Cardiovascular Disease Prediction Using Super Learner

OLUSANYA, Oyebanji, POPOOLA, Olusogo and SHENFIELD, Alex <http://orcid.org/0000-0002-2931-8077>

Cardiovascular Disease Prediction Using Super Learner

Oyebanji Olusanya[a]* Olusogo Popoola[a] Alex Shenfield[b]

[a]Department of Computing, Sheffield Hallam University, UK; Oyebanji.Olusanya@student.shu.ac.uk

[b]Department of Engineering and Math, Sheffield Hallam University, UK

ABSTRACT

This project addresses the global health challenge presented by cardiovascular disease (CVD), with a specific focus on Ischaemic Heart Disease (IHD), commonly known as coronary heart disease (CHD). CHD involves the narrowing of coronary arteries due to arterial plaque buildup, contributing significantly to substantial mortality rates worldwide. The project recognizes the importance of early and accurate detection of CVD, as demonstrated by clinical studies, to improve patient survival rates.

However, barriers such as the high cost of diagnosis and the financial burden of treating the disease hinder effective healthcare delivery. Existing studies often oversimplify CHD classifications, overlooking the full range of severity levels within the disease. This study seeks to overcome these limitations by employing Machine Learning (ML) algorithms, including Random Forest Classifier (RFC), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), etc, within an ML ensemble known as the Super Learner.

The research focuses on the urgency to accurately categorize patients into specific severity levels, optimizing investigation time and cost. The ML ensemble, Super Learner, combines diverse base learners to create a model that surpasses individual models, providing robust predictions across diverse scenarios. The achievements of the project include the development of a predictive model with the ability to classify CHD beyond binary classifications, achieving an unprecedented ROC score of 0.96. This performance underscores the model's potential as a valuable tool in the early diagnosis and management of CHD.

Keywords: Machine Learning, Super Learner, UCI Dataset Repository, Cardiovascular Disease

1. INTRODUCTION

Cardiovascular Disease (CVD) presents a formidable global health challenge, particularly evident in Ischaemic Heart Disease (IHD), commonly known as coronary heart disease (CHD). CHD involves the narrowing of coronary arteries due to arterial plaque buildup, contributing significantly to over 55,000 deaths in the United Kingdom in 2019 alone (ONS, 2019). On a global scale, CVD claims approximately 17.9 million lives annually (WHO, 2023).

Early and accurate detection of CVD carries immense potential to improve patient survival rates. Clinical studies, like Eckersley et al. (2016), have unequivocally demonstrated that timely identification of cardiovascular diseases can substantially enhance patient survival rates. However, barriers such as the high cost of diagnosis, especially prevalent in developing nations, and the substantial financial burden of treating this disease concern governments and families alike (Walker et al., 2016; He et al., 2022). Existing studies often overlook the full range of classifications within datasets, limiting understanding of disease severity and subsequent treatment approaches.

Recent efforts to aid in the disease prognosis employ Machine Learning (ML) algorithms like Random Forest Classifier (RFC), Support Vector Machine (SVM), and Multi-Layer Perceptron (MPL), among others. However, these studies predominantly categorize patients into two broad classes—those with and without CHD—overlooking the disease's multiple severity levels. Nonetheless, (Zghebi et al., 2021) emphasize that higher CVD severity heightens overall risk. Apart from the asymptomatic patients, the severity of CHD has been classified into two broad categories with one of the categories subdivided into three subcategories. (Shahjehan et al., 2023) classify the disease into Stable Ischemic Heart Disease (SIHD) and Acute Coronary Syndrome (ACS), with ACS encompassing ST-elevation Myocardial Infarction (STEMI), Non-ST-elevation Myocardial Infarction (NSTEMI), and unstable angina. The study recommends a 2-month rest or nitro-glycerine for patients suffering from SIHD which presents as a stable angina but with substernal chest pain that can worsen with exertion or emotional stress. Unstable angina is characterized by sudden and unexpected chest pain or pressure, even during periods of rest. It serves as a warning sign for an impending heart attack and typically arises when stable angina deteriorates. NSTEMI is a type of heart attack that can be identified through blood tests but not by an electrocardiogram (ECG). This indicates either partial

blockage of coronary arteries or a brief blockage period. On the other hand, STEMI represents a more severe heart attack. Health Providers can detect it through both blood tests and ECG. This condition occurs when blood flow to the heart is completely obstructed for an extended period, affecting a significant portion of the heart muscle (Cleveland, 2022). Distinguishing these categories is crucial, given their varying treatments and urgency levels.

The urgency to accurately categorize patients into specific severity levels arises to optimize investigation time and cost, particularly in SIHD cases. Employing the ML ensemble, Super Learner, becomes imperative. This approach combines diverse base learners, each specializing in unique data aspects or patterns, aiming to create a model that surpasses individual models and provides robust predictions across diverse scenarios.

The major contributions of this research to the body of knowledge are listed below:

- Finding and extracting a CHD-related dataset from a reliable data repository and accurately identifying the contributory features
- Identification of significant heart disease features using different feature selection methods
- Implementation of a super learner model, using five different ML models as base learners, that can efficiently segregate the various levels of the disease
- Finally, the model is evaluated using the hold-out dataset and an external dataset

2. BACKGROUND

Several machine learning specialists and data scientists have crafted models and frameworks utilizing the widely-used Cleveland Heart dataset. Despite the pioneering nature of these studies' models and systems, they all converge on a common conclusion: the binary classification of patients. These studies tend to group patients at various stages of CHD under the umbrella of simply having CHD. (Shahjehan et al., 2023) and (Cleveland, 2023) emphasize the critical importance of distinguishing between the four primary stages of this disease due to their distinct severities and treatments. The new study seeks to go beyond simply classifying the patients under the two categories and classify the patients based on the disease severity.

Authors	Research Title	Models	Result	Classification Count
Pouriyeh et.al	A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease	DT, NB, K-NN, SVM, MLP, RFC and SCRL with 10-fold CV	Accuracy: 84.15%; ROC: 83.6%	2
Ms. Ruqiya et.al	Review on Cleveland Heart Disease Dataset using Machine Learning	None	None	None
Kausar et.al	A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data	MLP, MLR, FURIA, C4.5 with 10-fold CV	Accuracy: 88.4%; ROC: None	2
Verma et.al	A Data Mining Model for Coronary Artery Disease Detection using Non-invasive Clinical Parameters	PNN, ADTree, RBFN	Accuracy: 96%; ROC: None	2
Randa El-Bialy et al	Feature Analysis of Coronary Artery Heart Disease Data Sets	C4.5, FDT	Accuracy: 78.6%; ROC: None	2
Malav et.al	Prediction of Heart Disease Using K-Means and Artificial Neural Network as Hybrid Approach to Improve Accuracy	K-means clustering, ANN	Accuracy: 93%; ROC: None	2
Burak Kolukisa	Development of Data Mining Methodologies and Machine Learning Models to Understand Cardiovascular Disease Mechanisms	KNN, LR, LDA, NB, SVM and ensemble method	Accuracy: 90.13%; ROC: 94.2%	2
Prashasti et. al	Prediction of Cardiovascular Diseases using Support Vector Machine and Bayesian Classification	SVM, NB	Accuracy: 60%; ROC: None	2
K.Subhadra et. al	Neural network-based intelligent system for predicting heart disease	MLPNN	Accuracy: 94%; ROC: None	2

Shylaja et. al	Hybrid SVM-ANN Classifier is used for Heart Disease Prediction System	Hybrid SVM-ANN	Accuracy: 88.54%; ROC: None	2
Kathleen et.al	Coronary Heart Disease Diagnosis using Deep Neural Networks	DNN	Accuracy: 83.67%; ROC: 89.22%	2
Latha et.al	Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques	Ensemble classifier	Accuracy: 85.48%; ROC: None	2
Safial et.al	Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques	LR, SVM, DNN, DT, NB, RF and KNN	Accuracy: 97.36%; ROC: None	2
Mustafa Jan et al.	Ensemble approach for developing a smart heart disease prediction system using classification algorithms	Ensemble approach	Accuracy: 98.17%; ROC: 90%	2
Sumit Sharma and Mahesh Parmar	Heart Diseases Prediction Using Deep Learning Neural Network Model	LR, KNN, SVM, NB, RF, and Hyperparameter optimization with Talos	Accuracy: 90%; ROC: 90%	2
Proposed	Cardiovascular Disease Prediction Using Super Learner	Super Learner	Accuracy: 88%; ROC:96%	5

Table 1: Studies Involving the Cleveland Heart Dataset

3. PROPOSED METHOD

This study makes use of secondary data from one of the reliable and popular sources which is known as UCI Machine Learning Repository. This data is relevant and contains all the required features (age, sex, cholesterol level, fasting blood sugar, resting blood pressure, etc) that explain the occurrence of CHD. It is made up of 303 instances and 13 features with categorical, integer, and real numbers. But more importantly, it complies with the necessary data regulations and ethical rules governing the use of data for research purposes by anonymizing sensitive data. Consequently, the collected data is processed and analyzed using Python programming. Firstly, the data is uploaded into the programming environment and its descriptive statistics are displayed to have a general idea of the whole data and each of its components. It is then pre-processed for quality and validation purposes to identify the missing data, wrong data, and outliers, and take appropriate measures to fix them. During this process, six rows containing wrong data are identified and deleted. The processed data is visualized to view the relationship among the features and form an initial opinion. The next phase is to prepare the data for modeling by ensuring imbalanced data is addressed through an oversampling technique and then perform data normalization. This leads to feature selection as not all the feature variations are significant in the predictive modeling of the disease. Hence, the features were sampled using Recursive Feature Elimination (RFE), SelectKBest, and Tree-based feature importance to identify the significant ones.

In this study, ML models such as RFC, XG Boost, SVM, MLP, and KNN are used as base learners for the super learner while RFC is the meta learner which in turn uses the k-fold cross-validation technique to estimate their performances and pick the best of them to predict the occurrence of the CHD.

3.1 Feature Selection

The dataset contains both numerical and categorical variables with some of the categorical variables made up of more than two variations. For instance, chest pain (cp) has four variations which are 1, 2, 3, and 4 representing typical angina, atypical angina, non-anginal pain, and asymptomatic respectively. In this case, one-hot encoding is employed to create dummy features to represent each of the variations as a way of converting them from categorical to numerical variables (Satya Sree et al., 2021). This is done as part of the data preprocessing and it is important because ML algorithms can only understand numerical variables. Consequently, the number of features increased from 13 to 22 which need to be subjected to significant tests to confirm their importance in predicting the disease. Hence, RFE, SelectKBest, and Tree-based feature importance are used to identify the significant ones as follows.

- Recursive Feature Elimination
  This method uses Random Forest Classifier to estimate the significance of each of the features. RFE selects 19 features as the significant ones to efficiently predict the occurrence of CHD except thal_6, restecg_ST-T_wave_abnormality and slope_downsloping. However, it is important to confirm this with the other methods.

	Feature	Feature ranking
0	age	True
1	sex	True
2	trestbps	True
3	chol	True
4	fbs	True
5	thalach	True
6	exang	True
7	oldpeak	True
8	ca	True
9	cp_asymptomatic	True
10	cp_atypical_angina	True
11	cp_non_anginal_pain	True
12	cp_typical_angina	True
13	thal_3	True
14	thal_6	False
15	thal_7	True
16	restecg_ST-T_wave_abnormality	False
17	restecg_left_ventricular_hypertrophy	True
18	restecg_normal	True
19	slope_downsloping	False
20	slope_flat	True
21	slope_upsloping	True

Table 2: Recursive Feature Elimination (RFE) Output

- SelectKBest

SelectKBest uses Pearson's Chi Square test which is a univariate method that considers separately each of the features, using P-value of 0.05 to select the significant features. This method shows that the last variables are insignificant with P-values greater than 0.05. However, the domain knowledge of the disease establishes the influence of sex in the occurrence of the disease (Hemal et al., 2016). Thus, sex is retained as one of the predictors. Nevertheless, thal_6, restecg_ST-T_wave_abnormality, and slope_downsloping are still confirmed as less significant by this method.

	Feature	Score	P-Value
5	thalach	476.214759	9.329791e-102
13	thal_3	282.795812	5.561262e-60
8	ca	200.952245	2.344975e-42
7	oldpeak	198.230199	9.023162e-42
21	slope_upsloping	196.604651	2.017446e-41
11	cp_non_anginal_pain	150.172043	1.870079e-31
10	cp_atypical_angina	118.448980	1.145260e-24
18	restecg_normal	112.837438	1.805989e-23
3	chol	75.509504	1.554693e-15
0	age	72.634909	6.301472e-15
2	trestbps	70.761593	1.567428e-14
9	cp_asymptomatic	58.354167	6.430296e-12
15	thal_7	48.515581	7.368360e-10
6	exang	45.315175	3.419048e-09
12	cp_typical_angina	37.250000	1.599834e-07
4	fbs	29.703704	5.623448e-06
20	slope_flat	27.050251	1.941925e-05
17	restecg_left_ventricular_hypertrophy	25.083095	4.841094e-05
14	thal_6	10.190476	3.733850e-02
19	slope_downsloping	8.285714	8.165532e-02
1	sex	5.338235	2.543189e-01
16	restecg_ST-T_wave_abnormality	1.000000	9.097960e-01

Table 3: SelectKBest Output

- Tree-based Feature Importance

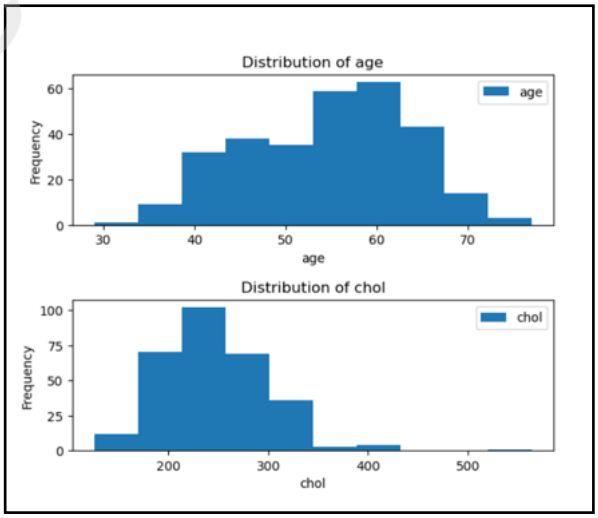
This method arranges in descending the features according to their importance with the three already marked insignificant features among the least four. Thus, this study does not consider the blood disorder due to fixed defect (thal_6), resting ECG measurement with ST-T wave abnormality (restecg_ST-T_wave_abnormality), and the peak exercise in the downward sloping category (slope_downsloping) in the prediction of the disease.

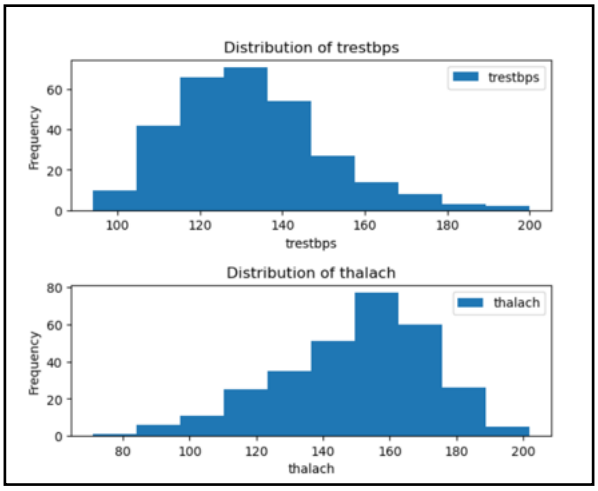
Feature Importance Scores:

	Feature	Importance
8	ca	0.131703
5	thalach	0.115309
3	chol	0.111753
7	oldpeak	0.110522
2	trestbps	0.101263
0	age	0.096839
13	thal_3	0.058358
21	slope_upsloping	0.032149
9	cp_asymptomatic	0.031501
6	exang	0.031334
17	restecg_left_ventricular_hypertrophy	0.027216
20	slope_flat	0.026307
15	thal_7	0.025853
18	restecg_normal	0.025721
1	sex	0.024781
11	cp_non_anginal_pain	0.018553
4	fbs	0.009625
10	cp_atypical_angina	0.008666
12	cp_typical_angina	0.004703
14	thal_6	0.004386
19	slope_downsloping	0.002917
16	restecg_ST-T_wave_abnormality	0.000541

Table 4: Tree-based feature importance Output

3.2 Heart Dataset





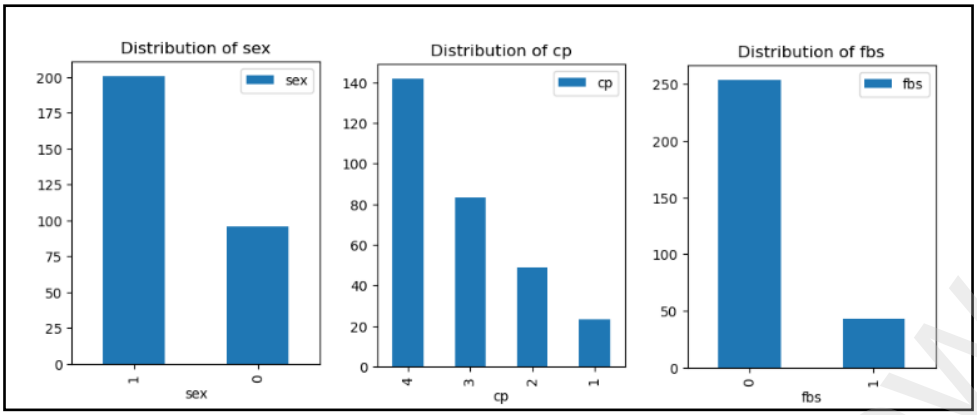


Figure 1: Distributions of Some Key Variables of the Datase

Variable Name	Role	Type	Demographic	Description	Units	Missing Values	Value Ranges
age	Feature	Integer	Age		years	no	
sex	Feature	Categorical	Sex			no	1: Male; 0: Female
cp	Feature	Categorical		Chest pain		no	— Value 1: typical angina — Value 2: atypical angina — Value 3: non-anginal pain — Value 4: asymptomatic
trestbps	Feature	Integer		resting blood pressure (on admission to the hospital)	mm Hg	no	
chol	Feature	Integer		serum cholesterol	mg/dl	no	
fbs	Feature	Categorical		fasting blood sugar > 120 mg/dl		no	1 = true; 0 = false
restecg	Feature	Categorical		Resting electrocardiographic measurement		no	— Value 0: showing probable or definite left ventricular hypertrophy by Estes' criteria — Value 1: Normal — Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
thalach	Feature	Integer		maximum heart rate achieved		no	
exang	Feature	Categorical		exercise-induced angina		no	1 = yes; 0 = no
oldpeak	Feature	Integer		ST depression induced by exercise relative to rest		no	
slope	Feature	Categorical		the slope of the peak exercise ST segment		no	0: downsloping; 1: flat; 2: upsloping
ca	Feature	Integer		number of major vessels (0-3) coloured by fluoroscopy		yes	0 - 3
thal	Feature	Categorical		A blood disorder called thalassemia Value		yes	3 = normal; 6 = fixed defect; 7 = reversible defect
num	Target	Integer		diagnosis of heart disease		no	

Table 5: Cleveland Heart Dataset Source: UCI Machine Learning Repository

The above figures show some of the key variables of the dataset. The age is distributed between 29 and 77 years with a good number of the tested patients between the ages 55 and 60 years. The trestpbs shows that the blood pressure of the majority of the people tested is between 120mmHg and 140mmHg while the chol shows that the cholesterol of the majority of the people is between 200mg/dl and 300mg/dl. Cholesterol below 200mg/dl is considered a safe zone (Cleveland Clinic, 2022) while an ideal normal blood pressure is between 80mmHg and 120mmHg (American Heart Association, 2023). The thalach distribution shows the maximum heart rate which stands at 160 beats per minute with the majority of the tested population lying between 160 and 170 beats per minute. However, this is age-dependent as an individual's age is subtracted from 220 (CDC, 2023) to arrive at the correct value; correlating this with the majority age which is 60 years, any value above 160 is considered a risk. The dataset features more males than females and this is still in line with the fact that the male is more at risk of CHD than females (Hemal et al., 2016). The cp plot refers to the chest pain type experienced by the tested population; it shows that most people experienced asymptomatic chest pain while the fbs plot shows the fasting blood sugar with the majority of the tested population having fasting blood sugar less than 120mg/dL while the normal fasting blood sugar is between 70mg/dL and 100mg/dL (WHO, 2024)

Moreover, the dataset contains 160 individuals with no CHD and 137 patients with CHD; people having the disease are filtered from the dataset, and below are some of the relationships among the features.

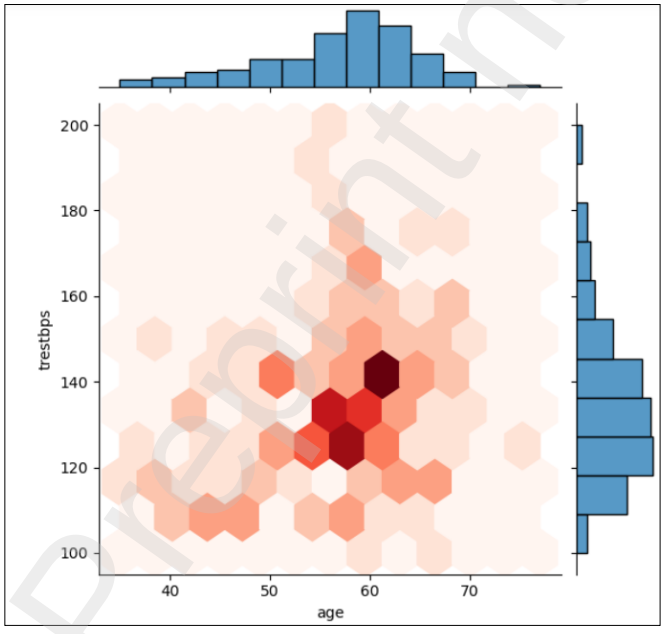


Figure 2: Resting Blood Pressure Versus Age

As the population advances in age, the resting blood pressure tends to increase as seen in figure 2. Although this is combined with the other factors, the figure shows the propensity of the older population to develop the disease.

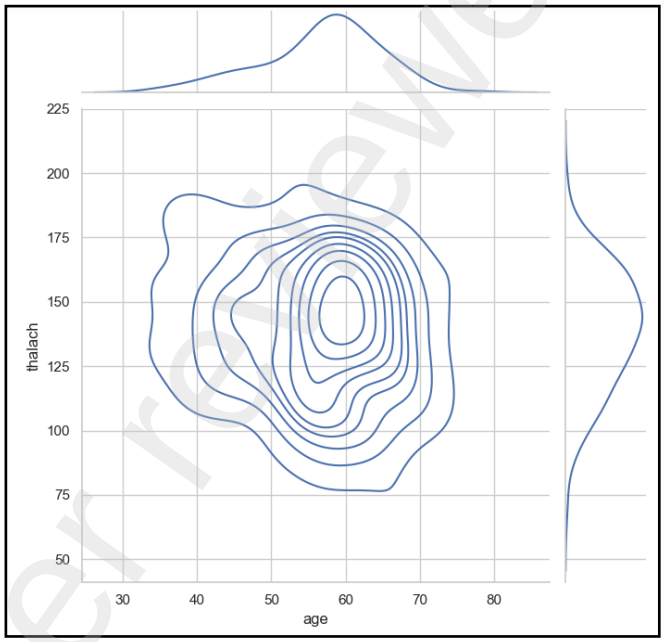


Figure 3: Maximum Heart Rate Versus Age

From Figure 3, the younger population has a higher maximum heart rate which decreases with an increase in age. This shows the tendency of the old population to develop the disease.

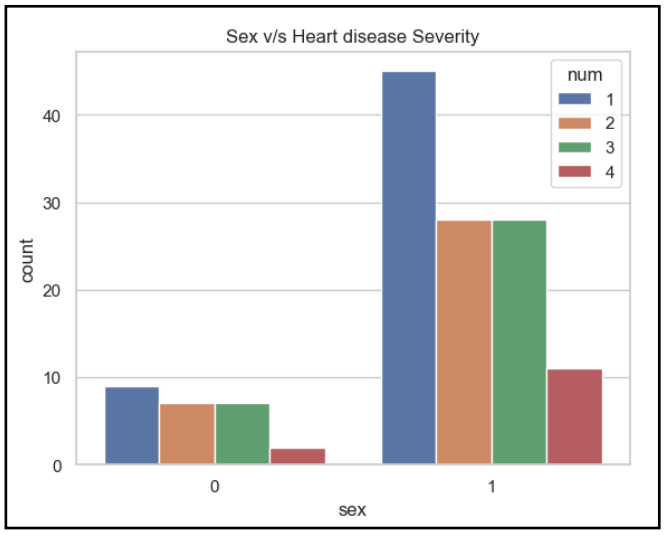


Figure 4: Sex Versus Heart Disease Severity

Figure 4 shows more cases and severity of CHD in males than females. Although there is a data imbalance in the dataset for this variable, (Hemal et al., 2016) confirm men are at more risk than women.

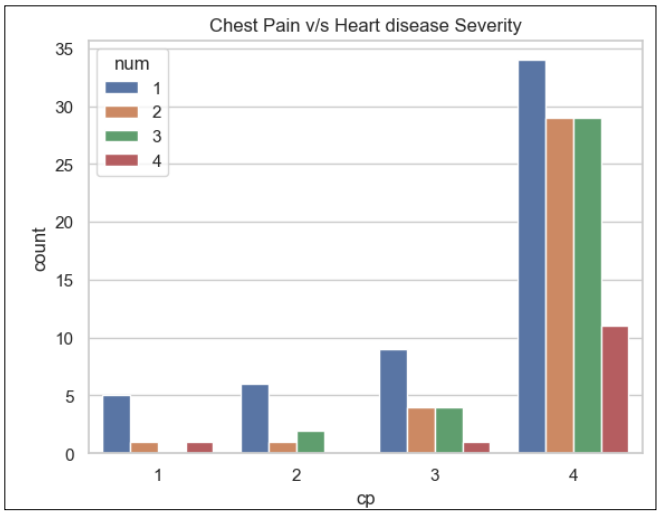


Figure 5: Chest Pain Versus Heart Disease Severity

Apart from the asymptomatic chest pain cases, individuals with non-anginal pain tend to have more severity of the disease.

Correlation Matrix

From Figure 6, a positive correlation is observed between all the features and the target except thalach which shows a slightly high negative correlation. oldpeak, cp, exang, and slope show a slightly strong correlation with the target. Nevertheless, there is no strong correlation among the predictors to suggest a redundant feature except for oldpeak and slope which are slightly correlated.

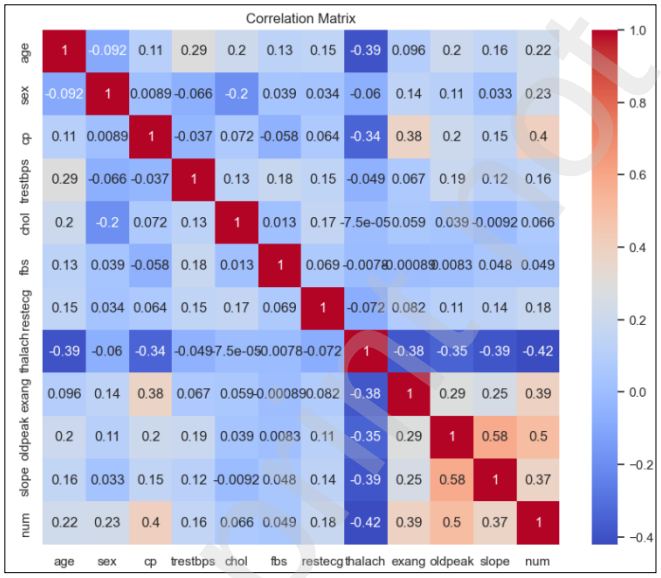


Figure 6: Correlation Matrix

3.3 Proposed Super Learner Building

An SL is an ML Ensemble that aggregates the predictive power of many ML algorithms known as the base learner to employ the best-performing model for prediction. It is also known as stacking and is more accurate than a single predictive model (Phillips et al., 2023).

In this study, RFC, KNN, XG Boost, MLP, and SVM are chosen as the BLs based on their diverse learning approaches and strengths in handling different aspects of the dataset. For instance, K-Nearest Neighbors (KNN) is known for its simplicity in classification, Random Forest Classifier (RFC) excels in handling complex relationships in data, XGBoost is effective in dealing with gradient boosting tasks, Support Vector Machine (SVM) can handle non-linear data, and Multi-Layer Perceptron (MLP) is adept at learning intricate patterns in data. A custom 'SuperLearner' class is implemented, designed to utilize the predictions made by the BLs to create a meta-learner (MeL). The MeL learns from the predictions generated by individual BL. The Super Learner is trained on a training dataset that comprises selected features relevant to predicting CHD, ensuring a robust understanding of various data patterns. During the training process, each BL is utilized to make predictions on the training set, and these predictions are aggregated. The MeL is then trained on these aggregated predictions to make the final classification decisions.

To evaluate the super learner (SL)'s performance, k-fold cross-validation is employed. This technique assesses the model's robustness by splitting the dataset into k subsets iteratively training the model on k-1 subsets and validating on the remaining subset. After validation, the Super Learner is tested on a separate test dataset to measure its predictive accuracy on unseen data, providing insights into its real-world applicability. Evaluation metrics such as accuracy, precision, recall, or area under the ROC curve (AUC-ROC) are computed to quantify the Super Learner's performance in predicting CHD.

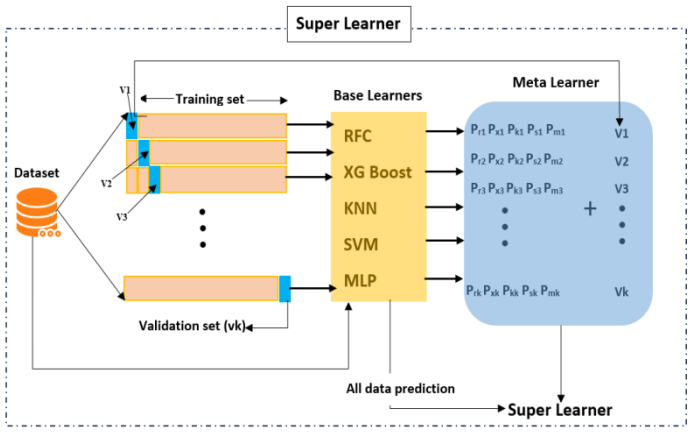


Figure 7: Proposed Super Learner Framework

4. RESULTS AND DISCUSSION

This section focuses on assessing the individual performances of the BLs and compares them with the SL's performance, crucial for validating the enhancements achieved through the ML ensemble. Additionally, it evaluates the SL's performance using

both hold-out and external datasets, along with performance metrics.

4.1 Models Performances

Machine Learning Model	Accuracy
Random Forest Classifier, RFC	**0.86**
XG Boost	**0.83**
Support Vector Machine, SVM	**0.65**
Multi-Layer Perceptron, MLP	**0.78**
K-Nearest Neighbour, KNN	**0.75**
Logistic Regression, LR	**0.61**
Super Learner, SL	**0.88**

Table 6: Models Accuracies

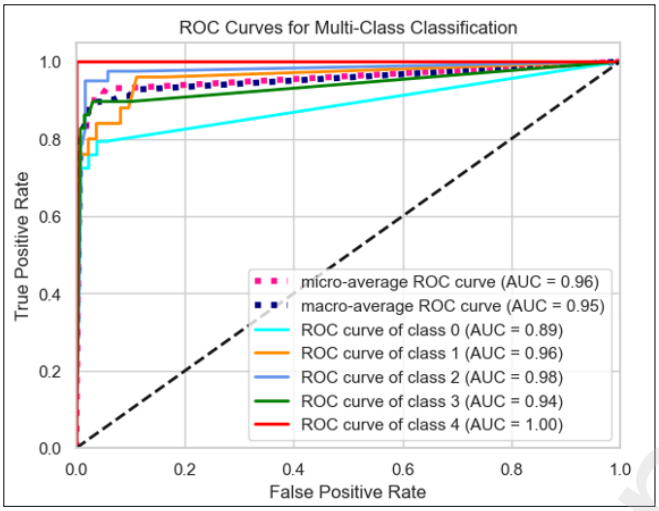


Figure 8: ROC Curves for Multi-Class Classification

Classifications	Precision	Recall	F1-Score
RFC			
0 - Normal	0.89	0.83	0.86
1 - SIHD	0.84	0.72	0.76
2 - Unstable Angina	0.79	0.88	0.83
3 - NSTEMI	0.83	0.91	0.87
4 - STEMI	0.92	0.94	0.93
XG Boost			
0 - Normal	0.88	0.73	0.8
1 - SIHD	0.74	0.72	0.73
2 - Unstable Angina	0.79	0.88	0.83
3 - NSTEMI	0.82	0.97	0.89
4 - STEMI	0.97	0.89	0.93
SVM			
0 - Normal	0.81	0.87	0.84
1 - SIHD	0.72	0.58	0.65
2 - Unstable Angina	0.48	0.52	0.5
3 - NSTEMI	0.62	0.45	0.53
4 - STEMI	0.65	0.86	0.74
MLP			
0 - Normal	0.83	0.8	0.81
1 - SIHD	0.74	0.64	0.69
2 - Unstable Angina	0.69	0.88	0.77
3 - NSTEMI	0.8	0.73	0.76
4 - STEMI	0.92	0.97	0.95
KNN			
0 - Normal	0.85	0.77	0.81
1 - SIHD	0.7	0.64	0.67
2 - Unstable Angina	0.59	0.76	0.67
3 - NSTEMI	0.82	0.7	0.75
4 - STEMI	0.85	0.94	0.89
LR			
0 - Normal	0.84	0.87	0.85
1 - SIHD	0.63	0.33	0.44
2 - Unstable Angina	0.43	0.52	0.47
3 - NSTEMI	0.5	0.45	0.48
4 - STEMI	0.66	0.92	0.77
SL			
0 - Normal	0.85	0.79	0.82
1 - SIHD	0.77	0.8	0.78
2 - Unstable Angina	0.92	0.9	0.91
3 - NSTEMI	0.87	0.9	0.88
4 - STEMI	0.95	0.97	0.96

Table 7: Models Performance

Model Validation

Model validation is a crucial step in assessing the performance and reliability of ML models like the SL. It involves techniques to estimate how well the model will generalize to new, unseen data. The validation process helps in selecting the best model, identifying potential issues, and ensuring the model's robustness. Evaluation metrics discussed in the previous section and performance testing on the hold-out data are employed to validate the model's performance. ML algorithms like RFC, SVM, XG boost, MLP, and KNN are the BLs of the SL, and below are their performances as seen in Table 6. The least-performing model is Logistic Regression (LR) and it is excluded as part of the BLs.

However, the multi-class tends to affect the overall accuracy of the model. Thus, the performance of the SL model is not based on its accuracy alone but on other metrics such as precision, Recall, F1-Score, and ROC. Figure 9 shows the overall ROC of this study which stands at 0.96. This indicates a high level of performance and accuracy in distinguishing between the classes in this multi-class classification problem. This score suggests that the model has excellent discrimination ability, with minimal misclassifications. A ROC score closes to 1 implies that the model has a strong capability to differentiate between positive and negative cases. Table 6 shows that the SL model outperforms all the BLs with an accuracy of 0.88, a

performance attributed to the k-fold cross-validation process. According to the table, classes 0, 1, 2, 3, and 4 have precisions of 0.85, 0.77, 0.92, 0.87, and 0.95 respectively; each value represents the true positively predicted percentage for each class, and a higher value means a lower false positive. Although this is not enough to evaluate the performance of the model, as far as ML is concerned, the model has high precisions for the classes. Similarly, the sensitivity or recall for the classes 0, 1, 2, 3, and 4 are 0.79, 0.80, 0.90, 0.90, and 0.97 respectively while their F1-Scores are 0.82, 0.78, 0.91, 0.88, and 0.96 respectively. Combining all these metrics shows the actual performance of the model.

4.1.1 Super Learner Performance on Hold-Out Dataset

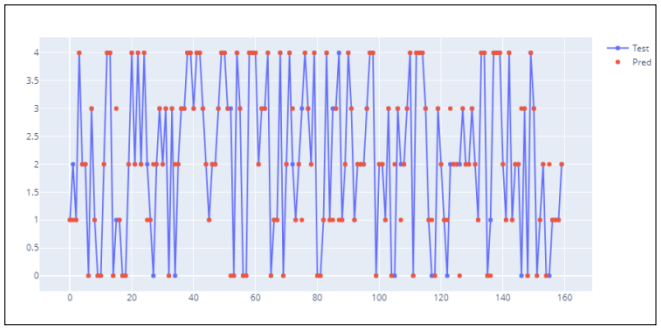


Figure 9: Prediction on Hold-Out Data

Figure 9 shows the performance of SL on the test dataset with very few inaccurate predictions.

4.2 Super Learner Evaluation with External Dataset

An external dataset called statlog dataset (publicly available at UCI Machine Learning Repository) contains the same set of attributes as the Cleveland Heart dataset except for the target variable that has only two variations – 1 for individuals without CHD and 2 for persons with CHD. Unlike the Cleveland dataset used to train the super learner which has four severities of the disease, this dataset uses 2 to represent all the classes. Thus, only manual testing is possible in this case.

A tool capable of predicting CHD is developed with the SL, which has a graphical user interface (GUI) for user interaction, the tool is called CHD Predictor (See Figure 10). The CHD predictor is tested using random ten instances from the statlog dataset and the results are as shown below.

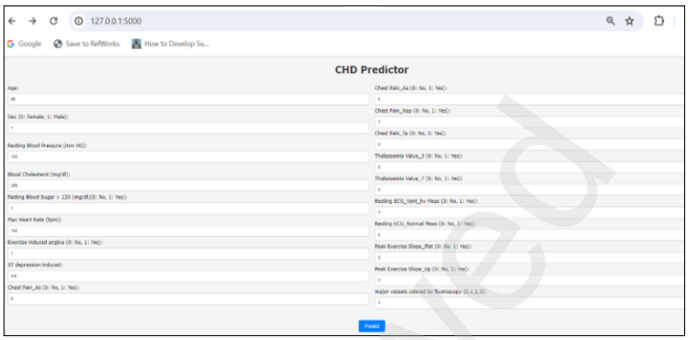


Figure 10: CHD Predictor

Instance No	External Dataset Class	CHD Predictor Prediction
8	2	Unstable Angina
14	1	Normal
26	1	Normal
62	1	Normal
83	2	SIHD
114	2	SIHD
154	1	Normal
203	2	SIHD
223	2	Unstable Angina
267	1	Normal

Table 8: CHD Predictor Result

CHD predictor accurately predicts all the ten instances. It should be noted that the classes are hard coded in the predictor as below:

0: Normal means the individual is normal

1: SIHD means the individual is suffering from Stable Ischemic Heart Disease

2: Unstable Angina

3: NSTEMI means the individual is suffering from Non-ST-elevation Myocardial Infarction

4: STEMI means the individual is suffering from ST-elevation Myocardial Infarction.

5. CONCLUSION

In conclusion, this project has made significant strides in addressing the global health challenge posed by CHD, a prevalent and life-threatening cardiovascular condition. The primary aim of the study was to develop an efficient model for classifying CHD into its various severity stages, considering the limitations of existing studies that often oversimplify the disease classifications. This aim has been met by developing a model called Super Learner which categorizes CHD according to its severity.

The objectives outlined in this project were effectively achieved, with the model showcasing a remarkable ability to classify CHD beyond the conventional binary classifications. A heart disease dataset from the UCI Machine Learning repository was used while credible ML models such as RFC, SVM, MLP, XG Boost, and KNN were identified as BLs. The identified BLs were used to train the SL and its performance was assessed via the hold-out dataset while the developed model was employed to implement the CHD prognostic system. The system was evaluated with an external dataset extracted from the same source as the heart disease dataset. The incorporation of diverse ML BLs within the SL ensemble contributed to robust predictions and improved accuracy. Notably, the achieved ROC score of 0.96 signifies an unprecedented level of performance in diagnostic accuracy, demonstrating the model's potential as a valuable tool for early diagnosis and disease management.

Furthermore, the outcomes of this project emphasize the importance of leveraging advanced ML techniques in healthcare for enhanced disease prognosis and patient care. The developed model holds promise not only for CHD but also as a framework for addressing other complex medical conditions. Future research endeavours can build upon these foundations, exploring additional features and refining the model to further enhance its accuracy and applicability in real-world clinical settings. Overall, this project contributes to the growing field of ML applications in healthcare, with the potential to revolutionize diagnostic approaches and improve patient outcomes in the realm of cardiovascular health.

REFERENCES

American Heart Association (2023). Available at https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

Ayon, S. I., Islam, M. M., & Hossain, M. R. (2022). Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques. *Journal of the Institution of Electronics and Telecommunication Engineers, 68*(4), 2488-2507. https://10.1080/03772063.2020.1713916

Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., & Singh, P. (2021). Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. Computational Intelligence and Neuroscience, 2021, 8387680. https://10.1155/2021/8387680

Burak Kolukisa (2020). Available at http://www.agu.edu.tr/userfiles/Fuarlar/GSES/51_BURAK_KOLUKISA_min.pdf

CDC (2023). Available at https://www.cdc.gov/physicalactivity/basics/measuring/heartrate.htm

Cleveland (2022). Available at https://my.clevelandclinic.org/health/diseases/22910-acute-coronary-syndrome

Cleveland Clinic (2024). Available at https://my.clevelandclinic.org/health/diseases/16753-atherosclerosis-arterial-disease

Eckersley, L., Sadler, L., Parry, E., Finucane, K., & Gentles, T. L. (2016). Timing of diagnosis affects mortality in critical congenital heart disease. Archives of Disease in Childhood, 101(6), 516-520.https://10.1136/archdischild-2014-3

El-Bialy, R., Salamay, M. A., Karam, O. H., & Khalifa, M. E. (2015). Feature Analysis of Coronary Artery Heart Disease Data Sets. *Procedia Computer Science, 65*, 459-

468. https://https://doi.org/10.1016/j.procs.2015.09.132

Jan, M., Awan, A. A., Khalid, M. S., & Nisar, S. (2018). Ensemble approach for developing a smart heart disease prediction system using classification algorithms. *Research Reports in Clinical Cardiology, 9*, 33-45. https://10.2147/RRCC.S172035

He, Q., Dou, Z., Su, Z., Shen, H., Mok, T., Zhang, C. J. P., Huang, J., Ming, W., & Li, S. (2023). Inpatient costs of congenital heart surgery in China: results from the National Centre for Cardiovascular Diseases. *The Lancet Regional Health. Western Pacific, 38*, 100884. https://https://doi.org/10.1016/j.lanwpc.2022.100623

Hemal, K., Pagidipati, N. J., Coles, A., Dolor, R. J., Mark, D. B., Pellikka, P. A., Hoffmann, U., Litwin, S. E., Daubert, M. A., Shah, S. H., Ariani, K., Bullock-Palmer, R. P., Martinez, B., Lee, K. L., & Douglas, P. S. (2016). Sex Differences in Demographics, Risk Factors, Presentation, and Noninvasive Testing in Stable Outpatients with Suspected Coronary Artery Disease: Insights from the PROMISE Trial. JACC. Cardiovascular Imaging, 9(4), 337-346. https://10.1016/j.jcmg.2016.02.001

Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve? *Emergency Medicine Journal: EMJ, 34*(6), 357-359. https://10.1136/emermed-2017-206735

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

Malav, A., Kadam, K., & Kamat, P. (2017). PREDICTION OF HEART DISEASE USING K-MEANS and ARTIFICIAL NEURAL NETWORK as HYBRID APPROACH to IMPROVE ACCURACY. *International Journal of Engineering and Technology (Chennai, India), 9*(4), 3081-3085. https://10.21817/ijet/2017/v9i4/170904101

Miao, K. H., & H., J. (2018). Coronary Heart Disease Diagnosis using Deep Neural Networks. *International Journal of Advanced Computer Science & Applications, 9*(10)https://10.14569/IJACSA.2018.091001

Mirjalili, S. R., Soltani, S., Heidari Meybodi, Z., Marques-Vidal, P., Kraemer, A., & Sarebanhassanabadi, M. (2023). An innovative model for predicting coronary heart disease

using triglyceride-glucose index: a machine learning-based cohort study. *Cardiovascular Diabetology, 22*(1), 200. https://10.1186/s12933-023-01939-9

ONS (2019). Ischaemic heart diseases deaths including comorbidities, England and Wales - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/ischaemicheartdiseasesdeathsincludingcomorbiditiesenglandandwales/2019registrations

Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022a). Risk prediction of cardiovascular disease using machine learning classifiers. Open Medicine, 17(1), 1100-1113. https://10.1515/med-2022-0508

Phillips, R. V., van der Laan, M. J., Lee, H., & Gruber, S. (2023). Practical considerations for specifying a super learner. *International Journal of Epidemiology, 52*(4), 1276-1285. https://10.1093/ije/dyad023

Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017). A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. *2017 IEEE Symposium on Computers and Communications (ISCC),* , 204-207. https://10.1109/ISCC.2017.8024530

Prashasti K, Disha R S. (2016). Prediction of Cardiovascular Diseases using Support Vector Machine and Bayesian Classification. Available at https://www.ijcaonline.org/archives/volume156/number2/kanikar-2016-ijca-912368.pdf

Salimi, A., Zolghadrasli, A., Jahangiri, S., Hatamnejad, M. R., Bazrafshan, M., Izadpanah, P., Dehghani, F., Askarinejad, A., Salimi, M., & Bazrafshan Drissi, H. (2023). The potential of HEART score to detect the severity of coronary artery disease according to SYNTAX score. *Scientific Reports, 13*(1), 7228. https://10.1038/s41598-023-34213-9

Satya Sree, K. P. N. V., Karthik, J., Niharika, C., Srinivas, P. V. V. S., Ravinder, N., & Prasad, C. (2021). Optimized Conversion of Categorical and Numerical Features in Machine Learning Models. *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC),* 294-299. https://10.1109/I-SMAC52330.2021.9640967

Ruqiya. (2023). Review on Cleveland Heart Disease Dataset using Machine Learning. *Quaid-E-Awam University Research Journal of Engineering, Science & Technology, 21*(1), 87-98. https://https://doi.org/10.52584/QRJ.2101.11

Shahjehan RD, Bhutta BS. Coronary Artery Disease (2023). In: StatPearls [Internet]. Treasure Island (FL). https://www.ncbi.nlm.nih.gov/books/NBK564304/

Shylaja S, Muralidharan R (2019). Hybrid SVM-ANN Classifier is used for Heart Disease Prediction System. Available at https://www.researchgate.net/publication/334965479_Hybrid_SVM-ANN_Classifier_is_used_for_Heart_Disease_Prediction_System

Subhadra K, Vikas B (2019). Neural network based intelligent system for predicting heart disease. Available at https://www.researchgate.net/publication/332035370_Neural_network_based_intelligent_system_for_predicting_heart_disease

Sharma, S., & Parmar, M. (2020). Heart Diseases Prediction using Deep Learning Neural Network Model. *International Journal of Innovative Technology and Exploring Engineering, 9*(3), 2244-2248. https://http://dx.doi.org/10.35940/ijitee.C9009.019320

Verma, L., & Srivastava, S. (2016). A Data Mining Model for Coronary Artery Disease Detection Using Noninvasive Clinical Parameters. *Indian Journal of Science and Technology, 9*(48)https://10.17485/ijst/2016/v9i48/105707

Walker, S., Asaria, M., Manca, A., Palmer, S., Gale, C. P., Shah, A. D., Abrams, K. R., Crowther, M., Timmis, A., Hemingway, H., & Sculpher, M. (2016). Long-term healthcare use and costs in patients with stable coronary artery disease: a population-based cohort using linked health records (CALIBER). *European Heart Journal. Quality of Care & Clinical Outcomes, 2*(2), 125-140. https://https://doi.org/10.1093/ehjqcco/qcw003

WHO (2024) https://www.who.int/data/gho/indicator-metadata-registry/imr-details/2380

Zghebi, S. S., Mamas, M. A., Ashcroft, D. M., Rutter, M. K., Van Marwijk, H., Salisbury, C., Mallen, C. D., Chew-Graham, C. A., Qureshi, N., Weng, S. F., Holt, T., Buchan, I., Peek, N., Giles, S., Reeves, D., & Kontopantelis, E. (2021). Assessing the severity of cardiovascular disease in 213 088 patients with coronary heart disease: a retrospective cohort study. *Open Heart, 8*(1)https://10.1136/openhrt-2020-001498