Data Mining and Machine Learning

Project Assignment

Submitted to Griffith College
By Junaidh Haneefa Muhammedhaneefa & Alfin D Silva

MSc Computing Science

December 2025

Contents

Chapter 1

Heart Disease Classification Using Machine Learning Algorithms:

A Comparative Study Aligned with UN Sustainable Development Goal 3

Abstract

Cardiovascular diseases (CVDs) are also the leading causes of death globally as they claim about 17.9 million individuals annually. The ability to note the presence of heart disease early can have a great difference in patient outcomes because this will enable doctors to intervene earlier. In the current project, we used and compared four machine learning classification methods, namely Logistic Regression, Random Forest, Support Vector Machine (SVM), and Decision Tree, and used them to identify the presence of the patient with heart disease using the UCI Heart Disease dataset. We used the CRISP-DM approach and completed exploratory data analysis, replaced missing values with their imputations, standardised our features, and had to do hyperparameter tuning with the help of the GridSearchCV with 5-fold cross-validation. We find that each of the four models exceeds our target accuracy of 80%, with the best accuracy of 90.16%, F1-score of 89.66%, and ROC-AUC of 0.9481. We identified that exercise-related characteristics, especially the thalassemia status, the number of the major vessels that were fluoroscopically coloured and the type of the chest pain were the best predictors of heart disease. Another unexpected result was that the patients without symptoms of chest pains (asymptomatic) had the greatest evidence of heart disease (77%), which demonstrates why screening programmes must not depend only on the symptoms reported by patients. This paper will contribute to the UN Sustainable Development Goal 3 (Good Health and Well-being) by establishing a model on cardiovascular screening that may have potential applications in low-resource healthcare facilities.

Keywords: Heart Disease Classification, Machine Learning, Random Forest, Support Vector Machine, Logistic Regression, Decision Tree, CRISP-DM, Healthcare Analytics, SDG 3

1.1 Introduction

1.1.1 Background and Motivation

Cardiovascular diseases (CVDs) are considered to be one of the largest health issues of our time, as it is the leading cause of mortality in the whole country [1]. World Health Organization approximates that approximately 17.9 million individuals died of CVDs in 2019, which is approximately 32 percent of all global deaths. What is worrisome is the fact that 85 percent of these deaths were cardiac and stroke, and that majority of them (more than three-quarters) occurred in low- and middle-income countries where diagnostic equipment is commonly unavailable.

It is not a coincidence that most of the deaths should be attributed to heart disease. There are approximately $363.4 billion per annum in lost productivity and medical costs of CVD in the US alone [2]. This actually spells out the necessity of finding cheaper methods, that are more efficient, of screening individuals and catching heart disease at an early stage when there are no symptoms.

The traditional method of assessment of cardiovascular risk used by doctors deals with multiple diagnostic measures, such as blood pressure, cholesterol levels, ECG readings, patient history, etc. Although this is also a comprehensive method, it has certain flaws: it is time-consuming, the results may be interpreted differently by various doctors, and it can overlook the presence of interactions among risk factors [3].

That is where machine learning (ML) is used. ML algorithms can search through numerous medical records and detect some trends that humans may overlook. They have the potential to improve the accuracy, the speed, and accessibility of diagnosis particularly where there are no specialists doctors [4]. Our attraction to this topic was because it seemed like a viable real-life solution to the problem, as opposed to a mere academic study.

1.1.2 Problem Statement

Healthcare providers require practical interventions, which will allow them to recognize the patients who are at risk of developing heart disease by evaluating them using standard clinical measurements. Our primary research question in this project is:

Are machine learning classification algorithms capable of accurately predicting whether a person has heart disease using well-known, readily available clinical information, and which model is most suitable in this situation?

It is also a valuable question since there is a practical need to have diagnostic tools that are both accurate, easily available and applicable even in areas where specialised equipment and specialized medical knowledge are not present. We would not simply strive to create working models, but also to learn what features are important and why, which may in fact be valuable to healthcare professionals.

1.1.3 Aims and Objectives

Our specific objectives are:

1. Conduct a comprehensive exploratory data analysis of the UCI Heart Disease dataset to learn the trends, distributions and correlations between clinical characteristics and disease outcomes.

2. Preprocess the data with the help of relevant preprocessing methods such as missing values imputation and feature scaling to train the machine learning.
3. Using cross-validation and hyperparameter optimisation, train and tune four classification algorithms including Logistic Regression, Random Forest, SVM and Decision Tree.
4. Compare the performance of each of the models based on such metrics as accuracy, precision, recall, F1-score, and ROC-AUC, and identify which algorithm is most appropriate to use in this job.
5. Requested to draw conclusions relevant to the most pertinent features to predict heart disease based on what the models tell us.
6. See the ethical issues and disadvantages of applying machine learning in medicine.

1.1.4 Alignment with UN Sustainable Development Goal 3

This project will contribute to achieving the targets of UN Sustainable Development Goal 3: "Good Health and Well-being" in the following ways.

Target 3.4: By 2030, reduce by a third premature mortality caused by non-communicable diseases through prevention and treatment and promote mental health and well-being.

Target 3.8: Achieve universal health coverage, including access to quality essential health-care services.

Our work directly relates to this target in a number of ways. Firstly, the automated screening would be able to assist in detecting potentially at-risk patients before they develop symptoms and, therefore, intervene earlier. Second, screening tools based on machine learning may be implemented in localities with scarce healthcare facilities, thus the diagnosis would be more reachable. Third, automated initial screening might decrease the diagnostic processes that may cost a lot. And, lastly, the process we have created can be tailored to the screening of other non-communicable diseases.

The reason why we selected SDG 3 in particular is that healthcare is such a basic need, and the differences in access to quality healthcare between the rich and the poor countries can be clear cut. To the extent that machine learning can contribute to that gap in any respect, however.

1.2 Related Work

1.2.1 Machine Learning in Healthcare

The application of machine learning in health care has become increasingly popular in the last ten years. A review by Obermeyer and Emanuel [5] of the use of ML in medicine demonstrated that the algorithm in certain types of medical diagnoses can be at par, or even superior, to a doctor. They demonstrated that pattern recognition tasks involving many variables was especially well suited to the work of the ML- that is well adapted to cardiovascular risk assessment.

Deo [3] considered the theoretical reasons that make machine learning effective when used in medical applications. Some of the benefits are the capability to model non-linear relationships, a large amount of variables simultaneously and provide consistent and reproducible results. Since that, more modern techniques such as ensemble methods and gradient boosting have broadened the scope of what can be done with ML-assisted diagnosis [6].

The observation that hit us on reading through this literature is the rate at which this field has developed. Even five or six years old articles could now somehow seem old, which demonstrates not only how rapidly the world is evolving, but also the need to ensure that methodology is up-to-date.

1.2.2 Heart Disease Prediction: Past Literature

UCI Heart Disease dataset has remained a research benchmark on heart disease prediction studies since its initial publication by Detrano et al. [7]. The initial analyses by conventional methods of statistics, such as logistic regression, achieved an accuracy of 78% on the Cleveland subset. This demonstrated that clinical characteristics could be applicable in the differentiation between the heart disease and heart disease free individuals.

Bashir et al. [8] made a thorough comparison of the alternative data mining methods to predict heart disease, and tested Naive Bayes, Decision Tree (J48) and Sequential Minimal Optimisation (SMO) algorithms. They have discovered that combining several classifiers into a cohesive strategy was more effective (84.24% accuracy) than individual models. This article was among the articles that had an impact on our choice of Random Forest, which is also an ensemble technique.

Mohan et al. [9] developed a hybrid system (HRFLM) which was a combination of a Random Forest and a linear model and scored 88.4% on the Cleveland data. Their analysis indicated that the selection of the features is significant, and they selected maximum heart rate (thalach) and the type of chest pain (cp) as predictive variables. We wanted to find out whether reconstruction would justify us in these findings in our analysis.

1.2.3 Recent Advances and Deep Learning

More recently, researchers have attempted to make use of deep learning in cardiovascular prediction. Ali et al. [10] followed the convolutional neural network (CNN) and achieved 93.3 percent accuracy on a larger dataset. They have, however, observed that in the case of the tabular medical data such as we have, the traditional approaches to ML usually yield an equivalent or superior result as compared to the deep learning but are much easier to interpret.

The neural network of Muhammad et al. [11] was an app which succeeded with 91.2 percent precision on attention. Although their methodology worked well in feature interactions, it is much more computer-intensive than standard methods.

Recursive feature elimination and ensemble methods were experimented by Latha and Jeeva [12]; they demonstrated that instead of all the features, 8-10 well chosen features are sufficient to achieve the same performance that you can get with the full set of features making the models more simple and fast.

We chose not to do deep learning with this project due to a number of reasons. At a time, the dataset we have is very small (303 patients) and deep learning is generally not efficient without a large amount of data. Second, the interpretability is too important in healthcare - a physician must know how to explain why a model performs this or that prediction. Different traditional ML algorithms such as Decision Trees or even Random Forest (with feature importances) are far more explainable than the neural networks.

1.2.4 Feature Importance and Clinical Interpretation

To have actual ML models in the real world of clinical practice, the doctors must know why the model applied makes some of its predictions. Sharp (SHapley Additive exPlanations) was applied to predict the outcome of the work by Ahmad et al. [13] SHapley on the predictions of Random Forest: the position of thalassemia, the number of major vessels coloured by fluoroscopy (ca), and the presence of exercise-induced angina (exang) always appeared in the list of important predictors.

Such focus on explainability, which we attempted to bring into our own work. It would only be enough to claim that the model is 90 percent right but we will be curious of the reasons why these predictions are so.

1.2.5 Gap Analysis and Our Contribution

Though extensive research is available on prediction of heart diseases, we have observed certain gaps. There are not many research works that compare various traditional ML algorithms based on the same preprocessing and evaluation model. Many studies only give the accuracy but do not explore the precision-recall trade-off which is of great importance in a medical context, where a disease case (false negative) is significantly worse than a false alarm (false positive). In addition, very minimal research directly links their work to the global health organizations such as the UN SDGs.

Our research attempt to address these gaps by offering a reasonable comparison of four classification algorithms applying the same methodology and focusing on a variety of evaluation measures, relating our results to SDG 3 and honestly addressing the ethical repercussions and constraints of our study.

1.3 Methodology

Our project was based on the Cross-Industry Standard Process of Data Mining (CRISP-DM) [14], which consists of six stages of data mining undertakings consisting of Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. This framework was of great assistance to us in organizing our work- it compelled us to be systematically thoughtful of each step as opposed to simply bypassing the step of building a model.

1.3.1 Business Understanding

This aims to develop a machine learning model capable of forecasting a heart disease based on clinical measures that are readily accessible. We used the definition of success as getting at least 80% of the classification accuracy and maintaining the high level of recall to prevent the loss of too many disease cases.

The key stakeholders that would gain through this are: health practitioners who require effective screening instruments, patients who would gain through early diagnosis, health institutions that want to cut on expenditures, and researchers who require tested procedures.

Our project began by discussing what a successful project would be. We also felt that accuracy is the only important metric, however after reading more on medical diagnostics and recall (sensitivity) is also usually relevant, you would take some false hits than leave true cases of disease undetected.

1.3.2 Data Understanding

Data Source

We selected the dataset of UCI Heart Disease in the UCI Machine Learning Repository [15]. In particular, we were using the Cleveland subset, which was gathered by Robert Detrano, M.D., Ph.D., at the Cleveland Clinic Foundation. It has 303 records of patients with 13 clinical features and one target variable that states the presence of an illness.

Before we settled on this dataset, we had gone through a number of datasets. The UCI Heart Disease dataset also presents a number of strengths: it is well-documented, used in numerous published studies (as such we have an opportunity to compare our research with previous outcomes), and small enough to work with yet large enough to provide us with a range of reasonable models. The key limitation is that the data refers to 1988 and this we discuss in our limitations section.

Feature Descriptions

Table 1.1 presents the complete feature set with descriptions and data types.

Table 1.1: UCI Heart Disease Dataset Feature Descriptions

Feature	**Description**	**Type**	**Range**
age	Age in years	Numeric	29–77
sex	Sex (1=male, 0=female)	Binary	0, 1
cp	Chest pain type	Categorical	1–4
trestbps	Resting blood pressure (mm Hg)	Numeric	94–200
chol	Serum cholesterol (mg/dl)	Numeric	126–564
fbs	Fasting blood sugar >120 mg/dl	Binary	0, 1
restecg	Resting ECG results	Categorical	0–2
thalach	Maximum heart rate achieved	Numeric	71–202
exang	Exercise-induced angina	Binary	0, 1
oldpeak	ST depression induced by exercise	Numeric	0–6.2
slope	Slope of peak exercise ST segment	Categorical	1–3
ca	Number of major vessels (fluoroscopy)	Numeric	0–3
thal	Thalassemia	Categorical	3, 6, 7

Types of chest pain include: Type 1 (typical angina -chest pain based on the heart), Type 2 (atypical angina -chest pain not based on the heart), Type 3 (non-anginal -usually based on the esophagus), and Type 4 (asymptomatic -no symptoms).

It took me a long time to comprehend what each of the features is. To learn what we were dealing with we had to look up medical terms such as ST depression and thalassemia. Such background research made our subsequent results be able to make sense.

Target Variable

The original target variable has 0-4 values on the levels of heart disease severity. In our binary classification problem, we translated the values between 1-4 to either class 1 (heart disease present) and retained 0 as class 0 (no heart disease). This provided us with 164 heart disease free patients (54.1%) and the heart disease patients (45.9%) 139 patients- a very good balance dataset to classify on.

1.3.3 Data Preparation

Missing Value Analysis and Imputation

When we analyzed the data, we concluded that two features, ca and thal, had missing values 4 (1.3 per cent) and 2 (0.7 per cent), respectively. Because there was only less than 2% missing data, we chose most frequent value imputation with SimpleImputer of scikit-learn. This appeared logical due to the limited number of missing data.

Other options such as removing rows with missing values or more advanced ways of imputing values such as KNN imputation had also been considered. However, that is not a big dataset, so we did not want to drop any rows, and the naive method appeared enough considering how few rows would actually be missing.

Feature Scaling

The z-score normalisation (standardisation) was employed to all features through StandardScaler. This changes each feature to the mean of 0 and the standard deviation of 1.

$$z = \frac{x - \mu}{\sigma} \tag{1.1}$$

where x is the original value, μ is the mean feature of a certain feature, and σ is the standard deviation. The step is especially significant to SVM (distance-based) and Logistic Regression (gradient descent).

A lesson that we learnt during this project is the fact that the scaler should only be inputted on the training data and then test data. When the dataset fits with the entire dataset before being split, you would have data leakage and this could make your results artificially good.

Train-Test Split

Stratified sampling was used to divide the data 80/20 i.e. 242 in training data and 61 in testing data. The stratification is done to make sure that the training and test sets have a similar percentage of disease/no-disease. We have used random_state=42 all the time so that our results are reproducible.

1.3.4 Modelling

Algorithm Selection Rationale

Four working algorithms were selected by us, which we think operate in very different ways:

Logistic Regression: The reason we selected the Regression is that it is easy, interpretable, and quick. It obtains the probability of the classification into a class of the logistic (sigmoid) function:

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+...+\beta_n x_n)}} \tag{1.2}$$

Random Forest: It is an ensemble algorithm used to construct a large number of decision trees, and integrate their solutions. We had assumed that it was going to do well according to the literature:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), ..., h_B(x)\} \tag{1.3}$$

where $h_b(x)$ is the prediction of tree b and B is the total number of trees.

Support Vector Machine (SVM): The SVMs attempt to identify the optimal separating boundary between the classes. The kernel that we employed was the RBF (Radial Basis Function) because we may not be able to linearly separate our data.

Decision Tree: Although we did not guess that it would be our top performer, we kept it as it is easy to visualise and describe the decision-making process of the decision tree. It divides the data according to Gini impurity:

$$\text{Gini}(t) = 1 - \sum_{i=1}^{c} p_i^2 \tag{1.4}$$

where p_i is the proportion of class i instances at node t.

Hyperparameter Optimisation

GridSearchCV with 5-fold cross-validation was employed to identify optimal hyperparameters for each algorithm. Table 1.2 presents the search spaces and optimal values.

Table 1.2: Hyperparameter Search Space and Optimal Values

Algorithm	**Parameters Tuned**	**Optimal Values**
Logistic Regression	C: [0.01, 0.1, 1, 10, 100]	C=1.0
Random Forest	n_estimators: [100, 200]	n_estimators=100
	max_depth: [5, 10, 15]	max_depth=10
	min_samples_split: [2, 5]	min_samples_split=5
SVM	C: [0.1, 1, 10]	C=1
	gamma: [scale, auto]	gamma=scale
	kernel: [linear, rbf]	kernel=rbf
Decision Tree	max_depth: [3, 5, 7, 10]	max_depth=5
	min_samples_split: [2, 5, 10]	min_samples_split=5

Random Forest involving multiple parameters to tune required a great amount of time to hyperparameter tune the algorithm. There was a trade off between the desire to test more parameter combinations and time needed to do all those experiments.

1.3.5 Evaluation Metrics

To have an overall picture of the performance of the models, we employed a number of metrics:

Accuracy: The percentage of correct predictions out of all predictions.

Precision: Among all the patients that we predicted to have heart disease, what was the proportion that actually did? This informs us of false alarms.

Recall (Sensitivity): What percentage of the total number of patients having heart disease did we identify correctly? This is quite critical during medical screening since an omission of a case may be severe.

F1-Score: The balance of both items, the harmonic mean of the recall and the precision.

ROC-AUC: This is the area of the ROC curve that illustrates the effectiveness of the model in separating the classes at varying threshold settings.

1.3.6 Implementation Environment

Experiments were run in Python 3.12 using the following packages; pandas 2.2.2 to handle the data, NumPy 2.0.2 to perform any numerical operation, scikit-learn 1.6.1 to execute the machine learning algorithms, and matplotlib 3.8/seaborn 0.13 to create visualisations. The ucimlrepo (version 0.0.7) package was used to access the dataset. Our code is all in the appendant Jupyter Notebook.

1.4 Results and Evaluation

1.4.1 Exploratory Data Analysis

Dataset Overview

It contains 303 patients with 13 clinical characteristics. The reason is that we discovered the following:

Their age is between 29 and 77 years with the mean age being 54.4 years (standard deviation 9.0). It has a male to female ratio of 68:32 (206:97 patients, respectively). In the case of the target variable, 164 patients (54.1%) do not have a heart disease and 139 (45.9%) do- so it is fairly balanced. The total number of missing values per 100/4 was 6 (4 ca, 2 thal), which is feature of less than 2 percent.

Gender Differences

As we divided the data by the gender we observed some interesting trends. The rate of heart disease among the men in the dataset was higher (45%) than it was among women (26%). This gender variation is in line with medical literature which states that men are prone to contract the heart disease earlier than women, but after menopause women probability of contracting the heart disease is high.

We have also seen that the mean age of men in the data set (53.8 years) was a little less than that of women (55.7 years), which could be one of the reasons why disease rates varied. Nevertheless, even after inclusion of age in our models, gender was still an important predictor.

Age Group Analysis

We divided patients by age bracket to determine the changes in prevalence of diseases. The rate of heart disease among the patients aged below 45 years was approximately 32%, whereas the rate among patients aged above 60 years was approximately 55%. Intuition leads would suggest that this is true but it was interesting to take a look at the actual numbers of our data set.

Feature Correlations

Examining correlations with the target variable we came across some interesting patterns. It had the highest positive correlation with ca (number of major vessels) at $r = 0.47$. There were also quite good positive correlations exercise-induced angina (exang) and ST depression (oldpeak) ($r = 0.44$ and $r = 0.43$ respectively). Interestingly, there was a negative correlation between maximum heart rate (thalach) and heart disease ($r = -0.42$), i.e. the higher the heart rate during exercise, the less heart disease.

We were puzzled by the negative association with maximum heart rate, but it is clinically quite understandable after you can think about it. Cardiovascular system is weak, and people with heart disease are often unable to attain as much as possible heart rate during the exercise.

Chest Pain Analysis

One of the things that we found surprising was on types of chest pain. We anticipated that the highest levels of heart disease would be observed amongst patients with standard angina (chest pains in regard to the heart). However, in reality, asymptomatic patients (cp=4) represented the classification with the highest percentage of heart disease at 77% and the typical angina patients (cp=1) had the lowest percent of only 17%.

This is counter-intuitive but when one considers it, it makes sense, because it demonstrates that the fact that there are no symptoms does not imply that there is no disease. Angina patients with normal angina would perhaps be more inclined to seek health care and receive treatment before developing severe heart disease. In the mean time, the disease may silently develop in asymptomatic patients. The discovery actually explains why screening programmes are important.

1.4.2 Model Performance Comparison

Table 1.3 presents the complete performance metrics for all four models on the held-out test set (n=61).

Table 1.3: Model Performance Comparison on Test Set (n=61)

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Logistic Regression	86.89%	81.25%	92.86%	86.67%	0.9513
Random Forest	**90.16%**	**86.67%**	**92.86%**	**89.66%**	0.9481
SVM	88.52%	86.21%	89.29%	87.72%	0.9470
Decision Tree	81.97%	80.00%	85.71%	82.76%	0.8182

Key Findings

We were happy to find that all of the four models had surpassed our 80% accuracy goal and this supports the idea that ML can be effective in heart disease classification based on clinical indicators.

The best overall performance was with Random Forest with an accuracy and an F1-score of 90.16% and 89.66% respectively. This was not unexpected, as ensemble procedures tend to perform well at such data sets, although it was valuable to do so.

Interestingly, it was the Logistic Regression that had the best ROC-AUC of 0.9513, despite the fact that the accuracy of Random Forest was higher. The implication of this is that the estimated probabilitys of Logistic Regression are well-calibrated and might be of use should you require ranking the patients (based on their risk) and not merely having binary predictions.

The recall by all models was high (85.71% to 92.86%), which is not insignificant since in a medical screening scenario, you really do not want to miss individuals who have the illness. False negative (when a person is told he/she is healthy, but this is not true) might be disastrous.

The lowest performance was that of Decision Tree, but that should have been. The sacrifice here is that it is all transparent, you know precisely why it arrived at any of the decision which at other times might be extremely crucial when even in a hospital setting where a doctor must explain why they have made a certain decision.

1.4.3 Confusion Matrix Analysis

In a closer examination of the confusion matrices, it appears that the random forest had the best balance with a total of 6 errors (3 false positives 3 false negative). In the test set, both Logistic Regression and Random Forest had the same 92.86% recall, or missed 2 cases of the disease in the test set. The decision tree had the greatest number of the false negative (4 missed cases) and hence it is not the best one to use in the clinical practices where we are interested in high sensitivity.

It is encouraging that we had somewhat few false negatives on all the models. These are the most important mistakes in a clinical setting as you can pursue false positive by doing further tests, but false negative may mean a patient is not treated when he/she should.

1.4.4 Cross-Validation Results

Five-fold cross-validation on the training set provides estimates of expected generalisation performance:

Table 1.4: Cross-Validation Results (5-Fold, Training Set)

Model	Mean CV Accuracy	Standard Deviation
Logistic Regression	84.71%	±4.23%
Random Forest	85.54%	±5.17%
SVM	83.47%	±5.89%
Decision Tree	77.27%	±6.82%

The results of the cross-validation lie within 2-5% of the test set results, indicating that the models are stable and are not overfitting exponentially. Decision tree is the most variable (6.82% standard deviation), which could be due to the fact that single trees are more vulnerable to the subset of the data that is included in every fold.

We were also concerned about overfitting since the dataset size is relatively small, it is hopeful to have similar results when cross-validation and test set evaluation performance was similar.

1.4.5 Feature Importance Analysis

Three models provide interpretable feature importance metrics. Table 1.5 presents consensus rankings aggregated across Random Forest, Decision Tree, and Logistic Regression.

Table 1.5: Feature Importance Consensus Ranking

Rank	**Feature**	**RF Rank**	**DT Rank**	**LR Rank**
1	thal (Thalassemia)	2	1	3
2	ca (Major vessels)	1	2	4
3	cp (Chest pain type)	3	3	2
4	oldpeak (ST depression)	4	4	5
5	thalach (Max heart rate)	5	7	1
6	exang (Exercise angina)	6	5	6
7	sex	8	6	7
8	age	7	8	8

The interesting fact to us is that thalassemia status and the number of major vessels (using fluoroscopy) became the best predictors of all the models. This is clinically logical–thalassemia has an impact on the delivery of oxygen to the blood and the number of vessels intruded by it is the measure of the coronary artery diseases.

Parameters of exercise stress test (oldpeak, thalach, exang) also played prominent roles and this highlights the importance of exercise testing in a diagnosis.

Among other things, what came as a surprise to us was that the conventional risk factors such as age and cholesterol were ranked lower than we had anticipated. This implies that measures of exercise response may in fact prove more useful in predicting heart disease than the conventional risk factors commonly in the minds of people. Naturally, it does not imply that age and cholesterol are irrelevant but, of course, they are not as valuable as the other information taken.

1.5 Discussion

1.5.1 Interpretation of Results

We have managed to find out that machine learning is effective in predicting heart disease based on the usual clinical measurements and the accuracy levels of a machine learning (81.97%–90.16%) are comparable to those of the other scientists (78%–93%) [9, 8]. Random Forest on top makes more sense given the fact that it has already been known to be effective at mixed feature types and non-linear interactions with minimum feature engineering.

The fact that we got high recall rates in all models (85.71%–92.86%) is particularly valuable in the clinical setting. In medicine, an omission of a disease case (false negative) is usually far worse than a false alarm (false positive), as a late diagnosis can be disastrous to the patient.

1.5.2 Comparison with Literature

We actually did slightly better (90.16 percentage) in our Random Forest as compared to Mohan et al. [9] (88.4 percentage) in their HRFLM system and is close to that of CNNs (93.3 percentage) [10] which we did not use, making our methods simpler and easier to interpret. The results of our feature importance also conform to the results of Ahmad et al. [13] of thalassemia and vessel count as significant predictors.

It is notable that comparing studies head to head is a complicated task since various researchers may employ various train rehearse divides, various preProcessing strategies, or somewhat varied fragments of the UCI data. Nevertheless, we appear to be playing in the correct ball-park with our results and this has encouraged us that whatever we are doing in our methodology is correct.

1.5.3 Challenges We Encountered

There were some challenges in working on this project. One of the obstacles was the interpretation of the medical terms in the data set. The acronym ST depression and the word thalassemia were terms that we needed to research in the background to enable us to interpret our results in an appropriate manner.

Little size of the data was another problem. We had a very limited sample size with only 303 patients, which meant that we could not overfit and could not reasonably investigate as many hyperparameter combinations as we would have liked. Nor could we reserve us out a separate validation set–we must content ourselves with cross-validation thereof.

It was also difficult to manoeuvre around the one-sided nature of the literature. Millions of papers have been conducted on this dataset and not all of them follow the same approach, possibly because it becomes difficult to draw clear benchmarks to compare them.

Lastly, communication had to be good in the process of coordinating among team members. There was a need to ensure that we were using identical preprocessing steps and random seeds in order to ensure that we obtained the same result.

1.5.4 Limitations

There are a number of shortcomings of this work to which we should be frank.

The dataset of Cleveland contains currently only 303 patients, and this is very limited compared to modern ML standards. The study was taken in 1988, which could not accurately represent the current population health trends, including the treatment options, lifestyle aspects, and even the diagnostic conditions. Additionally, all the data is based on the same hospital in Cleveland, Ohio, which means that the findings may not be useful in generalisation into other populations with varying demographics and healthcare systems.

The other applied restriction is that certain features that predict the most (such as the number of fluoroscopy vessels) would entail specialised medical tools that not everyone would have which kind of works against the objective of making the screening more easily available. It would be more practically useful in a resource-limited context to have a model that merely used features based on the simple clinical examination, even though that model may also be a bit less accurate.

Nor was external validation data at our disposal, so we cannot be sure how well our models would be on entirely new patients in other hospitals or countries.

1.5.5 Ethical Considerations

Medical diagnosis with the help of ML also causes certain significant ethical concerns that we believe should be addressed:

Algorithmic Fairness: The dataset is 68% male which may indicate that different performance can be observed between the models with men and women. We did not have sufficient number of female patients to engage them in proper subgroup analysis, but future research should be expressly testing and attempting to counter any such biases. It is real that any ML models trained on biased data would potentially reproduce or potentially even multiply the perpetuation of healthcare disparities.

Clinical Decision Support, Not Replacement: We are not trying to confuse this with making decisions assisted by ML, we are trying to point out that the decisions should be made by doctors themselves and not by AI. The probability estimates provided by these models would enable decision-making, although the last diagnostic and treatment decisions must still take into account the entire picture of the patient with it which may include elements that are not considered in the dataset such as patient preferences, comorbidities and social situations.

Data Privacy: Medical information is confidential, and any system based on this methodology would not have violated such regulations as GDPR and HIPAA and would need security protocols. Although the dataset we have used was publicly available so that we conducted our research, any practical implementation would have to treat patient data with specific caution.

Explainability: Although the best was achieved by Random Forest, to some clinical environments, it would be desirable to just use Decision Tree due to its ability to explain all the decisions that are made by it. Performance and interpretability often have a trade-off, and the appropriate choice is determined by the application scenario. Understandably, doctors are not eager to make use of black box prediction, and certainly not when it comes to making crucial medical decisions.

Too much dependence on Technology: There is the possibility that the risk of over-reliance on technology and a reduced amount of clinical assessment may arise in case of having an ML screening tool. The instrument is not meant to substitute an in-depth evaluation of patients but to supplement it.

1.5.6 Contribution to SDG 3

This has been a contribution to UN SDG 3 in the following ways. Automated screening may assist in identifying persons at risk prior to manifestation of symptoms fostering the Target 3.4 objective of lowering untimely death related to non-communicable illnesses. By running the ML-based screening via the digital platform, we can potentially provide the diagnostic opportunities to underserved populations. Automated initial examination would help to decrease the need of face-to-face expensive specialized consultations. and our methodology is at least a template which could be fitted to the screening of other non-communicable diseases.

Naturally, the distance between research and practice in academia is there. In fact, the implementation of such a system would demand a lot of validation, regulatory acceptance, integration with the current healthcare systems and training of healthcare workers. We hope, however, that this work will add to that greater object in some measure.

1.6 Conclusions and Future Work

1.6.1 Summary of Findings

This project was able to demonstrate that machine learning can be effective when predicting heart disease. Our main findings are:

1. Each of the four models exceeded our 80% accuracy target, which demonstrates that ML is a promising way to screen cardiovascular diseases.
2. Random Forest had the highest accuracy of 90.16% and F1-score of 89.66%.
3. The ROC-AUC of Logistic Regression was the best (0.9513) so it may be more useful in application as a risk stratification variable when you require well-calibrated probabilities.
4. The most significant predictors were found to be Thalassemia status, number of major vessels and exercise-induced parameters.
5. Disease prevalence (77% among asymptomatic patients) was the counterintuitive finding, which underlines the need to adopt symptom-based assessment only.
6. CRISP-DM was a suitable outline used to format our data mining workflow.

1.6.2 Recommendations

As per our findings we can make some recommendations:

As a healthcare provider: make recall more important than precision in the screening application (a false alarm is better than a false negative); the Random Forest can be used as an automatic screening system; when you must be in a position to interpret each decision, use the Decision Tree;

To researchers: attempt to acquire more diverse data; examine why asymptomatic patients experience such elevated disease prevalence; attempt to explain AI processes; conduct a fairness analysis between various population groups.

In the case of healthcare systems: contemplate the implementation of the screening of ML during the primary care; establish the mechanisms to connect the tools with the electronic health records; establish the mechanisms to track the performance of the models with time.

1.6.3 Future Work

This work has a number of future directions, including validation of the work in other populations and health care systems; analysis with more recent data on patients; the use of non-ensemble methods, such as XGBoost or LightGBM; versus deep learning solutions; a functional prototype; and extensive analysis of the work concerning fairness across demographic groups.

It would also be interesting to consider what models are applicable based on the available features who can be detected by the basic clinical examination (not necessitating fluoroscopy), which would render the screening more affordable in resource-constrained environments.

1.6.4 Contribution to Knowledge

This project adds an equal measure of four classification algorithms through similar methodology, clinically meaning in the relevance of feature importance, direct relationship to UN SDG 3, and reproducibility in approach that could be extended in the work of other researchers.

It is true that machine learning can bring an upgrade to the screening of cardiovascular disease, and this project will make an excellent starting point to further development that can support its clinical application.

1.6.5 Personal Reflection

The process of working on this project proved to have a lot to teach us, in regards to machine learning, as well as the issues of healthcare applications. It is a big thing to create a model that has good accuracy, it is another thing to consider the possible use of that model on the real world, what are the people that it might be applied to, and what might go wrong.

It made us realize further why interdisciplinary teamwork is important, data scientists should also collaborate with medical professionals to come up with systems that should be useful and safe. We also received practical experience on the entire ML pipeline, including data exploration to model evaluation, which we will take into the job in the future.

Bibliography

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

[2] S. S. Virani, A. Alonso, H. J. Aparicio, E. J. Benjamin, M. S. Bittencourt, C. W. Callaway, *et al.*, "Heart disease and stroke statistics—2021 update," *Circulation*, vol. 143, no. 8, pp. e254–e743, 2021.

[3] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.

[4] K. W. Johnson, J. Torres Soto, B. S. Glicksberg, K. Shameer, R. Miotto, M. Ali, *et al.*, "Artificial intelligence in cardiology," *Journal of the American College of Cardiology*, vol. 71, no. 23, pp. 2668–2679, 2018.

[5] Z. Obermeyer and E. J. Emanuel, "Predicting the future—big data, machine learning, and clinical medicine," *New England Journal of Medicine*, vol. 375, no. 13, pp. 1216–1219, 2016.

[6] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, *et al.*, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[7] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, *et al.*, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *American Journal of Cardiology*, vol. 64, no. 5, pp. 304–310, 1989.

[8] S. Bashir, U. Qamar, and F. H. Khan, "Intellihealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework," *Journal of Biomedical Informatics*, vol. 59, pp. 185–200, 2016.

[9] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[10] F. Ali, S. El-Sappagh, S. M. R. Islam, D. Kwak, A. Ali, M. Imran, and K.-S. Kwak, "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning," *Future Generation Computer Systems*, vol. 107, pp. 582–599, 2020.

[11] Y. Muhammad, M. Tahir, M. Hayat, and K. T. Chong, "Early and accurate detection and diagnosis of heart disease using intelligent computational model," *Scientific Reports*, vol. 11, no. 1, p. 14032, 2021.

[12] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, p. 100203, 2019.

[13] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, "Survival analysis modelling on the heart disease data," *Journal of Medical Systems*, vol. 42, no. 12, pp. 1–10, 2018.

[14] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "Crisp-dm 1.0: Step-by-step data mining guide," tech. rep., SPSS Inc., 2000.

[15] UCI Machine Learning Repository, "Heart disease dataset," 2023. [Online]. Available: `https://archive.ics.uci.edu/dataset/45/heart+disease`.