Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 1267-1275
Received 25 April 2021; Accepted 08 May 2021.

Identification of Heart Disease Using Fs Algorithm

Mrs.S.Ramya [1], Balachandar Raju [2], Prabha.R [3], Srimathi.C [4], Sabarishni.K [5]

[1] Assistant Professor (Sr.G), Department of CSE, Kongu Engineering College, Perundurai.
Email Id: sramya.cse@kongu.edu

[2] Engineer, Target Corporation India, Bangalore, Email Id: balachandar.raju@gmail.com

[3,4,5] UG Scholar, Department of Computer Science and Engineering,
Kongu Engineering College, Perundurai.
Email Id: prabhar2000@gmail.com,srimathichenniyappan@gmail.com,sabarishni99@gmail.com

ABSTRACT

Among many other diseases**,** heart disease is a leading ailment of today, and it affects many people all over the world. As a result, early detection and diagnosis of this disease is an essential factor in health care field. The health sector benefits greatly from an effective and reliable diagnosis of this disease. Centred on various machine learning techniques, we formulated an accurate and efficient recognition method for diagnosing heart disease in this paper. Linear Regression, Support vector machine, Linear SVC, MLP Classifier, Stochastic Gradient, Decision Tree Classifier, Random forest classifier, XGB Classifier, LGBM Classifier, Gradient Boosting Classifier, Ridge Classifier, Bagging Classifier, Extra Trees Classifier, AdaBoost Classifier, Logistic Regression, K-nearest neighbour (KNN), Naive Bayes, Neural network(NN) with Keras, Gaussian Process Classification and Voting classifier are some of the classification algorithms used in this machine learning system, while, on the other hand, for eliminating obsolete and redundant features, standard feature selection techniques like Recursive Feature Elimination (RFE),Feature Selection with the Pearson Correlation and Chi-squared method have been introduced. Feature selection algorithms are employed to enhance classification performance and minimize classification system execution time. Of all the classifiers used, Stochastic Gradient Descent classifier performed well and obtained an accuracy of 93.44%. Furthermore, the proposed scheme can be effectively applied in the health sector to correctly assess the HD.

Keywords: Heart disease, diagnosis, feature selection, recursive feature elimination

I. INTRODUCTION

Heart disease (HD) is a severe and complicated health condition to medicate. And this disease causes hardship for a large number of people all over the world [1]. Heart disease symptoms include abnormal heartbeats, nausea, chest discomfort, jaw or throat pain, shortness of breath, dizziness, light-headedness and fainting frequently [2]. Traditional heart disease diagnostic approaches are ineffective in identifying heart disease for a variety of reasons, including execution time and accuracy [3]. As a result, many researchers are working to develop an efficient framework for heart disease diagnosis. The diagnosis and treatment of this ailment would be exceedingly exhausting without the assistance of advanced technologies and medical experts [4]. A proper treatment and correct diagnosis will certainly

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 1267-1275
Received 25 April 2021; Accepted 08 May 2021.

help healthcare providers in identifying this disease [5].According to the Centres for Disease Control and Prevention, HD is one of the strongest predictors of death in the United States (CDC). Heart disease is also responsible for the demise of 1 in 4 adults in the US in particular. Traditionally, the diagnosis of HD is based on a synopsis of the patient's previous clinical encounters, a physical examination report, and an examination of the patient's symptoms by a physician. However, the findings of this diagnostic process have not been very reliable in determining the patient's HD. It is also difficult to evaluate and it is much expensive [6]. A selection of studies used the Cleveland HD data set to detect heart disease. Various researchers have suggested various diagnostic techniques in different studies, but these techniques can be further improved by implementing feature selection techniques which has been proposed in this paper to obtain successful outcomes.

II. LITERATURE REVIEW

Many researchers in the field have proposed various ML approaches for the detection of HD. To emphasize the significance of the proposed framework, several available ML procedures have been taken into account here. Detrano et al.[8] used ML classification techniques to construct a heart disease classification approach and accuracy of the system was 77%. For the analysis, the Cleveland dataset was used. In another study, Humar et al.[10] used a computational model with probabilistic reasoning incorporation to describe a heart disease categorization and achieved the best performance of 87.4 percent. Artificial Neural Network diagnosis system for heart disease diagnosis that incorporates multiple base algorithms into a single integrated predictive algorithm was proposed in previous work by Resul et al.[11]. Additionally, an extra predictive measuring system, with a precision of 89.01 percent, a recall of 80.09 percent, and a reliability of 95.91 percent was generated. Yun et al.[7] suggested a related set of strategies for varying sorts of feature selection, as with large-scale data, limited sample sized data with high dimensions, and also stable ones. They have as-well focused on several key topics in feature selection, namely multi-view, stable, distributed, multi-label, adversarial and online. The feature selection complexities for big data were explored by Jundong et al.[13]. For different learning activities, it is crucial to minimize data's dimensionality due to the Hughes phenomenon. Olaniyi et al.[12] used an artificial neural network technique on a three-phase approach to anticipate heart disease in patients with angina with an accuracy of 88.89 percent. For the early stage diagnosis of heart disease, all current approaches used a range of methods. Despite the fact that all of those techniques lacked predictability in their system accuracy and needed a considerable amount of computation time for disease detection, they were all successful. As a consequence, the major disadvantages of previously existing methods are that they generate a comparatively low accuracy rate and take a longer time for computation. Much of this is due to the inclusion of insignificant features in the dataset they used. To resolve these barriers, new and reliable methods for detecting the occurrence of heart disease are needed.

III. RESOURCES AND APPROACH

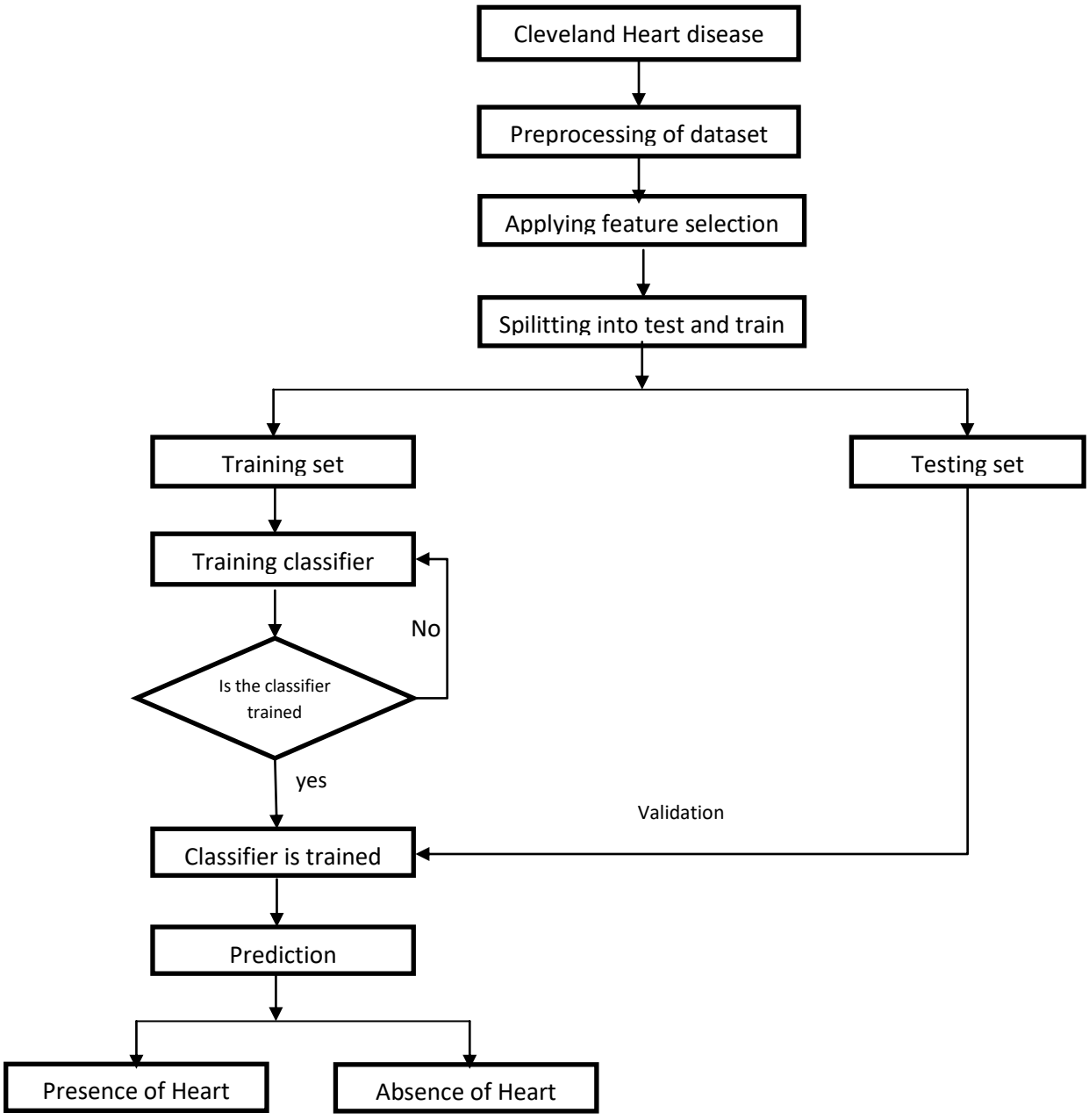
Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 1267-1275
Received 25 April 2021; Accepted 08 May 2021.

The following sub sections describe all of the study approaches and methodologies that were used. The entire procedure has been depicted in the form of flowchart as shown in Figure 1.

A. DATA SET

The dataset used in this analysis for research purposes is the Cleveland HD [14] dataset from Kaggle released on April 2, 2018. During the development of this dataset, there were 303 instances and 75 attributes. However, some researchers claim to have used a subset of 14 of them. The remaining 297 samples and 13 features dataset are left, along with a single output mark indicating whether HD exists or not. As a result, a 303*14 features matrix of extracted features is generated [9].

Figure 1: Recommended Heart disease detection strategy [9]



Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 1267-1275
Received 25 April 2021; Accepted 08 May 2021.

B. PREPROCESSING OF DATA SET

Before anything, it is essential to perform data preprocessing for data normalization. The preprocessing step involves many procedures such as extracting missing feature, eliminating recurred instances from the dataset. It has already been performed by Detrano et al.[8] during his work. Hence feature selection is proceeded with in the forthcoming procedures.

C. FEATURE CREATION

With the help of 13 attributes available in the dataset, feature creation is implemented by forming different combinations. After this procedure, a total of 61 data columns were formed.

D. FEATURE SELECTION

Following the completion of feature creation, the collection of appropriate features are needed for the subsequent processes. Since big data has several dimensions, choosing features from it is a daunting task that sometimes might result in huge problems. It is generally the most essential step in building a classification model. Feature selection has a greater effect on a variety of applications, including improving learning efficiency, producing clean and understandable data and making building model easier.

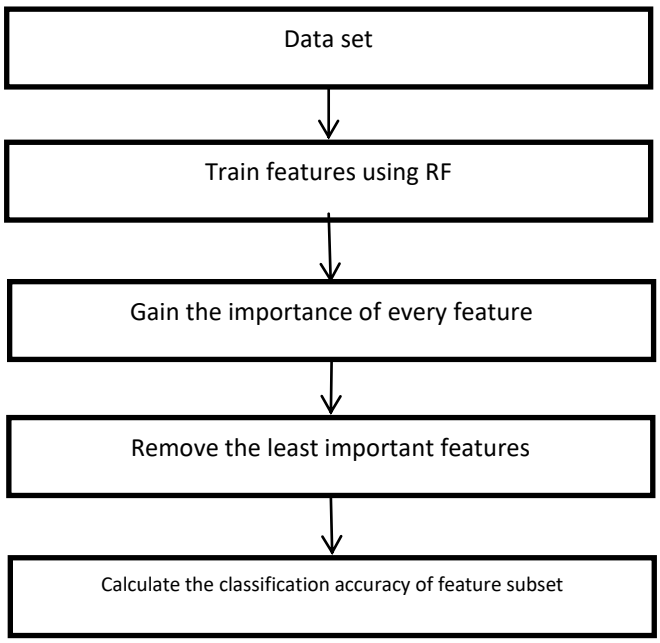
E. IMPLEMENTATION OF FS TECHNIQUES

For this purpose, we made use of three different methods, which include Recursive Feature Elimination, Pearson correlation coefficient and the results of these are also compared with the results of Chi-squared test to derive at an input dataset.

i. Recursive Feature Elimination(RFE)

It helps in removing the weakest feature that are available within the data until the required number of features is reached.

Figure 2: Proposed RFE algorithm



Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 1267-1275
Received 25 April 2021; Accepted 08 May 2021.

ii. Pearson Correlation Coefficient

Since it is based on the method of covariance, it is considered as the best method for calculating the relation among variables of interest.

iii. Chi-squared Test

A chi-square statistic is a test that assesses how well a model matches real data. It tests the independence of two variables.

Algorithm 2: Chi-squared Test approach

Step 1: Defining the hypothesis

Step 2: Construction of the contingency table

Step 3: Identification of expected values.

Step 4: Computation of chi-square scores.

Step 5: Acceptance/ rejection of null hypotheses.

The techniques for feature selection and abstraction aid in improving the model's reliability. But at the other hand, a suitable machine learning model is needed in order to obtain successful outcomes.

F. CLASSIFIERS

Some of the classifiers taken into account include Linear Regression, Support vector machine, Linear SVC, MLP Classifier, Stochastic Gradient, Decision Tree Classifier, Random forest classifier, XGB Classifier, LGBM Classifier, Gradient Boosting Classifier, Ridge Classifier, Bagging Classifier, Extra Trees Classifier, AdaBoost Classifier, Logistic Regression, K-nearest neighbour (KNN), Naive Bayes, Neural network(NN) with Keras, Gaussian Process Classification and Voting Classifier. The results after applying feature selection techniques to different classifiers has been continued in Section IV.

IV. EXPERIMENTAL RESULTS

A. RESULTS OF DATA PREPROCESSING

The generated dataset contains 303 examples, 13 input attributes and one classification tag. Visualization of data is the process of projecting data in a graph-based manner. Through recapping and exhibiting vast volumes of data in an easily readable format, it facilitates the comprehension of the data's significance. Figure 3 depicts the correlation between the dataset's attributes as depicted by a heatmap. A bi-dimensional data where the measurements are expressed

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 1267-1275
Received 25 April 2021; Accepted 08 May 2021.

in colors is a heatmap. A heatmap accurately provides a short visual overview of the information. Furthermore, complex datasets can easily be understood by the heat maps.

Figure 3: The correlation heat map for the Cleveland heart disease dataset

B. FEATURES SELECTED BY FEATURE SELECTION TECHNIQUES

The features that are selected after the implementing RFE, Pearson Correlation and Chi-squared method have been reported in Table 1 along with their ranking. In Pearson correlation, with a threshold value set to 0.9, 29 features were selected. Then, RFE was performed which eliminated the weakest features according to the problem statement and generated 35 features as fit. Finally, the chi-square values of all the features were sorted according to their descending order. The optimal features from the results of all the 3 FS techniques formed 25 features as a part of the input dataset.

Table 1: Features selected by FS techniques

#	Column(DType – float64)	Non-Null Count
0	thal_oldpeak2	241
1	Age	241
2	Thalach	241
3	Ca	241
4	sex_ca	241
5	thal_chol2	241
6	Trestbps	241
7	Chol	241
8	sex_oldpeak2	241
9	age2_ca	241
10	fbs_cp	241
11	restecg_cp	241
12	exang_ca	241
13	age2_cp	241
14	age2_oldpeak2	241
15	fbs_oldpeak2	241
16	fbs_ca	241
17	exang_oldpeak2	241
18	thal_cp	241
19	thal_trestbps2	241
20	thal_slope	241
21	thal_ca	241
22	Cp	241
23	Oldpeak	241
24	oldpeak2	241

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 1267-1275
Received 25 April 2021; Accepted 08 May 2021.

C. CLASSIFIERS RESULTS BASED ON PROPOSED FEATURE SELECTION METHODOLOGY

The proposed FS techniques performed well in terms of precision and lowering of the execution time of classification in comparison with previous methods in the literature of HD diagnosis. Stochastic Gradient Descent outperformed the other 20 classifiers, with an accuracy of about 93.44 percent, which is mentioned in the Figure 4.The performances of the various classifiers can be seen in Table 2.

Figure 4: Result of proposed analysis

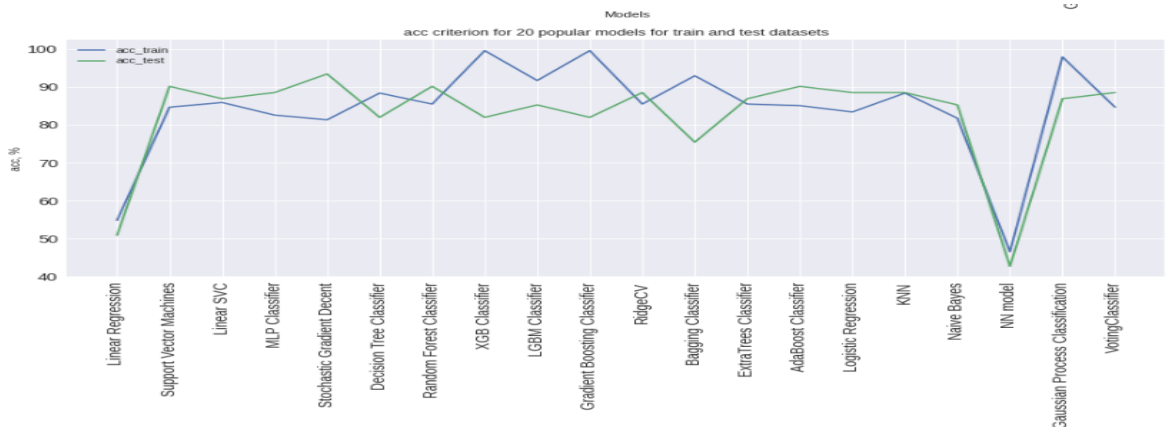


Table 2: Prediction accuracy of models after applying proposed FS techniques

Model	acc-train	acc-test
Stochastic Gradient Descent	81.33	93.44
Random Forest Classifier	85.48	90.16
AdaBoost Classifier	85.06	90.16
Support Vector Machines	84.65	90.16
KNN	88.38	88.52
Ridge CV Classifier	85.48	88.52
Voting Classifier	84.65	88.52
Logistic Regression	83.40	88.52
MLP Classifier	82.57	88.52
Gaussian Process Classifier	97.93	86.89
Linear SVC	85.89	86.89
Extra Trees Classifier	85.48	86.89
LGBM Classifier	91.70	85.25
Naïve Bayes Classifier	81.74	85.25
XGB Classifier	99.59	81.97
Gradient Boosting Classifier	99.59	81.97
Decision Tree Classifier	88.38	81.97
Bagging Classifier	92.95	75.41
Linear Regression	54.77	50.82
NN Model	46.47	42.62

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 1267-1275
Received 25 April 2021; Accepted 08 May 2021.

V. CONCLUSION

In this project, a system to diagnose heart disease has been implemented with the help of ML classifiers. Although earlier studies produced a pretty good accuracy, it could be improved as healthcare is a very critical field to play around with data. The dataset used was Cleveland Heart Disease dataset. We performed feature creation and selection to first prepare an input dataset with the most relevant and appropriate features. Feature creation produced 61 total features by making different possible combinations. Then we performed selection approaches to cut down to the most essential attributes. Initially, correlation coefficient was done to find the highly associated features. Chi-squared test for all the features were performed and in the result all the features were sorted according to their descending value of chi scores. Further, Recursive Feature Elimination was used to eliminate the weakest features. In this, it produced 35 attributes. Combining the 3 results, the derived input dataset consisted of 25 attributes in total. In that, the 25 columns were considered more appropriate and more accurate for the identification. These 25 features were tested with 20 different classifiers which are stated above, out of which, Stochastic Gradient Descent classifier gained an accuracy of about 93.44%.Hence the proposed system's accuracy is greater than the previous proposed methods. Further research can be done by giving a thorough analysis on the performances of different classifiers used in this study to arrive at sophisticated conclusions.

REFERENCES

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 1267-1275
Received 25 April 2021; Accepted 08 May 2021.

[1] Bui, Anh L., Tamara B. Horwich, and Gregg C.Fonarow."Epidemiology and risk profile of heart failure."Nature Reviews Cardiology 8.1 (2011): 30.

[2] Durairaj,M., and Nandhakumar Ramasamy."A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate."Int. J. Control Theory Appl 9.27 (2016): 255-260.

[3] Allen, Larry A., et al."Decision making in advanced heart failure: a scientific statement from the American Heart Association."*Circulation* 125.15 (2012): 1928-1952.

[4] Ghwanmeh, Sameh, Adel Mohammad, and Ali Al-Ibrahim."Innovative artificial neural networks-based decision support system for heart diseases diagnosis." (2013).

[5] Amato, Filippo, et al."Artificial neural networks in medical diagnosis." (2013): 47-58.

[6] Tsanas,Athanasios, et al."Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity."Journal of the royal society interface 8.59 (2011): 842-855.

[7] Li, Yun, Tao Li, and Huan Liu."Recent advances in feature selection and its applications."Knowledge and Information Systems 53.3 (2017): 551-577.

[8] Detrano, Robert, et al."International application of a new probability algorithm for the diagnosis of coronary artery disease."The American journal of cardiology 64.5 (1989): 304-310.

[9] Li, Jian Ping, et al."Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare."IEEE Access 8 (2020): 107562-107582.

[10] Kahramanli, Humar, and Novruz Allahverdi."Design of a hybrid system for the diabetes and heart diseases."Expert systems with applications 35.1-2 (2008): 82-89.

[11] Das, Resul, Ibrahim Turkoglu, and Abdulkadir Sengur."Effective diagnosis of heart disease through neural networks ensembles."Expert systems with applications 36.4 (2009): 7675-7680.

[12] Olaniyi, Ebenezer Obaloluwa, Oyebade Kayode Oyedotun, and Khashman Adnan. "Heart diseases diagnosis using neural networks arbitration."International Journal of Intelligent Systems and Applications 7.12 (2015): 72.

[13] Li, Jundong, and Huan Liu."Challenges of feature selection for big data analytics."IEEE Intelligent Systems 32.2 (2017): 9-15.

[14]https://www.kaggle.com/