Using Machine Learning Approach to Identify and Analyze High Risks Patients with Heart Disease

Wenbo Sun*

Weatherhead School of Management, Case Western Reserve University, Cleveland,

*Corresponding author: OH wxs336@case.edu

Keywords: Heart disease, Random Forest, ANOVA Analysis, Classification.

Abstract: Cardiovascular disease is one of the most threatening diseases to human health today. Exploring the performance of different models in predicting cardiovascular diseases will help medical practitioners to make more accurate medical diagnoses using non-invasive means to save lives. In this paper, a comparative analysis of different classification prediction models was applied to predicting heart disease cases using heart disease data from the UCI machine learning Repository. This data source contains 14 dimensions of data for 303 patients. The classifiers applied in this study were decision trees, random forests, support vector machines (SVM) and logistic regression. To examine the performance of each classifier, criteria such as accuracy, sensitivity, and specificity were used, and a 10-fold cross-validation method was used to measure the unbiased estimates of these prediction models. According to our results, SVM can make predictive judgments for suspected cardiovascular disease cases to the maximum extent possible.

1. Introduction

Heart disease is the leading cause of death before cancer and traffic accidents [1]. Data from the CDC reveal that heart disease causes about 655,000 deaths each year, accounting for 25% of all deaths in the United States [2]. According to the National Cancer Institute's definition of heart disease, heart disease refers to a category of diseases that affect the heart or blood vessels. In fact, the risk of cardiovascular disease is associated with other factors as well; smoking, high blood pressure, high cholesterol, unhealthy diet, lack of exercise and obesity all contribute to the potential risk of heart disease. However, identifying heart disease can be difficult due to the uncertainty of clinical practice. Due to this limitation, scientists have turned to modern methods, such as data mining and machine learning, to predict disease. In this regard, it is particularly important to compare the performance of various techniques and algorithms and determine the best approach.

This study builds classification models to make predictions on typical data from heart disease patients. It compares the performance of different algorithms to gain insight into the ability of different models to reveal hidden patterns in medical data. Understanding the ability of models can empower non-invasive means in cardiovascular disease diagnosis and assist medical professionals in making more accurate judgments. This can help save lives through early diagnosis of heart problems and save money by avoiding costly invasive treatments. Existing classifier algorithms, such as decision trees [3], logistic regression, random forests, and support vector machines [4], and hybrid data mining [5], have been used to explore different types of cardiac problems. Medical data mining has great potential in exploring hidden patterns in datasets in the clinical domain.

2. Related Works

The extensive use of classifiers to model and diagnose cardiovascular diseases has promoted this work. This section will describe the results of a brief literature survey. In reference [6], the authors used an open-heart disease dataset from the Cleveland Clinic and classified 303 patients by different methods of decision trees (CART, ID3, DT) and concluded that the CART classifier had the best performance with an accuracy of 83.49%. A similar study was conducted in [7] using the same dataset,

where researchers trained a deeper neural network on the feature scaling data and improved the accuracy significantly to 96%. In reference [8], the authors improve the accuracy from 98.97% to 100% by balancing the original data, and in the context of the proposed hybrid prediction model [9], Luxmi and Sangeet tested the performance of eight different classifiers(SVM, neural network, decision tree, generalized linear model, Lasso, Bayesian regularized neural network, classification and regression tree) for the prediction of the UCI heart disease database showing that for a single model, support vector machine, the logistic regression classifier and random forest performed better, with accuracies of 86%, 84% and 83%, respectively. In a similar study in [10], the authors compared the performance of different models using the same dataset, and this time, logistic regression classifier and support vector machine proved to be the best methods. Another study [11] compared the effect of different feature selection methods on the prediction effectiveness of the models. The experimental results proved that using the combination of CFS and PSO was effective in improving the model's prediction, and the combination improved the correct rate of the MLP algorithm by almost 7%. The following table summarizes and describes in detail all the methods previously analyze. The obvious conclusion is that the hybrid algorithm-based methods provide more accurate results than those using a single algorithm. Also, feature selection can improve the performance of the model. The pre-analysis of the existing literature helped the study to adopt the best algorithm based on the results obtained in previous papers. For this purpose, the study will optimize the input data through feature selection and train the test set on support vector machines, decision trees, and logistic regression models.

3. Data& Methods

The following section is a brief discussion of the method and materials applied in this research.

3.1 Data Source

The dataset used in the study is obtained from UCI machine learning Repository [12], contains 303 instances and 14 attributes, which is also the most commonly used by researchers. After processing the missing values, six samples will be removed, and the sample used for this study is composed of 13 characteristics from 297 patients. The output field, which is defined as angiographic disease status, has a value range of 1 to 4. To simplify the prediction, the new target will appear as a binary value, value 0 for cases without risk of heart disease, and other value (value 1,2,3,4) means the presence of cardiovascular disease. Complete information and descriptions of the 297 instances of the 13 features in the dataset are given in Table 1.

Table 1. Attributes of UCI heart disease dataset

Index	Attribute	label	Description	Domain range of values
1	Patient's age	age	age in years	29-77
2	Patient's gender	sex	1 = male 0 = female	0,1
3	Chest pain type	cp	1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 = asymptomatic	1,2,3,4
4	Resting blood pressure	trestbps	in mm Hg on admission to the hospital	94-200
5	Cholesterol measurement	chol	in mg/dl	126-564
6	Fasting blood sugar	fbs	1 = fbs> 120 mg/dl 0 = fbs≤120mg/dl	0,1
7	A blood disorder called thalassemia	thal	3 = normal 6 = fixed defect 7 = reversible defect	3,6,7
8	Resting electrocardiographic results	resting	0 = normal 1 = having ST-T wave abnormality 2 = probable or definite left ventricular hypertrophy	0,1,2
9	maximum heart rate	Thalach		71-202
10	Exercise induced angina	exang	1 = yes 0 = no	0,1
11	ST depression induced by exercise relative to rest.	Oldpeak		0-6.2
12	The slope of peak exercise ST segment	slope	0 = downsloping 1 = flat 2 = upsloping	0,1,2
13	Number of major vessels (0-3) colored by fluoroscopy	ca		0,1,2,3

3.2 Proposed model

The following section describes how the proposed model has been developed to predict the occurrence of cardiovascular disease and how the performance of different feature selection methods and machine learning algorithms are tested. Feature selection algorithms, such as ANOVA and LASSO, were used to pick up important features. Then the performance of the famous classifiers applied in related works, such as KNN, decision tree, Naive Bayes, random forest, SVM and logistic regression, were tested Cross-validation method will be applied in the process of training data. To evaluate the results of the statistical analysis, this study also uses different performance metrics. The proposed model consists of four functional modules, in the order of data preprocessing, feature selection, machine learning, and evaluation. Figure p.1 shows the main workflow of the whole process.

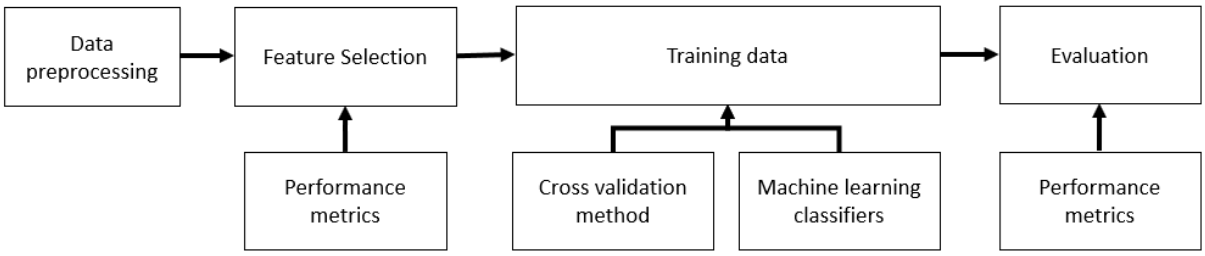

Figure 1. The main workflow of predicting heart disease

3.3 Data preprocessing

Low-quality data will lead to low quality mining results. Besides that, the noise in the clinical data set increases the computational effort and computation time. Thus, pre-processing of data is necessary to build effective machine learning classifiers. This process involves dealing with missing values and standard scalars.

(1) platform

With respect to [14], python provides a good platform for data analysis and machine learning due to its object-oriented, open-source nature, and the scikit-learn package enables a comprehensive list of machine learning methods and is a handy toolkit for statisticians. Scholars have found the python environment to be concise and accurate and have used it to predict heart disease in medical data [15].

In this paper, several packages, especially scikit-learn, are used in the python environment to apply various statistical models for data analysis, data visualization, feature selection and training and testing using machine learning methods.

(2) Feature Selection

Medical experts face several problems when using algorithms to make diagnoses on clinical datasets: because clinical datasets are often complex, unintuitive, and contain subjective data, some features in the dataset may be redundant or irrelevant, which can lead to degradation of the classifier's performance [16]. At this point, feature selection will become our concern. Feature selection is an integral part of data analysis and mining. By selecting features in the data that contribute more to the final result, overfitting can be reduced in a way that reduces data noise. At the same time, clinical datasets are usually high-dimensional, limiting the medical experts from manually removing features that contribute very little to the results, leading us to turn to automatic means of feature selection. According to related studies comparing different means of feature selection [17], SelectKBest paired with a classifier has the highest performance in processing high-dimensional data, yielding 97% accuracy in 0.11 seconds.

This study uses three feature methods for selecting the most informative features.

(a)Random Forest Feature selection

The main idea of feature importance assessment using random forests is that by calculating the contribution made by each feature on each CART tree in a random forest and averaging the sum, a comparison of the contribution size of different features can be made. The measures of contribution include Gini index and oob error (out-of-bag error). The Gini index will be used as the evaluation criterion in this study.

(b) ANOVA Method for Feature selection

The basic idea of ANOVA is that it allows the contribution of different sources of variation to the total variation to be assessed and thus to objectively determine the magnitude of the influence of controllable factors on the study results [18]. The method can select a subset of features with significant influence on the system state from many sample features with strong correlation and redundancy as a new sample of features reflecting the system state.

(c) Least Absolute Shrinkage and Selection Operator

The least absolute shrinkage and selection operator, also Lasso or LASSO, is a commonly used feature selection method in machine learning. It makes the training solving parameter process take into account the magnitude of the coefficients by adding a penalty term to the loss function (i.e.,

optimization objective), and by setting the scaling factor (penalty coefficient), it will make the coefficients of the less influential features decay to zero and only retain the important features. Lasso feature selection method suffers from low stability, i.e., it may lead to large model differences even when there are small changes in the data. Also, it cannot handle data with multicollinearity [19]: when there is a high correlation between the data, it tends to select one from each group and ignore the others.

(3) classifiers

Machine learning classifiers are good at identifying hidden patterns and regularities within data. In this study, a number of machine learning classifiers were used to predict cardiovascular disease incidence events. In the following, their theoretical background will be briefly described.

(a) Logistic Regression

Logistic regression is a generalized linear regression model used to deal with classification problems. The probability of a patient having the cardiovascular disease is 1 when the event occurs and 0 when healthy. p is the predicted probability of disease occurrence, then 1-p is the probability that the patient has no risk of heart disease. Suppose that the independent variables are x1, x2…, xn, then the logistic regression formula can be expressed as

$ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$, while β_i (i = 0, 1, 2, …, k) is the regression coefficient.

(b) SVM

The basic idea of the support vector machine is to map the sample data to a higher-dimensional space and to build a hyperplane in such a higher dimensional space so that the distance between the hyperplane and different class sample sets is maximized for the purpose of classification. In this study, the svc function in the sklearn package is used to implement the model prediction work of the support vector machine, and the penalty parameter is 0.5 when the kernel function is radial basis kernel function (RBF) and 1 when the kernel function is linear, as confirmed by the hyperparameter auto search module GridSearchCV.

(c) Naive Bayes

Naive Bayes is a simple but very powerful linear classifier. It uses the training data set to find the conditional probability value of each given class vector based on data point xi in the feature data set. After calculating the conditional probability value of each vector, it calculates the class of the new vector-based on its conditional probability. The Naive Bayes approach has a stable classification efficiency in prediction.

(d) Decision Tree

Decision trees usually have three steps: feature selection, decision tree generation, and decision tree pruning. When building a decision tree, a feature of the instance will be tested starting from the root node. The instance will be assigned to its child nodes according to the test result, at which time each child node corresponds to a value taken for that feature. So, on recursively, the instance will be tested and assigned until it reaches the leaf node, and finally, the instance will be assigned to the class of the leaf node.

(e) Random Forest

A random forest is an algorithm that integrates multiple decision trees through the idea of integrated learning. To explain it intuitively, each decision tree is a classifier, then for one input sample, N trees will have N classification results. And the random forest integrates all the classification votes, designating the category with the most votes as the final output. In the process of building the random forest model, after parameter tuning, the maximum depth of the tree is 4, and the minimum number of samples of leaf nodes is 2, with calculating Gini as the criterion.

(f) K-Nearest Neighbor

The principle of KNN is that when predicting a new value x's, it determines which class x belongs to base on what class it is from the nearest K points. Compared to other algorithms, the model training time is fast, and it is easy to obtain higher accuracy. When comparing the performance of different knn

algorithms, the model has good performance when k equals 6. The following section will not show the rest of the knn model training results.

(d) Cross Validation

To avoid data bias from a single partitioned dataset, k-fold cross-validation is applied. The following picture shows its basic idea. By dividing the whole data into 10 copies and taking one copy at a time randomly without duplication as the test set, while using the other 9 copies as the training set to train the model, ten biases will be obtained. Their mean value will be the evaluation result, which will be the closest to the real performance of the model [20].

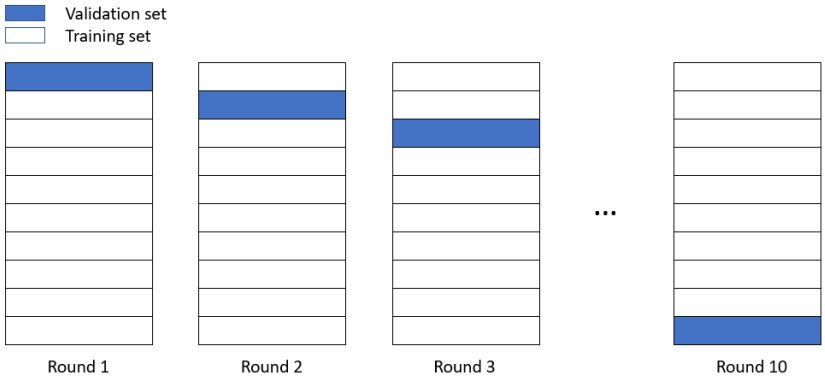

Figure 2. The main idea of 10-fold cross validation

(4) Performance Metrics

This study will evaluate the performance of different models using confusion matrices. The confusion matrix can help us to quickly visualize the model performance to help us further adjust the parameters or choose a better model. Confusion matrix consists of an N × N matrix (N category), where each row represents the true category to which the data belongs, and the total amount of data in each row represents the number of data instances in that category. The values in each column indicate the number of real data instances predicted to be in that category. The following figure shows its basic form.

Table 2. Confusion matrix

	Predicted: At the risk of heart disease	Predicted: Healthy
Actual: Cases diagnosed with heart disease based on angiographic findings (target = 1)	TP	FN
Actual: healthy (target = 0)	FP	TN

In this case, if a patient (target = 1) is correctly labeled as the person at risk of heart disease, this case will be classified as TP. Those subjects who were correctly classified as healthy (target = 0) by the model are classified as TN. At the same time, FP, also called type I error, means that the model has made a wrong decision to label a healthy person as a cardiac patient. FN shows the cases when cardiac patients are incorrectly predicted to be healthy by predictive models (Type II error).

Based on the different classifications, several evaluation criteria are calculated:

Accuracy: The percentage of correctly predicted cases out of all cases, which is calculated as:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \times 100\% \quad (1)$$

True positive rate: True positive rate (TPR), also Sensitivity (Sn) and recall, measures how apt the model is to detect events in the positive class. In this case, Sn quantifies how many of the actual heart disease patients are correctly predicted as at the risk of heart disease. Sensitivity is calculated as:

$$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}} \times 100\%\% \quad (2)$$

False positive rate: False positive rate (FPR) measures how exact the model is to the actual false category. In this case, Sp quantifies how many of the healthy ones are incorrectly predicted in the actual healthy category. FPR is calculated as:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN+FP}} \times 100\%\% \tag{3}$$

Matthews' correlation coefficient (MCC): The Matthews correlation coefficient (MCC) or phi coefficient is a simple and efficient measurement for binary classification problems [21]. The output value from -1 to +1 represents a perfect match between the predicted and true values of the model from a perfect mismatch to a perfect match.

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TN+FN})(\mathrm{TP+FP})(\mathrm{TN}+\mathrm{FP})(\mathrm{TP+FN})}} \times 100\%.\% \tag{4}$$

Area Under the Curve (AUC): AUC is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the model's performance at distinguishing between the positive and negative classes.

Processing time: this criterion measures the time efficiency of the model. The processing time of the training part, which includes using 10-fold cross-validation to evaluate the classifiers, will be recorded using the time package in python.

4. Experimental Results

The following section discusses the classifiers' performance and the influence of different kinds of feature selection methods. First, the six classifiers (Logistic regression, SVM, Naïve Bayes, decision tree, random forest, k-nearest neighbor) were applied on heart disease data on full features. Next, three kinds of feature selection methods that have been mentioned below (Random Forest, ANOVA, LASSO) were applied to select the most informative features, respectively, and the classifiers will train the filtered input to obtain new results. All features were normalized and standardized before applying to classifiers. Also, 10-fold cross validation was applied to make sure these models were stable and reliable.

4.1 Comparison of results for different feature selection methods

The comparison of rankings of informative features by different feature selection methods is shown as follows. Among them, the ANOVA method will be implemented using the sklearn method on feature selection. For both random forest and lasso feature selection methods, the determination of the optimal parameters is achieved by GridSearchCV referencing. Random forest will be applied with n = 80, i.e., the number of decision trees is 80. The optimal alpha for lasso feature selection is 0.1.

It can be seen that different feature selection methods agree on the importance of some features, such as thal (blood disorder), thalach (heart rate), exang (exercise induced angina), oldpeak (ST depression) and ca (angiography results). These features are commonly admitted as important. As for fbs(blood sugar),chol(cholesterol), restecg (electrocradiograpic results) and trestbps (blood pressure), although their rankings have fluctuated somewhat, they are still generally among the less important characteristics.

Some disagreement emerged between the different methods regarding the importance of chest pain features. The chest pain feature was the most important feature in the random forest feature selection method, at the same time, the other two methods came to relatively more different conclusions.

Figure 3. Comparison of results for different feature selection methods (in rankings, 14 as not selected)

4.2 Results of 10-Fold Cross-Validation for Classifiers Performance on Full Features

In this section, the six classifiers mentioned below (logistic regression, SVM, Naïve Bayes, decision tree, random forest, k-nearest neighbor) were applied to the UCI dataset with full features. The input data has been standardized before putting into use. In this experiment, different parameters have been tested using GridSearchCV, and the result from only the best classifiers is shown in the table. The performance of the different models is rated by comparing different performance metrics. Additionally, 10-fold cross validation is applied, which means that in each training round, 90% of data will be used as a training set, and 10% of the data will be test data.

In the table, the SVM RBF has a relatively good performance with an accuracy rate of 84.5%, 78.4% TPR,69.5%MCC and 9.7% FPR, 90.5% AUC. For diagnostic accuracy, SVM has the best TPR (True positive rate) and best FPR (False positive rate), representing a smarter ability to identify suspicious cases to the greatest extent possible and avoid wasting medical resources. A relatively good model is Naïve Bayes, which gives out 75.7% accuracy, 72.6% TPR, 20.7% FPR, 52.2% MCC and 75.9% AUC. A significant advantage of this model is that it takes the shortest time, which is almost a third of SVM's processing time. In the dimension of effectiveness, Naïve Bayes has the better AUC and MCC rate, which means the prediction matches real labels.

Table 5. Performance metrics for different classifiers on full features

	Performance Metrics					
classifiers	Accuracy	TPR	FPR	MCC	AUC	processing time (second)
DT	0.7578	0.7267	0.207	0.5225	0.7597	0.048
Naive Bayes	0.8419	0.7974	0.1206	0.6801	0.8935	0.048
random forest (n_estimators = 48)	0.8286	0.7588	0.1089	0.657	0.9075	0.4471
SVM rbf(C = 0.5)	0.8455	0.7841	0.0972	0.6958	0.9056	0.161
SVM linear (C = 1)	0.825	0.7842	0.1341	0.6526	0.9078	0.1465
K-NN (n_neighbors = 6)	0.8118	0.723	0.1065	0.6277	0.8892	0.0621
Logistic Regression (C= 0.001, penalty= 'l2')	0.7784	0.5692	0.0347	0.5898	0.9084	0.061

4.3 Results of K-Fold Cross-Validation (k = 10) Classifier Performance on Nine Selected Features by Random Forest Feature Selection Algorithm

Table 6. Performance metrics for different classifiers on features selected by random forest

	Performance Metrics					
classifiers	Accuracy	TPR	FPR	MCC	AUC	processing time (second)
DT	0.7411	0.7084	0.2206	0.4868	0.7439	0.046
Naive Bayes	0.8385	0.7982	0.1269	0.6719	0.8939	0.048
random forest	0.8184	0.7777	0.1413	0.6388	0.9075	0.4295
SVM rbf	0.8423	0.7675	0.0935	0.6849	0.9074	0.124
SVM linear	0.8388	0.7531	0.0863	0.6813	0.9034	0.161
6-NN	0.8188	0.7254	0.0977	0.6452	0.8886	0.0612
Logistic Regression	0.7917	0.5915	0.028	0.6186	0.8963	0.059

4.4 Results with K-Fold Cross-Validation of Classifiers Performance on Nine Selected Features by ANOVA Feature Selection Algorithm.

Table 7. Performance metrics for different classifiers on features selected by ANOVA

	Performance Metrics					
classifiers	Accuracy	TPR	FPR	MCC	AUC	processing time (second)
DT	0.7411	0.7073	0.2255	0.4865	0.741	0.0619
Naive Bayes	0.8454	0.8074	0.1242	0.6853	0.8952	0.0609
random forest	0.8388	0.7843	0.1118	0.6825	0.9035	0.5451
SVM rbf	0.8453	0.7801	0.0999	0.692	0.9063	0.1695
SVM linear	0.842	0.7912	0.1142	0.6844	0.9081	0.1466
6-NN	0.8254	0.7618	0.1153	0.6588	0.8761	0.0828
Logistic Regression	0.7782	0.5634	0.028	0.5956	0.9056	0.0735

4.5 Results with K-Fold Cross-Validation of Classifiers Performance on Selected Features (n = 9) by LASSO Feature Selection Algorithm.

Table 8. Performance metrics for different classifiers on features selected by LASSO

	Performance Metrics					
classifiers	Accuracy	TPR	FPR	MCC	AUC	processing time (second)
DT	0.7585	0.763	0.2402	0.5214	0.7606	0.046
Naive Bayes	0.8285	0.774	0.1242	0.6555	0.8871	0.0441
random forest	0.8284	0.7886	0.1371	0.6552	0.9001	0.4059
SVM rbf	0.8354	0.7775	0.1142	0.6719	0.9027	0.1849
SVM linear	0.8457	0.7641	0.0819	0.6952	0.9042	0.1496
6-NN	0.8321	0.7372	0.0859	0.668	0.872	0.0983
Logistic Regression	0.7751	0.5604	0.0288	0.591	0.9003	0.4723

5. Conclusion

Different data mining techniques can be used to identify and prevent cardiovascular disease in patients, potentially being put into new artificial intelligence devices and assisting professionals in their judgment. This paper compares the performance of different classifiers for predicting cardiovascular disease in patients: decision trees, random forests, support vector machines, and logistic regression. These techniques are compared based on sensitivity, specificity, accuracy, error rate, true positive rate and false positive rate. Our study shows that the support vector machine model is the best classifier for cardiovascular disease prediction. In the future, we intend to improve the performance of

these basic classification techniques by creating metamodels, which will be used to predict cardiovascular disease in patients.

References

[1] M. Kirmani, "Cardiovascular disease prediction using data mining techniques," Oriental Journal of Computer Science and Technology, vol. 10, pp. 520-528, 2017.

[2] Heart Disease | cdc.gov. [online]https://www.cdc.gov/heartdisease/index.htm

[3] Randa El-Bialy, Mostafa A. Salamay, Omar H. Karam, M. Essam Khalifa,Feature Analysis of Coronary Artery Heart Disease Data Sets,Procedia Computer Science,Volume 65, 2015, Pages 459-468, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.09.132.

[4] J. S. Sonawane and D. R. Patil, "Prediction of heart disease using multilayer perceptron neural network," International Conference on Information Communication and Embedded Systems (ICICES2014), 2014, pp. 1-6, doi: 10.1109/ICICES.2014.7033860.

[5] M. Saini, N. Baliyan and V. Bassi, "Prediction of heart disease severity with hybrid data mining," 2017 2nd International Conference on Telecommunication and Networks (TEL-NET), 2017, pp. 1-6, doi: 10.1109/TEL-NET.2017.8343565.

[6] Chaurasia, Vikas and Pal, Saurabh, Early Prediction of Heart Diseases Using Data Mining Techniques (2013). Caribbean Journal of Science and Technology, Vol. 1, 208-217, 2013, Available at SSRN: https://ssrn.com/abstract=2991237

[7] Darmawahyuni, A. Coronary Heart Disease Interpretation Based on Deep Neural Network. Comput. Eng. Appl. J. 2019, 8.

[8] Durgadevi Velusamy, Karthikeyan Ramasamy, Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset, Computer Methods and Programs in Biomedicine, Volume 198, 2021,105770, ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2020.105770.

[9] A. UL HAQ, J. PING 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), 1–21

[10] M. A. JABBAR, B. L. DEEKSHATULU, P. CHANDRA:Prediction of heart disease using random forest and feature subset selection , Advances in intelligent systems and computing, 424 (2016), 187–196

[11] L. VERMA, S. SRIVASTAVA, P. C. NEGI: A Hybrid Data MiningModel to Predict Coronary Artery DiseaseCases Using Non-Invasive Clinical Data, Journal of Medical Systems, 40 (2016), 1–7

[12] UCI Machine Learning Repository: Heart Disease Data Set

[13] Darmawahyuni, A. Coronary Heart Disease Interpretation Based on Deep Neural Network. Comput. Eng. Appl. J. 2019, 8.

[14] Hao J, Ho TK. Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. Journal of Educational and Behavioral Statistics. 2019;44(3):348-361. doi:10.3102/1076998619832248

[15] A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.

[16] M. Shouman, T. Turner and R. Stocker, "Using data mining techniques in heart disease diagnosis and treatment," 2012 Japan-Egypt Conference on Electronics, Communications and Computers, 2012, pp. 173-177, doi: 10.1109/JEC-ECC.2012.6186978.

[17] Powell, A., Bates, D., van Wyk, C. & Darren de Abreu, A. A cross-comparison of feature selection algorithms on multiple cyber security data-sets. CEUR Workshop Proc. 2540, 196–207 (2019).

[18] George A Morgan. SPSS for introductory statistics [M]. Lawrence Erlbaum Associates, 2004.

[19] Fonti, V., and Belitser, E. (2017). Feature selection using LASSO. Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S., and Turner, R. (1996). Movement-related effects in fMRI time-series. Magn. Reson. M 35, 346–355. doi: 10.1002/mr2910350312

[20] Kahramanli, H. & Allahverdi, N. Design of a hybrid system for the diabetes and heart disease, Expert systems with applications, pp. 82-89, 2008.

[21] Boughorbel S., Jarray F., El-Anbari M. (2017). Optimal classifier for imbalanced data using Matthew's correlation coefficient metric. PLoS ONE 12: e0177678. 10.1371/journal.pone.0177678