International Journal of Engineering Research & Management Technology
ISSN: 2348-4039
December-2019 Volume 6, Issue-6
Email: editor@ijermt.org www.ijermt.org

STUDY ON THE UTILIZATION OF MACHINE LEARNING CLASSIFICATION FOR E-HEALTHCARE-BASED IDENTIFICATION OF HEART DISEASE

Pulloori Prathibha

Research Scholar, Dept of Computer Science, Himalayan University

Dr. Manish Saxena

Research Guide, Dept of Computer Science, Himalayan University

ABSTRACT

Heart disease is a complex and widespread health issue affecting many people globally. Timely and effective identification of heart disease is crucial in healthcare, especially in the field of cardiology. In this paper, we present a precise and efficient system for diagnosing heart disease using machine learning techniques. The system is based on classification algorithms, such as Support Vector Machine, Logistic Regression, Artificial Neural Network, K-Nearest Neighbor, Naïve Bayes, and Decision Tree. We have utilized standard feature selection algorithms, such as Relief, Minimal Redundancy Maximal Relevance, Least Absolute Shrinkage Selection Operator, and Local Learning, to remove irrelevant and redundant features to enhance classification accuracy and reduce execution time. Additionally, we propose a novel Fast Conditional Mutual Information Feature Selection Algorithm to address feature selection challenges. We have employed the leave-one-subject-out cross-validation method for hyper parameter tuning and optimal model selection. The performance of the classifiers has been evaluated using various metrics on the selected features. The experimental results demonstrate the feasibility of the proposed feature selection algorithm (FCMIM) in combination with the Support Vector Machine classifier for designing an intelligent system to identify heart disease. The proposed diagnosis system (FCMIM-SVM) achieves high accuracy compared to previous approaches, and it can be easily implemented in healthcare for heart disease identification.

Keywords: Artificial neural network, healthcare, FCMIM

1. INTRODUCTION

Heart disease is a major health concern worldwide and affects a large number of people. It is characterized by common symptoms such as shortness of breath, physical weakness, and swollen feet. Current diagnostic techniques for heart disease are not always effective in early identification due to various reasons, such as accuracy and execution time. This can make the diagnosis and treatment of heart disease extremely challenging, especially in situations where modern technology and medical experts are not available. The European Society of Cardiology estimates that approximately 26 million people are diagnosed with heart disease annually, with 3.6 million new cases diagnosed every year. In the United States, heart disease is one of the leading causes of death.

Traditionally, heart disease diagnosis is done by analyzing the medical history of the patient, conducting a physical examination, and analyzing the relevant symptoms. However, the results obtained from this diagnostic method are not always accurate in identifying patients with heart disease, and it can be expensive and computationally difficult to analyze. To address these issues, researchers are exploring the use of machine learning classifiers to develop a non-invasive diagnosis system. Expert decision systems based on machine learning classifiers and the application of artificial fuzzy logic have been shown to be effective in diagnosing heart disease, resulting in a decrease in the death rate.

Various machine learning predictive models have been proposed, with the Cleveland heart disease dataset being commonly used for the identification of heart disease. However, for machine learning models to perform optimally, they need proper training and testing data. The performance of machine learning models can be increased by using a balanced dataset for training and testing and by using relevant features from the data. Therefore, data balancing and feature selection are significantly important for improving model performance.

Various pre-processing techniques such as removal of missing feature value instances, Standard Scalar (SS), Min-Max Scalar, etc., can help standardize the data for machine learning models. Feature extraction and selection techniques can also improve model performance. Feature selection techniques such as Least-absolute-shrinkage-selection-operator (LASSO), Relief, are commonly used for important feature selection. Additionally, challenges of feature selection for structured, heterogeneous and streaming data as well as its scalability and stability issues must be addressed, especially for big data analytics.

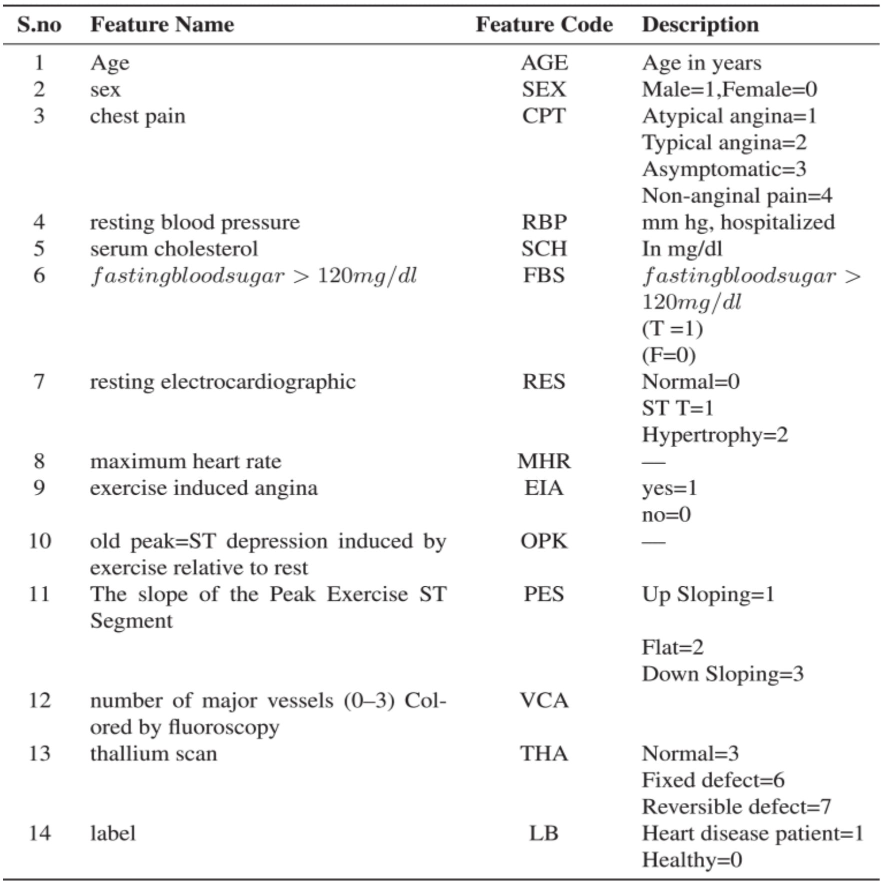
2. MATERIALS AND METHODS

All the techniques of this research are discussed in this part

2.1 Data Set

For this study, the Cleveland Heart Disease dataset was chosen for testing purposes. The original dataset consisted of 303 instances and 75 attributes; however, previous experiments have only used a subset of 14 attributes. To prepare the dataset for analysis, pre-processing was performed, and 6 samples were removed due to missing values. This left a dataset of 297 samples with 13 features and 1 output label, which describes the presence or absence of heart disease. Therefore, a feature matrix of size 297*13 was extracted. More information about the dataset matrix can be found in Table 1.

Table 1 Heart Disease Dataset



S.no	Feature Name	Feature Code	Description
1	Age	AGE	Age in years
2	sex	SEX	Male=1,Female=0
3	chest pain	CPT	Atypical angina=1 Typical angina=2 Asymptomatic=3 Non-anginal pain=4
4	resting blood pressure	RBP	mm hg, hospitalized
5	serum cholesterol	SCH	In mg/dl
6	$fastingbloodsugar > 120mg/dl$	FBS	$fastingbloodsugar > 120mg/dl$ (T =1) (F=0)

7	resting electrocardiographic	RES	Normal=0
			ST T=1
			Hypertrophy=2

8	maximum heart rate	MHR	—

9	exercise induced angina	EIA	yes=1
			no=0

10	old peak=ST depression induced by exercise relative to rest	OPK	—
11	The slope of the Peak Exercise ST Segment	PES	Up Sloping=1 Flat=2 Down Sloping=3
12	number of major vessels (0–3) Colored by fluoroscopy	VCA	
13	thallium scan	THA	Normal=3 Fixed defect=6 Reversible defect=7
14	label	LB	Heart disease patient=1 Healthy=0

2.2. Pre-processing of Dataset

To obtain a good representation, it is necessary to pre-process the dataset. Various techniques have been employed for this purpose, including the removal of missing attribute values, as well as the use of Standard Scalar (SS) and Min-Max Scalar techniques.

2.3 Standard State of the art Algorithm

In order to construct a classification model, feature selection is a crucial step that comes after data pre-processing. The purpose of this step is to reduce the number of input features in a classifier to generate more accurate predictions and to construct computationally efficient models. For our study, we employed four standard state-of-the-art FS algorithms in addition to one FS algorithm that we developed ourselves.

2.4 Relief

The Relief algorithm assigns weight values to each feature in the dataset and automatically updates those weights. The high-weight features are selected, while the low-weight ones are discarded. The Relief algorithm uses the same process as the K-NN algorithm to determine feature weights. The algorithm involves repeating Relief m times on random training samples (R_k) without substitution, where m is a parameter. At each iteration k, R_k is the "target" sample, and the weight W of that sample is updated accordingly. To provide a better understanding of the Relief algorithm, we present algorithm 1, which outlines the Pseudo-code for the Relief feature selection algorithm.

Algorithm 1- Pseudo Code

```
S Training data (feature vectors with class labels), Parameter m: number of random training samples out of total samples used to W.
W weights for each feature
n← total number of training samples
d← number of features (dimensions)
W[A]←0.0 ; ▹ Feature weights set for k←1 to m do
    Randomly choose a ‘Target’ sample Rk
    Find a nearest hit H and nearest miss M
    for A←1 to a do
        W[A]←W[A]−diff(A,Rk,H)/m+diff(A,Rk,M)/m
    end for
end for
Return W; ▹ weight vector of features that calculate the quality of features
```

2.5 Proposed Heart Diagnosis Methodology

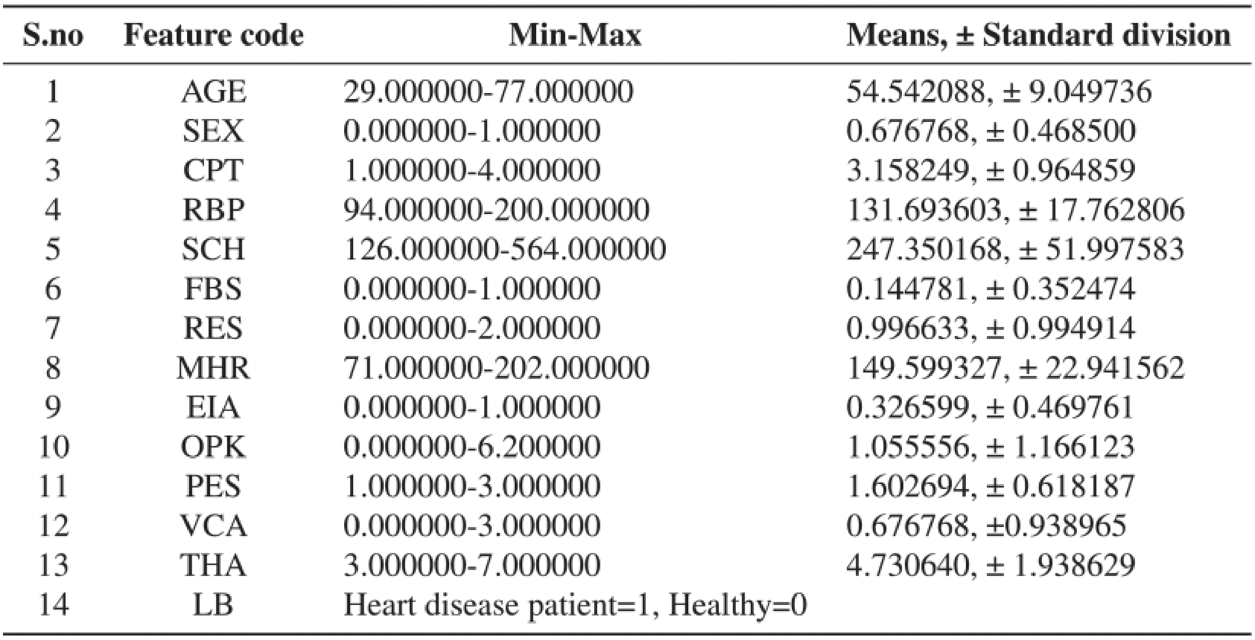
The aim of this study was to design a system for identifying heart disease using machine learning classifiers. The study evaluated the performances of various classifiers on selected features, using both standard state-of-the-art algorithms for feature selection such as Relief, MRMR, LASSO, and LLBFS, as well as a newly proposed FCMIM algorithm. The performance of the classifiers was evaluated using the selected feature sets, and the LOSO cross-validation technique was used to determine the best model. Performance was measured using several metrics, including accuracy, specificity, sensitivity, MCC, and processing time. The proposed methodology for the system was organized into several steps, including pre-processing of the dataset, feature selection algorithms, cross-validation methods, machine learning classifiers, and evaluation metrics. Pseudo-code for the proposed system algorithm is presented in Algorithm 2.

3. RESULTS AND DISCUSSIONS

Results of Data Pre-Processing Techniques

The dataset underwent various statistical operations, including removing missing attribute values, applying Standard Scalar (SS), Min-Max Scalar, calculating means and standard deviations. Table 2 contains the results of these operations. After processing, the dataset contained 297 instances, 13 input attributes, and one output label. Data visualization is a graphical representation of data that simplifies and presents large amounts of information in an understandable format, making it easier to comprehend and communicate. Figure 2 displays the dataset's histogram, which shows the frequency of occurrence of specific phenomena within a particular range of values arranged in consecutive intervals. Figure 3 depicts the relationship among the dataset's features using a heat map, which is a two-dimensional representation of data in which colors indicate values. A heat map provides a quick visual summary of information, and more complex heat maps can help the viewer understand complex datasets. Additionally, heat maps are useful when determining which intersections of categorical values have higher data concentrations compared to others.

Table 2 Results of Statistical Operations on the dataset



S.no	Feature code	Min-Max	Means, ± Standard division
1	AGE	29.000000-77.000000	54.542088, ± 9.049736
2	SEX	0.000000-1.000000	0.676768, ± 0.468500
3	CPT	1.000000-4.000000	3.158249, ± 0.964859
4	RBP	94.000000-200.000000	131.693603, ± 17.762806
5	SCH	126.000000-564.000000	247.350168, ± 51.997583
6	FBS	0.000000-1.000000	0.144781, ± 0.352474
7	RES	0.000000-2.000000	0.996633, ± 0.994914
8	MHR	71.000000-202.000000	149.599327, ± 22.941562
9	EIA	0.000000-1.000000	0.326599, ± 0.469761
10	OPK	0.000000-6.200000	1.055556, ± 1.166123
11	PES	1.000000-3.000000	1.602694, ± 0.618187
12	VCA	0.000000-3.000000	0.676768, ±0.938965
13	THA	3.000000-7.000000	4.730640, ± 1.938629
14	LB	Heart disease patient=1, Healthy=0	

Figures 1 Proposed heart disease identification system.

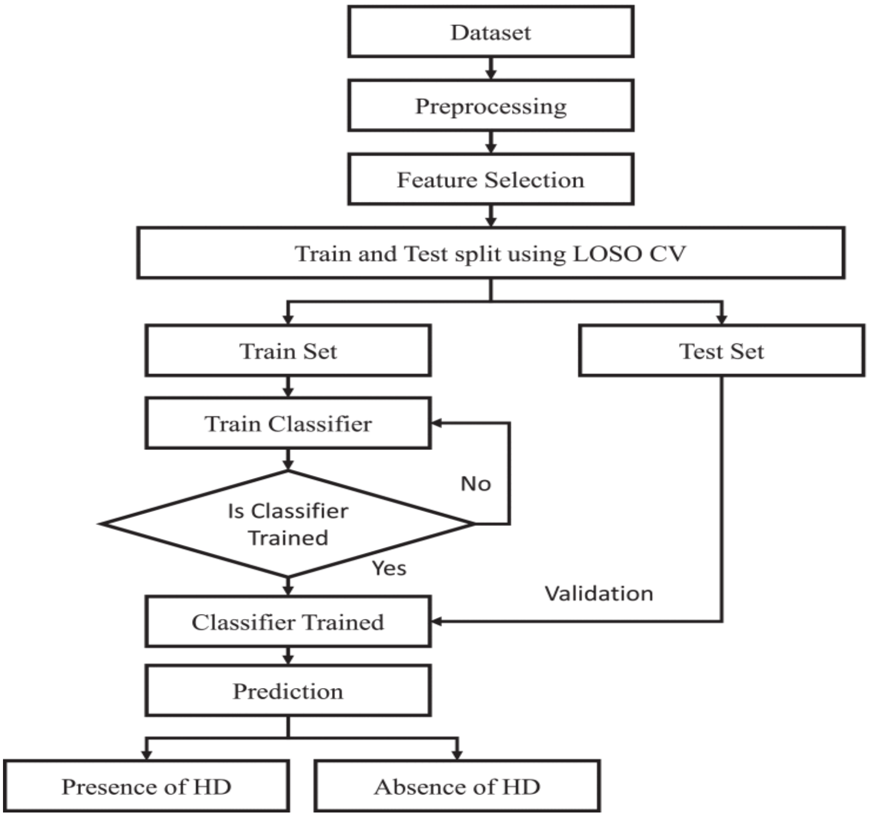


Figure 2 Histograms of heart disease dataset

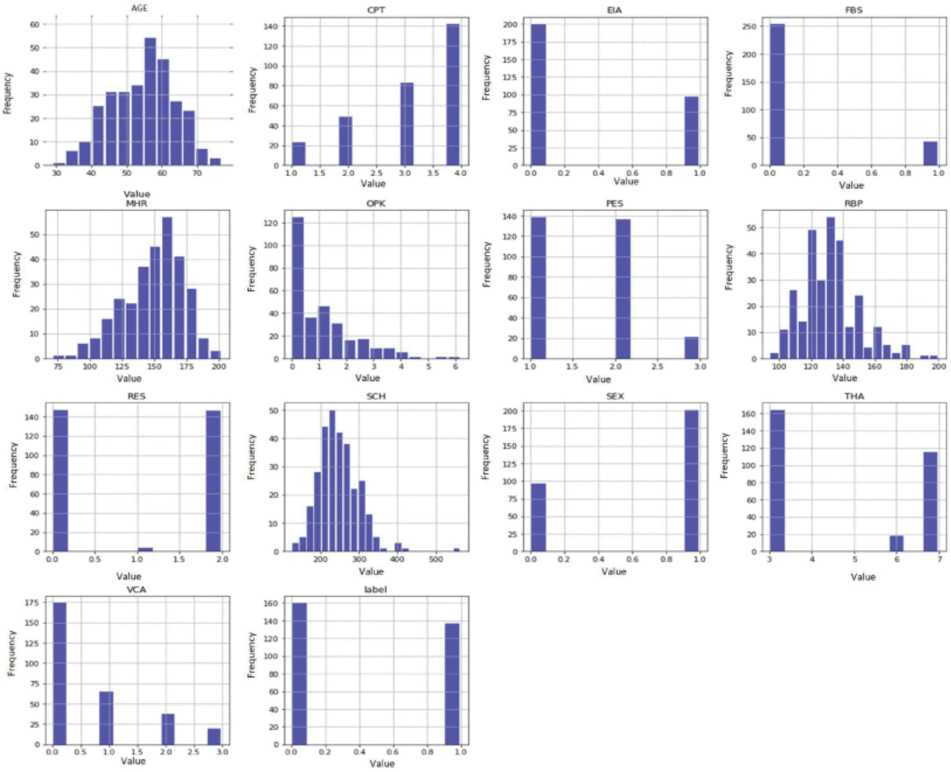
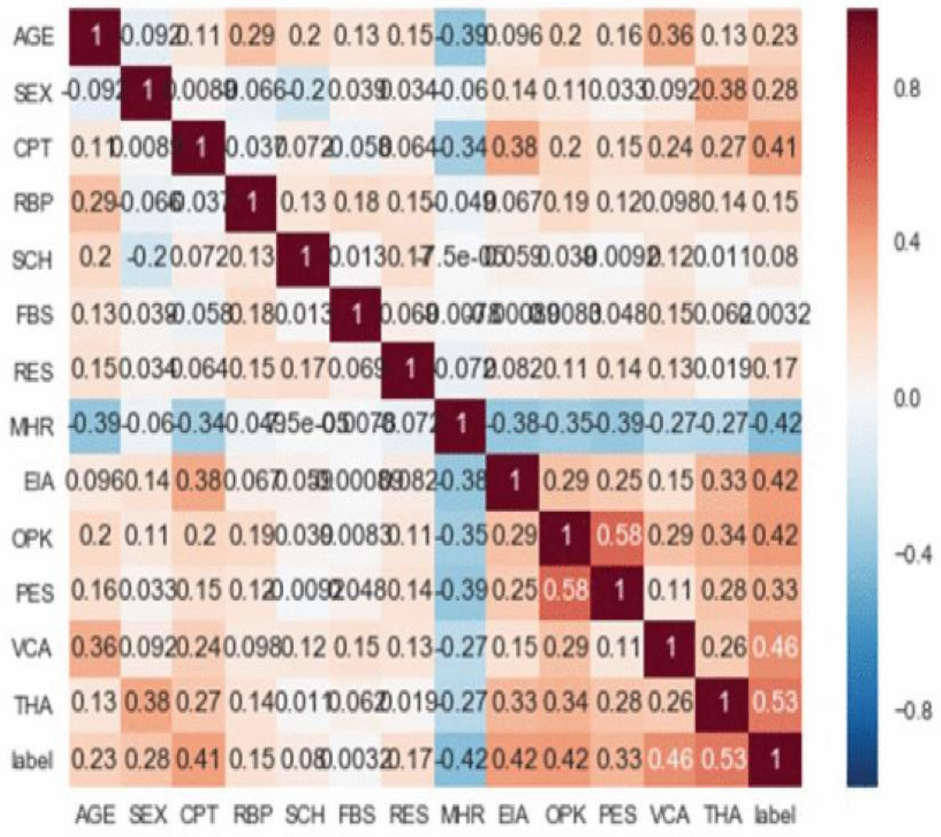


Fig 3 The heat map for correlation features of heart disease dataset.

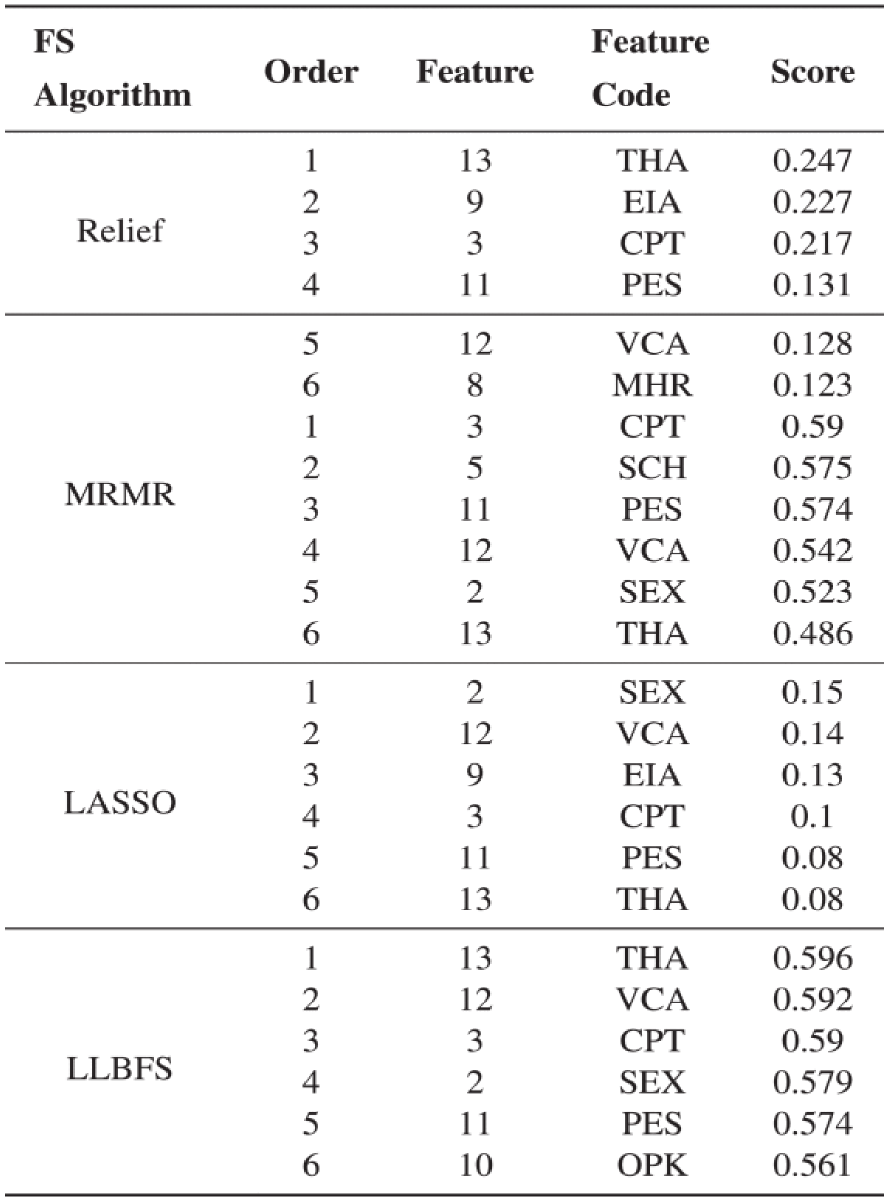


Features Selected by Standard State of the Arts Algorithms

Table 3 presents the data preprocessing and important features that were selected using four different feature selection (FS) algorithms, namely Relief, MRMR, LASSO, and LLBFS. The table includes feature scores and rankings, and according to the Relief algorithm, THA, EIA, and CPT are the most important features for identifying heart disease. These features were also identified as important by the other FS algorithms, including THA, CPT, SEX, VCA, and EIA. On the other hand, FBS received a low score in feature selection. Figure 4 provides a graphical representation of the important feature scores and rankings generated by the four FS algorithms.

The LASSO FS algorithm, which performs binary classification, selected five out of the 13 features that were labeled as true for the output target class. These selected features are also reported in Table 3. Additionally, the LASSO cross-validation mean square error results are presented in Figure 5, where the weight parameter lambda (ranging between 0 and 1) is plotted on the x-axis and the validation mean square error (MSE) is plotted on the y-axis. A total of 100 different models of feature subsets were generated by LASSO using different lambda values, and the highest point on the graph (at index 60) represents the minimum MSE of the generated model. The vertical line on the left side of the graph represents the highest value of lambda.

Table 3 Selected Features by Relief, MRMR, LASSO, and LLBFS



FS Algorithm	Order	Feature	Feature Code	Score
Relief	1	13	THA	0.247
	2	9	EIA	0.227
	3	3	CPT	0.217
	4	11	PES	0.131
MRMR	5	12	VCA	0.128
	6	8	MHR	0.123
	1	3	CPT	0.59
	2	5	SCH	0.575
	3	11	PES	0.574
	4	12	VCA	0.542
	5	2	SEX	0.523
	6	13	THA	0.486
LASSO	1	2	SEX	0.15
	2	12	VCA	0.14
	3	9	EIA	0.13
	4	3	CPT	0.1
	5	11	PES	0.08
	6	13	THA	0.08
LLBFS	1	13	THA	0.596
	2	12	VCA	0.592
	3	3	CPT	0.59
	4	2	SEX	0.579
	5	11	PES	0.574
	6	10	OPK	0.561

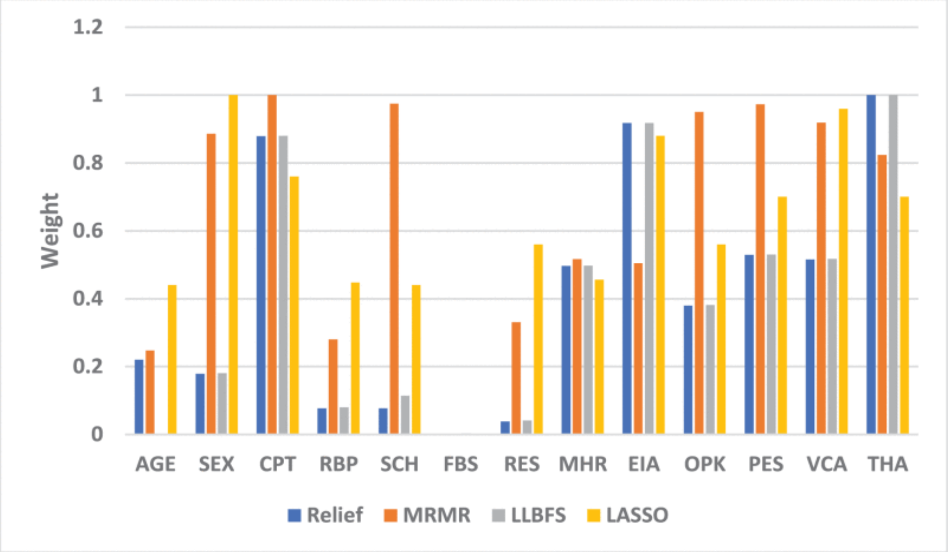


Fig 4 The score of features and Rankings selected by FS algorithms

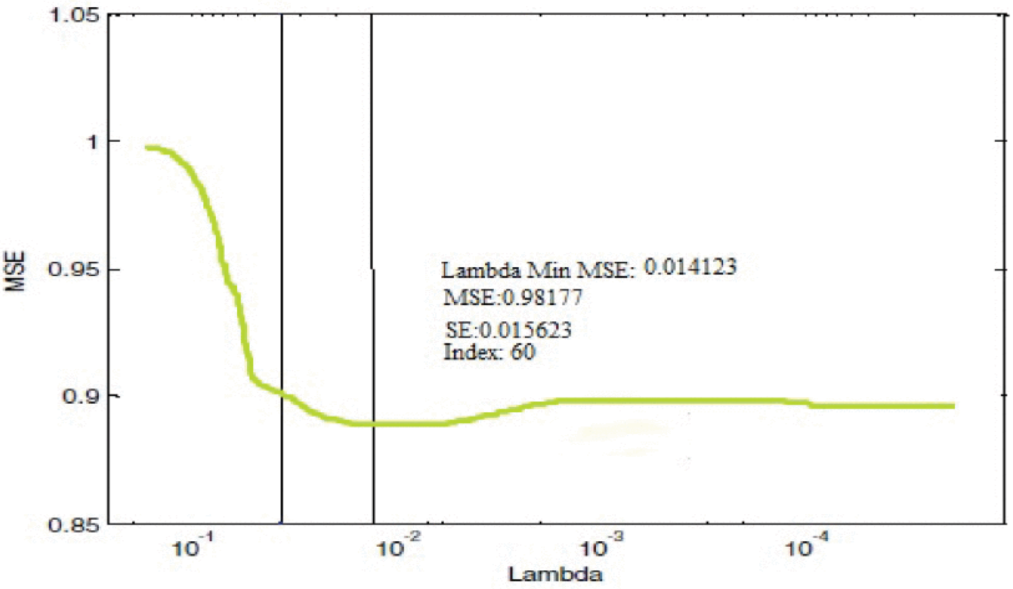


Fig 5 Cross Validation MSE

4. CONCLUSION

The use of machine learning classification algorithms for the identification of heart disease has proven to be a promising approach in healthcare. This paper presents a comprehensive study of feature selection algorithms, including a novel Fast Conditional Mutual Information Feature Selection Algorithm, to enhance classification accuracy and reduce execution time. The proposed diagnosis system achieves high accuracy compared to previous approaches and can be easily implemented in healthcare for heart disease identification. The Cleveland Heart Disease dataset was used for testing purposes, and various pre-processing techniques were employed to prepare the data for analysis. The study emphasizes the significance of proper feature selection and dataset balancing for machine learning models to perform optimally. The results of this study can be applied in the development of non-invasive and efficient diagnosis systems for heart disease.

REFERENCES

- L. Bui, T. B. Horwich and G. C. Fonarow, "Epidemiology and risk profile of heart failure", Nature Rev. Cardiol., vol. 8, pp. 30, 2020.
- M. Durairaj and N. Ramasamy, "A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate", Int. J. Control Theory Appl., vol. 9, no. 27, pp. 255-260, 2016.
- L. A. Allen, L. W. Stevenson, K. L. Grady, N. E. Goldstein, D. D. Matlock, R. M. Arnold, et al., "Decision making in advanced heart failure: A scientific statement from the American heart association", Circulation, vol. 125, no. 15, pp. 1928-1952, 2020.
- S. Ghwanmeh, A. Mohammad and A. Al-Ibrahim, "Innovative artificial neural networks-based decision support system for heart diseases diagnosis", J. Intell. Learn. Syst. Appl., vol. 5, no. 3, 2013.
- Q. K. Al-Shayea, "Artificial neural networks in medical diagnosis", Int. J. Comput. Sci. Issues, vol. 8, no. 2, pp. 150-154, 2011.
- J. Lopez-Sendon, "The heart failure epidemic", Medicographia, vol. 33, no. 4, pp. 363-369, 2011.
- P. A. Heidenreich, J. G. Trogdon, O. A. Khavjou, J. Butler, K. Dracup, M. D. Ezekowitz, et al., "Forecasting the future of cardiovascular disease in the united states: A policy statement from the American heart association", Circulation, vol. 123, no. 8, pp. 933-944, 2011.
- Tsanas, M. A. Little, P. E. McSharry and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson’s disease symptom severity", J. Roy. Soc. Interface, vol. 8, no. 59, pp. 842-855, 2011.
- S. I. Ansarullah and P. Kumar, "A systematic literature review on cardiovascular disorder identification using knowledge mining and machine learning method", Int. J. Recent Technol. Eng., vol. 7, no. 6S, pp. 1009-1015, 2019.
- S. Nazir, S. Shahzad, S. Mahfooz and M. Nazir, "Fuzzy logic based decision support system for component security evaluation", Int. Arab J. Inf. Technol., vol. 15, no. 2, pp. 224-231, 2018.
- R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease", Amer. J. Cardiol., vol. 64, no. 5, pp. 304-310, Aug. 1989.
- J. H. Gennari, P. Langley and D. Fisher, "Models of incremental concept formation", Artif. Intell., vol. 40, no. 1, pp. 11-61, Sep. 1989.

- Y. Li, T. Li and H. Liu, "Recent advances in feature selection and its applications", Knowl. Inf. Syst., vol. 53, no. 3, pp. 551-577, Dec. 2017.
- J. Li and H. Liu, "Challenges of feature selection for big data analytics", IEEE Intell. Syst., vol. 32, no. 2, pp. 9-15, Mar. 2017.
- L. Zhu, J. Shen, L. Xie and Z. Cheng, "Unsupervised topic hypergraph hashing for efficient mobile image retrieval", IEEE Trans. Cybern., vol. 47, no. 11, pp. 3941-3954, Nov. 2017.
- S. Raschka, "Model evaluation model selection and algorithm selection in machine learning", arXiv:1811.12808, 2018, [online] Available: http://arxiv.org/abs/1811.12808.
- S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques", Proc. IEEE/ACS Int. Conf. Comput. Syst.