International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(Review Article)

Boruta based feature selection model for heart disease prediction

Yutika Agarwal [1], Rita Chhikara [1, *] and Sanjeev Rana [2]

[1] *School of Engineering and technology, The Northcap University, India.*

[2] *VISA Worldwide Inc, Singapore.*

International Journal of Science and Research Archive, 2023, 10(01), 768–774

Publication history: Received on 04 September 2023; revised on 10 October 2023; accepted on 13 October 2023

Article DOI: https://doi.org/10.30574/ijsra.2023.10.1.0830

Abstract

In today's time the rate of heart disease is increasing at a very fast pace and because of that it is becoming the reason for major cause of deaths worldwide. It is very important to give treatment for heart disease or predict any such disease beforehand but there are some medical centers where experts lack appropriate or fair expertise to diagnose and treat the patient on time. So often they assume their readings and as a result, poor outcome is shown which sometimes lead to death of the patient. This paper identifies the relevant attributes of heart diseases using Boruta, Lasso and Ridge feature selection method. It also presents valuable insight on effectiveness of various machine learning algorithms to predict heart disease. The feature selection method reduces number of features and at the same time maintaining comparable accuracy of the model. Experimental results demonstrate that Boruta feature selection with Random Forest classifier outperforms all the other state-of-art methods used in this study.

Keywords: Boruta; Lasso; Ridge; Random Forest; XGBoost; Logistic Regression; Naïve Bayes

1. Introduction

According to statistics from the World Health Organization (WHO), heart disease is the major cause of the mortality worldwide, resulting in around 17.9 deaths annually [1]. Heart attacks occur due to blockage in blood flow or an imbalance in certain health parameters. Individuals who have a high level of danger to get exposed to heart disease exhibit signs of elevated blood pressure, glucose and lipid levels as well as stress. The symptoms related to heart problems are somewhat similar or have same type of characteristics when compared with other illnesses and age-related factors may further complicate the diagnosis of healthcare professionals, leading to delays in treatment. The timely and accurate prediction of heart disease, combine with early detection plays a crucial role in improving patient survival rates [2].

When the heart and the blood vessels are affected, there is a possibility that it can lead to certain heart disease conditions. This includes how the fluid circulates in the body when it enters the bloodstream. The accurate diagnosis of such diseases is crucial and it is a difficult task which should be done efficiently and effectively. Medical experts play a vital role in making correct/accurate decisions which are essential for providing quality treatment to the patients [3]. Therefore, medical centers must provide training and guidance to healthcare professionals who may lack sufficient expertise in diagnosing these diseases. This training is necessary to ensure accuracy of all the important readings related to heart and other body parameters.

The existing methods for the diagnosis and prediction of heart disease have certain limitations including the challenge of accurately predicting the diseases [4]. In order to address this issue, this paper aims to improve upon these constraints by utilizing the Boruta algorithm and machine learning algorithms to identify relevant features and enhance accuracy and predictability for heart disease. Boruta algorithm helps in identifying the most significant features from

[*] Corresponding author: Rita Chhikara; Email: ritachhikara@ncuindia.edu

the dataset, providing valuable insights into the factors that contribute to heart disease. To achieve this, a heart dataset was taken from UCI Repository, which serves as the basis for training and testing the machine learning models.

Through this research, the goal is to overcome the limitations of existing methods by leveraging machine learning and feature selection techniques. By doing so, it is expected that the proposed approach will enhance the ability to predict specific conclusions in the diagnosis and treatment of heart disease, ultimately leading to more effective healthcare interventions.

2. Related Work

Numerous studies and analyses have been conducted to predict heart disease using various methods and machine learning algorithms. These studies have explored the accuracy of different algorithms and compared their performance. Avinashi Golande et al. [5] examined several machine learning algorithms, including Decision Tree, KNN, and k-means, for classifying heart disease. Their findings suggested that Decision Tree had the highest accuracy among the algorithms tested. They also concluded that by adjusting parameters and indicators, the efficiency of the Decision Tree algorithm could be further improved. Fahd Saleh Alotaibi [6] developed a machine learning model to compare five different methods. The classification algorithms Decision Tree, Logistic Regression, Random Forest, Naive Bayes, and SVM were evaluated, with Decision Tree being identified as the most accurate.

Theresa Princy R et al. [7] employed Naive Bayes, KNN, Decision Tree, and Neural Network algorithms and evaluated the accuracy of these classifiers with different attribute counts. Nagaraj M Lutimath et al. [8] focused on the Naive Bayes and SVM algorithms for heart disease prediction. The study concluded that SVM outperformed Naive Bayes in terms of accuracy. Anna Karen et. Al. [9] applied combination of chi-square and Principal component analysis for identifying relevant features for heart disease prediction. Best results were obtained by random forest classifier after reducing the dimensions.

In another study [10] authors have applied correlation and ANOVA feature selection method to improve the accuracy of classifier. Some of the feature selection methods applied for heart disease in literature are ReliefF [11], Minimum Redundancy Maximum Relevance Feature Selection (MRMR) [12], Genetic Algorithm [13], Cuckoo Search [14], chi-square, Relief and Correlation model combined together [15] etc. It has been shown that feature selection [16] plays a vital role in enhancing the performance of the machine learning models as redundant features tend to degrade the performance of the model. To the best of our knowledge Boruta wrapper feature selection method has not been applied for heart disease prediction.

3. Proposed Methodology

The whole object about the implementation and analysis of the proposed ideology is to predict the major symptoms for the early detection of the existence of heart disease. In this approach, different machine learning algorithms, Logistic Regression, Naive Bayes, Random Forest Classifier, Extreme Gradient Boost (XGB), K-Nearest Neighbor (KNN), Decision Tree and Support Vector Machine (SVM) are used to predict the heart disease based on some health parameters. In this prediction analysis Receiver Operating Characteristic Area Under the Curve is also used as a performance metrics [9] which is used to determine the performance of the classification models. The ROC curve helps to showcase a plot that represents the relationship between the true positive rate (TPR) and the false positive rate (FPR) at various threshold settings. ROC AUC score in this study has helped to compare the performance of different models that are used in prediction for a better conclusion as this approach is a threshold-independent, reliable, and comprehensive metric that takes false positives and false negatives into consideration.

To have a better accuracy and better result feature selection methods like Lasso [17], Ridge [18] Boruta [19] have been applied. Data is split into two categories that is training and testing data. It is split in such a way that for testing purpose 25% data is utilized while for training purpose 75% data is taken into consideration.

Different classification algorithms were analyzed namely Decision Tree, Random Forest, Logistic Regression and Naive Bayes based on their Accuracy, Precision, Recall and f-measure scores and identified the best classification algorithm which can be used in the heart disease prediction. The general framework for proposed methodology is given in Figure 1.

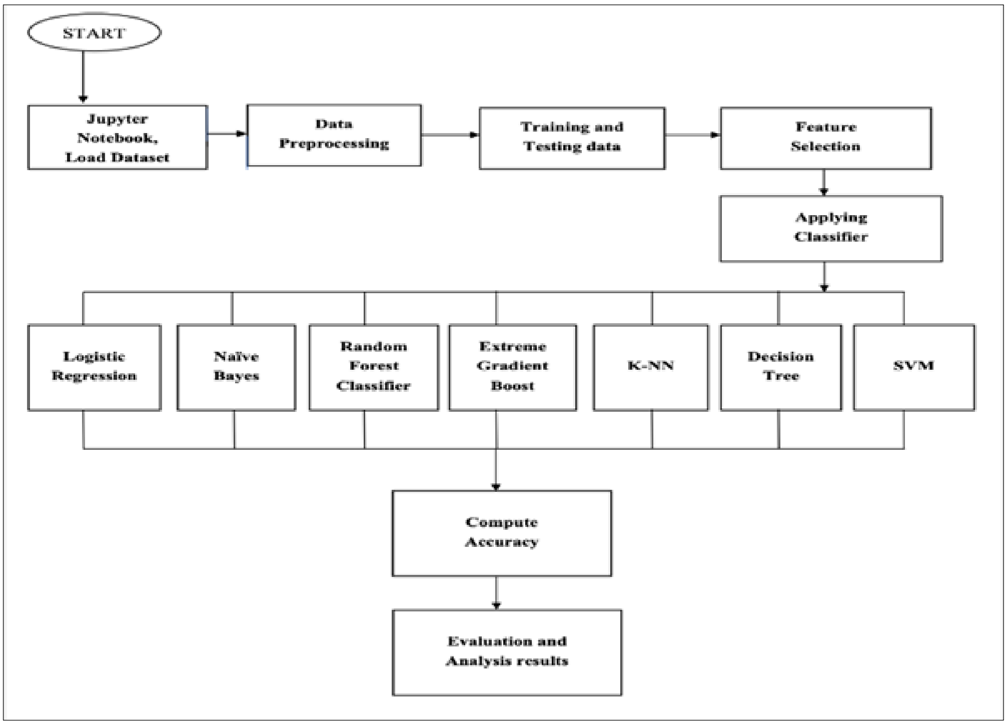
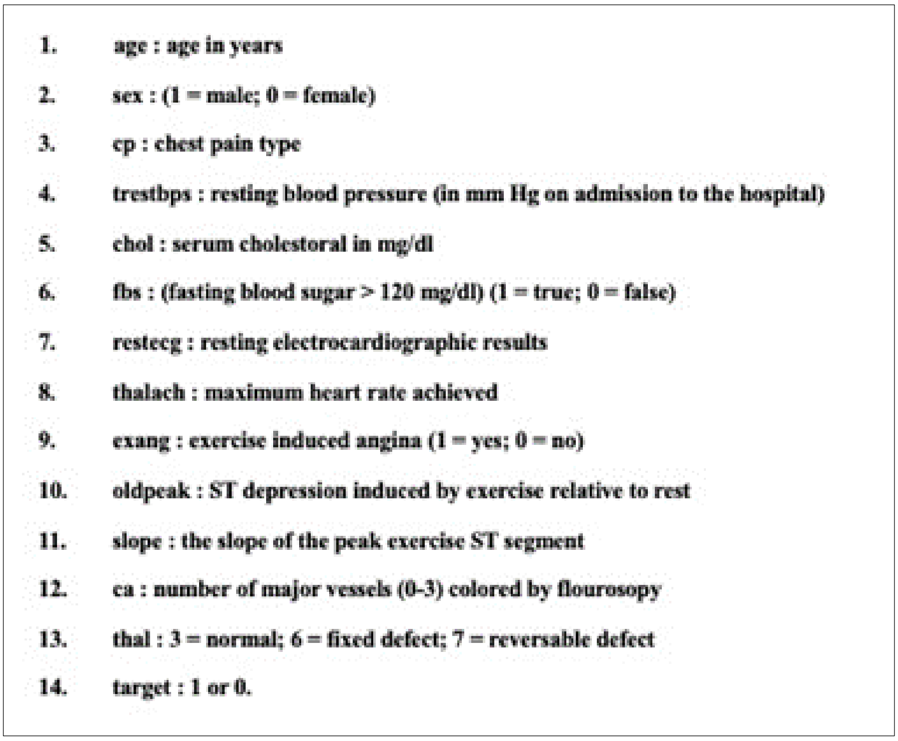


Figure 1 General Framework of Proposed Methodology

3.1. Dataset

The heart dataset from UCI repository [20] has been taken for predicting heart disease. The dataset contains 14 parameters as given below in Figure 2.



1. age : age in years
2. sex : (1 = male; 0 = female)
3. cp : chest pain type
4. trestbps : resting blood pressure (in mm Hg on admission to the hospital)
5. chol : serum cholestoral in mg/dl
6. fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. restecg : resting electrocardiographic results
8. thalach : maximum heart rate achieved
9. exang : exercise induced angina (1 = yes; 0 = no)
10. oldpeak : ST depression induced by exercise relative to rest
11. slope : the slope of the peak exercise ST segment
12. ca : number of major vessels (0-3) colored by flourosopy
13. thal : 3 = normal; 6 = fixed defect; 7 = reversable defect
14. target : 1 or 0.

Figure 2 List of Parameters in the Dataset

3.2. Feature Selection Methods

Three Feature Selection Methods have been applied on heart disease dataset in this study to identify the relevant features; namely Lasso, Ridge and Boruta.

3.2.1. Lasso

Lasso has been used to perform feature selection and improve model performance. It is particularly useful when dealing with datasets that have many features or when dealing with multicollinearity between features. The use of Lasso allows to select only the most important features for the model, improving model interpretability and reducing the risk of overfitting. The cross-validation is used to get the best alpha value for the model.

During cross-validation, the data is divided into subgroups, and the model is trained on various combinations of these subsets to discover the alpha value that produces the greatest overall performance on the data.

3.2.2. Ridge

This method works by assigning weights to each feature, which helps in determining how much impact it has on the output. This method is particularly useful in cases where there are a large number of features, as it helps in reducing the number of features to only those that are most relevant to the problem at hand.

3.2.3. Boruta

This method is based on comparing the significance of original attributes with that of randomly attainable attributes and gradually eliminating unimportant features to balance the test. Boruta is a wrapper built around a random forest classification algorithm, which captures all important or interesting features of a dataset concerning the output variable. In this method duplicate copies of all independent variables are created and the values are shuffled to remove their correlations with the target variable. The resulting shuffled copies are called shadow features or permuted copies. Then, the original variables are combined with the shuffled copies, and a random forest classifier is run on the combined dataset to perform a variable importance measure. The algorithm then computes a 'z-score', finds the maximum 'z-score' among shadow attributes, and tags variables as 'unimportant' or 'important' depending on whether they have importance significantly lower or higher than the maximum 'z-score'. These steps are repeated for a predefined number of iterations or until all attributes are either tagged 'unimportant' or 'important'.

4. Experimental Result and Analysis

The whole idea of this analysis was to use different type of machine learning algorithms for heart disease in healthcare. To take out the classification algorithm that is the best for predicting heart disease, this study was conducted by implementing ML models. In order to identify the algorithm giving best accuracy a comparison on the different machine learning algorithms is performed in subsequent sections.

4.1. Results and Analysis of Machine Learning Algorithms

The accuracy, Precision, Recall, F1 -Score and Support obtained by applying various machine learning algorithms like Logistic Regression, Naive Bayes, Random Forest Classifier, Extreme Gradient Boost (XGB) , K-Nearest Neighbor(KNN) , Decision Tree and Support Vector Machine(SVM) classification techniques is shown in table 1.

Table 1 Analysis of machine learning algorithm

Algorithms	**Accuracy**	**Precision**		**Recall**		**F1 -Score**		**Support**	
		0	**1**	**0**	**1**	**0**	**1**	**0**	**1**
Logistic Regression	0.88	0.88	088	0.86	0.89	0.88	0.88	29	32
Naïve Bayes	0.86	0.84	0.90	0.90	0.84	0.87	0.87	29	32
Random Forest Classifier	0.88	0.89	0.88	0.86	0.91	0.88	0.89	29	32
Extreme Gradient Boost	0.85	0.83	0.87	0.86	0.84	0.85	0.86	29	32
KNN	0.70	0.68	0.73	0.72	0.69	0.70	0.71	29	32
Decision Tree	0.86	0.84	0.90	0.90	0.84	0.87	0.87	29	32
SVM	0.70	0.76	0.68	0.55	0.84	0.64	0.75	29	32

According to the table 1 the best accuracy on the given dataset is of 88% by Random Forest and Logistic Regression and lowest accuracy of 70% by KNN and SVM. Precision, Recall, F1-Score are best with Random Forest. Hence Random Forest is applied to evaluate importance of feature set selected in subsequent section.

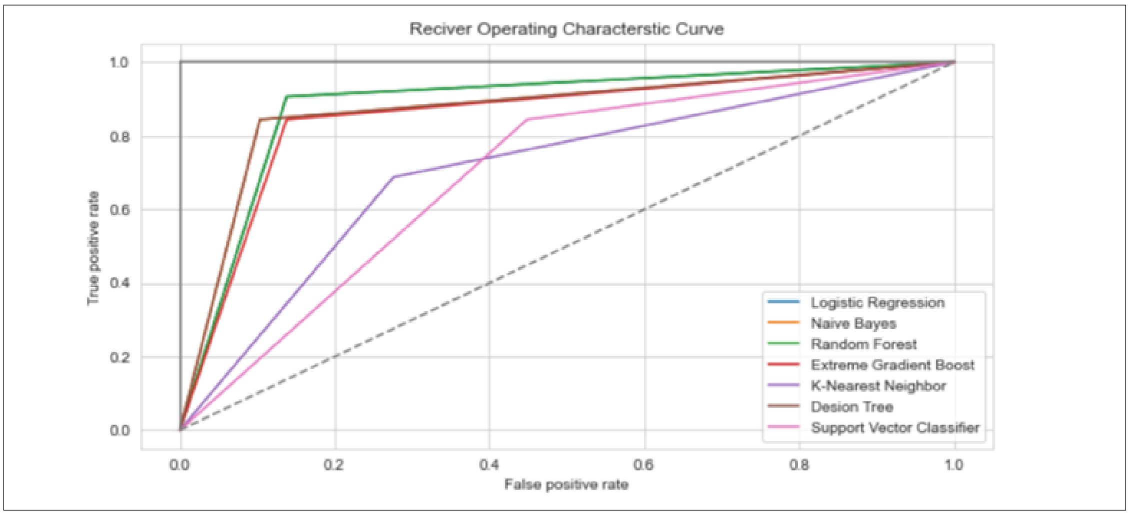


Figure 3 ROC Curve

As shown by Fig 3 the ROC AUC score of the Logistic Regression and Random Forest Classifier is 0.8841 which is the highest. The lowest AUC Score is of 0.6977 reported by Support Vector Machine(SVM). So according to the above analysis if the disease is diagnosed correctly/accurately and if it is timely detected then it is extremely beneficial to cure patients and treat the suffering from the disease.

4.2. Results and Analysis of Feature Selection Methods

4.2.1. Lasso Feature Selection method

The ten features that are selected by Lasso are as follows; ‘cp', 'trestbps', 'chol', 'restecg' , 'thalach', 'exang','oldpeak', 'slope', 'ca', 'thal'. The accuracy obtained through these features with Logistic regression is 86%.

4.2.2. Ridge Feature Selection method

Based on feature importance depicted in figure 4, the top six features are -Features ‘cp’, ‘thalach’, ‘exang’, ‘oldpeak’, ‘ca’, ‘thal’.

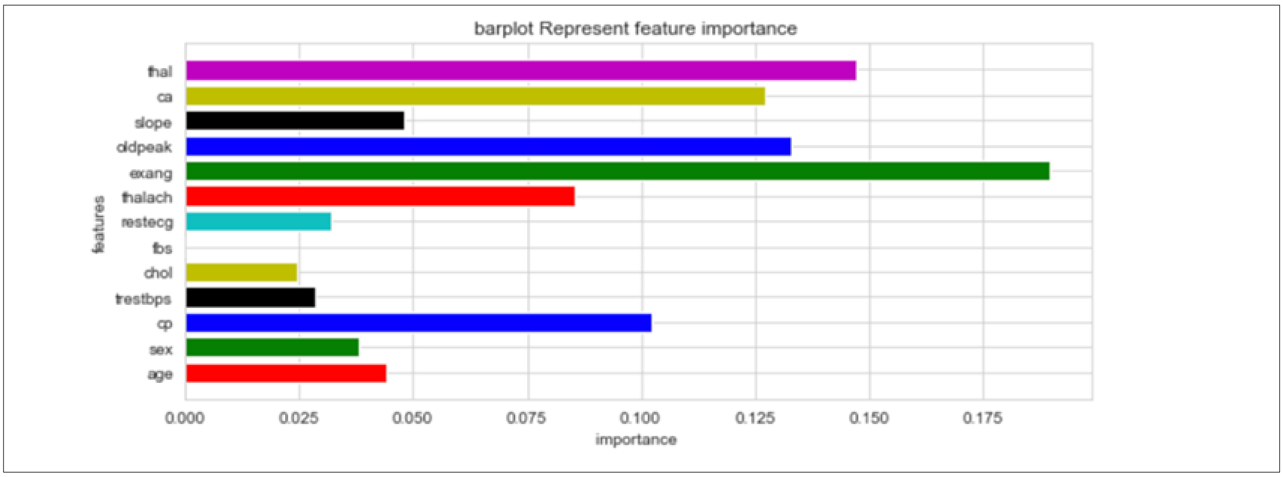


Figure 4 Feature Importance

As per observation from Figure 4, symptom of exercise induced angina (exang) is one of the worrying cause of the heart disease with more that 0.175 score. Angina in this is type of a pain in the chest which is caused by exercise, stress, or other things that exerts the heart to a greater extent. In exang it is recorded as 1 if there is pain and 0 if there is no pain. The accuracy obtained with these features and Logistic regression is 85%.

4.2.3. Boruta Feature Selection Method

The Boruta Feature Selection method selects following features ‘age’, ‘cp’ , ‘ thalach’ , ‘exang’, ‘oldpeak’ , ‘slope’ , ‘ca’, ‘thal’ as relevant features with an accuracy of 88% with Random Forest.

To conclude, this study indicates the Machine Learning models with feature selection techniques can be useful in understanding the important parameters for early prediction and timely diagnoses of the heart disease.

5. Conclusion and Future Scope

This study mainly focused on the use of data machine learning and feature selection algorithms on the proposed dataset which contained parameters related to the area of healthcare so that it can be easy to detect the heart disease. Heart disease is a serious disease which may cause death. The algorithms that were implemented are Logistic Regression , Naive Bayes, Random Forest Classifier, Extreme Gradient Boost(XGB) , K-Nearest Neighbour(KNN) , Decision Tree and Support Vector Machine(SVM). According to the analysis and implementation of these algorithms, it is that Logistic regression and Random Forest classifiers are the algorithms that has resulted with the highest accuracy of 88% and with the highest ROC AUC score of 0.8841. Based on this feature selection techniques (Lasso, Ridge, Boruta) which applies these two algorithms are experimented with to identify relevant and non-redundant feature set for heart disease prediction. Boruta feature selection method with Random Forest classifier has been found to outperform all other methods applied in this work.

Further research work could focus on increasing or improving accuracy of classification models by adding more features. Datasets related to EEGs can also be analysed in future.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Who link: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[2] K. Karthick, S. K. Aruna, Ravi Samikannu, Ramya Kuppusamy, Yuvaraja Teekaraman, Amruth Ramesh Thelkar, Implementation of a Heart Disease Risk Prediction Model Using Machine Learning, Computational and Mathematical Methods in Medicine, Article ID 6517716, 14 pages, 2022. https://doi.org/10.1155/2022/6517716

[3] Say R. E., Thomson R. The importance of patient preferences in treatment decisions-challenges for doctors. BMJ. , 327(7414), 542-555, 2003

[4] Cook CE, Décary S. Higher order thinking about differential diagnosis. Braz J Phys Ther., 24(1):1-7, 2020. doi: 10.1016/j.bjpt.2019.01.010.

[5] Avinash Golande, Pavan Kumar T, Heart Disease Prediction Using Effective Machine Learning Techniques, International Journal of Recent Technology and Engineering, Vol 8, 944-950, 2019.

[6] Fahd Saleh Alotaibi, Implementation of Machine Learning Model to Predict Heart Failure Disease, (IJACSA) International Journal of Advanced Computer Science and Applications, 10(6), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100637.

[7] J. Thomas and R. T. Princy, Human heart Disease Prediction System using Data Mining Techniques, *2016* International Conference on Circuit, Power and Computing Technologies (ICCPCT), Nagercoil, 2016, pp. 1-5, doi: 10.1109/ICCPCT.2016.7530265.

[8] Nagaraj M Lutimath,Chethan C, Basavaraj S Pol., Prediction Of Heart Disease using Machine Learning, International journal Of Recent Technology and Engineering,Vol. 8, Issue2S10), 474-477, 2019.

[9] Anna Karen Gárate-Escamila, Amir Hajjam El Hassani, Emmanuel Andrès, Classification models for heart disease prediction using feature selection and PCA, Informatics in Medicine Unlocked, Volume 19, 100330, 2020, https://doi.org/10.1016/j.imu.2020.100330

[10] Spencer, Robinson, et al. Exploring feature selection and classification methods for predicting heart disease. Digital health, 29:6:2055207620914777, 2020, doi: 10.1177/2055207620914777.

[11] H. Takci, Improvement of Heart Attack Prediction by the Feature Selection Methods, Turkish Journal of Electrical Engineering & Computer Sciences, Vol. 26, No. 1, 1-10, 2018,.

[12] S. Bashir, Z.S. Khan, F.H. Khan, A. Anjum, and K. Bashir, Improving Heart Disease Prediction Using Feature, 16th International Bhurban Conference on Applied Sciences & Technology (IBCAST), 619-623, 2019.

[13] Chandra Reddy, N. S., Shue Nee, S., Zhi Min, L., & Xin Ying, C., Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction, International Journal of Innovative Computing, 9(1), 2019, https://doi.org/10.11113/ijic.v9n1.210.

[14] A.M. Usman, U.K. Yusof, and S. Naim, Cuckoo Inspired Algorithms for Feature Selection in Heart, International Journal of Advances in Intelligent Informatics, Vol. 4, No. 2, 95-106, 2018.

[15] Muhammad Salman Pathan, Avishek Nag, Muhammad Mohisn Pathan, Soumyabrata Dev, Analyzing the impact of feature selection on the accuracy of heart disease prediction, Healthcare Analytics, Volume 2, 100060, 2022.

[16] Deepika Bansal, Kavita Khanna, Rita Chhikara, Rakesh Kumar Dua, Rajeev Malhotra, Analysis of Classification & Feature Selection Techniques for Detecting Dementia, SUSCOM, India, February 26-28, 2019

[17] Ghosh, Pronab, et al. Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access, 9, 2021, 19304-19326,

doi: 10.1109/ACCESS.2021.3053759

[18] Saurabh Paul, Petros Drineas, Feature Selection for Ridge Regression with Provable Guarantees, Neural Computation, Vol 28, Issue 4, 2016, https://doi.org/10.48550/arXiv.1506.05173.

[19] Neeyati Anand, Riya Sehgal, Sanchit Anand and Ajay Kaushik, Feature selection on educational data using Boruta algorithm, International Journal Computational Intelligence Studies, Vol. 10, No. 1, pp 27-35, 2021.

[20] Dataset UCI Repository - https://archive.ics.uci.edu/dataset/45/heart+disease