Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare

JIAN PING LI [1], **AMIN UL HAQ** [1], **SALAH UD DIN** [2], **JALALUDDIN KHAN** [1], **ASIF KHAN** [1], **AND ABDUS SABOOR** [1]

[1] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2] Data Mining Laboratory, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

Corresponding authors: Jian Ping Li (jpli2222@uestc.edu.cn) and Amin Ul Haq (khan.amin50@yahoo.com)

ABSTRACT Heart disease is one of the complex diseases and globally many people suffered from this disease. On time and efficient identification of heart disease plays a key role in healthcare, particularly in the field of cardiology. In this article, we proposed an efficient and accurate system to diagnosis heart disease and the system is based on machine learning techniques. The system is developed based on classification algorithms includes Support vector machine, Logistic regression, Artificial neural network, K-nearest neighbor, Naïve bays, and Decision tree while standard features selection algorithms have been used such as Relief, Minimal redundancy maximal relevance, Least absolute shrinkage selection operator and Local learning for removing irrelevant and redundant features. We also proposed novel fast conditional mutual information feature selection algorithm to solve feature selection problem. The features selection algorithms are used for features selection to increase the classification accuracy and reduce the execution time of classification system. Furthermore, the leave one subject out cross-validation method has been used for learning the best practices of model assessment and for hyperparameter tuning. The performance measuring metrics are used for assessment of the performances of the classifiers. The performances of the classifiers have been checked on the selected features as selected by features selection algorithms. The experimental results show that the proposed feature selection algorithm (FCMIM) is feasible with classifier support vector machine for designing a high-level intelligent system to identify heart disease. The suggested diagnosis system (FCMIM-SVM) achieved good accuracy as compared to previously proposed methods. Additionally, the proposed system can easily be implemented in healthcare for the identification of heart disease.

INDEX TERMS Heart disease classification, features selection, disease diagnosis, intelligent system, medical data analytics.

I. INTRODUCTION

Heart disease (HD) is the critical health issue and numerous people have been suffered by this disease around the world [1]. The HD occurs with common symptoms of breath shortness, physical body weakness and, feet are swollen [2]. Researchers try to come across an efficient technique for the detection of heart disease, as the current diagnosis techniques of heart disease are not much effective in early time identification due to several reasons, such as accuracy and execution time [3]. The diagnosis and treatment of heart disease is extremely difficult when modern technology and medical experts are not available [4]. The effective diagnosis

and proper treatment can save the lives of many people [5]. According to the European Society of Cardiology, 26 million approximately people of HD were diagnosed and diagnosed 3.6 million annually [6]. Most of the people in the United States are suffering from heart disease [7]. Diagnosis of HD is traditionally done by the analysis of the medical history of the patient, physical examination report and analysis of concerned symptoms by a physician. But the results obtained from this diagnosis method are not accurate in identifying the patient of HD. Moreover, it is expensive and computationally difficult to analyze [8]. Thus, to develop a non-invasive diagnosis system based on classifiers of machine learning (ML) to resolve these issues. Expert decision system based on machine learning classifiers and the application of artificial fuzzy logic is effectively diagnosis the HD as a

The associate editor coordinating the review of this manuscript and approving it for publication was Navanietha Krishnaraj Krishnaraj Rathinam.

result, the ratio of death decreases [9] and [10]. The Cleveland heart disease data set was used by various researchers [11] and [12] for the identification problem of HD. The machine learning predictive models need proper data for training and testing. The performance of machine learning model can be increased if balanced dataset is use for training and testing of the model. Furthermore, the model predictive capabilities can improved by using proper and related features from the data. Therefore, data balancing and feature selection is significantly important for model performance improvement. In literature various diagnosis techniques have been proposed by various researchers, however these techniques are not effectively diagnosis HD. In order to improve the predictive capability of machine learning model data preprocessing is important for data standardization. Various Preprocessing techniques such removal of missing feature value instances from the dataset, Standard Scalar (SS), Min-Max Scalar etc. The feature extraction and selection techniques are also improve model performance. Various feature selection techniques are mostly used for important feature selection such as, Least-absolute-shrinkage-selection-operator (LASSO), Relief, Minimal-Redundancy-Maximal-Relevance (MRMR), Local-learning-based-features-selection (LLBFS), Principle component Analysis (PCA), Greedy Algorithm (GA), and optimization methods, such as Anty Conley Optimization (ACO), fruit fly optimization (FFO), Bacterial Foraging Optimization (BFO) etc. Similarly Yun *et al.* [13] presented different techniques for different type of feature selection, such as feature selection for high-dimensional small sample size data, large-scale data, and secure feature selection. They also discussed some important topics for feature selection have emerged, such as stable feature selection, multi-view feature selection, distributed feature selection, multi-label feature selection, online feature selection, and adversarial feature selection. Jundong *et al.* [14] discussed the challenges of feature selection (FS) for big data. It is necessary to decrease the dimensionality of data for various learning tasks due to the curse of dimensionality. Feature selection has great influence in numerous applications such as building simpler, increasing learning performance, creating clean and understandable data. The feature selection from big data is challenging job and create big problems because big data has many dimensions. Further, challenges of feature selection for structured, heterogeneous and streaming data as well as its scalability and stability issues. For big data analytics challenges of feature selection is very important to resolved. In [15] designed unsupervised hashing scheme, called topic hyper graph hashing, to report the limitations. Topic hypergraph hashing effectively mitigates the semantic shortage of hashing codes by exploiting auxiliary texts around images. The proposed Topic hyper graph hashing can achieve superior performance equaled with numerous state-of-the-art approaches, and it is more appropriate for mobile image retrieval. The feature selection algorithms are classified into three type such as filter based, wrapper based and embedded based. All these feature selection mechanisms have some

advantages and limitations in certain cases. The filter based method measures the relevance of a feature by correlation with the dependent variable while the wrapper feature selection algorithm measure the usefulness of a subset of features by actually training the classifier on it. The filter method is less computationally complex than wrapper method. The feature set selected by the filter is general and can be applied to any model and it is independent of a specific model. In feature selection global relevance is of greater importance.

On another hand suitable machine learning model is necessary for good results. Obviously, a good machine learning model is a model that not only performs well on data seen during training (else a machine learning model could simply learn the training data), but also on unseen data. To evaluate all classifiers on data and find that they get, on average, 50% of the cases right [16]. Furthermore, appropriate cross validation techniques and performance evaluation metrics are critical necessary for a model when model is train and test on dataset.

We proposed a machine learning based diagnosis method for the identification of HD in this research work. Machine learning predictive models include ANN, LR, K-NN, SVM, DT, and NB are used for the identification of HD. The standard state of the art features selection algorithms, such as Relief, mRMR, LASSO and Local-learning-based-features-selection (LLBFS) have been used to select the features. We also proposed fast conditional mutual information (FCMIM) features selection algorithm for features selection. Leave-one-subject-out cross-validation (LOSO) technique has been applied to select the best hyper-parameters for best model selection. Apart from this, different performance assessment metrics have been used for classifiers performances evaluation. The proposed method has been tested on Cleveland HD dataset. Furthermore, the performance of the proposed technique have been compared with state of the art existing methods in the literature, such as NB [17], Three phase ANN (Artificial neural Network) diagnosis system [18], Neural network ensembles (NNE) [19], ANN-Fuzzy-AHP diagnosis system (AFP) [20], Adaptive-weighted-Fuzzy-system-ensemble (AWFSE) [21]. The research study has the following contributions.

- Firstly, the authors try to address the problem of features selection by employing pre-processing techniques and standard state of the art four features selection algorithms such as Relief, mRMR, LASSO, and LLBFS for appropriate subset of features and then applied these features for effective training and testing of the classifiers that identify which feature selection algorithm and classifier gives good results in term of accuracy and computation time.

- Secondly, the authors proposed fast conditional mutual information (FCMIM) FS algorithm for feature selection and then these features are input to classifiers for improving prediction accuracy and reducing computation time. The classifiers performances have been compared on features selected by the standard state

of the art FS algorithms with the selected features of the proposed FS algorithm.

- Thirdly, identify weak features from the dataset which affect the performance of the classifiers.
- Finally, suggests that heart disease identification system (FCMIM-SVM) effectively identify the HD.

The paper remaining sections are structured as follows. The literature related to the problem has been discussed in section 2. In section 3 the dataset and the theoretical and mathematical knowledge of feature selection and classification algorithms are discussed in details. Additionally, discuss the technique of cross-validation and performance measuring metrics. In section 4 results of all experiments are analyzed and discussed in details. The last section 5 the conclusion and future direction of the research work have been explored in details.

II. LITERATURE REVIEW

In literature various machine learning based diagnosis techniques have been proposed by researchers to diagnosis HD. This research study present some existing machine learning based diagnosis techniques in order to explain the important of the proposed work. Detrano *et al.* [11] developed HD classification system by using machine learning classification techniques and the performance of the system was 77% in terms of accuracy. Cleveland dataset was utilized with the method of global evolutionary and with features selection method. In another study Gudadhe *et al.* [22] developed a diagnosis system using multi-layer Perceptron and support vector machine (SVM) algorithms for HD classification and achieved accuracy 80.41%. Humar *et al.* [23] designed HD classification system by utilizing a neural network with the integration of Fuzzy logic. The classification system achieved 87.4% accuracy. Resul *et al.* [19] developed an ANN ensemble based diagnosis system for HD along with statistical measuring system enterprise miner (5.2) and obtained the accuracy of 89.01%, sensitivity 80.09%, and specificity 95.91%. Akil *et al.* [24] designed a ML based HD diagnosis system. ANN-DBP algorithm along with FS algorithm and performance was good. Palaniappan *et al.* [17] proposed an expert medical diagnosis system for HD identification. In development of the system the predictive model of machine learning, such as navies bays (NB), Decision Tree (DT), and Artificial Neural Network were used. The 86.12% accuracy was achieved by NB, ANN accuracy 88.12% and DT classifier achieved 80.4% accuracy. Olaniyi *et al.* [18] developed a three-phase technique based on the artificial neural network technique for HD prediction in angina and achieved 88.89% accuracy. Samuel *et al.* [20] developed an integrated medical decision support system based on artificial neural network and Fuzzy AHP for diagnosis of HD. The performance of the proposed method in terms of accuracy was 91.10%. Liu *et al.* [25] proposed a HD classification system using relief and rough set techniques. The proposed method achieved 92.32% classification accuracy. In [26] proposed a HD identification method using feature selection

and classification algorithms. Sequential Backward Selection Algorithm (SBS FS) for Features Selection. The classifier K-Nearest Neighbor (K-NN) performance has been checked on full and on selected features set. The proposed method obtained high accuracy. In another study MOHAN *et al.* [27] designed a HD prediction method by using hybrid machine learning techniques. He also proposed a new method for significant feature selection from the data for effective training and testing of machine learning classifier. They have been recorded 88.07% classification accuracy. Geweid *et al.* [28] designed HD identification techniques by using improved SVM based duality optimization technique. In the above literature the proposed HD diagnosis methods limitation and advantages have been summarized in Table 1 for better understanding the important of our proposed approach. All these existing techniques used numerous methods to identify the HD at early stages. However, all these techniques have lack of prediction accuracy and high computation time for prediction of HD. According to Table 1 the prediction accuracy of HD detection method need further improvement for efficient and accurate detection at early stages for better treatment and recovery. Thus, the major issues in these previous approaches are low accuracy and high computation time and these might be due the use of irrelevant features in dataset. In order to tackle these problems new methods are needed to detect HD correctly. The improvement in prediction accuracy is a big challenge and research gap.

III. MATERIALS AND METHOD

All the research materials and techniques background are discussed in the following subsections.

A. DATA SET

Cleveland Heart Disease [29] dataset is considered for testing purpose in this study. During the designing of this data set there were 303 instances and 75 attributes, however all published experiments refer to using a subset of 14 of them. In this work, we performed pre-processing on the data set, and 6 samples have been eliminated due to missing values. The remaining samples of 297 and 13 features dataset is left and with 1 output label. The output label has two classes to describe the absence of HD and the presence of HD. Hence features matrix 297*13 of extracted features is formed. The dataset matrix information's are given in Table 2.

B. PRE-PROCESSING OF DATA SET

The pre-processing of dataset required for good representation. Techniques of pre-processing such as removing attribute missing values, Standard Scalar (SS), Min-Max Scalar have been applied to the dataset.

C. STANDARD STATE OF THE ART FEATURES SELECTION ALGORITHMS

After data pre-processing, the selection of feature is required for the process. In general, FS is a significant step in constructing a classification model. It works by reducing the number of input features in a classifier, to have good

TABLE 1. Summary of the previous methods.

Ref	Technique	Limitations	Advantages	Acc(%)
[11]	HD diagnosis using ML classifiers	The Proposed method accuracy is very low.	Computationally less complex.	77
[22]	MLP+SVM	Computationally complex.	The performance of the proposed method is high in terms of prediction accuracy.	80.41
[23]	ANN+Fuzzy Logic	More execution time required to generate results.	Accuracy is high.	87.4
[19]	ANN ensemble based diagnosis system	Computationally complex.	High accuracy.	89.01
[17]	HD diagnosis system based on NB, DT and ANN	The NB and DT performance are low.	ANN achieved high performance in term of accuracy	88.12
[18]	Three phase technique based on ANN	High computation time.	High accuracy.	88.89
[20]	ANN-FUZZY-AHP	Computationally complex.	Achieved high accuracy.	91.1
[25]	Relief-Rough set based method for HD detection	Computation time is high.	High accuracy due to selection of appropriate feature for training and testing of the model.	92.32
[27]	Hybrid ML method	Low accuracy.	Low computation time.	88.07

predictive and short computationally complex models [30]. We have been used four standard state of the art FS algorithms and one our proposed FS algorithm in this study.

1) RELIEF
Relief [31] algorithm assigns weights to each data set features and updated weights automatically. The features having high weight values should be selected and low weight will be discarded. Relief and K-NN algorithm process to determine the weights of features are the same [32]. The algorithm relief repeated through m random training samples (R_k), without selection substitution, and m is the parameter. Each k, R_k is the 'target' sample and weight W of the is updated [33]. The algorithm 1 is the Pseudo-code for Relief FS algorithm.

2) MINIMAL-REDUNDANCY-MAXIMAL–RELEVANCE
MRMR algorithm chooses features that are suitable for the prediction and selected features that are non redundant. It does not take care of the combination of features [32]. The MRMR pseudo code is given in algorithm 2 [34].

3) LEAST-ABSOLUTE-SHRINKAGE-SELECTION-OPERATOR ALGORITHM
LASSO choose feature based on modifying the absolute coefficient value of the features. Then these features coefficient values set to zero and finally zero coefficient features are eliminated from the features set. In the selected features set

---

Algorithm 1 Pseudo-Code for Relief FS Algorithm

Input: S: Training data (feature vectors with class labels), Parameter m: number of random training samples out of total samples used to W.

Output: W: weights for each feature

1: $n \leftarrow$ total number of training samples
2: $d \leftarrow$ number of features (dimensions)
3: $W[A] \leftarrow 0.0$; $\triangleright$ Feature weights set
4: **for** $k \leftarrow 1$ to m **do**
5:     Randomly choose a 'Target' sample R_k
6:     Find a nearest hit H and nearest miss M
7:     **for** $A \leftarrow 1$ to a **do**
8:         $W[A] \leftarrow W[A] - diff(A, R_k, H)/m + diff(A, R_k, M)/m$
9:     **end for**
10: **end for**
11: **Return** W; $\triangleright$ weight vector of features that calculate the quality of features

---

those features to include who coefficient have a high value. Sometime LASSO selects irrelevant features and includes in the subset of feature [35].

4) LOCAL LEARNING BASED FEATURES SELECTION ALGORITHM
LLBFS assigns weights to features and reduced the complexity of non-linear problems into linear. Features having large

TABLE 2. Cleveland heart disease dataset 2016 [11] and [12].

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	$fasting blood sugar > 120 mg/dl$	FBS	$fasting blood sugar > 120 mg/dl$ (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

Algorithm 2 Pseudo Code for MRMR Algorithm

Input: *CF*: Set of initial candidate features, *numR*: number of reduced features wanted.

Output: *SF*: Selected features

1: **for** each feature $f_i \in CF$ **do**
2:      *relevance* ← *mutualInfo(fI , class)*
3:      *redundancy* ← 0
4:      **for** each feature $f_j \in CF$ **do**
5:          *redundancy* ← *redundancy* + *mutualInfo(f_i, f_j)*
6:      **end for**
7:      *mrmrValues[fi]* ← *relevance − redundancy*
8: **end for**
9: *SF* ← *sort(mrmrValues).take(numR)*
10: **return** *SF*   ▷ The set of selected features

---

weighted values are selected and features weights are small discarded from a subset of features [36].

D. PROPOSED FEATURE SELECTION ALGORITHM

In order to tackle the feature selection problem, we proposed Fast conditional mutual information (FCMIM) feature selection algorithm [37] in this study. It is an efficient feature selection method which is designed from conditional mutual information (CMI). The "FCMIM" algorithm designing having the following procedures. Let us consider a dataset $O(X, Y)$, where X instances and Y is output labels. As written in Eq. 1.

$$O(X, Y) = \{(X_i, Y_i) | X_i \in R^n, Y_i \in \{0, 1\}\}_{i=1}^{k} \quad (1)$$

where x_i can be written as in Eq. 2.

$$X_i = \{X_1, X_2, \ldots, X_n\} \quad (2)$$

We apply pre-processing statistical techniques, such as Min-Max normalization on the dataset O(X, Y) as expressed in Eq. 3.

$$V^- = \frac{v - min}{max - min}(new_{max} - new_{min}) + new_{min} \quad (3)$$

Now we apply FCMIM FS technique to select the subset of feature $O(x_i, y_i)$. The FCMIM deploy the CMI to compute the value of feature relevance and duplication in the dataset. FCMIM algorithm chooses features that maximize their mutual information with the target class, conditionally

---

Algorithm 3 Pseudo Code for the Proposed FS Algorithm

Input: load the HD dataset, where O(X, Y) as a data matrix, X is instances and Y output labels. Maxnumerfeatures, selectedfeaturesubset, MI(Uutual Information), CMI(Conditional mutual Information), L(least used index), p(partial score)

Output: selected featuresubset $O(x_i, y_i)$

1: Pre-process the dataset
2: Initialize selected features $= \phi$
3: **for** features $o_i \in O$ **do**
4:     ComputeM_i
5:     set $p_i \leftarrow M_i$
6:     set $L_i \leftarrow 0$
7: **end for**
8: **for** $k \leftarrow 1$ *to* K **do** Initialize $score_i \leftarrow 0$
9:     **for** features $o_i in O$ **do**
10:         **while** $P_i > score_k$ And $L_i < k - 1$ **do**
11:             set $L_i \leftarrow L_i + 1$
12:             Calculate VU_i between o_k and o_i
13:             Set $p_i \leftarrow min(p_i CM_{ik})$
14:         **end while**
15:         **if** $p_i > score_k$ **then**
16:             Set $score_k = p_i$
17:             Selected featuressubset $\leftarrow$ Selected features subset $U o_i$
18:         **end if**
19:     **end for**
20: **end for**

---

to the result of any feature selected before (O). This condition selects features to vary from ones that have selected already even if they are separately correct as they don't more information about output class. It will be good condition between relevance and duplication [37]. The FCMIM high value shows that feature X_n is more relevant to output Y and is highly compatible with another selected feature X_j where $j \in O$ [38]. Mathematically the stated condition is expressed in Eq. 4.

$$CMIM(X_n) = min_{j \in S} I(X_i; Y | X_j)) \qquad (4)$$

The FCMIM algorithm tries to obtain a balance between separate power and independence between the comparison of each new feature with features that elected already. The feature X_0 will be good consideration only if $I(Y, X_0 | X)$ is large for every X already selected. The fast implementation applied feature score during the selection process and evaluate CMI only for those features which give more information and less redundant. FCMIM keep a partial score P_i for every feature O_i which is minimum out of the FCMI that appears in the min in the algorithm equation number 4. The L_i vector store the index of the selected feature based on the calculation of P_i. The "FCMIM" pseudocode is given in algorithm 3.

E. CLASSIFIERS

For the identification of the heart disease classifiers are utilized in this paper and shortly discussed in Table 3.

F. LEAVE-ONE-SUBJECT-OUT CROSSES VALIDATION TECHNIQUE

In this LOSO validation strategy, one sample is separated as test data and remaining subjects to train the model. The test subject is predicted as HD otherwise, the subject is classified as healthy.

G. PERFORMANCE EVALUATION METRICS

Different performance evaluation metrics have been used for classifiers performance evaluation [50], [51]. These metrics are calculated with the help of the confusion matrix. Table 4 shows the binary classification matrix. From Table 4 we computed the following performance evaluation metrics and mathematically shown in Eq. 5-9 respectively.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \times 100 \qquad (5)$$

$$Sensitivity = \frac{TP}{(TP + FN)} \times 100 \qquad (6)$$

$$Specificity = \frac{TN}{(TN + FP)} \times 100 \qquad (7)$$

$$Precision = \frac{TP}{(TP + FP)} \times 100 \qquad (8)$$

$$MCC = \frac{T_1}{\sqrt{T_2 \times T_3 \times T_4 \times T_5}} \times 100 \qquad (9)$$

Here MCC is Matthews correlation coefficient, $T_1 = (TP \times TN - FP \times FN)$, $T_2 = (TP + FP)$, $T_3 = (TP + FN)$, $T_4 = (TN + FP)$, and $T_5 = (TN + FN)$.

H. PROPOSED HEART DISEASE DIAGNOSIS METHODOLOGY

The system has been designed for the identification of heart disease. The performances of various machine learning classifiers for HD identification have been checked on selected features. The standard state of art algorithms of features selection includes Relief, MRMR, LASSO, and LLBFS are utilized for features selection. We also proposed FCMIM algorithm for features selection. The performance of the classifiers evaluated on selected features sets which are selected by the state of the art FS algorithms and proposed FCMIM algorithm. The LOSO technique of cross-validation also used for best model evaluation. The model's performance measuring metrics include accuracy, specificity, sensitivity, MCC and processing time is automatically calculated for classifiers evaluation. The proposed system methodology is organized

TABLE 3. Machine learning algorithms.

Reference	Classifier	Description
[36], [39]–[41]	logistic regression	LR is applied for binary classification problem to predict the value of determining y variable when y [0, 1], negative class is 0 and positive class is 1. It also uses for multi-classification to determine the value of y when y [0,1,2,3].
[26], [41]–[46]	Support Vector Machine	The SVM algorithm mostly used for classification problems. Due to the excellent performance of SVM in classification, various applications mostly used it.
[47]	Naive Bayes	The NB algorithm which is used for classification concerned problem.The training dataset used by NB to compute the value of the conditional probability of vectors for a given class. Conditional Probability value evaluated for each vector, then the new vectors class is evaluated based on its conditionality probability.
[41]	Artificial Neural Network	The ANN algorithm that combines neurons that message passing. ANN classifier has input layer, hidden layer, and output layer. The input layer takes input values, which are used for the training process of the network. The ANN output is calculated for the known class. The weight is re-evaluated using the error margin between the predicted and actual class value.
[41], [48], [49]	Decision Tree	A DT shape is just like a tree in which have a leaf or a decision node. A DT has internal and external nodes linked to each other. The decision making part is internal nodes that takes a decision and child node to visit next nodes.
[41]	K-Nearest Neighbor	K-NN predicts the class label of new input, k-NN used the similarity of new input to its inputs samples in the training set. If the new input is same the samples in the training set. The K-NN classification performance is not good.

TABLE 4. Confusion matrix [29], [52]–[54].

	Predicted healthy person 0(Negative case)	Predicted HD person 1(Positive case)
Actual healthy person 0(negative case)	TN(True Negative)	FP(False Positive)
Actual HD person 1(positive case)	FN(False Negative)	TP(True Negative)

Algorithm 4 Pseudo-Code of Proposed Heart Disease Diagnosis System

---

1: Begin
2: The pre-processing of heart disease dataset using pre-processing methods
3: Features selection using standard state of the art and proposed FCMIM FS algorithms
4: Train the classifiers using training dataset
5: Validate using testing dataset
6: Computes performance evaluation metrics
7: End

---

into these steps such as preprocessing of the dataset, feature selection algorithms, cross-validation method, machine learning classifiers, and classifiers performance evaluation metrics. The algorithm 4 is pseudo-code of the proposed system.

IV. ANALYSIS AND DISCUSSIONS OF EXPERIMENTAL RESULTS

A. EXPERIMENTAL DESIGN SETUP

Supervised classification experiments have been conducted in order to evaluate the classification performance of classifiers. In the first phase, standard features selection algorithms are applied such as Relief, MRMR, LASSO and LLBFS for selection of appropriate features. Then in the second phase of experiments, the proposed FS algorithm was used for features selection. Then the classifiers performances were evaluated on selected features. Furthermore, LOSO CV method is applied with each classifier. To test the performances of the classifiers, various performance evaluation metrics are computed. All the experiments have been performed in a python environment using different machine learning libraries on an Intel(R) C i7-2400 CPU @3.10 GHz system.

B. EXPERIMENTAL RESULTS

1) RESULTS OF DATA PRE-PROCESSING TECHNIQUES

The different statistical operations such as removing attributes missing values, Standard Scalar (SS), Min-Max Scalar, means, standard division have been applied to the dataset. The results of these operations are reported in Table 5. The processed dataset has 297 instances and 13 inputs attribute with one output Label. Data Visualization is the presentation of data in graphical format. It helps people understand the significance of data by summarizing and presenting huge amount of data in a simple and easy-to-understand format and helps communicate information
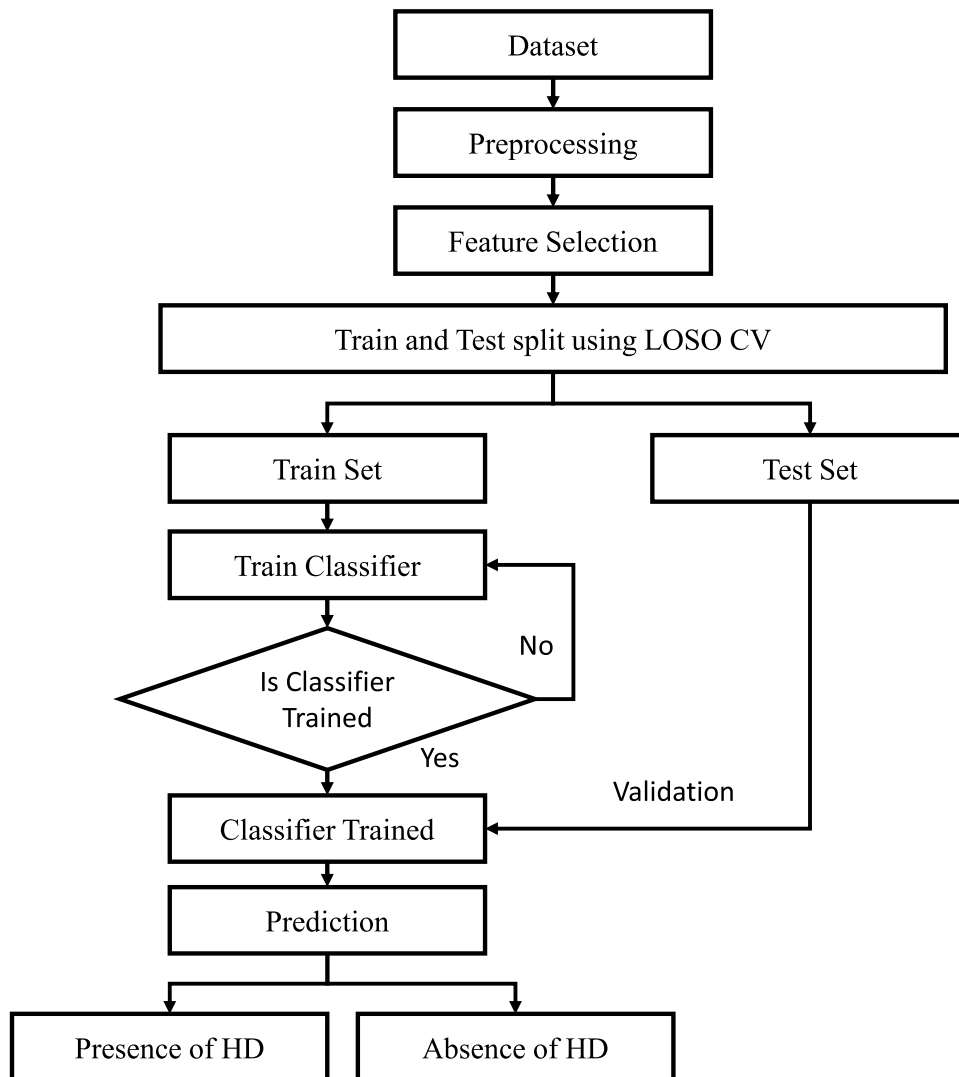
FIGURE 1. Proposed heart disease identification system.

clearly and effectively. Figure 2 is the histogram of the data set represents the frequency of occurrence of specific phenomena which lie within a specific range of values and arranged in consecutive and fixed intervals and Figure 3 describes the co-relation among the features of the dataset using heat map. The heat map, which is a two-dimensional representation of data in which col- ors represent values. A single heat map provides a quick visual summary of information. More elaborate heat maps allow the viewer to understand complex datasets. Furthermore, Heatmap can be super useful when we want to see which intersections of the categorical values have higher concentration of the data compared to the others.

2) FEATURES SELECTED BY STANDARD STATE OF THE ARTS ALGORITHMS

The data preprocessing and important features selected by Relief, MRMR, LASSO, and LLBFS FS algorithms have

been reported in Table 6 along with the features scores and their ranking. According to the results of relief algorithm, the most important features for the identification of heart disease are THA, EIA, and CPT. Other FS algorithms are also selecting these important features such as THA, CPT, SEX, VCA, and EIA. These features are more appropriate for the identification of heart disease. Moreover, FBS has a low score in features scores. All the FS algorithms select some features that mostly selecting by every FS algorithm. Figure 4 shows the important features of scores and ranking graphically for a better understanding of four FS algorithms. The LASSO FS algorithm makes binary classification. LASSO create most realted features to output target class as true and the reminder as false. From 13 features 5 features have been true labeled by LASSO. The selected features have been reported in Table 6. LASSO cross validation mean square results are shown in Figure 5 Lambda is weight parameter and the value of lambda lies between [0,1]. In Figure 5 y-axis is a validation

TABLE 5. The results of statistical operation 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	
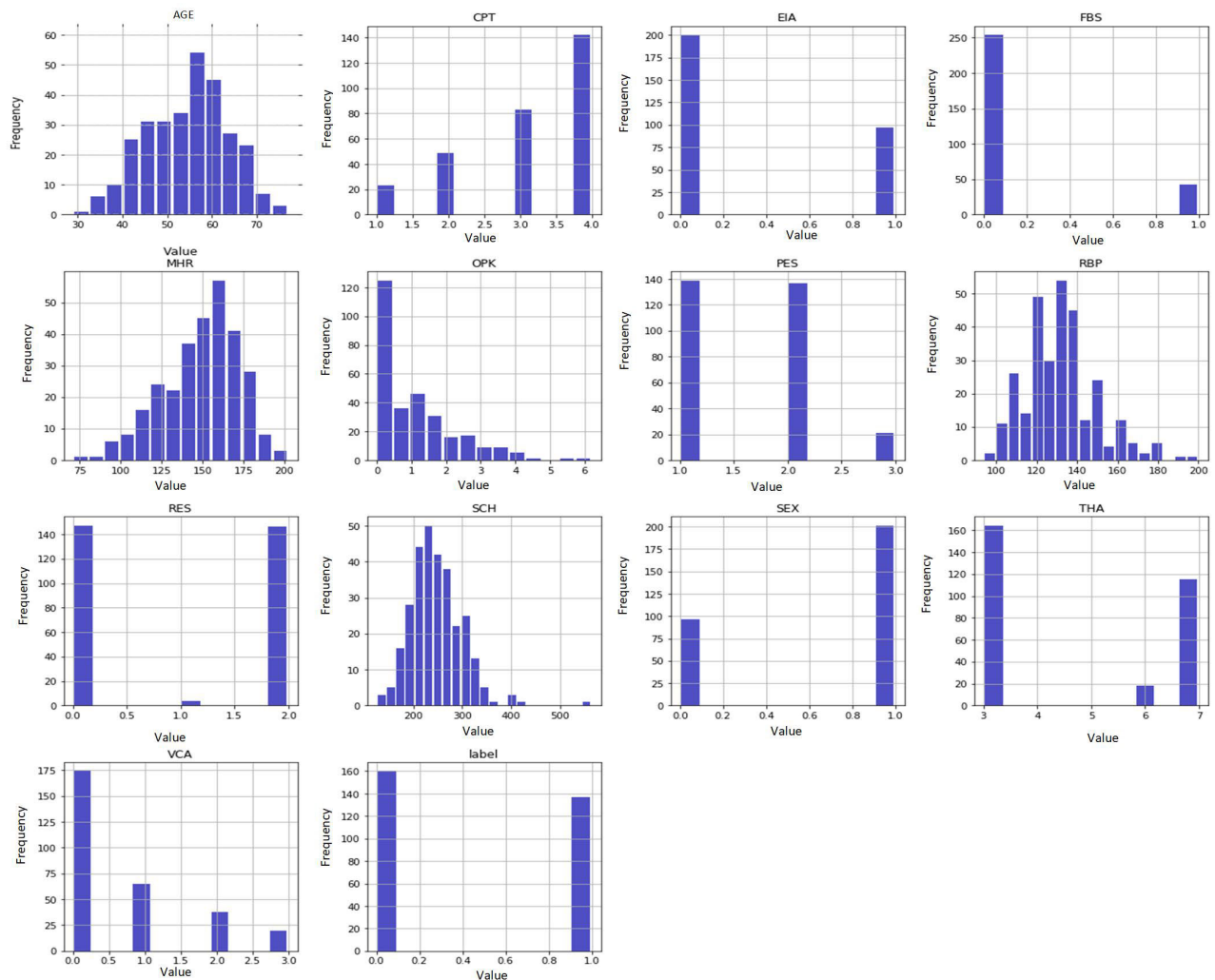
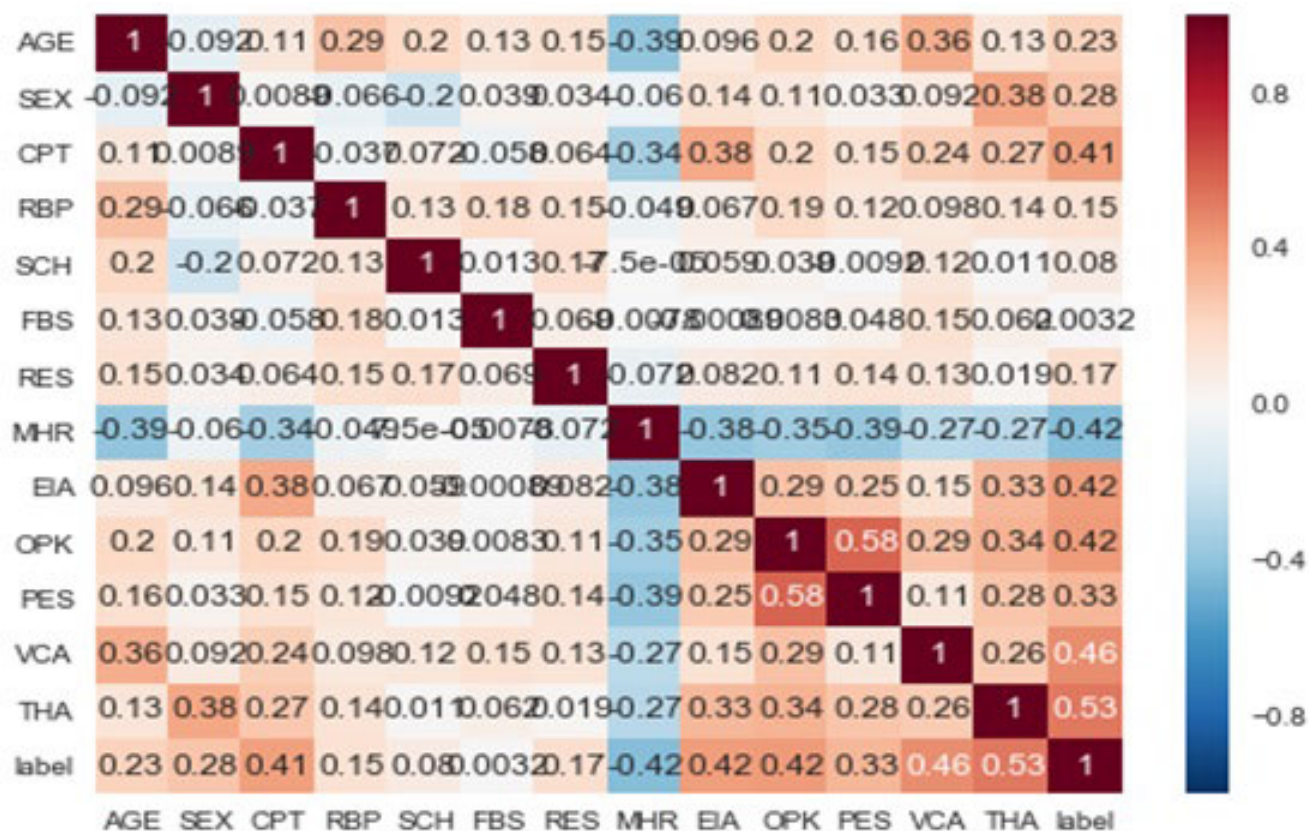**FIGURE 2.** Histgrams of heart disease dataset.

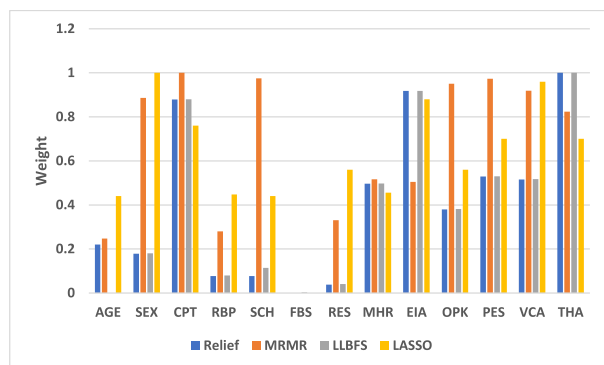**FIGURE 3.** The heat map for correlation features of heart disease dataset.

FIGURE 4. The score of features and ranking selected by FS algorithms.

FIGURE 5. Cross validation MSE of LASSO fit.

measn square error (MSE). 100 various models of subsets of feature created by LASSO using different lambda values. In the Figure 5 the highest point at index 60 shows minimum MSE of the generated model. The vertical line on the left represts the hight value of lambda.

3) RESULTS OF THE PROPOSED FS ALGORITHM
The proposed fast conditional mutual information (FCMIM) feature selection algorithm chooses features on the basis of features mutual information. The FCMIM FS algorithm selected features are SEX, CPT, RBP, SCH, RES, MHR,
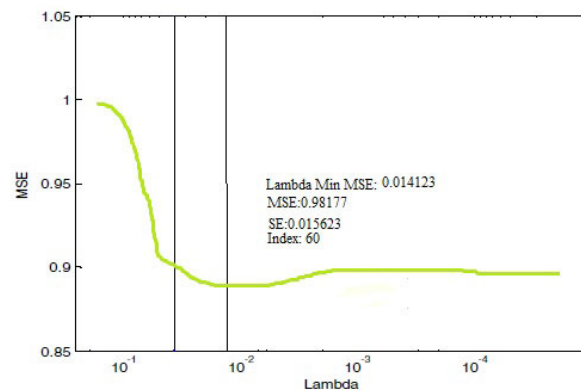
EIA, OPK, PES, and THA. The classifiers classification performances on these selected features are very good. The AGE and FBS features are not selected by this algorithm. In Table 7, we report the features selected by FCMIM FS algorithm along with feature score and graphical describes in Figure 6.

4) RESULTS LOSO CV FOR CLASSIFIERS PERFORMANCE ON FULL FEATURES SET
In this section, on full features set the classifiers performances are measured with the LOSO validation method. Classifiers

TABLE 6. 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
	5	12	VCA	0.128
	6	8	MHR	0.123
MRMR	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

TABLE 7. Features selected by FCMIM FS algorithm.

S.no	Feature code	score
1	SEX	0.523
2	CPT	0.217
3	RBP	0.165
4	SCH	0.575
5	RES	0.696
6	MHR	0.123
7	EIA	0.298
8	OPK	0.561
9	PES	0.574
10	THA	0.486

other parameters values also passed during the training process. Table 8 represents the performance evaluation of classifiers with LOSO CV. According to Table 8, the classifier logistic regression has good performance that obtained 84% accuracy, 93% specificity, and 75% sensitivity and MCC was 84%, and processing time was 0.003 seconds at C = 10 as compared with others values of parameter C. The K-NN, different experiments conducted with different values of k. However, at k = 7 the performance of K-NN was excel-
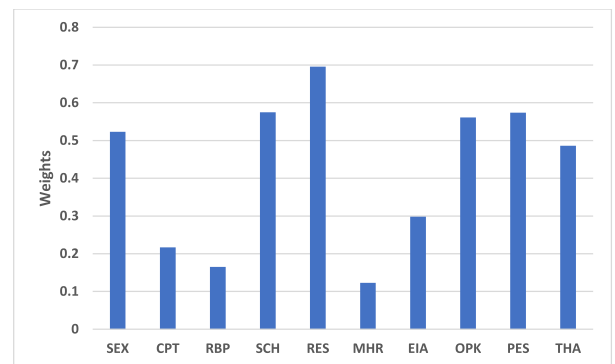
FIGURE 6. Features selected by FMIM FS algorithm.
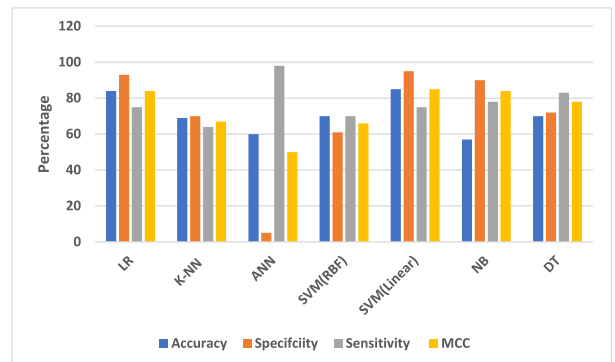
FIGURE 7. Classifiers performance with LOSO CV on set of full features.

lent. ANN was trained with hidden neurons but at 10 hidden neurons give better performance result with accuracy 60%, specificity 100%, and sensitivity 0%. SVM (RBF) with C = 100, g = 0.001 has 61% specificity, 70% sensitivity and 70% accuracy. The SVM linear kernel has 95% specificity, 75% sensitivity, and 85% accuracy. The NB was third good classifiers which have 90% specificity, 78% sensitivity and 80%accuracy. DT has 72% specificity, 83% sensitivity, and 70% accuracy. Figure 7 shows that the SVM outperformed as compared to the other five classifiers. The accuracy of SVM (linear) is 85%, sensitivity 77%, and specificity 95%, and 85% accuracy. Logistic regression is second good classifier has 84% accuracy. The third important classifier is NB and its specificity is 90%, sensitivity is 78%, and classification accuracy is 80%. The worst classifiers were K-NN at k = 1 with LOSO cross-validation. The MCC of SVM is 85% pretty good and SVM is good classifier for heart disease prediction. In Figure 11, we have been shown the execution time of each algorithm in which classifier Svm (linear) on C = 100 and g = 0.009 processing time is 30.145 seconds and logistic regression at C = 10 is 0.003 seconds very fast exaction time as compared to others classifiers with LOSO cross-validation method. Table 8 shows the LOSO cross validation classifiers performance with full features.

In the following sections, the classifiers performances evaluated on features selected by the standard FS algorithm.

TABLE 8. Performance of classifiers on full features set.

Predictive Model	Parameter	Acc(%)	Spe(%)	Sen(%)	MCC (%)	P-Time(s)
LR	C=10	84	93	75	84	0.003
	C=100	83	95	75	85	0.005
K-NN	K=1	59	62	50	56	0.112
	K=3	64	67	59	63	2.511
	K=7	69	70	64	67	6.601
ANN	10	60	5	98	50	8.731
	20	55	0	100	50	9.777
	40	60	100	0	50	7.876
SVM(rbf)	C=100,g=0.0009	57	54	53	54	0.008
	C=10, g=0.001	70	61	70	66	0.01
	C=1, g=0.01	59	70	44	57	0.005
SVM(linear)	C=100, g=0.009	85	95	75	85	30.145
	C = 10, g =0.001	84	94	75	85	9.035
	C=1, g=0.010	83	95	75	85	0.359
NB	100	75	90	78	84	7.49
DT	50	70	72	83	78	1.121

5) CLASSIFIERS PERFORMANCES ON FEATURES SET SELECTED BY RELIEF FS ALGORITHM

In this experiment, features selected by Relief in Table 6 are used with different classifiers with LOSO CV method. Additionally, various parameters values are used with classifiers. First, the classifiers are trained and tested with the selected 3 numbers of features subset, second time 4 feature, than 6 features subset, Similarly 8, 10 subsets were used and lastly used 12 features. The classifiers performance are pretty good on 6 number of features set. Thus 7 tables with LOSO were constructed however, we report the performance of classifiers on 6 important features set as shown in Table 9. Additionally, for a better understanding of the results, some graphs have been created.

According to 9, the results of logistic regression with C = 10 was very good and obtained 85% accuracy, 98% specificity, 72% sensitivity along with 88% MCC. And with low processing time 0.001seconds on reduced 6 features set as compared to other values of hyperparameter C. It is clear that the performance of logistic regression improves with on features selection. There are significant improvements in all evaluation metrics. The classification of Logistic Regression whole features was 84% and on reduced features 85%. We used various values of K however with k = 7 the K-NN show good performance in all metrics 80% accuracy and computation time 4.266 seconds on selected features with LOSO validation methods. However, at the K-NN performances were not good on full features on the same values of k = 7. The processing time K-NN with k = 7 on full and selected features 6.601 seconds and 4.266 seconds respec-

tively. This one of the advantages of features selected for the classification problem. The performance of ANN was designed as MLP and used a various of hidden neurons units. The MLP on 20 units neurons the MLP gives high results on selected features with LOSO validation method and obtained classification accuracy 80% and on full features, the accuracy was 55%. It clears the difference of performance improvement with features selection. Also, the computation time of the ANN algorithm also reduced from 9.777 seconds to 1.867 seconds. The SVM (RBF) at C = 100, and g = 0.0009 were high performance as compared C and g other values as shown in Table 9. SVM (kernel = RBF) obtained accuracy 81%, on selected features and 57% accuracy value on full features with LOSO validation method. The computational time was 0.003 seconds on selected features as compared to the time on full features which was 0.008 seconds. SVM (linear) at C = 100, and g = 0.0009 achieved accuracy 86%, with a computational time 11.569 seconds on reducing selected features by relief with LOSO validation method. The NB accuracy on full feature was 75% and on reduce feature the 1% improvement in performance. Similarly, the DT performance improved from 70% accuracy to 73% with reducing feature. As shown in Figure 7 that in term of accuracy SVM performance was better as compared to other classifiers on selected features. The greater value of ANN specificity 100% which good for detecting healthy people. The sensitivity of SVM linear is 76% on selected features which is good then the sensitivity of full features so SVM good for detecting people with heart disease. The performance evaluation of different classifiers with relief features selection algorithm
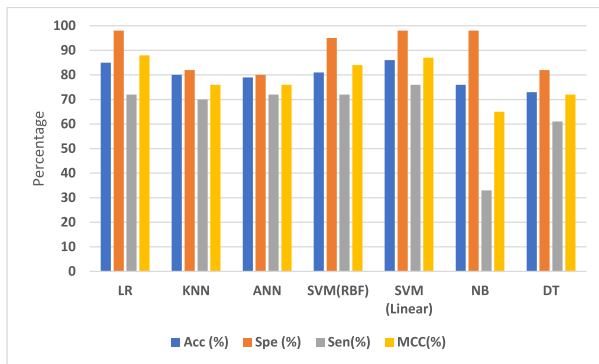
FIGURE 8. Classifiers performances on selected features by Relief.

FIGURE 9. Classifiers performances on 6 important features selected set by mRMR.

has been shown in Figure 8 for a better demonstration for the results.

6) CLASSIFIERS PERFORMANCES ON SELECTED FEATURES BY MRMR

In this section, selected features by mRMR was used in classifiers with LOSO CV. Furthermore, various parameters values were used classifiers. Initially, train- test the classifiers with 3 features subset, second time 5 features, than 7 features subset, similarly used 9, 11 features subsets and lastly used 12 subset of features. The classifiers results were pretty good on 6 numbers of features subset. Total, 8 tables with LOSO CV were formed but we only report the results of classifiers on 6 features set in Table 10 because the overall results of classifiers at 6 features set was high as compared to the performance on experiments on 3, 5, 9, 11, 12 features sets. In Table 10, the results of LR on C = 10 gives high performances. Achieved 86% accuracy, 97% specificity, and 73% sensitivity along with 87% MCC. There are significant improvements in all evaluation metrics. The classification of LR whole features was 84% and on selected features 86% on the same parameter value. The K-NN on k = 7 gives high results in all metrics 82% accuracy and computation time 2.376 seconds on selected features with LOSO validation methods. However, at the K-NN performances were not good on full features on the same values of k = 7. The processing time K-NN at k = 7 on whole and selected features 6.601 seconds and 3.276 seconds respectively. The performance of ANN was created as multilayer perceptron and used a various number of hidden neurons. On 20 hidden neurons the MLP gives high results on selected features and obtained classification accuracy 80% and on full features, the accuracy was 55%. It clears the difference of performance improvement with features selection. Also, the computation time of the ANN algorithm also reduces from 9.777 seconds to 2.867 seconds. The SVM (RBF) results at C = 100 and g = 0.0009 was high as compared to other values of C and g as shown in Table 9. SVM (RBF) achieved accuracy 83%, on selected features and 57% accuracy value on full features with LOSO validation method. The computational time was 0.103 seconds on selected features as compared to the time on full features which was 0.008 seconds. SVM (Linear) with
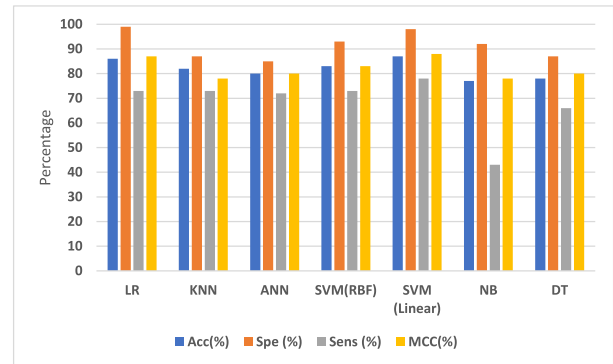
C = 100, and g = 0.0009 achieved accuracy 87%, with a computational time 7.509 seconds on reduced 6 important selected features set by MRMR. The NB accuracy on full feature was 77% and on reduce feature the 2% improvement in performance. Similarly, the DT performance improved from 70% accuracy to 78% with reducing feature with LOSO validation method. As shown in Figure 8 that in term of accuracy SVM performance was better as compared to other classifiers on selected features. The greater value of specificity 99% of logistic regression which good for detecting healthy people. The sensitivity of SVM linear is 79% on selected features which is good then the sensitivity of full features so SVM good for detecting of people with heart disease. The performance evaluation of different classifiers with MRMR features selection algorithm has been shown in Figure 9 for a better demonstration for the results.

7) CLASSIFIERS RESULTS ON FEATURES SELECTED BY LASSO ALGORITHM

The features selected by LASSO were used by classifiers with LOSO CV. We used 3 features set, second time 4 features set, and then 6 features set, similarly 8, 10 features sets were used and lastly used 12 features set. The classifiers performances were high on 6 features set. Hence, 8 tables were constructed on these results but we only described the results of classifiers on 6 features set in Table 11 because the overall results of classifiers at 6 t features set was good as compared to the results of (3, 4, 8, 10, 12) features sets. According to Table 11 results show that the logistic regression on hyperparameter C = 10 was very good performances and obtained 87% accuracy, 95% specificity, 74% sensitivity along with 86% MCC. And with low processing time 0.001 seconds on reduced 6 features set as compared to other values of hyper parameter C with LOSO validation methods. It is clear that the performances of logistic regression improve with on features selection and there were significant improvements in all evaluation metrics. The 95% specificity shows that logistic regression is very best detecting algorithm for healthy people. And 74% sensitivity of logistic regression used for detecting of people with heart disease. The K-NN on k = 7 shows high results. The performance of ANN was formed as multilayer

TABLE 9. Performance of classifiers on features selected by Relief.

Predictive model	Parameter	Acc (%)	Spe (%)	Sen(%)	MCC(%)	P- Time(s)
LR	C=10	85	98	72	88	0.001
	C=100	83	98	72	85	0.002
K-NN	K=1	74	88	59	73	0.002
	K=3	79	80	72	76	2.358
	K=7	80	82	70	76	4.266
	K=9	75	82	70	76	6.217
ANN	10	75	75	70	72	10.06
	20	79	80	72	76	1.867
	40	60	100	75	72	4.121
SVM(RBF)	C=100, g= 0.0009	81	95	72	84	0.003
	C=10, g=0.001	80	85	75	80	0.003
SVM (Linear)	C=100, g=0.0009	86	98	76	87	11.569
	C=10, g=0.001	84	98	75	87	0.359
	C=1, g=0.01	83	98	75	87	0.043
NB	100	76	98	33	65	13.86
DT	50	73	82	61	72	1.397

TABLE 10. Results of classifiers on features selected by MRMR.

Predictive model	Parameter	Acc(%)	Spe (%)	Sens (%)	MCC(%)	P- Time(s)
LR	C=1	83	98	77	84	0.011
	C=10	86	99	73	87	0.031
	C=100	84	95	75	85	0.072
K-NN	K=1	78	87	63	73	0.232
	K=3	80	83	75	79	3.356
	K=7	82	87	73	78	3.276
	K=9	75	82	72	78	5.207
ANN	10	78	70	77	77	5.06
	20	80	85	72	80	2.867
	40	64	91	70	72	1.181
SVM(RBF)	C=100, g= 0.0009	83	93	73	83	0.103
	C=10, g=0.001	81	88	76	82	0.033
	C=1, g=0.01	79	30	79	80	0.202
SVM (Linear)	C=100, g=0.0009	87	98	78	88	7.509
	C=10, g=0.001	85	96	75	87	0.159
	C=1, g=0.01	82	88	45	86	0.003
NB	-	77	92	43	78	9.1
DT	-	78	87	66	80	0.327

perceptron and in MLP were used a various number of hidden neurons. The ANN on 20 hidden neurons the MLP gives high results on selected features set with LOSO validation method and obtained classification accuracy 82% and on full features, the accuracy was 55%. It clears the difference of performance improvement with features selection. Also, the computation time of the ANN algorithm also reduced from 9.777 seconds to 5.931 seconds. The specificity of ANN was 94% at 20 hidden neurons. Therefore, the ANN is good for detection of healthy people. The results of SVM (rbf) at C = 100 and
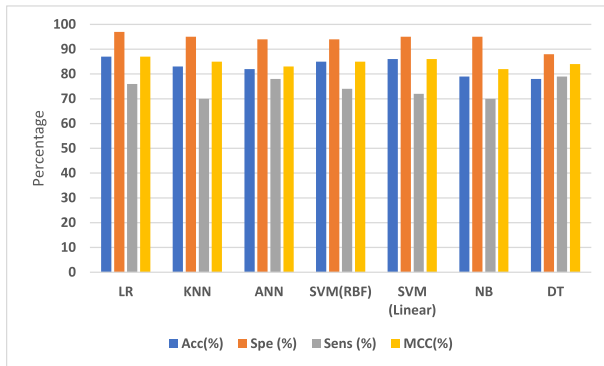
FIGURE 10. Classifiers performances on 6 important features selected set by LASSO.

FIGURE 11. Classifiers performances on 6 important features selected set by LSBFS.

g = 0.0009 were good as compared to other values of C and g as shown in Table 11. SVM (rbf) obtained accuracy 85%, on selected features and 57% accuracy value on full features set. The computational time was 0.007seconds on selected features as compared to the time on full features which was 0.008 seconds. SVM (Linear) on C = 100, and g = 0.0009 achieved accuracy 86% with a computational time 0.021 seconds on 6 selected features set by LASSO with LOSO validation method. The NB accuracy on full feature was 75% and on reduces features set was 76% only 1% improvement in performance. Similarly, the DT performance improved from 70% accuracy to 79% with a reduced feature with LOSO validation method. The decision tree accuracy 78% on selected. As shown in Figure 10 that in term of accuracy logistic regression performance was better as compared to other classifiers on selected features. The greater value of specificity 97% of logistic regression which good for detecting healthy people. The sensitivity of DT 78% on the selected features set and identification of people with heart disease. The performance evaluation of different classifiers with LASSO features selection algorithm has been shown in Figure 10 for the better demonstration for the results.

8) CLASSIFIERS RESULTS ON FEATURES SELECTED BY LLBFS

In these experiments, the features selected by LLBFS FS algorithm were used by classifiers with LOSO CV. Furthermore, various parameters values were used with classifiers. Hence, 8 tables were constructed however we only report the results of 6 features subset in Table 12. Table 12 shows the results described that the logistic regression on hyperparameter C = 10 was very good performances and obtained 88% accuracy, 93% specificity, 75% sensitivity along with 89% MCC. The execution time as 0.001seconds on selected features as compared to other values of hyperparameter C with LOSO validation methods. The 93% specificity shows that logistic regression is very best detecting algorithm for healthy people and 75% sensitivity of logistic regression used detecting of people with heart disease. We used different values of k but k = 7 the K-NN show high results on selected features set. The ANN was created as MLP was used a various number of hidden neurons. At 40 hidden neurons the MLP
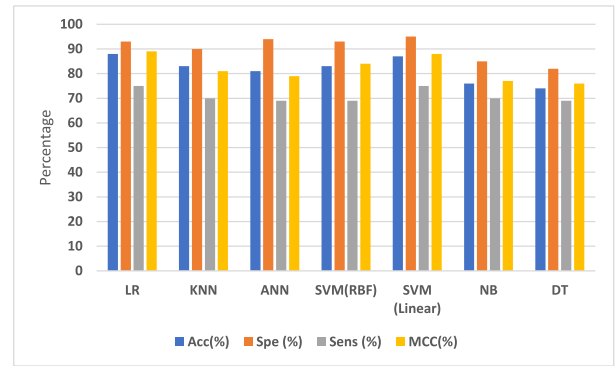
gives high results on selected features with LOSO validation method and obtained classification accuracy 81% and on full features, the accuracy was 55%. The computation time of the ANN algorithm also reduced from 9.777 seconds to 2.501 seconds. The results of SVM (RBF) on C = 100 and g = 0.0009 were high as compared to other values of C, and g as shown in Table 9. SVM (RBF) achieved 82% accuracy on selected features and 57% accuracy value on full features set. The computational time was 0.002 seconds on selected features as compared to the time on full features which was 0.008 seconds. SVM (Linear) at C = 100 and g = 0.0009 achieved accuracy 87%, with a computational time 0.032 seconds on reduced 6 important selected features set. The NB accuracy on full feature was 75% and on reduces, features set 76% only 1% improvement in performance. Similarly, the DT performance improved from 70% accuracy to 74% with reduces features set with LOSO validation method. The performance evaluation of different classifiers with LASSO features selection algorithm has been shown in Figure 11 for a better demonstration for the results.

9) CLASSIFIERS RESULTS ON FEATURES SELECTED BY PROPOSED FS ALGORITHM (FCMIM)

The classifiers results have been evaluated on selected features. The FCMIM FS algorithm selected features are shown in Table 7. The LOSO CV is used and different parameters values are used with classifiers. To demonstrate the results some graphs have been designed for better understanding. In Table 13 report the classification performances of the classifier on selected features LOSO validation. In figure 12 and 13, performance of proposed method graphically shown.

10) CLASSIFIERS PERFORMANCE COMPARISON ON SELECTED FEATURES SELECTED BY PROPOSED FS ALGORITHM (FCMIM) AND STANDARD STATE OF ART FS ALGORITHMS

To determine the best classifiers result with best features selection algorithms using LOSO validation method. According to the results of four states of the art features selection algorithms and proposed FCMIM algorithm, the results of best classifiers with their evaluation metrics have been given in Table 14. According to Table 15, the performance of SVM

TABLE 11. Results of classifiers on s features selected by LASSO.

Predictive Model	Parameter	Acc(%)	Spe (%)	Sen (%)	MCC(%)	P- Time(s)
LR	C=1	85	95	74	86	0.001
	C=10	87	97	76	87	0.001
	C=100	84	95	72	84	0.001
K-NN	K=1	79	95	72	84	0.184
	K=3	82	90	72	80	2.408
	K=7	83	95	70	85	4.519
ANN	16	79	95	70	82	1.364
	20	82	94	78	83	5.931
	40	78	88	64	79	4.543
SVM(RBF)	C=100, g=0.0009	85	94	74	85	0.007
	C=10, g=0.001	80	93	70	81	0.002
SVM (Linear)	C=100, g=0.0009	86	95	72	86	0.021
	C=10, g=0.001	83	96	73	84	0.003
	C=1, g=0.01	79	95	74	82	0.001
NB	-	79	95	70	82	4.134
DT	-	78	88	79	84	0.102

TABLE 12. Results of classifiers on features selected by LLBFS.

Predictive Model	Parameter	Acc(%)	Spe (%)	Sen (%)	MCC(%)	P- Time(s)
LR	C=10	88	93	75	89	0.001
	C=100	83	94	75	85	0.002
K-NN	K=1	77	82	66	75	0.001
	K=3	80	92	67	80	1.471
	K=7	83	90	70	81	3.148
ANN	16	80	93	67	80	8.093
	20	79	88	84	86	8.07
	40	81	94	69	79	2.501
SVM(RBF)	C=100, g=0.0009	82	93	75	85	0.002
	C=10, g=0.001	83	93	69	84	0.003
	C=1, g=0.01	82	92	70	84	0.002
SVM (Linear)	C=100, g=0.0009	87	95	75	88	0.032
	C=10, g=0.001	83	93	75	84	0.004
NB	-	76	85	70	77	2.13
DT	-	74	82	69	76	0.925

in term of accuracy is good and achieved 92.37%accuarcy on selected features selected by proposed FS algorithm (FCMIM) as compared to the state of the arts FS algorithms (Relief, MRMR, LASSO, LLBFS) with LOSO CV. Hence in term of accuracy FCMIM, FS algorithm best for features selection and SVM is suitable classifier for HD diagnosis. LASSO and MRMR performances in term of accuracy with

LOSO validation are also good for heart diagnosis. The specificity of classifiers as reported in Table 13 that specificity of ANN classifier is best on Relief FS algorithm as compared to the specificity of MRMR, LASSO, LLBFS, and FCMIM feature selection algorithms. Therefore, Relief FS algorithm with classifier ANN the specificity is good and best diagnosis system for correct classification of healthy people. The

TABLE 13. Classifiers results on features selected by FCMIM.

Predictive Model	Parameter	Acc(%)	Spe (%)	Sen (%)	MCC(%)	P- Time(s)
LR	C=10	88.67	96	98	91	0.001
K-NN	K=7	82.11	87	74	79	3.276
ANN	40 hidden neuron	75.23	88	59	74	23.361
SVM (Linear)	C=10, g=0.0009	92.37	98	89	90	5.697
NB	-	86.01	89	79	85	2.13
DT	-	79.12	92	89	80	0.925
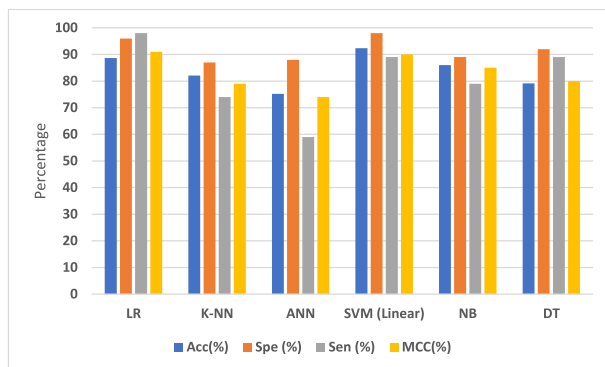
FIGURE 12. Classifiers accuracy on features selected by FCMIM FS algorithm.
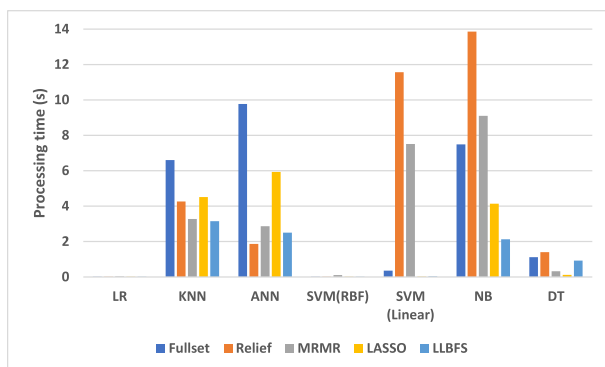
FIGURE 13. Processing time of different algorithm.

specificity of Logistic Regression with MRMR is also best for the correct prediction of healthy people. The sensitivity of the classifier logistic regression is 98% on features selected by FCMIM FS algorithm and correctly classify the people with heart disease. The sensitivity of classifier NB on selected features set by LASSO FS algorithm also give the best result as compared to the sensitivity values of Relief FS algorithm with classifier SVM (linear). In the case of MCC, FCMIM chooses appropriate features with classifier LR and achieved best MCC 91% as compared to the MCC values of MRMR, LASSO, LLBFS, and Relief FS algorithm. The computation time of Relief, LASSO, LLBFS, and FCMIM FS algorithms

with classifier logistic regression are good low as compared to MRMR FS algorithms. Table 15 shows the accuracy of LR improved from 84% to 88% on reduces features with LLBFS algorithm. Similarly, SVM (linear) accuracy improved from 85% to 92.37% on reduces features set with FCMIM. Thus, the performance of classifiers improved with selected features. Finally, we concluded that the diagnosis system for heart disease using FCMIM FS algorithm with classifier SVM is good for effective diagnosis for heart disease. The proposed system (FCMIM + SVM) accuracy is high and achieved 92.37% accuracy as compared to other features selection algorithms and classifiers.

11) PERFORMANCE OF BACKWARD PROPAGATION DEEP NEURAL NETWORK (BPDNN) FOR DETECTION OF HD

In order to compare the performance of machine learning models with deep learning models, we use BPNN for classification problem. The training parameters are updated of BPNN in order to generate high classification results. Therefore, different number of hidden layers, hidden neurons, learning rate and epochs are applied for producing excellent result in our experiments. In Table 16 the BPNN architectures of different networks are given such as BPNN1, BPNN2, and BPNN3. These networks are trained and validated with full features set. According to Table 16 the performance of BPNN2 is high and achieved 91.10% classification accuracy. Thus deep neural network performance is not high as compared to transitional machine learning classifiers. The FCMIM-SVM according to Table 15 achieved 92.37% accuracy. However, deep backward neural network no need feature selection for classification. Deep neural network automatically select import features for improving the result of classification. These are great advantages of deep neural network. However, in our experiments, the Deep neural network performances are not good comparatively to Machine learning models because DNN require more numbers of instances for training the model effectively. The dataset used in our experiments have 297 instances which are insufficient for training the model of DNN to achieved good results. Therefor, ML models are more suitable in case of Small dataset.

**TABLE 14.** Best performances metrics results and best classifiers with Features selection algorithms.

FS Algorithm	The best Acc (%) and classifier	The best Spe (%) and best classifier	The best Sen (%) and classifier	The best MCC and classifier	The best processing time (s) and classifier
Relief	86 SVM (linear) C=100 and g=0.0009	100 ANN with 40 hidden neurons	76 SVM(linear) C=100 and g=0.0009	88 LR with C=10	0.001 LR with C=10
mRMR	87 SVM (linear) C=100 and g=0.0009	99 LR C=10	79 SVM(RBF)with C=1 and g=0.01	88 SVM (linear) C=100 and g=0.0009	0.031 LR with C=10
LASSO	87 LR with C=10	97 LR with C=10	79 DT	87 LR with C=10	0.001 LR with C=10
LLBFS	88 LR with C=10	95 SVM(linear) with C=100 and g=0.0009	84 ANN with 20	89 LR with C=10	0.001 LR with C=10
Proposed Algorithm	92.37 SVM (linear) with C=10,g=0.0009	98 92.5 SVM (linear) with C=10,g=0.0009	98 LR with C=10	91 LR with C=10	0.001 LR with C=10

TABLE 15. Performance comparison of best classifiers before and after feature selection using standers features selection algorithm and proposed FCMIM FS algorithm.

Best Classifier	Acc on Full set	Relief	mRMR	LASSO	LLBFS	FCMIM
LR with C=10	84	85	86	87	88	88.67
SVM(linear) with C=10 and g=0.0009	85	86	87	86	87	92.37

TABLE 16. Training parameters for BPDNNs.

Nework	BPN1	BPN2	BPN3
Training instances	200	200	200
Validating instances	97	97	97
Learning rate	0.0110	0.0001	0.0101
Activation function	relu	relu	relu
Epochs	400	800	900
Training Time(s)	130	180	200
Accuracy (%)	89.50	91.10	90.30
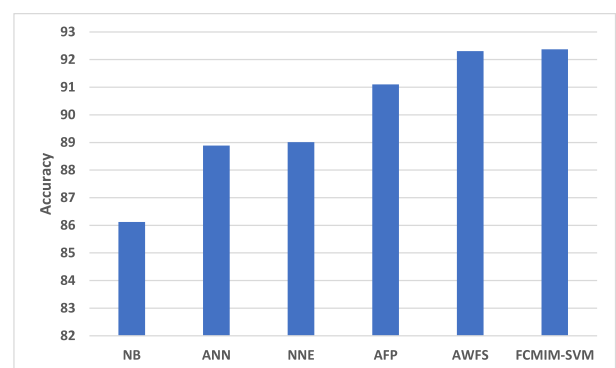
FIGURE 14. Performance comparison of the proposed method with previously proposed methods.

12) PERFORMANCE COMPARISON OF PROPOSED METHOD WITH PREVIOUSLY PROPOSED METHODS

The proposed method (FCMIM-SVM) performance in term of accuracy compared with existing methods in the literature for heart disease diagnosis. The proposed method achieved accuracy of 92.37% as compared to the previous method. The accuracy of the proposed method and existing reported in Table 17 and graphically described in Figure 14 for better understanding. Furthermore, the proposed method suggested for heart disease detection due to an accurate diagnosis. The proposed system can be easily incorporated into health care organization.

TABLE 17. Proposed method performance comparison with existing methods.

Method	Accuracy %)	Reference
NB based HD diagnosis system	86.12	[17]
Three phase ANN diagnosis system	88.89	[18]
Neural network ensembles (NNE)	89.01	[19]
ANN_Fuzzy_AHP diagnosis system (AFP)	91.1	[20]
Adaptive-weighted-Fuzzy-system-ensemble (AWFSE)	92.31	[21]
proposed method (FCMIM-SVM)	92.37	2020

V. CONCLUSION

In this study, an efficient machine learning based diagnosis system has been developed for the diagnosis of heart disease. Machine learning classifiers include LR, K-NN, ANN, SVM, NB, and DT are used in the designing of the system. Four standard feature selection algorithms including Relief, MRMR, LASSO, LLBFS, and proposed a novel feature selection algorithm FCMIM used to solve feature selection problem. LOSO cross-validation method is used in the system for the best hyperparameters selection. The system is tested on Cleveland heart disease dataset. Furthermore, performance evaluation metrics are used to check the performance of the identification system. According to Table 15 the specificity of ANN classifier is best on Relief FS algorithm as compared to the specificity of MRMR, LASSO, LLBFS, and FCMIM feature selection algorithms. Therefore for ANN with relief is the best predictive system for detection of healthy people. The sensitivity of classifier NB on selected features set by LASSO FS algorithm also gives the best result as compared to the sensitivity values of Relief FS algorithm with classifier SVM (linear). The classifier Logistic Regression MCC is 91% on selected features selected by FCMIM FS algorithm. The processing time of Logistic Regression with Relief, LASSO, FCMIM and LLBFS FS algorithm best as compared to MRMR FS algorithms, and others classifiers. Thus the experimental results show that the proposed features selection algorithm select features that are more effective and obtains high classification accuracy than the standard feature selection algorithms. According to feature selection algorithms, the most important and suitable features are Thallium Scan type chest pain and Exercise-induced Angina. All FS algorithms results show that the feature Fasting blood sugar (FBS) is not a suitable heart disease diagnosis. The accuracy of SVM with the proposed feature selection algorithm (FCMIM) is 92.37% which is very good as compared previously proposed methods as shown in Table 17. Further,

the performance of machine learning based method FCMIM-SVM is high then Deep neural network for detection of HD. A little improvement in prediction accuracy have great influence in diagnosis of critical diseases. The novelty of the study is developing a diagnosis system for identification of heart disease. In this study, four standard feature selection algorithms along with one proposed feature selection algorithm is used for features selection. LOSO CV method and performance measuring metrics are used. The Cleveland heart disease dataset is used for testing purpose. As we think that developing a decision support system through machine learning algorithms it will be more suitable for the diagnosis of heart disease. Furthermore, we know that irrelevant features also degrade the performance of the diagnosis system and increased computation time. Thus another innovative touch of our study to used features selection algorithms to selects the appropriate features that improve the classification accuracy as well as reduce the processing time of the diagnosis system. In the future, we will use other features selection algorithms, optimization methods to further increase the performance of a predictive system for HD diagnosis. The controlling and treatment of disease is significance after diagnosis, therefore, i will work on treatment and recovery of diseases in future also for critical disease such as heart, breast, Parkinson, diabetes.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

AVAILABILITY OF DATA AND MATERIAL

The data set used in this study available on UCI machine learning repository.

AUTHORS' CONTRIBUTIONS

All authors equally contributed in the article.

REFERENCES

[1] A. L. Bui, T. B. Horwich, and G. C. Fonarow, "Epidemiology and risk profile of heart failure," *Nature Rev. Cardiol.*, vol. 8, no. 1, p. 30, 2011.

[2] 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.

[3] L. A. Allen, L. W. Stevenson, K. L. Grady, N. E. Goldstein, D. D. Matlock, R. M. Arnold, N. R. Cook, G. M. Felker, G. S. Francis, P. J. Hauptman, E. P. Havranek, H. M. Krumholz, D. Mancini, B. Riegel, and J. A. Spertus, "Decision making in advanced heart failure: A scientific statement from the American heart association," *Circulation*, vol. 125, no. 15, pp. 1928–1952, 2012.

[4] 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, Art. no. 35396.

[5] Q. K. Al-Shayea, "Artificial neural networks in medical diagnosis," *Int. J. Comput. Sci. Issues*, vol. 8, no. 2, pp. 150–154, 2011.

[6] J. Lopez-Sendon, "The heart failure epidemic," *Medicographia*, vol. 33, no. 4, pp. 363–369, 2011.

[7] 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, and Y. J. Woo, "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.

[8] A. 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.

[9] 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.

[10] 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.

[11] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "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.

[12] J. H. Gennari, P. Langley, and D. Fisher, "Models of incremental concept formation," *Artif. Intell.*, vol. 40, nos. 1–3, pp. 11–61, Sep. 1989.

[13] 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.

[14] 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.

[15] 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.

[16] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," 2018, *arXiv:1811.12808*. [Online]. Available: http://arxiv.org/abs/1811.12808

[17] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl.*, Mar. 2008, pp. 108–115.

[18] E. O. Olaniyi, O. K. Oyedotun, and K. Adnan, "Heart diseases diagnosis using neural networks arbitration," *Int. J. Intell. Syst. Appl.*, vol. 7, no. 12, p. 72, 2015.

[19] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7675–7680, May 2009.

[20] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," *Expert Syst. Appl.*, vol. 68, pp. 163–172, Feb. 2017.

[21] A. V. S. Kumar, "Diagnosis of heart disease using fuzzy resolution mechanism," *J. Artif. Intell.*, vol. 5, no. 1, pp. 47–55, Jan. 2012.

[22] M. Gudadhe, K. Wankhade, and S. Dongre, "Decision support system for heart disease based on support vector machine and artificial neural network," in *Proc. Int. Conf. Comput. Commun. Technol. (ICCCT)*, Sep. 2010, pp. 741–745.

[23] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Syst. Appl.*, vol. 35, nos. 1–2, pp. 82–89, Jul. 2008.

[24] M. A. Jabbar, B. Deekshatulu, and P. Chandra, "Classification of heart disease using artificial neural network and feature subset selection," *Global J. Comput. Sci. Technol. Neural Artif. Intell.*, vol. 13, no. 3, pp. 4–8, 2013.

[25] X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang, and Q. Wang, "A hybrid classification system for heart disease diagnosis based on the RFRS method," *Comput. Math. Methods Med.*, vol. 2017, pp. 1–11, Jan. 2017.

[26] A. U. Haq, J. Li, M. H. Memon, M. H. Memon, J. Khan, and S. M. Marium, "Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection," in *Proc. IEEE 5th Int. Conf. Converg. Technol. (ICT)*, Mar. 2019, pp. 1–4.

[27] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[28] G. G. N. Geweid and M. A. Abdallah, "A new automatic identification method of heart failure using improved support vector machine based on duality optimization technique," *IEEE Access*, vol. 7, pp. 149595–149611, 2019.

[29] A. U. Haq, J. Li, M. H. Memon, J. Khan, and S. U. Din, "A novel integrated diagnosis method for breast cancer detection," *J. Intell. Fuzzy Syst.*, vol. 38, no. 2, pp. 2383–2398, 2020.

[30] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," *Inf. Sci.*, vol. 282, pp. 111–135, Oct. 2014.

[31] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *J. Biomed. Informat.*, vol. 85, pp. 189–203, Sep. 2018.

[32] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[33] A. M. D. Silva, *Feature Selection*, vol. 13. Berlin, Germany: Springer, 2015, pp. 1–13.

[34] A. Unler, A. Murat, and R. B. Chinnam, "Mr2PSO: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification," *Inf. Sci.*, vol. 181, no. 20, pp. 4625–4641, Oct. 2011.

[35] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc., B, Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.

[36] F. E. Harrell, Jr., "Ordinal logistic regression," in *Regression Modeling Strategies*. Cham, Switzerland: Springer, 2015, pp. 311–325.

[37] F. Fleuret, "Fast binary feature selection with conditional mutual information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, Nov. 2004.

[38] R. Alzubi, N. Ramzan, H. Alzoubi, and A. Amira, "A hybrid feature selection method for complex diseases SNPs," *IEEE Access*, vol. 6, pp. 1292–1301, 2018.

[39] K. Larsen, J. H. Petersen, E. Budtz-Jørgensen, and L. Endahl, "Interpreting parameters in the logistic regression model with random effects," *Biometrics*, vol. 56, no. 3, pp. 909–914, 2000.

[40] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 2013.

[41] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.

[42] A. V. D. Sánchez, "Advanced support vector machines and kernel methods," *Neurocomputing*, vol. 55, nos. 1–2, pp. 5–20, Sep. 2003.

[43] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[44] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011.

[45] H.-L. Chen, B. Yang, J. Liu, and D.-Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 9014–9022, Jul. 2011.

[46] J. Mourão-Miranda, A. L. W. Bokde, C. Born, H. Hampel, and M. Stetter, "Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data," *NeuroImage*, vol. 28, no. 4, pp. 980–995, Dec. 2005.

[47] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.*, vol. 29, no. 2, pp. 131–163, Nov. 1997.

[48] R. Sivaranjani, V. S. Naresh, and N. V. Murthy, "4 coronary heart disease prediction using genetic algorithm based decision tree," *Intell. Decis. Support Syst., Appl. Signal Process.*, vol. 4, p. 71, Oct. 2019.

[49] A. U. Haq, J. Li, J. Khan, M. H. Memon, S. Parveen, M. F. Raji, W. Akbar, T. Ahmad, S. Ullah, L. Shoista, and H. N. Monday, "Identifying the predictive capability of machine learning classifiers for designing heart disease detection system," in *Proc. 16th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process.*, Dec. 2019, pp. 130–138.

[50] A. Ul Haq, J. Li, Z. Ali, J. Khan, M. H. Memon, M. Abbas, and S. Nazir, "Recognition of the Parkinson's disease using a hybrid feature selection approach," *J. Intell. Fuzzy Syst.*, vol. 39, pp. 1–21, May 2020, doi: 10.3233/JIFS-200075.

[51] A. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad, G. A. Khan, and A. Ali, "Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data," *Sensors*, vol. 20, no. 9, p. 2649, May 2020.

[52] A. U. Haq, J. Li, M. H. Memon, J. Khan, S. U. Din, I. Ahad, R. Sun, and Z. Lai, "Comparative analysis of the classification performance of machine learning classifiers and deep neural network classifier for prediction of parkinson disease," in *Proc. 15th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP)*, Dec. 2018, pp. 101–106.

[53] A. U. Haq, J. P. Li, M. H. Memon, J. Khan, A. Malik, T. Ahmad, A. Ali, S. Nazir, I. Ahad, and M. Shahid, "Feature selection based on L1-norm support vector machine and effective recognition system for Parkinson's disease using voice recordings," *IEEE Access*, vol. 7, pp. 37718–37734, 2019.

[54] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Inf. Syst.*, vol. 2018, pp. 1–21, Dec. 2018.

JIAN PING LI is currently a Chairman of the Computer Science and Engineering College and the Model Software College, University of Electronic Science and Technology of China. He is also the Director of the International Centre for Wavelet Analysis and Its Applications. He is with the National Science and Technology Award Evaluation Committee, the National Natural Science Foundation Committee of China, and the Ministry of Public Security, China, such as Technical Adviser and a dozen academic and social positions. He serves as the Chief Editor for the International Progress on Wavelet Active Media Technology and Information Processing. He also serves as an Associate Editor for the *International Journal of Wavelet Multiresolution and Information Processing*.

JALALUDDIN KHAN received the M.S. degree in computer science from Aligarh Muslim University, Aligarh, India. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China. He has an impressive academic, research, and professional experience in Saudi Arabia. He was a Lecturer with the Deanship of Skills Development and a Researcher with the Center of Excellence in Information Assurance (COEIA), King Saud University, Riyadh, Saudi Arabia. He is accompanying with the Wavelets Active Media Technology and the Big Data Laboratory under the supervision of Prof. J. P. Li and with a collaborated way with other researchers in UESTC. He has authored some research articles. His research interests include the IoT, security and privacy, e-health and telemedicine, machine learning, medical big data concerned technologies, and the IoT security with medical data.

AMIN UL HAQ received the M.S. degree in computer science. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, China. He has a vast academic, technical, and professional experience in Pakistan. He is also a Lecturer with Agricultural University, Peshawar, Pakistan. He is associated with the Wavelets Active Media Technology and the Big Data Laboratory, as an International Student. He has been published high-level research articles in good journals. His research interests include machine learning, medical big data, the IoT, e-health and telemedicine, and concerned technologies and algorithms.

ASIF KHAN received the B.Sc. (Hons.) and Master of Computer Science and Application (M.C.A.) degrees from Aligarh Muslim University, India, and the Ph.D. degree (Hons.) in computer science and technology from the University of Electronic Science and Technology of China (UESTC), China, in 2016. He was an Adjunct Faculty with the University of Bridgeport, USA, for the China Program, in Summer 2016. He was a Visiting Scholar of big data mining and application with the Chongqing Institute of Green and Intelligent Technology (CIGIT), Chinese Academy of Sciences, Chongqing, China. He is currently a Postdoctoral Scientific Research Fellow with UESTC. He is also an Assistant Professor with BSA Crescent University, India. He is a contributor to many international journals with robotics and vision analysis about the contemporary world in his articles. His research interests include machine learning, robotics vision, and new ideas regarding vision-based information critical theoretical research. He received the Academic Achievement Award and the Excellent Performance Award, UESTC, from 2015 to 2016.

SALAH UD DIN received the master's degree from COMSATS University Islamabad, Pakistan. He is currently pursuing the Ph.D. degree in computer science and technology with the University of Electronic Science and Technology of China, China. His research interests include data stream mining, especially on data stream classification, novel class detection, and semi-supervised learning.

ABDUS SABOOR is currently pursuing the M.S. degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, China. He is also a Lecturer with Government University, Peshawar, Pakistan. His research interests include machine learning, medical big data, the IoT, e-health and telemedicine, and concerned technologies and algorithms.

• • •