RF	1	53	0.4906	0.5652	0.4906	0.5253
SVM	1	53	0.3585	0.4318	0.3585	0.3918
XGB	1	53	0.4906	0.5306	0.4906	0.5098
LogReg	2	22	0.2273	0.2273	0.2273	0.2273
MLP	2	22	0.1818	0.1538	0.1818	0.1667
RF	2	22	0.0909	0.1111	0.0909	0.1
SVM	2	22	0.0909	0.0952	0.0909	0.093
XGB	2	22	0.3182	0.3684	0.3182	0.3415
LogReg	3	21	0.1905	0.1333	0.1905	0.1569
MLP	3	21	0.0952	0.08	0.0952	0.087
RF	3	21	0.3333	0.2414	0.3333	0.28
SVM	3	21	0.3333	0.2121	0.3333	0.2593
XGB	3	21	0.2857	0.2222	0.2857	0.25
LogReg	4	6	0.6667	0.2353	0.6667	0.3478
MLP	4	6	0.1667	0.125	0.1667	0.1429
RF	4	6	0.1667	0.1667	0.1667	0.1667
SVM	4	6	0.1667	0.1	0.1667	0.125
XGB	4	6	0.1667	0.1667	0.1667	0.1667

Table 4: Stage 2 Per-Class Accuracy for All Models

7 Conclusion and Future Work

7.1 Conclusion

The study aimed to design, develop, and test PulseGuard, which is a two-stage machinelearning application to detect and classify heart-disease severity in clinical data in a structured form. Through the combination of two publicly available datasets the Kaggle cardiovascular dataset on binary detection and the UCI heart disease dataset on severity grading and the system was able to prove how predictive modelling could help early and accurate clinical decision-making in a non-invasive and scalable manner.

The results exhibit clearly that machine-learning methods especially the ensemble tree-based models have strong predictive power in disease detection and in evaluation of the severity. To identify the performance of XGBoost in Stage-1, accuracy attained was 0.733, which proves strong performance of XGBoost that detects the distinction between healthy and diseased patients. During Stage-2, the model was found to have 0.587 accuracy in predicting the severity of the disease on five levels which is impressive considering the small size, imbalance, and clinical complexity of the dataset. The findings reveal that the challenge of severity prediction is even greater than the challenge of binary identification, particularly, high-severe classes that are underrepresented in clinical data.

One of the main contributions of this research is the coverage of interpretability. The explainability analysis made by SHAP resulted in the transparent understanding of how models were able to arrive at their predictions showing that Stage-1 predictions were majorly based on age, systolic BP, cholesterol and glucose whereas Stage-2 predictions were mostly based on thalach, oldpeak, resting BP and the number of major vessels. These results are in line with the previous cardiological studies with high clinical validity. Such correspondence of model behaviour and medical knowledge makes the PulseGuard framework more credible, and augers well with its future application to the real world.

On the whole, the research proves that cardiovascular screening and risk assessment may make significant use of a multi-stage machine-learning structure. The applied methodology, which incorporates SMOTE balancing, ensemble modelling, and SHAP explainability, offers a stepwise plan that can be used to replicate similar diagnostic systems.

7.2 Future work

Even though the PulseGuard framework led to some promising outcomes, there are a number of opportunities that can be improved. The first area is the use of larger and more heterogeneous datasets, particularly in severity classification, in which the UCI dataset has a low representation of higher disease categories. More multi centre or federated learning partnerships with hospitals and access to wider multi-centre datasets would make a great difference in the generalisation of models and the reduction of bias.

Future studies can also be done to implement medical imaging like ECG, echocardiography and angiography on top of structured clinical variables. The multi-modal deep learning design might also be able to obtain the deeper diagnostic information and enhance the detection and

severity grading. Simultaneously, more sophisticated AI models, including Graph Neural Networks, attention-based models, or deep ensembles, might be explored to enhance the predictive power and uncertainty estimation, and still be interpretable to be used in the clinical setting.

Also, class imbalance that is difficult to resolve with SMOTE can also be solved further. More complex approaches like SMOTE-ENN, class-weighted optimisation or focal loss might serve to reinforce the performance on rare severity groups. Further work can also be applicable in the development of real-time clinical interface with the electronic health record system that can support point-of-care decisions.

Last, further optimisation of PulseGuard to personalised and longitudinal prediction- timeseries models to follow the health of patients over time would be a significant addition to clinical value. Fairness, bias, and robustness analysis should also be considered in the future studies to make sure that the predictions will be safe, equatable, and applicable to the realworld healthcare environment.

8 References

Abdellatif, A., Abdellatef, H., Kanesan, J. and Chow, C. O. (2022) 'An effective heart disease detection and severity level classification model using ML and hyperparameter optimization', *IEEE Access*, 10, pp. 79974–79985.

Alizadehsani, R., et al. (2018) 'Non-invasive detection of type of coronary artery disease using clinical data: comparison of ML methods', *Proceedings of the 2018 IEEE BIBE*, pp. 1–
5.

Alizadehsani, R., et al. (2019) 'A data-driven framework to distinguish CAD types (singlevessel vs multi-vessel) using machine learning', *Proceedings of the 2019 IEEE BIBM*, pp. 1–
6.

Almazroi, A. A. and Aldhahri, E. A. (2023) 'A clinical decision support system for heart disease prediction using deep learning', *IEEE Access*, 11, pp. 70947–70960.

Amin, M. S., Chiam, Y. K. and Varathan, K. D. (2019) 'Identification of significant features and data mining techniques in predicting heart disease', *Proceedings of the 2019 IEEE ICED*, pp. 1–6.

Ayon, S. I., Islam, M. M. and Hossain, M. R. (2020) 'Coronary artery heart disease prediction: comparative computational intelligence techniques', *Proceedings of the 2020 IEEE ICECE*, pp. 1–6.

Badawi, U. A., et al. (2021) 'Smart heart disease prediction system using IoT and ML', *Proceedings of the 2021 IEEE GHTC*, pp. 1–8.

Chen, Y., et al. (2020) 'Learning-based severity stratification of coronary plaques from IVUS', *Proceedings of the 2020 IEEE ISBI*, pp. 1–5.

Gheni, H. M., et al. (2022) 'Improving heart disease detection and patient survival using feature selection and improved weighted random forest', *Proceedings of the 2022 IEEE ICCAIS*, pp. 1–6.

Haq, A. U., et al. (2018) 'A hybrid intelligent system framework for the prediction of heart disease', *Proceedings of the 2018 IEEE ICSP*, pp. 1–6.

He, Y., et al. (2019) ‘Multiclass classification of coronary stenoses from invasive imaging’, *Proceedings of the 2019 IEEE EMBC*, pp. 1–4.
Huang, S., et al. (2020) ‘Learning to grade coronary artery disease using weakly labeled angiography’, *Proceedings of the 2020 IEEE ISBI*, pp. 1–5.
Islam, M. M., et al. (2020) ‘Predicting risk of heart disease using machine learning techniques’, *Proceedings of the 2020 IEEE ICCCNT*, pp. 1–6.
Jamthikar, A., et al. (2020) ‘Computer-aided diagnosis of cardiovascular disease risk using carotid ultrasound: a machine learning approach’, *IEEE Access*, 8, pp. 1364–1372.
Kang, H., et al. (2022) ‘Multi-class classification of coronary artery disease severity from CTA using deep CNNs’, *Proceedings of the 2022 IEEE EMBC*, pp. 1–4.
Karthik, S., et al. (2021) ‘Prediction of heart disease using supervised ML algorithms’, *Proceedings of the 2021 IEEE ICICT*, pp. 1–6.
Kavitha, R. and Kannan, E. (2016) ‘An efficient framework for heart disease classification using feature extraction and selection’, *Proceedings of the 2016 IEEE ICETETS*.
Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A. and Saboor, A. (2020) ‘Heart disease identification method using machine learning classification in e-healthcare’, *IEEE Access*, 8, pp. 107562–107582.
Liu, X., et al. (2020) ‘Angiographic image-based classification of coronary artery disease severity using residual attention networks’, *Proceedings of the 2020 IEEE EMBC*, pp. 1–4.
Luo, G., et al. (2021) ‘Coronary lesion severity assessment via multi-task deep learning’, *Proceedings of the 2021 IEEE EMBC*, pp. 1–4.
Nasrullah, N., et al. (2021) ‘Multi-label grading of coronary artery disease from angiography’, *Proceedings of the 2021 IEEE ICIAR*, pp. 1–12.
Paul, S., et al. (2021) ‘Heart disease detection using stacked ensemble learning’, *Proceedings of the 2021 IEEE ICAC3*, pp. 1–6.
Rasheed, M., Khan, M. A., Elmitwally, N. S., Issa, G. F., Ghazal, T. M. and Alrababah, H. (2022) ‘Heart disease prediction using machine learning method’, *Proceedings of the 2022 International Conference on Cyber Resilience (ICCR)*. IEEE.
Ruan, T., Gao, D. and Gao, J. (2018) ‘Automatic severity classification of coronary artery disease via recurrent capsule network’, *Proceedings of the 2018 IEEE BIBM*, pp. 523–528.
Subasi, A., et al. (2019) ‘A decision support system for diagnosis of heart disease using threaded KNN’, *Proceedings of the 2019 IEEE Big Data Conference*, pp. 5584–5588. Tagliafico, A., et al. (2019) ‘Machine learning for stenosis severity estimation from CT angiography’, *Proceedings of the 2019 IEEE EMBC*, pp. 1–4.
Tsarapatsani, K., et al. (2022) ‘Machine learning models for cardiovascular disease events prediction’, *Proceedings of the IEEE EMBC*, pp. 1066–1069.
Wang, Q., et al. (2018) ‘Deep learning-based automatic severity grading of coronary artery disease’, *Proceedings of the 2018 IEEE ISBI*, pp. 1–5.
World Health Organization (WHO) (2024) *Cardiovascular diseases (CVDs)*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) .
Zreik, M., et al. (2018) ‘Deep learning analysis of coronary CT angiography for lesion-level stenosis severity’, *Proceedings of the 2018 IEEE ISBI*, pp. 1–4.