RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt24313

Detect the Cardiovascular Disease's in Initial Phase using a Range of Feature Selection Techniques of ML

Prashant Maganlal Goad [a,*], Pramod J Deore [a]

[a] Department of Electronic & Telecommunication Engineering, R.C. Patel Institute of Technology, Shirpur-425405, Maharashtra, India

* Corresponding Author Email: prashantgoad@gmail.com

DOI: https://doi.org/10.54392/irjmt24313

Received: 28-02-2024; Revised: 13-04-2024; Accepted: 27-04-2024; Published: 14-05-2024

Abstract: Heart-related conditions remain the foremost global cause of mortality. In 2000, heart disease claimed around 14 million lives worldwide, a number that surged to approximately 620 million by 2023. The aging and expanding population significantly contribute to this rising mortality trend. However, this also underscores the potential for significant impact through early intervention, crucial for reducing fatalities from heart failure, where prevention plays a pivotal role. The aim of the present research is to develop a prospective ML framework that can detect important features and predict cardiac conditions as an early stage using a variety of choice of features strategies. The Features subsets that were chosen were designated as FST1, FST2, and FST3, respectively. Three distinct methods, including correlation-based feature selection, chi-square and mutual information, were used for picking features. Next, the most confident theory & the most appropriate feature selection were identified using six alternative machine learning models: Logistical Regression (LR) (AL1), the support vector Machine (SVM) (AL2), K-nearest neighbor (K-NN) (AL3), Random forest (RF) model (AL4), Naive Bayes (NB) model (AL5), and Decision Tree (DT) (AL6). Ultimately, we discovered that, with 95.25% accuracy, 95.11% sensitivity, 95.23% specificity, 96.96 area below receiver operating characteristic and 0.27 log loss, the random forest model offered the most excellent results for F3 feature sets. No one has investigated coronary artery disease forecasting in depth; however, our study evaluates multiple statistics (specificity, sensitivity, accuracy, AUROC, and log loss) and uses multiple attribute choices to improve algorithms success for important features. The suggested model has considerable promise for medical use to speculate CVD find in Precursor at a minimal cost and in a shorter amount of time as well as will assist limited experience physician to take right decision based on the results of the used model combined with specific criteria.

Keywords: ML Technique, Feature Selection Method, Categorization and Modelling, Data Interpretation, Random Forest.

1. Introduction

These days, data mining techniques are widely applied everywhere. Learning from data is a popular tool in the healthcare sector for early illness prediction. Around the world, it saves a great deal of deaths by early illness prediction. Yet coronary artery disease claims the lives of millions of individuals each year. If robots be able to forecast the initial phases of the ailment, this prognosis ought to lower the feasibility of a heart disease. Despite the heart being a crucial organ, heart failure is the leading cause for mortality in modern civilization. The cerebral cortex and many limbs stop working if it is able to function correctly, causing extremities to become occluded and a person to pass away in a matter of seconds. It represents one of the leading illnesses that primarily strikes middle-aged or older individuals and causes serious problems for the way the body functions [1]. Because there are so numerous risk variables for coronary heart disease, diagnosis can be challenging. Lack of breathing, chest discomfort, fast or rapid heartbeats, and fatigue are the primary signs and symptoms of heart disease [2]. In the USA (US), the prevalence of cardiovascular illness is significantly higher, while one adult passes away from heart diseases each thirty-four seconds [3]. Roughly 25 million individuals worldwide suffer from cardiovascular conditions [4]. A total of 17.9 million individuals worldwide suffer from coronary heart disease each year, and this illness accounts for 32% of all heart disease related deaths [5]. The United Nations Health Organisation (WHO) estimated that heart-related disorders cost India up to $237 billion between 2005 and 2015 [5]. It is anticipated that the yearly mortality toll from CVD will increase from 2.26 million in 1990 to 4.77 million in 2020. Cardiovascular disease (HD) affects

both sexes equally [6]. Coronary artery disease can also manifest in the middle decade and beyond due to prolonged being exposed to unhealthy lifestyles. Once this study is completed, we will be able to identify cardiac problems front on. The circulatory system is impacted by CVD both men and women . Due to decades of living hazardous lifestyles, heart illnesses can also manifest in middle age and later in life. We will be able to identify cardiovascular disease soon once this study has been completed.

Billions of people will be averted and hundreds of thousands of people with cardiovascular diseases will benefit globally from this forecast. Heart failure is known to generate significant losses in the overall economy, while early detection of the condition is expected to save thousands of us. Six ttechniques based on ML are used in prophesy to determine the optimal fidelity. Next, determine one of the algorithms is the most superior.

Furthermore, there hasn't been an in-depth exploration of forecasting coronary artery disease. Our study, however, delves into various statistics like specificity, sensitivity, accuracy, AUROC, and log loss, employing diverse attribute selection methods to enhance algorithmic performance for crucial features. In this research, we establish a forward-looking machine learning framework capable of identifying significant features, marking a novel aspect of this study.

2. Literature Survey

The second part discusses earlier research on heart disease that used automated learning techniques, which served as inspiration for the present research. As per Ramalingam et al. [7], supervised algorithms of machine learning were utilized in this study on various medical datasets and multiple data experimentation. This study advances several model-based approaches and methodologies. These academics use a variety of supervised techniques, including SVM, DT, RF (RF), KNN, and a NB model. The execution of several strategies employed in this study was assessed to determine efficiency. Using SVM-RFE chosen in the top 10 attributes, the correctness obtained from the NB outcomes is 84.1584%. Pouriyeh *et al.* [8] state that the NB approach achieved an error rate of 83.49% in this study utilising 13 characteristics. The logistic regression (KNN) rule, a nonparametric solution to pattern grouping, was first presented by Fix and Johnson [9] in 1951. DT and KNN had accuracy rates of 82.17% and 83.16%, respectfully. The artificially intelligent cardiovascular condition assessment in algorithms that use machine learning is predicted by Palaniappan and Awang [10]. Quality was suggested by the combined use of the techniques.

When DT, NB, and NN were used to perform HD death, their precision became 80.4%, 86.12%, and 85.68%. In order to demonstrate the correctness, by Palaniappan and Awang [10] classed all three approaches using the Cleveland average cardiovascular illness dataset. The computer programme predictions the algorithm's performance is predicted through the use of artificially generated neural networks (ANN), KNN, and SVM. In this precision KNN (82.963%) and ANN (73.3333%) is adopted. Authors suggested that SVM be used as the most accurate technique for classification to forecast cardiac events. In the present investigation, Haq et al. [11] developed seven classification system performance assessment measures, including classification precision, accuracy, The research of Matthews' relationship, awareness, and processing time, using the UCI dataset and well-known algorithms, the cross-validation method, and serving choosing features (FS) algorithms. Effect on reliability and time required for execution of the predictor is presented. Effectiveness, specificity, then responsiveness, and inaccuracy were developed by selecting the key attributes using three picking feature computations: mRMR, relief, and LASSO. Despite all the earlier research [6], a study on neural network algorithms for heart attack detection was conducted by Ramalingam *et al.* Each technique will function at its best with the best knowledge [7]. The contributor conducted a thorough analysis on the comparative effectiveness of computer-aided learning in the topic of congestive heart failure using the UCI information set. However, approaches to selecting features determine how well these approaches perform [8]. Using data from nearby 1000 individuals, Palaniappan and Awang used data analysis approaches to forecast the occurrence of heart failure. Yet when dealing with large volumes of the information, information mining becomes much more efficient [10]. Cheng *et al.* [12] claim that similar methods were used in their research, utilising multiple algorithms with a minimum of 90 percent performance. It could provide the best precision if it could deal with data with greater consideration. In summary, the researchers aimed to determine the optimal precision for heart disease prediction using patient’s health information taken from the UCI collection. Their results showed that fewer than 80% overall heart disease patients were accurately predicted. They looked for the optimum accuracy by using every attribute or by applying a particular feature selecting technique for an individual machine intelligence algorithm, but they failed to show a single feature correlation. The forecasting score of any method is the only thing that is displayed in the other studies; additional performance rating matrices, such as log loss, specificity, sensitivity and various another, haven't been described. Done Stojanov *et al.* [13] aimed to identify significant independent predictors relative to the outcome of heart failure versus chronic-ischemic heart disease. The author obtained data from 167 cardiac patients, of which 108 had Coronary Ischemic Heart Disease (CIHD) and 59 had Heart Failure (HF). These patients were hospitalized at the cardiology ward of Villa Scassi hospital in Genoa, Italy. Author recommends: Hb

+ Serum Creatinine+AST+hs-cTnI+CRP combination for accurate early detection of the outcome of HF versus CIHD in logistic regression-based model. author found that unit increase of AST, ALT or CRP increases the odds of HF against CIHD for3.43%, 2.46% and 4.11% respectively, p-value < 0.05. Rustem Yilmaz *et al.* [14] this research aims to utilize the concepts of explainable artificial intelligence (XAI) in analysing haematological indicators to diagnose Acute Heart Failure (AHF). XGBoost was used in conjunction with LASSO to diagnose AHF, the resulting model had an AUC of 87.9%, an F1 score of 87.4%, a Brier score of 0.036, and an F1 score of 87.4%. The findings of this study demonstrated that the combination of explainable artificial intelligence (XAI) with machine learning (ML) was effective in diagnosing Acute Heart Failure (AHF). P. Lakshmi Prabha *et al.* [15] in the proposed research, the Framingham risk score (FRS) parameter is calculated alongside CIMT for both diabetic and normal subjects. This approach aims to offer an accurate prediction of cardiovascular disease. To enhance the analysis, the user augmented the image data from 110 subjects to 1809 image data points. They then applied transfer learning techniques using VGG16. The results of the analysis showed highly significant relationships with a p-value < 0.001 between CIMT and various biochemical parameters including total cholesterol, HDL, LDL, FBS, and PPBS. The ROC curve indicated elevated CIMT values for diabetic subjects. The features extracted from VGG16 were utilized to train the classifier neural network, achieving an impressive accuracy of 99%.

In the present study, cardiovascular illness data are taken from the UCI ML database [13 characteristics]. It is pertinent to the subject of regulated AI. Heart failure have been extensively studied, yet several methods have been used in attempts to find a solution [15]. But a simple machine neural network cannot solve this kind of complex difficulty. Technologies like trees of decisions and regression techniques (LR) will be used to tackle the task at hand (DT). The collected data sets were subjected to a few feature selection techniques for these kinds of research. In cardiac illness, a number of classifiers have the best precision. Furthermore, algorithms using machine learning are essential for prompt illness prediction of a variety of medical conditions [16]. The following are the main contributions of the suggested investigation project:

a) The entire feature set of each classifier has been examined for effectiveness in terms of both implementation time and accuracy in classification.

b) The classifiers' efficacy has been assessed using specific characteristics selected using Chi Square, Mutual Information, and Correlation-Based methods with six different algorithms.

c) The research makes recommendations for what attribute algorithm works best with which predictor to create an advanced machine learning system of heart failure that can distinguish between individuals with coronary illness and those who are normal.

Figure 1 illustrates the procedure for forecasting heart disease research that was employed in this research in a visual format.

3. Methodology

Python 3.8 is increasingly available and facilitates rapid validation of techniques, which is why it was chosen for this research's trial.

3.1 Dataset

This research utilizes the UCI Cleveland database [11], which has been extensively analyzed and utilized in previous work.

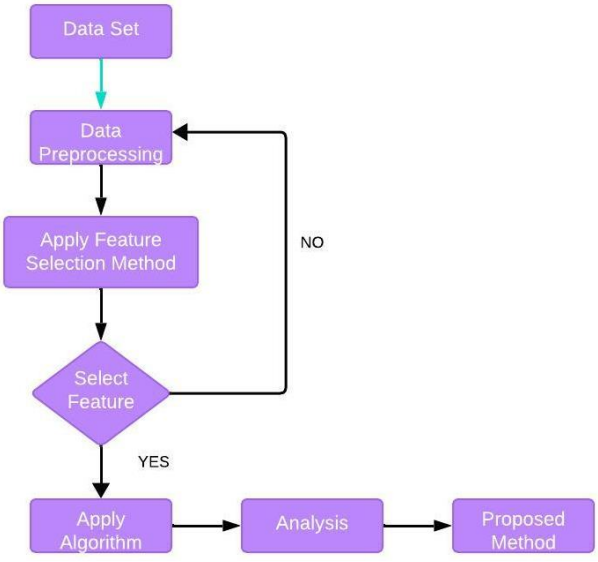

Figure 1. Work flow

Table 1. Data Set Description

Sr no.	Code	Description
1	age	The Forbearing Age
2	sex	Men = 1, Women = 0
3	cp	Chest pain type: 0,1,2,3
4	trestbps	Resting BP (in mm)
5	chol	Cholesterol Measurement in mg/dl
6	fbs	Before Meal Blood Sugar(0 &1)
7	restecg	ECG (0,1,2)
8	thalach	Highest Heart Bits
9	exang	Exercise-Induced Angina: (0,1)
10	oldpeak	ST Depression
11	slope	The highest point of the physical activity's slope (0,1,2)
12	ca	Number of Primary Vessels
13	thal	Thallium Stress

We use this database to predict cardiovascular disease, containing thirteen variables in each of the 303 medical files comprising the UCI heart attack collection. In my labeling scheme, both categories represent patients with cardiac disease and ordinary individuals. Table 1 contains information about the dataset matrices.

3.2 Data Preprocessing

Preprocessing of data was prepared for this investigation once they'd been gathered. In the Cleveland sample established, there are a total of two TS entries and four NMV entries that are erroneous false. The ideal values are used for substituting every single record that have suitable amounts.

4. Feature Selection

Choosing features is crucial to the method of machine learning since, on occasion, an information set has a large number of extraneous characteristics that have an impact on the systems' accuracy. The choice of feature aids in the reduction of these disconnected variables and enhances algorithmic performance [17]. In order to rank the features with the greatest importance according to their importance, it employed a variety of feature selection approaches [18]. Three popular methods for selecting features are employed in this research to determine the salient characteristics according to the corresponding grade.

4.1 Correlation-Based

Since it is a filtering strategy, the interconnection -based feature selection method is unaffected by the last categorization system. As its title clearly suggests, it simply considers data inherent properties correlations when evaluating subsets of features. To prevent repetition, the objective is to identify a subset of features that has minimal feature-feature association and high characteristic-class connection, which will either retain or improve the ability to predict. It goes without saying that we'll look at the subgroup with the greatest quality. Both a small feature-feature correlation in the denominator and an elevated feature-class connection in the sum of the features can lead to a higher score [19].

$$Score = \frac{N\overline{Ffc}}{\sqrt{N+N(N-1)\overline{Fff}}} \quad (1)$$

4.2 Chi Square Test

This metric is suitable for assessing categorical variables in a classification scenario. It helps identify the n_features feature with the highest values based on the chi-squared statistic derived from X, which should exclusively include non-negative features like booleans or frequencies. As a reminder, the chi-square test gauges the interdependence between random variables; utilizing this method helps filter out features that are likely independent of the class, hence not significant for classification [20]. The Chi-Square statistic is a widely employed tool for testing relationships among categorical variables.

$$X_i^2 = \sum \frac{(O-P)^2}{P} \quad (2)$$

4.3 Mutual Information

Mutual Knowledge evaluates mutual data across established groups such in an organization

question or a variable that is continuous in prediction situations. Mutual Sharing operates based on the parameters' complexity [21]. The following is a formal statement of the shared knowledge among the two unknown variables A and B:

$$M(A,B) = H(A) - H(A|B) \quad (3)$$

Where A and B is mutual information, while H [A] is the entropy for A. The outcome is expressed in bits (0 to 1).

The following step involves using the traditional scaling to ensure that each value has a means of 0 and an average of 1, and also to align each data with its appropriate value.

5. Categorization and Modelling

The following is a chronological description of cardiac forecasting systems. That order is followed when applying every method. For analyzing data, a variety of categorization techniques are offered. Six different kinds of categorization techniques are applied in this research. Following is a quick explanation of each algorithm.

5.1 Logistic Regression

A classified independent variable's result is predicted by the method of logistic regression. As a result, an isolated or category value is required for the result. Instead of providing precisely beliefs, which are 0 and 1, it provides the stochastic standards, which fall somewhere in the range of 0 to 1[15]. It may correspond to whether Yeah or No, 0 or 1, true or False, etc.

Since y is restricted to being from 0 to 1 in logistic regression, let's scale the following equation by (1-y):

$$\frac{Y}{1-Y}; 0 \; for \; Y = 0 \; \& \; \infty \; for \; Y = 1 \quad (4)$$

However, we require an interval from - ∞ to + ∞ after taking the expression's logarithms, it becomes:

$$\frac{Y}{1-Y} = a_0 + a_1x_1 + a_2x_2 + \ldots \quad (5)$$

5.2 Support Vector Machine

Among each of the categories, SVM establishes an efficient choice boundary, or hyper plane [22]. The closest data point's closest distance for both groups. SVM lowers the top limit of the predicted test error and calculates door gradually volume, and centroid even when its radial baseline product is employed as a kernel. We treat the support vector function of this research's instance as a radial baseline value. There, p denotes the vector's total height. Choice the primary divider used to classify the points is called a boundary. A hyper plane solution is the formula for the primary divider axis. Now let's examine the calculation for an uninterrupted line having an endpoint of c and an incline of m.

This is the final equation:

$$mx + c = 0 \quad (6)$$

Currently, it is simple to write a two-dimensional plane solution that separate the locations (for classification) as follows:

$$H: b + wT(x) = 0 \quad (7)$$

5.3 K-Nearest Neighbor

KNN classifies the test information using only an initial set of data. It speaks of the KNN's identity. It computes every bit of the simulated information and differentiates from it in order to assess every value. The separation between the query location and the remainder of the information points must be computed to be able to ascertain what information points are closest to a particular query point. The choices for limits that divide query items into various areas are formed in part by these measures of distance. Limits of decisions are frequently represented using Voronoi graphs. In this instance, b represents the surface of the equation's interception and biases element.

$$D(a,b) = \sum_i^n (b_i - a_i)^2 \quad (8)$$

5.4 Random Forest

The random forest method builds around the bagged approach by producing a non-relation forest of tree for choices by utilising feature variability in alongside bagged. Featured unpredictability is sometimes referred to as "the arbitrary domain method" or featured bagged. Ensures minimal association between decisions by producing a random collection of attributes. There is a significant difference in random forest structures and decision forests. Since decision trees believe any potential have divides, random forests just pick certain portions of those attributes.

5.5 Naive Bayes

It is referred to as naïve since it assumes that the appearance of one characteristic is unconnected to being a part of multiple qualities. For instance, if a vegetable is defined based on its color, form and flavor, then an apple can be recognized as a red, cylindrical, juicy fruit. Thus, each characteristic works independently of the others to help distinguish that it is an apple. It is referred to as the Bayes algorithm because it is based on the Bayesian hypothesis.

5.6 Decision Tree

The most effective method for categorizing situations is through decision trees. The technique

divides an information set according to least entropy or the greatest data gain after calculating the degree of entropy for each attribute in at least two similar sets using more accurate readings. A variety of multimodal rating measures, including AUROC, logarithmic loss, awareness, precision and precision, were assessed in order to compare the effectiveness of various algorithms and display the findings. The False Positive (FP), True Positive (TP), False Negative (FN) & True Negative (TN) indices were computed for displaying the values of these matrices. These parameters are covered in more detail in the section following. The optimal technique that yields the best results will be displayed once the study is complete.

The majority of the results obtained from a decision tree is "1" or "0."

The entropy we can calculate by

$$Entropy = -P\ log\ (P) - N\log(N) \qquad (9)$$

Accuracy, sensitivity, specificity, AUROC, and log loss are examples of multiplexed assessment indicators that were assessed in order to collate the outcomes of various techniques and display the results. The true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values were computed for displaying these matrices of data.

These parameters are covered in more detail in the section following. The optimal algorithm that yields the greatest results can be seen once the examination is complete.

6. Experiment Analysis

The present investigation employs Python Scikit-learn bundle for choosing and categorising features applications. The analysed data sets was subjected to a variety of techniques, including AL1, AL2, AL3, AL4, AL5, and AL6. All of which were used to check the performance of the data set. In the following example, matrix correlation heat maps as well as additional relationships between various features are visualised using Matlobplit as and the Python Seaborn package [23, 24]. Finally, a variety of ways to choose features were applied, including the chi-square , MI boolean selecting approach and relational based decision-making. These techniques are listed in Table 2, which and are marked as FST1, FST2, and FST3, accordingly. Fourth, various algorithmic capabilities were assessed in relation to the chosen features.

6.1 Outcome of Various Feature Selection Methods

Applying the attribute assets, the relational-based attribute selection evaluation strategy calculates the coefficient of variance among the factors in question. Table 3 displays the correlation-based results for the F rating. The aspects of this evaluation that are less significant are RES, CM, and FBS, whereas the three of its most significant features are EIA, CPT, and OP. Chi-square is an additional technique that determines the chi square value connecting each attribute and the goal. Table 4 displays the chi-square scores. The three most crucial aspects in this method are MHR, OP, and NMV, while TS, REC, and FBS are the 3 least important characteristics. Figure 2 displays the position qualities in the FST1 and FST2 algorithms. Two characteristics are free if their combined rating is nothing, while the more heavily reliant the traits are on each other, the higher the score will be. Table 5 displays the shared data scores. Here, fbs and restecg are the independent characteristics, and the three most heavily reliant features are CPT, TS, and NMV.

Table 2. Feature Selection Technique

FST	Description	Code
correlation-based	Find the Correlation between attributes	FST1
Chi-square	Calculate Chi square value	FST2
MI	Calculate Mutual Information value	FST3

Table 3. FST1 Score

Sr No	Code	Scores
1	age	16.12
2	sex	25.79
3	cp	69.77
4	trestbps	6.46
5	chol	2.20

6	fbs	0.24
7	restecg	5.78
8	thalach	65.12
9	exang	70.95
10	old peak	68.55
11	slope	40.90
12	ca	64.05
13	thal	31.80

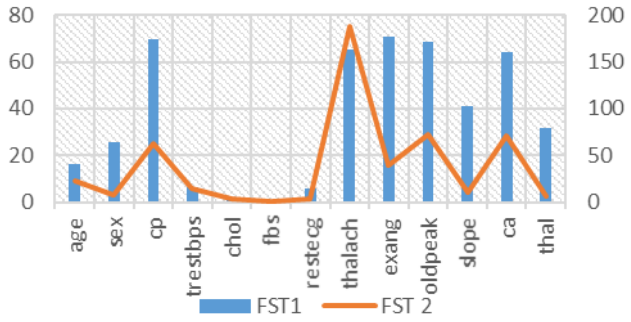

Figure 2. Feature Score

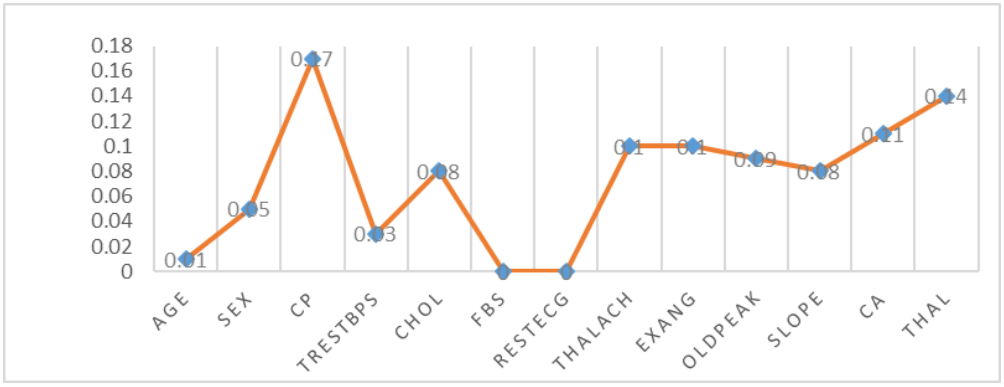

Figure 3. FST3 Score

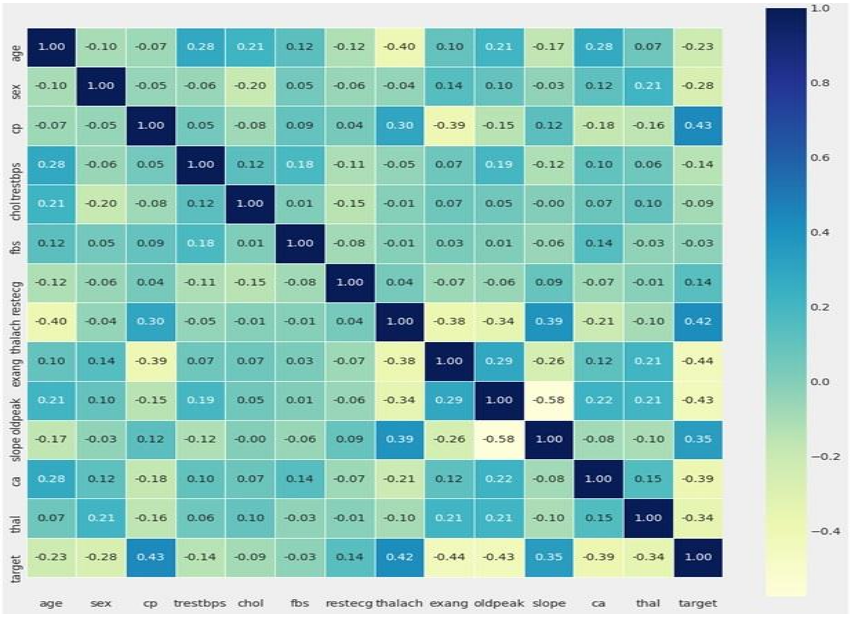

Figure 4. Correlation Matrix

Table 4. FST2 Score

Sr No	Code	Scores
1	age	23.29
2	sex	7.58
3	cp	62.60
4	trestbps	14.82
5	chol	2.94
6	fbs	0.20
7	restecg	2.98
8	thalach	188.32
9	exang	38.91
10	oldpeak	72.64
11	slope	9.80
12	ca	70.89
13	thal	5.90

Table 5. FST3 Score

Sr No	Code	Scores
1	age	0.01
2	sex	0.05
3	cp	0.17
4	trestbps	0.03
5	chol	0.08
6	fbs	0.00
7	restecg	0.00
8	thalach	0.10
9	exang	0.10
10	oldpeak	0.09
11	slope	0.08
12	ca	0.11
13	thal	0.14

Table 6. Selected attributes

Method	Selected Attributes
F1	Age, Sex, exang, Oldpeak , CP, ca, thal ,trestbps, chol, restecg, thalach, Slope,
F2	Age, Sex, Slope, ca, thal ,CP, Thalach, exang, Oldpeak, chol
F3	Age, Sex,Oldpeak, Slope, ca, thal,cp, Thalach, exang

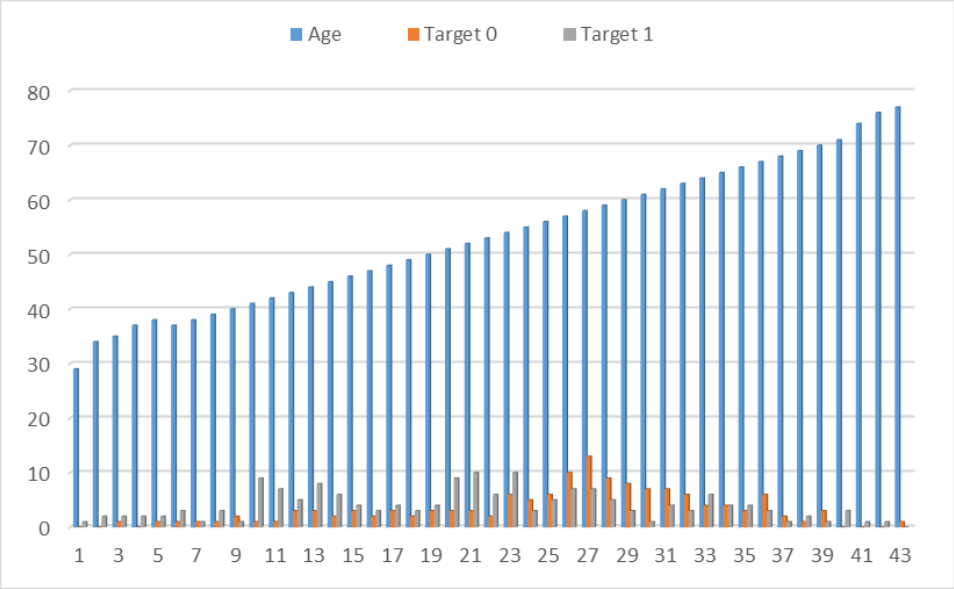

Figure 5. Correlation between Age & Target

Figure 3 displays the feature's overall rank according to the FST3 approach. Important characteristics for predicting of heart disease are shown in those three tables. Additionally, all three FSTs had lower total scores for FBS, REC, RBP, and CM, and none of the investigation's algorithms exploit those qualities. Based upon their score, three distinct sets of characteristics are chosen from every one of those attributes. F1, F2, and F3, in that order, represented each three distinct set of features. Table 6 displays the feature sets that were chosen.

6.2 Features Visualization

First, a heat map is shown, as shown in Figure 4. This heat map shows the association among the different characteristics in the information set. Nearly every characteristic in this dataset has substantially less association with every other feature, according to the correlation values. This suggests that the characteristics that can be removed are limited. The goal characteristic in this heat map exhibits the most negative connection with EIA, OP, and NMV, and has a significant positive correlation with CPT, MHR, and PES. The targets with smallest association scores, yet are FBS, CM, RBP, and REC. These attributes are deleted in separate SF, and this is comparable to another choice of feature method attribute rating. Furthermore, a relationship is displayed given the intended property in Figure 5 and aged. It reveals that about nine patients ages 41, 51, and 52 as well as eleven patient's age 54 had cardiac disease. It implies that heart disease primarily affected middle-aged persons between the ages of 41 and 54. Lastly, Figure 6 illustrates a relationship among MHR and goal. It demonstrates that the heart rates of the elderly are less than those of the young. Furthermore, there is a modest increase in the risk of cardiovascular disease with a greater pulse.

6.3 Accuracy Analysis

Table 7 displays the degree of precision of each technique that was used to analyse the information that was set. Regarding the precision for every method, AL4 determined that F3 had the greatest reliability (95.25 %); for F1 and F2, AL4 assessed efficiency of 91.20% and 90.12%, respectively. AL1 estimated with the second-greatest efficiency (93.51%) among all three SFs. Nevertheless, AL2 computed the low precision (76.64%) for F3. Moreover, AL2's efficiency for F1 and F2 was poor (80.25% and 80.24%). Furthermore, the outcome demonstrates that AL4 for F3 is the optimal technique for the set of data. Figure 7 displays all of the algorithmic accuracy for each of the several SFs.

6.4 Sensitivity Analysis

The degree of sensitivity of each of the techniques was examined in this investigation. Table 8 displayed the sensitivity rating for each of the algorithms in question. Al2 has the lowest reactivity (70.25) for F2. Additionally, Al2 provided F1 and F2 ratings of (70.83 and 71.42). Furthermore, Al4 reported the greatest accuracy of 95.11 for F3 as well; Al1 reported the next-highest sensitivities of 94.74 for all SFs. Furthermore, the outcome demonstrates that Al4 awarded F3 the highest score. Figure 8 displays all of the sensitivity's ratings for various algorithms for various SFs.

6.5 Specificity Analysis

Every of these methods' specific was investigated, and Table 9 displays the sensitivity ratings for each algorithm. AL2 received the lowest score (79.57) for F3 throughout the examination, while Al4 received the greatest rating (95.23). Out of all those SFs, Al1 had the second-highest rating (91.45). Furthermore, the outcome demonstrates that Al4 provided the highest rating for F3.

6.6 AUROC Analysis

In machine learning, the Receiver Operating Characteristic (ROC) curve is a popular visual tool for assessing how well binary classifiers operate. At different threshold values, it shows the true positive rate (sensitivity) against the false positive rate (1 - specificity). This curve is useful for evaluating a classifier's performance in differentiating between classes, especially in situations when the distribution of classes is not balanced. To assess the accuracy of the predictions provided for the heart disease dataset, AUROC analysis was done. The AUROC ratings for the different methods were displayed in Table 10. In this study, Al2 for F2 had the lowest AUROC score (77.27). AL2 also provided F1 and F3 ratings of 82.54 and 80.48, respectively. For F3, Al1 provided the highest possible rating (96.96). Moreover, Al1 provided AUROC values for F1 and F2 of 93.77 and 94.41. Al5 provided a second-best F2 score of 95.50. The AUROC ratings of the other methods ranged from 91.81 to 96.96. Furthermore, the outcome demonstrates that Al1 provided the highest rating for F3. Figure 9 display each of the AUROC scores of various techniques for various SFs.

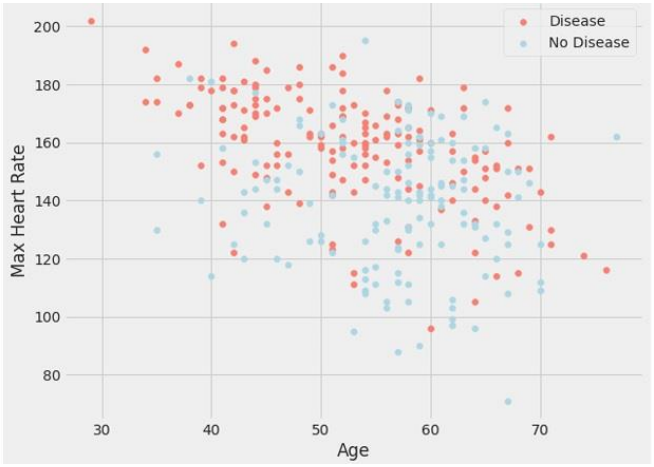

Figure 6. Correlation between Age & Heart Rate

Table 7. Algorithm Accuracy

Features	AL1	AL2	AL3	AL4	AL5	AL6
F1	92.6	80.25	87.91	91.2	89.01	83.52
F2	93.41	80.24	86.81	90.12	90.11	92.31
F3	93.51	76.64	84.61	95.25	90.11	92.54

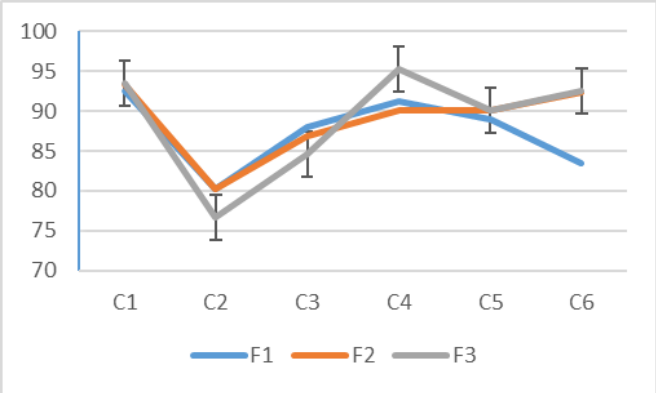

Figure 7. Algorithm Accuracy

Table 8. Algorithm Sensitivity

Features	AL1	AL2	AL3	AL4	AL5	AL6
F1	94.74	70.83	88.56	94.28	87.5	80.25
F2	94.56	70.25	83.33	91.66	87.6	93.65
F3	94.68	71.42	84.25	95.11	87.8	91.25

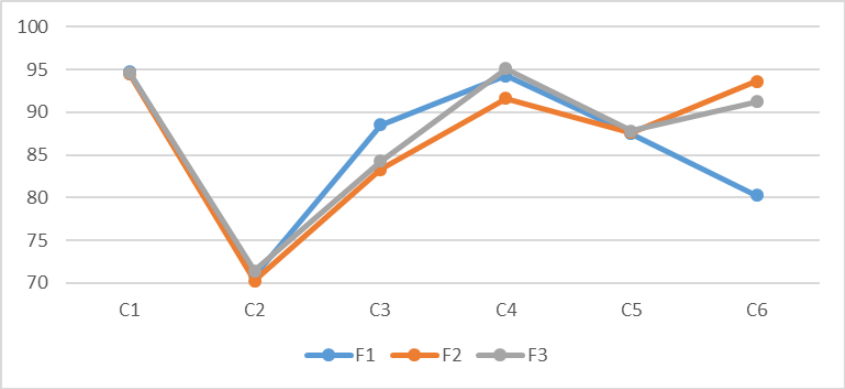

Figure 8. Algorithm Sensitivity

Table 9. Specificity Analysis

Features	AL1	AL2	AL3	AL4	AL5	AL6
F1	93.45	80.05	88.46	87.5	90.2	85.10
F2	90.45	85.71	90.79	87.27	91.0	90.50
F3	91.45	79.59	88.75	95.23	92.0	90.55

Table 10. AUROC Analysis

Features	AL1	AL2	AL3	AL4	AL5	AL6
F1	94.56	82.54	94.09	93.77	94.1	91.89
F2	93.03	77.27	93.43	94.41	95.5	91.81
F3	96.08	80.48	92.87	96.96	95.5	93.80

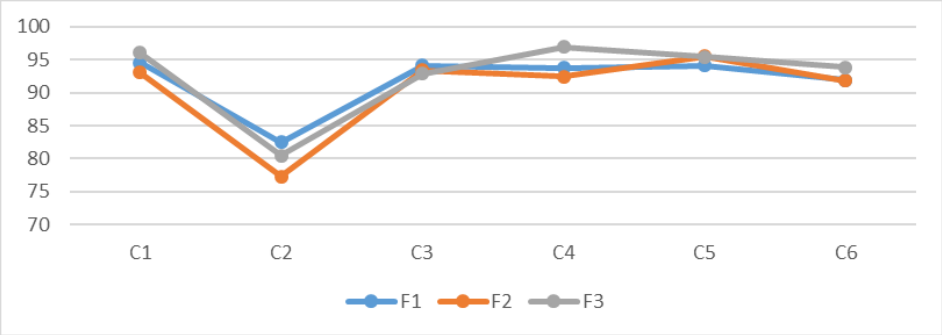

Figure 9. AUROC Analysis

6.7 Log Loss Analysis

Log loss was looked into in this study. Table 11 presents the outcomes produced by several techniques. Al2 received the best F3 score (8.56) in the current study. Al2 also provided F1 and F2 scores (7.59 and 8.01). Thus, between F2 and F3, Al1 provided the smallest log loss (0.27). Log loss ratings from 0.29 to 1.02 were obtained by the remaining algorithms. Figure 10 displays every log loss rating for every method for every SF.

7. Analysis

This study utilised the data set from UCI Cleveland in both training and testing, and a variety of ML methods were employed for precursor identification of cardiovascular disease. Six popular algorithms—LR, DT, RF, SVM, Gaussian NB, and KNN, among others—were used in particular to classify key characteristics that are more crucial when forecasting cardiovascular disease. Univariate analysis selection algorithms, Mutual information (MI), chi-square, ANOVA F value, and selection algorithms are also used. Specificity, Precision, Sensitivity, AUROC and log loss are among the criteria for valuation that were used to assess the execution of the various algorithms. According to the testing results, for all three of the SFs displayed in, strategy Al4 achieves the best accuracy (95.25%) for F3, and algorithm Al1 gets second-best accuracy (93.51%) shown in Table 7. As demonstrated in Tables 8 and 9,

All4 also had the greatest scores for F3 with regard to of specific (95.23) and sensitive (95.11). Then, according to Table 10, Al1 provided F3 with the greatest AUROC rating (96.96). Figure 11 display AUROC Curve. Following that, as indicated in Table 11, Al1 provides the smallest log loss value (0.27) for both F2 and F3. Al4 is the greatest model for prediction in terms of precision, specificity, and sensitivity since it performs most effectively with F3. Furthermore, Al1 outperforms F2 and F3, that is the second-greatest model for prediction overall, in terms of AUROC and log loss.

7.1 Contrasting with other pieces of work

By contrasting our study with other research, we discovered that Mohan *et al.* [25] used the HRFLM approach to create a heart disease prediction model. Using each of the thirteen attributes, the model we developed predicted (95.25%) precision, (95.11%) sensibility, and (95.23%) precision for the UCI cardiovascular disease data. Amin *et al*. [26] used both the logistic regression and 87.41% is an accurate prognosis for heart attack based on naive Bayes systems. A prior study [26] that used the J48 reduced error trimming technique yielded a precision of 56.76%. Table 12 presents further prior research, with a total precision ranging from 87.41 to 83.70%. Furthermore, nobody has thoroughly examined the forecasting of coronary artery disease; in contrast, our research assesses a variety of statistics (specificity, sensitivity, accuracy, AUROC, and log loss) and employs various techniques for attribute selection to enhance algorithmic achievement for key features

Table 11. Log Loss of Algorithms

Feature		AL1	AL2	AL3	AL4	AL5	AL6
Dataset	F1	0.31	7.59	0.38	0.33	0.31	1.02
	F2	0.27	8.01	0.36	0.34	0.3	0.67
	F3	0.28	8.56	0.4	0.31	0.31	0.62

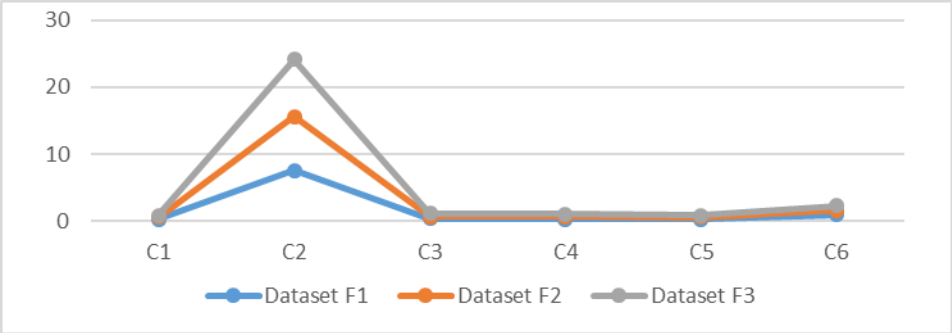

Figure 10. AUROC for F3

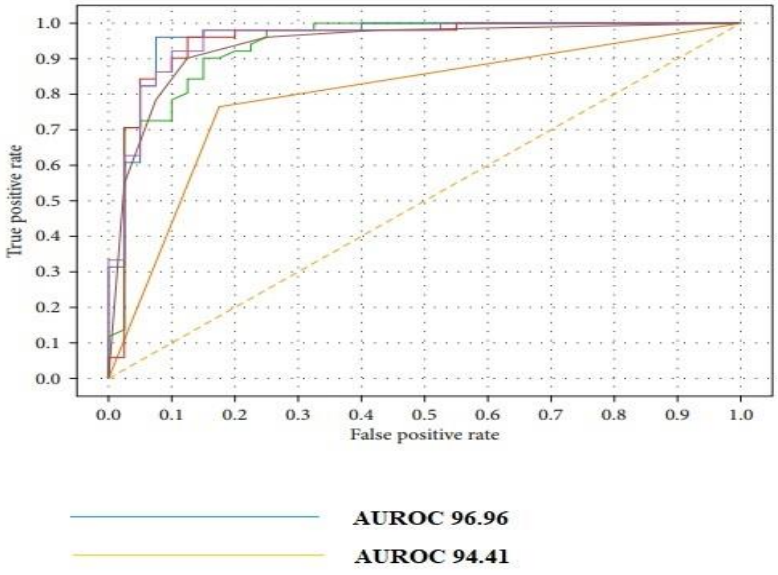

Figure 11. AUROC Curve

Table 12. Comparison of research

Authors	Method	Accuracy	Sensitivity	Specificity	AUROC	Log Loss
Our Research	RF	95.25	95.11	95.23	96.96	0.31
Mohan *et al.* [25]	HRFLM	88.47	92.8	82.6	-	-
Amin *et al.* [26]	NB, LR	87.41	-	-	-	-
Latha & Jeeva [27]	NB, B, RF	85.48	-	-	-	-
Patel *et al*. [28]	J48 reduced	56.78	-	-	-	-

8. Conclusion

In conclusion, three feature selection strategies were applied to identify the characteristics that were almost all useful in the diagnosis of heart failure . The chosen features were then subjected to six different prospective machine learning algorithms. Each algorithm utilized a distinct set of selected features to execute its core functions. Among all the algorithms evaluated in table 12, Random Forest (RF) demonstrated superior efficiency compared to the rest. Our research yielded an accuracy of 95.25%, a sensitivity of 95.11%, specificity of 95.23%, and an AUROC of 96.96%.

Yet, the lack of sufficient data on cardiovascular disease hindered the development of a more precise forecasting model. This study would yield more dependable outcomes by analysing a substantial amount of authentic medical data using a comparable approach. Future research will explore enhanced methods to enhance this prediction and boost the efficacy of techniques through the adoption of more robust feature selection strategies, including the integration of deep learning approaches. Alternatively, applying cross-validation techniques in the current method could lead to improved results.

References

[1] A.S. Deepika, N. Jaisankar, Detecting and Classifying Myocardial Infarction in Echocardiogram Frames with an Enhanced CNN Algorithm and ECV-3D Network. IEEE Access, 12, (2024) 51690-51703. https://doi.org/10.1109/ACCESS.2024.3385787

[2] M.S. Amin, Y.K. Chiam, K.D. Varathan, Identification of significant feature and data mining techniques in predicting heart disease. Telematics and Informatics, 36, (2019) 82-93. https://doi.org/10.1016/j.tele.2018.11.007

[3] T. Ullah, S.I. Ullah, K. Ullah, M. Ishaq, A. Khan, Y.Y. Ghadi, A. Algarni, Machine Learning-Based Cardiovascular Disease Detection Using Optimal Feature Selection. IEEE Access, 12, (2024) 16431-16446. https://doi.org/10.1109/ACCESS.2024.3359910

[4] P.A. Heidenreich, J.G. Trogdon, O.A. Khavjou, J. Butler, K. Dracup, M.D. Ezekowitz, E.A. Finkelstein, Y. Hong, S.C. Johnston, A. Khera, D.M. Lloyd-Jones, S.A. Nelson, G. Nichol, D. Orenstein, P.W.F. Wilson, Y.J. Woo, Forecasting the future of cardiovascular disease in the United States A Policy Statement From the American Heart Association. Circulation, 123(8), (2011) 933–944. https://doi.org/10.1161/CIR.0b013e31820a55f5

[5] G. Savarese, L.H. Lund, Global public health burden of heart failure. Cardiac failure review, 3(1), (2017) 7–11. https://doi.org/10.15420/cfr.2016:25:2

[6] C. Beyene, P. Kamat, Survey on prediction and analysis the occurrence of heart disease using data mining techniques. International Journal of Pure and Applied Mathematics,118(8), (2018) 165-174,

[7] V.V. Ramalingam, A. Dandapath, M.K. Raja, Heart dis ease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology, 7(2.8), (2018) 684–687. https://doi.org/10.14419/ijet.v7i2.8.10557

[8] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, J. Gutierrez, (2017) A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. IEEE Symposium on Computers and Communications, IEEE, Greece. https://doi.org/10.1109/ISCC.2017.8024530

[9] E. Fix, J.L. Hodges, Discriminatory analysis. Nonpara metric discrimination: consistency properties. International Statistical Review / Revue Internationale de Statistique, 57(3), (1989) 238–247. https://doi.org/10.2307/1403797

[10] S. Palaniappan, R. Awang, (2008) Intelligent heart disease pre diction system using data mining techniques. IEEE/ACS International Conference on Computer Systems and Applications, IEEE, Qatar. https://doi.org/10.1109/AICCSA.2008.4493524

[11] U. Haq, J.P. Li, M.H. Memon, S. Nazir, R. Sun, A hybrid intelligent system framework for the

prediction of heart disease using machine learning algorithms. Mobile Information Systems, 2018, (2018) 1-22. https://doi.org/10.1155/2018/3860146

[12] J. Cheng, G. Li, X. Chen, Research on travel time prediction model of freeway based on gradient boosting decision tree, IEEE Access, 7, (2019) 7466-7480. https://doi.org/10.1109/ACCESS.2018.2886549

[13] D. Stojanov, E. Lazarova, E. Veljkova, P. Rubartelli, M. Giacomini, Predicting the outcome of heart failure against chronic-ischemic heart disease in elderly population–Machine learning approach based on logistic regression, case to Villa Scassi hospital Genoa, Italy. Journal of King Saud University-Science, 35(3), (2023) 102573. https://doi.org/10.1016/j.jksus.2023.102573

[14] R. Yilmaz F.H. Yagin, C. Colak, K. Toprak, N. Abdel Samee, N.F. Mahmoud, A.A. Alshahrani, Analysis of hematological indicators via explainable artificial intelligence in the diagnosis of acute heart failure: a retrospective study. Frontiers in Medicine, 11, (2024) 1285067. https://doi.org/10.3389/fmed.2024.1285067

[15] P. Lakshmi Prabha, A.K. Jayanthy, C. Prem Kumar, B. Ramraj, Prediction of cardiovascular risk by measuring carotid intima media thickness from an ultrasound image for type II diabetic mellitus subjects using machine learning and transfer learning techniques. The Journal of Supercomputing, 77, (2021) 10289-10306. https://doi.org/10.1007/s11227-021-03676-w

[16] P. Ghosh, S. Azam, M. Jonkman, A. Karim, F.J.M. Shamrat, E. Ignatious, S. Shultana, A.R. Beeravolu De F. Boer, Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access, 9, (2021) 19304-19326. https://doi.org/10.1109/ACCESS.2021.3053759

[17] S.S. Yadav, S.M. Jadhav, Detection of common risk factors for diagnosis of cardiac arrhythmia using machine learning algorithm. Expert systems with applications, 163, (2021) 113807. https://doi.org/10.1016/j.eswa.2020.113807

[18] S. Razia, J.C. Babu, K.H. Baradwaj, K.S.S.R. Abhinay, M. Anusha, Heart disease prediction using machine learning techniques. International Journal of Recent Technology and Engineering, 8(4), (2019) 10316–10320. http://www.doi.org/10.35940/ijrte.D4537.118419

[19] A. Degerli, M. Zabihi, S. Kiranyaz, T. Hamid, R. Mazhar, R. Hamila, M. Gabbouj, Early Detection of Myocardial Infarction in Low-Quality Echocardiography. IEEE Access, 9, (2021) 34442–34453. https://doi.org/10.1109/ACCESS.2021.3059595

[20] D. Mienye, Y. Sun, Effective feature selection for improved prediction of heart disease. In Pan-African Artificial Intelligence and Smart Systems Conference, 405, (2022) 94–107. https://doi.org/10.1007/978-3-030-93314-2_6

[21] V. Vakharia, V.K. Gupta, P.K. Kankar, A comparison of feature ranking techniques for fault diagnosis of ball bearing, Soft Computing, 20(4), (2015) 1601–1619. https://doi.org/10.1007/s00500-015-1608-6

[22] N. Carrara, J. Ernst, On the estimation of mutual information. Proceedings, 33(1), (2020) 31. https://doi.org/10.3390/proceedings2019033031

[23] T. Akter, M.S. Satu, M.I. Khan, M.H. Ali, S. Uddin, P. Lio, J.M.W. Quinn, M.A. Moni, Machine learning-based models for early stage detection of autism spectrum disorders. IEEE Access, 7, (2019) 166509-166527. https://doi.org/10.1109/ACCESS.2019.2952609

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, É. Duchesnay, (2011). Scikit-learn: Machine learning in Python. Journal of machine Learning research, 12, 2825-2830.

[25] S. Mohan, C. Thirumalai, G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques, IEEE Access, 7, (2019) 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

[26] M.S. Amin, Y.K. Chiam, K.D. Varathan, Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, (2019) 82–93. https://doi.org/10.1016/j.tele.2018.11.007

[27] [27] C.B.C. Latha, S. C. Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16, (2019) 1-9, https://doi.org/10.1016/j.imu.2019.100203

[28] J. Patel, D. Tejal Upadhyay S. Patel, Heart disease prediction using machine learning and data mining technique. Heart Disease, 7(1), (2015) 129-137.

Acknowledgement

The Authors would like to thank the Cleveland Heart Disease Data Set for this research. It helps to get a better result.

Authors Contribution Statement

Prashant Maganlal Goad: Conceptualization, Methodology, Data curation, Formal analysis, Validation, Writing – original draft, Writing – review & editing. Pramod J Deore: Formal analysis, Validation, Writing – original draft, Writing – review & editing. Both the authors read and approved the final version of the manuscript.

Funding

The Authors has not received any funding support from any institute.

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The Dataset generated during the study will be made available on request.

Has this article screened for similarity?

Yes