IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Effective Heart Disease Prediction using Hybrid Machine Learning Techniques

SENTHILKUMAR MOHAN[1], CHANDRASEGAR THIRUMALAI[2], AND GAUTAM SRIVASTAVA[3,4] (Member, IEEE)

[1]School of Information Technology and Engineering, VIT University, Vellore. (e-mail:senthilkumar.mohan@vit.ac.in)
[2]School of Information Technology and Engineering, VIT University(e-mail: chandrasegar.t@vit.ac.in)
[3]Department of Mathematics and Computer Science, Brandon University, Brandon, MB R7A 6A9, Canada (e-mail: srivastavag@brandonu.ca)
[4]Research Center for Interneural Computing, China Medical University, Taichung 40402, Taiwan, Republic of China

Corresponding author: Senthilkumar Mohan (e-mail: senthilkumar.mohan@vit.ac.in)
Corresponding author: Gautam Srivastava (e-mail: srivastavag@brandonu.ca)

ABSTRACT Heart disease is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. We have also seen machine learning (ML) techniques being used in recent developments in different areas of Internet of Things (IoT). Various studies give only a glimpse into predicting heart disease with machine learning techniques. In this paper, we propose a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features, and several known classification techniques. We produce an enhanced performance level with accuracy level of 88.7% through the prediction model for heart disease with Hybrid Random Forest with Linear Model (HRFLM).

INDEX TERMS Machine Learning; Heart disease prediction; Feature selection; Prediction model; Classification algorithms; Cardiovascular disease (CVD);

I. INTRODUCTION

IT is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Various techniques in data mining and neural networks have been employed to find out the severity of heart disease among humans. The severity of the disease is classified based on various methods like K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naive Bayes (NB) [11], [13]. The nature of heart disease is complex and hence, the disease must be handled carefully. Not doing so may affect the heart or cause premature death. The perspective of medical science and data mining are used for discovering various sorts of metabolic syndromes. Data mining with classification plays a significant role in the prediction of heart disease and data investigation.

We have also seen decision trees be used in predicting the accuracy of events related to heart disease [1]. Various methods have been used for knowledge abstraction by using known methods of data mining for prediction of heart disease. In this work, numerous readings have been carried out to produce a prediction model using not only distinct techniques but also by relating two or more techniques. These amalgamated new techniques are commonly known as hybrid methods [14]. We introduce neural networks using heart rate time series. This method uses various clinical records for prediction such as Left bundle branch block (LBBB), Right bundle branch block (RBBB), Atrial fibrillation (AFIB), Normal Sinus Rhythm (NSR), Sinus bradycardia (SBR), Atrial flutter (AFL), Premature Ventricular Contraction (PVC)), and Second degree block (BII) to find out the exact condition of the patient in relation to heart disease. The dataset with a radial basis function network (RBFN) is used for classification, where 70% of the data is used for training and the remaining 30% is used for classification [4], [15].

We also introduce Computer Aided Decision Support System (CADSS) in the field of medicine and research. In previous work, the usage of data mining techniques in the healthcare industry has been shown to take less time for the prediction of disease with more accurate results [16]. We

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2923707, IEEE Access

propose the diagnosis of heart disease using the GA. This method uses effective association rules inferred with the GA for tournament selection, crossover and the mutation which results in the new proposed fitness function. For experimental validation, we use the well known Cleveland dataset which is collected from a UCI machine learning repository. We will see later on how our results prove to be prominent when compared to some of the known supervised learning techniques [5], [17]. The most powerful evolutionary algorithm Particle Swarm Optimization (PSO) is introduced and some rules are generated for heart disease. The rules have been applied randomly with encoding techniques which result in improvement of the accuracy overall [2]. Heart disease is predicted based on symptoms namely, pulse rate, sex, age, and many others. The ML algorithm with Neural Networks is introduced, whose results are more accurate and reliable as we have seen in [8], [12].

Neural networks are generally regarded as the best tool for prediction of diseases like heart disease and brain disease. The proposed method which we use has 13 attributes for heart disease prediction. The results show an enhanced level of performance compared to the existing methods in works like [3]. The Carotid Artery Stenting (CAS) has also become a prevalent treatment mode in the medical field during these recent years. The CAS prompts the occurrence of major adverse cardiovascular events (MACE) of heart disease patients that are elderly. Their evaluation becomes very important. We generate results using a Artificial Neural Network ANN, which produces good performance in the prediction of heart disease [6], [18]. Neural network methods are introduced, which combine not only posterior probabilities but also predicted values from multiple predecessor techniques. This model achieves an accuracy level of up to 89.01% which is a strong results compared to previous works. For all experiments, the Cleveland heart dataset is used with a Neural Network NN to improve the performance of heart disease as we have seen previously in [9], [19].

We have also seen recent developments in machine learning ML techniques used for Internet of Things (IoT) as well [43]. ML algorithms on network traffic data has been shown to provide accurate identification of IoT devices connected to a network. Meidan *et al.* collected and labeled network traffic data from nine distinct IoT devices, PCs and smartphones. Using supervised learning, they trained a multi-stage meta classifier. In the first stage, the classifier can distinguish between traffic generated by IoT and non-IoT devices. In the second stage, each IoT device is associated with a specific IoT device class. Deep learning is a promising approach for extracting accurate information from raw sensor data from IoT devices deployed in complex environments [44]–[47]. Because of its multilayer structure, deep learning is also appropriate for the edge computing environment [48], [49].

In this work, we introduce a technique we call the Hybrid Random Forest with Linear Model (HRFLM). The main objective of this research is to improve the performance accuracy of heart disease prediction. Many studies have been conducted that results in restrictions of feature selection for algorithmic use. In contrast, the HRFLM method uses all features without any restrictions of feature selection. Here we conduct experiments used to identify the features of a machine learning algorithm with a hybrid method. The experiment results show that our proposed hybrid method has stronger capability to predict heart disease compared to existing methods.

The rest of the paper is organized as follows, Section II discusses heart related works, existing methods and techniques available. We also provide an overview of our results in Section III. Section IV discusses HRFLM Data pre-processing followed by feature selection, classification modeling and performance measure. Section V gives the algorithms used and the experiemental setup. Section VI shows the evaluation of datasets and experimental setup. It also shows how the experiment was conducted and the results that were achieved. Section VII contains a discussion about the HRFLM method results and benchmarking of the proposed model. Finally, Section VIII ends with a conclusion of current work and some notes on future enhancement.

II. RELATED WORK

There is ample related work in the fields directly related to this paper. ANN has been introduced to produce the highest accuracy prediction in the medical field [6]. The back propagation multilayer perception (MLP) of ANN is used to predict heart disease. The obtained results are compared with the results of existing models within the same domain and found to be improved [10]. The data of heart disease patients collected from the UCI laboratory is used to discover patterns with NN, DT, Support Vector machines SVM, and Naive Bayes. The results are compared for performance and accuracy with these algorithms. The proposed hybrid method returns results of 86.8% for F-measure, competing with the other existing methods [7]. The classification without segmentation of Convolutional Neural Networks (CNN) is introduced. This method considers the heart cycles with various start positions from the Electrocardiogram (ECG) signals in the training phase. CNN is able to generate features with various positions in the testing phase of the patient [22], [41]. A large amount of data generated by the medical industry has not been used effectively previously. The new approaches presented here decrease the cost and improve the prediction of heart disease in an easy and effective way. The various different research techniques considered in this work for prediction and classification of heart disease using ML and deep learning (DL) techniques are highly accurate in establishing the efficacy of these methods [27], [42].

III. OVERVIEW OF METHOD AND RESULTS

In HRFLM, we use a computational approach with the three association rules of mining namely, apriori, predictive and Tertius to find the factors of heart disease on the UCI Cleveland dataset. The available information points to the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2923707, IEEE Access

deduction that females have less of a chance for heart disease compared to males. In heart diseases, accurate diagnosis is primary. But, the traditional approaches are inadequate for accurate prediction and diagnosis.

HRFLM makes use of ANN with back propagation along with 13 clinical features as the input. The obtained results are comparatively analysed against traditional methods [20], [23]. The risk levels become very high and a number of attributes are used for accuracy in the diagnosis of the disease [24]. The nature and complexity of heart disease require an efficacious treatment plan. Data mining methods help in remedial situations in the medical field. The data mining methods are further used considering DT, NN, SVM, and KNN. Among several employed methods, the results from SVM prove to be useful in enhancing accuracy in the prediction of disease [25]. The nonlinear method with a module for monitoring heart function is introduced to detect the arrhythmias like bradycardia, tachycardia, atrial, atrial-ventricular flutters, and many others. The performance efficacy of this method can be estimated from the accuracy in the outcome results based on ECG data. ANN training is used for the accurate diagnosis of disease and the prediction of possible abnormalities in the patient [26], [34].

Diverse data mining approaches and prediction methods, such as KNN, LR, SVM, NN, and Vote have been rather popular lately to identify and predict heart disease [23]. The novel method Vote in conjunction with a hybrid approach using LR and NB is proposed in this paper. The UCI dataset is used for conducting the experiments of the proposed method, which resulted in 87.4% accuracies in the prediction of heart disease [28], [36]. The Probabilistic Principal Component Analysis (PPCA) method is proposed for evaluation, based on three data sets of Cleveland, Switzerland, and Hungarian in UCI respectively. The method extracts the vectors with high covariance and vector projection used for minimizing the feature dimension. The feature selection with minimizing dimension is provided to a radial basis function, which supports kernel-based SVM. The results of the methods are 82.18%, 85.82% and 91.30% of UCI data sets of Cleveland, Switzerland and Hungarian respectively [29]. The hybrid method combining Linear regression (LR), Multivariate Adaptive Regression Splines (MARS) and ANN is introduced with rough set techniques and is the main novel contribution of this paper. The proposed method effectively reduced the set of critical attributes. The remaining attributes are input for ANN subsequently. The heart disease datasets are used to demonstrate the efficacy of the development of the hybrid approach [30], [38]. The heart disease prediction with multilayer perception of NN is proposed. This method uses 13 clinical attribute features as the input and trained by back propagation are very accurate results in identifying whether the patient has heart disease or not [39].

We also introduce the Apriori algorithm with SVM and compare it with nine other classification methods to predict heart disease more accurately. The results of the classification method have proved a higher degree of accuracy and performance in the prediction of heart disease compared to the other existing methods [32]. The feature selection plays a prominent role in the prediction of heart disease. ANN with back propagation is proposed for better prediction of the disease. The results obtained from the application of ANN are highly accurate and very precise [33]. The genetic algorithm with fuzzy NN known as Recurrent Fuzzy Neural Network (RFNN) is introduced for the diagnosis of heart disease.

In the UCI data set 297 instances of patient records, in total, are considered of which 252 records are used for training and the remaining for testing. The results have been located to be satisfying based on the assessment [35]. Heart disease prediction with SVM and ANN is proposed. In this approach, two methods are used for the premise of the accuracy and time of testing. The proposed model arranges the data records into two classes in SVM as well as ANN for further analysis as shown in [37]. The Back Propagation Neural Network (BPNN) with classification method is introduced, where the hypertension gene sequence is generated and then, thereafter the exact gene sequence. The performance of the BPNN techniques has been measured in the training phase as well as the testing phase with the various numbers of samples. The accuracy of this technique has improved in correspondence to the number of records [40].

IV. PROPOSED METHOD HRFLM

In this study, we have used an R studio rattle to perform heart disease classification of the Cleveland UCI repository. It provides an easy-to-use visual representation of the dataset, working environment and building the predictive analytics. ML process starts from a pre-processing data phase followed by feature selection based on DT entropy, classification of modeling performance evaluation, and the results with improved accuracy. The feature selection and modeling keep on repeating for various combinations of attributes. Table 1 shows the UCI dataset detailed information with attributes used. Table 2 shows the data type and range of values.The performance of each model generated based on 13 features and ML techniques used for each iteration and performance are recorded. Section A summarizes the data pre-processing, Section B discusses the feature selection using entropy, Section C explains the classification with ML techniques and Section D presented for the performance of the results.

A. DATA PRE-PROCESSING

Heart disease data is pre-processed after collection of various records. The dataset contains a total of 303 patient records, where 6 records are with some missing values. Those 6 records have been removed from the dataset and the remaining 297 patient records are used in pre-processing. The multi-class variable and binary classification are introduced for the attributes of the given dataset. The multi-class variable is used to check the presence or absence of heart disease. In the instance of the patient having heart disease, the value is set to 1, else the value is set to 0 indicating the absence of heart disease in the patient. The pre-processing of data is carried

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2923707, IEEE Access

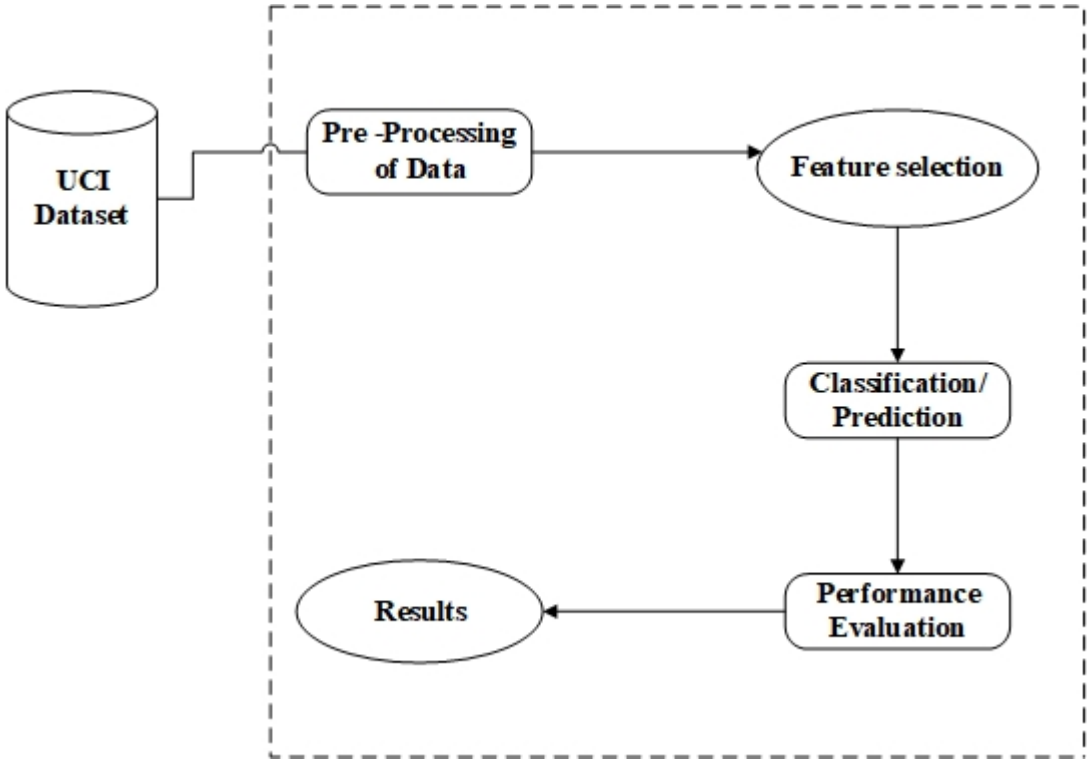

FIGURE 1. Experiment workflow with UCI dataset.

out by converting medical records into diagnosis values. The results of data pre-processing for 297 patient records indicate that 137 records show the value of 1 establishing the presence of heart disease while the remaining 160 reflected the value of 0 indicating the absence of heart disease.

B. FEATURE SELECTION AND REDUCTION

From among the 13 attributes of the data set, two attributes pertaining to age and sex are used to identify the personal information of the patient. The remaining 11 attributes are considered important as they contain vital clinical records. Clinical records are vital to diagnosis and learning the severity of heart disease. As previously mentioned in this experiment, several (ML) techniques are used namely, NB, GLM, LR, DL, DT, RF, GBT and SVM. The experiment was repeated with all the ML techniques using all 13 attributes. Figure 2 shows the prediction method of HRFLM.

C. CLASSIFICATION MODELLING

The clustering of datasets is done on the basis of the variables and criteria of Decision Tree (DT) features. Then, the classifiers are applied to each clustered dataset in order to estimate its performance. The best performing models are identified from the above results based on their low rate of error. The performance is further optimized by choosing the DT cluster with a high rate of error and extraction of its corresponding classifier features. The performance of the classifier is evaluated for error optimization on this data set.

1) Decision Trees

For training samples of data D, the trees are constructed based on high entropy inputs. These trees are simple and fast constructed in a top down recursive divide and conquer (DAC) approach. Tree pruning is performed to remove the irrelevant samples on D.

$$Entropy = -\sum_{j=1}^{m} p_{ij} \log_2 p_{ij} \quad (1)$$

2) Language Model

For given input features x_i, y_i with input vector x_i of data D the linear form of solution $f(x) = mx + b$ is solved by subsequent parameters:

$$m = \frac{(\sum_i x_i y_i) - n\overline{x_i y_i}}{(\sum_i x_i^2) - n\overline{x_i}^2} \quad (2)$$

$b = \bar{y} - m\bar{x}$ where $\bar{x}, \bar{y}$ are the means.

3) Support Vector Machine

Let the training samples having dataset $Data = \{y_i, x_i\}; i = 1, 2, \ldots, n$ where $x_i \in R^n$ represent the i^{th} vector and $y_i \in R^n$ represent the target item. The linear SVM finds the optimal hyperplane of the form $f(x) = w^T x + b$ where w is a dimensional coefficient vector and b is a offset. This is done by solving the subsequent optimization problem:

$$Min_{w,b,\xi_i} \frac{1}{2} w^2 + C \sum_{i=1}^{n} \xi_i \quad (3)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2923707, IEEE Access

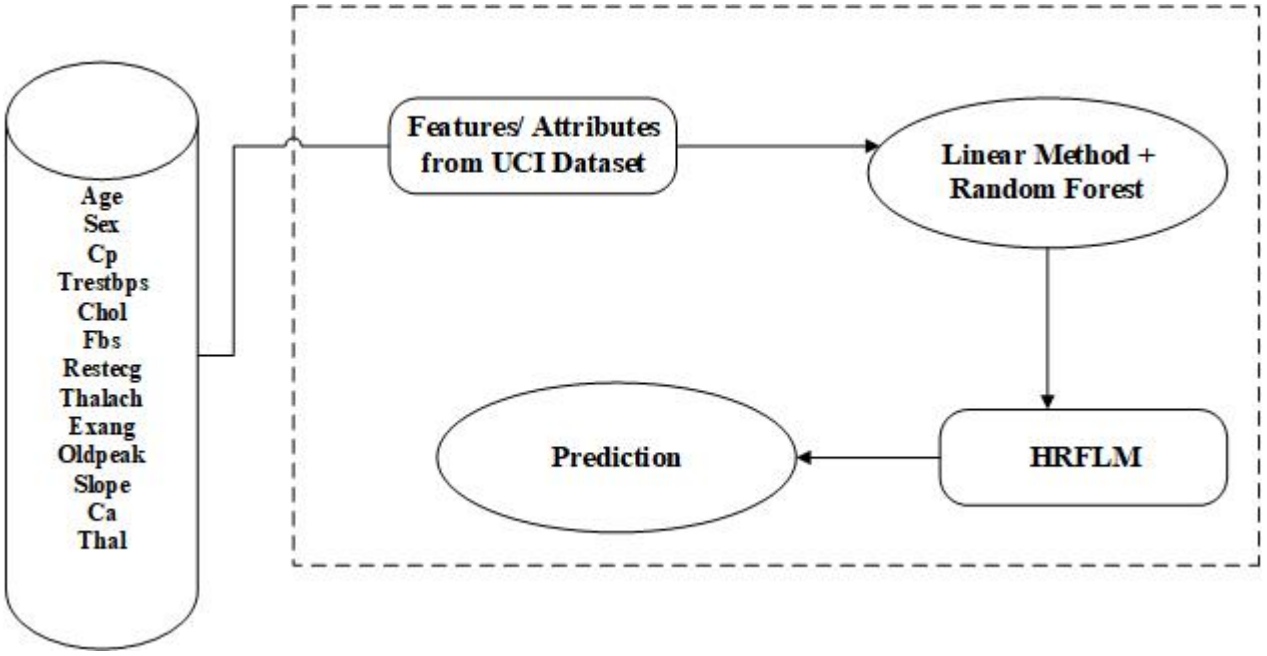

FIGURE 2. Prediction of heart disease with HRFLM.

TABLE 3
RESULT OF VARIOUS MODELS WITH PROPOSED MODEL

Models	Accuracy	Classification error	Precision	F-measure	Sensitivity	Specificity
Naive Bayes	75.8	24.2	90.5	84.5	79.8	60.0
Generalized Linear Model	85.1	14.9	88.8	91.6	94.9	20.0
Logistic Regression	82.9	17.1	89.6	90.2	91.1	25.0
Deep Learning	87.4	12.6	90.7	92.6	95	33.3
Decision Tree	85	15.0	86	91.8	98.8	0.0
Random Forest	86.1	13.9	87.1	92.4	98.8	10.0
Gradient Boosted Trees	78.3	21.7	94.1	86.8	80.7	60.0
Support Vector Machine	86.1	13.9	86.1	92.5	100	0.0
VOTE	87.41	12.59	90.2	84.4	-	-
HRFLM (proposed)	**88.4**	**11.6**	**90.1**	**90**	**92.8**	**82.6**

$$s.t.. \quad y_i \left(w^T x_i + b\right) \geq 1 - \xi_i, \xi_i \geq 0, \forall_i \in \{1, 2, \ldots, m\}$$

4) Random Forest

This ensemble classifier builds several decision trees and incorporates them to get the best result. For tree learning, it mainly applies bootstrap aggregating or bagging. For a given data,$X = \{x_1, x_2, x_3, ..., x_n\}$ with responses $Y = \{x_1, x_2, x_3, ..., x_n\}$which repeats the bagging from $b = 1$ to B. The unseen samples $x^{'}$ is made by averaging the predictions $\sum_{b=1}^{B} fb(x')$ from every individual trees on$x^{'}$:

$$j = \frac{1}{B} \sum_{b=1}^{B} fb\left(x'\right) \quad (4)$$

The uncertainty of prediction on these tree is made through its standard deviation,

$$\sigma = \sqrt{\frac{\sum_{\mathrm{b=1}}^{\mathrm{B}} \left(fb\left(x'\right) - \hat{f}\right)^2}{B - 1}} \quad (5)$$

5) Naive Bayes

This learning model applies Bayes rules through independent features. Every instance of data D is allotted to the class of highest subsequent probability. The model is trained through the Gaussian function with prior probability $P\left(X_f\right) = priority \in (0:1)$

$$P\left(X_{f1}, X_{f2}, \ldots, X_{f_n} | c\right) = \prod_{i=1}^{n} P\left(X_{fi} | c\right) \quad (6)$$

$$P\left(X_f | c_i\right) = \frac{P\left(c_i | X_f\right) P\left(X_f\right)}{P\left(c_i\right)} \quad c \in \{\, benign,\ malignant \,\}$$

At last, the testing data is categorized based on the probability of association:

$$c_{nb} = \arg\max P\left(c_k\right) \prod_{i=1}^{n} P\left(X_{fi} | c_k\right), \ for\ k = 1, 2$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2923707, IEEE Access

TABLE 1
UCI DATASET ATTRIBUTES DETAILED INFORMATION

Attribute	Description	Type
Age	Patient's age in completed years	Numeric
Sex	Patient's Gender (male represented as1 and female as 0)	Nominal
Cp	The type of Chest pain categorized into 4 values: 1. typical angina, 2. atypical angina, 3. non-anginal pain and 4. asymptomatic	Nominal
Trestbps	Level of blood pressure at resting mode (in mm/Hg at the time of admitting in the hospital)	Numeric
Chol	Serum cholesterol in mg/dl	Numeric
FBS	Blood sugar levels on fasting > 120 mg/dl; represented as 1 in case of true, and 0 in case of false	Nominal
Resting	Results of electrocardiogram while at rest are represented in 3 distinct values: Normal state is represented as Value 0, Abnormality in ST-T wave as Value 1, (which may include inversions of T-wave and/or depression or elevation of ST of > 0.05 mV) and any probability or certainty of LV hypertrophy by Estes' criteria as Value 2	Nominal
Thali	The accomplishment of the maximum rate of heart	Numeric
Exang	Angina induced by exercise. (0 depicting 'no' and 1 depicting 'yes')	Nominal
Oldpeak	Exercise-induced ST depression in comparison with the state of rest	Numeric
Slope	ST segment measured in terms of the slope during peak exercise depicted in three values: 1. unsloping, 2. flat and 3. downsloping	Nominal
Ca	Fluoroscopy coloured major vessels numbered from 0 to 3	Numeric
Thal	Status of the heart illustrated through three distinctly numbered values. Normal numbered as 3, fixed defect as 6 and reversible defect as 7.	Nominal
Num	Heart disease diagnosis represented in 5 values, with 0 indicating total absence and 1 to 4 representing the presence in different degrees.	Nominal

TABLE 2
UCI DATASET RANGE AND DATATYPE

AGE	Numeric [29 to 77;unique=41;mean=54.4;median=56]
SEX	Numeric [0 to 1;unique=2;mean=0.68;median=1]
CP	Numeric [1 to 4;unique=4;mean=3.16;median=3]
TESTBPS	Numeric [94 to 200;unique=50;mean=131.69;median=130]
CHOL	Numeric [126 to 564;unique=152;mean=246.69;median=241]
FBS	Numeric [0 to 1;unique=2;mean=0.15;median=0]
RESTECG	Numeric [0 to 2;unique=3;mean=0.99;median=1]
THALACH	Numeric [71 to 202;unique=91;mean=149.61;median=153]
EXANG	Numeric [0 to 1;unique=2;mean=0.33;median=0.00]
OLPEAK	Numeric [0 to 6.20;unique=40;mean=1.04;median=0.80]
SLOPE	Numeric [1 to 3;unique=3;mean=1.60;median=2]
CA	Categorical [5 levels]
THAL	Categorical [4 levels]
TARGET	Numeric [0.00 to 4.00;unique=5;mean=0.94;median=0.00]

6) Neural Networks

The neuron components includes inputsx_i, hidden layers and outputy_i. The final result is produced through the activation function like sigmoid and a bias constant b .

$$f\left(b+\sum_{i=1}^{n} x_i u_i\right) \qquad (7)$$

7) K-Nearest Neighbour

It extract the knowledge based on the samples Euclidean distance function $d(x_i, x_j)$ and the majority of k-nearest neighbors.

$$d\left(x_{i,x_i},\right) = \sqrt{(x_{i,1}-x_{j,1})^2 + \cdots + (x_{i,m}-x_{j,m})^2} \qquad (8)$$

D. PERFORMANCE MEASURES

Several standard performance metrics such as accuracy, precision and error in classification have been considered for the computation of performance efficacy of this model. Accuracy in the current context would mean the percentage of instances correctly predicting from among all the available instances. Precision is defined as the percentage of corrective prediction in the positive class of the instances. Classification error is defined as the percentage of accuracy missing or error available in the instances. To identify the significant features of heart disease, three performance metrics are used which will help in better understanding the behavior of the various combinations of the feature-selection. `ML` technique focuses on the best performing model compared to the existing models. We introduce `HRFLM`, which produces high accuracy and less classification error in the prediction of heart disease. The performance of every classifier is evaluated individually and all results are adequately recorded for further investigation.

E. HRFLM ALGORITHMS

Algorithm 1 Decision tree-based partition

Require: Input: D dataset – features with a target class
  for ∀features **do**
    for Each sample **do**
      Execute the Decision Tree algorithm
    end for
    Identify the feature space f_1, f_2, …, f_x of dataset UCI. (9)
  end for
  Obtain the total number of leaf nodes l_1, l_2, l_3,…, l_n with its constraints (10)
  Split the dataset D into d_1, d_2, d_3,…, d_n based on the leaf nodes constraints. (11)
  Output: Partition datasets d_1, d_2, d_3,…, d_n

Algorithm 2 Apply `ML` to find less error rate

Require: Input: Datasets with partition – d_1, d_2, d_3,…, d_n
  for ∀apply the rules **do**
    On the dataset R(d1,d2,d3,…… dn)
  end for
  Classify the dataset based on the rules C(R(d_1),R(d_2)……R(d_n)) (12)
  Output: Classified datasets with rules C(R(d_1),R(d_2)……R(d_n))

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2923707, IEEE Access

TABLE 4
CLASSIFICATION RULES FOR HRFLM

Rule No	Attributes	Values	Range
1.	root	212 97 0	(0.54 0.18 0.13 0.12 0.033)
2.	CA=0	123 29 0	(0.76 0.15 0.049 0.024 0.0081)
3.	CA=1,2,3	89 67 3	(0.24 0.21 0.24 0.25 0.067)
4.	EXANG	< 0.5 97 11 0	(0.89 0.1 0 0.01 0)
5.	EXANG	>=0.5 26 17 1	(0.31 0.35 0.23 0.077 0.038)
6.	SEX	< 0.5 24 11 0	(0.54 0 0.17 0.25 0.042)
7.	SEX	>=0.5 65 46 1	(0.12 0.29 0.26 0.25 0.077)
8.	THAL	=3 11 3 0	(0.73 0.18 0.091 0 0)
9.	THAL	=3,6,7 15 8 1	(0 0.47 0.33 0.13 0.067)
10.	THAL	=3 15 2 0	(0.87 0 0.067 0.067 0)
11.	THAL	=6,7 9 4 3	(0 0 0.33 0.56 0.11)
12.	THALACH	>=139.5 33 20 1	(0.15 0.39 0.33 0.061 0.061)
13.	RESTECG	< 1 17 10 2	(0.29 0.29 0.41 0 0)
14.	RESTECG	>=1 16 8 1	(0 0.5 0.25 0.12 0.12)
15.	THALACH	< 139.5 32 18 3	(0.094 0.19 0.19 0.44 0.094)

Algorithm 3 Feature Extraction using less error Classifier

Require: Input: Classified datasets C(R(d_1),R(d_2)..... R(d_n))

for ∀Find out min error rate from the input **do**

Min(C(R(d_1),R(d_2)..... R(d_n))) (13)

end for

Find out max(min) error rate from the classifier.

Output: Features with classified attributes F(d_1, d_2, d_3,..., d_n)

Algorithm 4 Apply Classifier on extracted features

Apply the hybrid method based on the error rate

$$\sum_0^n F(n) = d + m_1x_1 + m_2x_2 + ... + m_nx_n \quad (14)$$

$$\sum_0^n F(0) = Gain + \sum_0^n w_ix_i \quad (15)$$

V. EXPERIMENTATIONAL ENVIRONMENT

A. DATASETS

Heart disease data was collected from the UCI machine learning repository. There are four databases (i.e. Cleveland, Hungary, Switzerland, and the VA Long Beach). The Cleveland database was selected for this research because it is a commonly used database for ML researchers with comprehensive and complete records. The dataset contains 303 records. Although the Cleveland dataset has 76 attributes, the data set provided in the repository furnishes information for a subset of only 14 attributes. The data source of the Cleveland dataset is the Cleveland Clinic Foundation. Table 1 depicts the description and type of attributes. There are 13 attributes that feature in the prediction of heart disease, where only one attribute serves as the output or the predicted attribute to the presence of heart disease in a patient.

The Cleveland dataset contains an attribute named num to show the diagnosis of heart disease in patients on different scales, from 0 to 4. In this scenario, 0 represents the absence of heart disease and all the values from 1 to 4 represent patients with heart disease, where the scaling refers to the severity of the disease (4 being the highest). Figure 1 shows the distribution of the num attribute among the identified 303 records.

TABLE 6
RESULTS GENERATED BASED ON HRFLM

	Overall error rate		Best Model	Overall classification error rate		Best Model
Data Split	**RF**	**LM**		**RF**	**LM**	
1	4	6.7	RF	14.9	16.2	DT /RF
2	12.2	22.6	RF	37.7	38.7	RF
3	11.1	16.6	RF	27.8	50	RF
4	20.9	29.2	RF	54.1	54.2	RF
5	13.3	13.3	RF/ LM	53.4	53.3	LM
6	12	18.1	RF	57.6	54.6	LM
7	0	28.5	RF	28.5	42.8	RF
8	18.2	9.1	LM	27.3	27.3	RF/ LM

B. EXPERIMENTAL SETUP FOR EVALUATION

We have used an R studio rattle to perform the classification of heart disease from Cleveland UCI repository. Figure 1 depicts the evaluation of the experiment by step-by-step stages. In the first step, the UCI dataset is loaded and the data becomes ready for pre-processing. The subset of 13 attributes (Age, sex, cp, treetops, chol, FBS, restecg, thalach, exang, olpeak, slope, ca, that, target) is selected from the pre-processed data set of heart disease. The three existing models for heart disease prediction (DT, RM, LM) are used to develop the classification. The evaluation of the model is performed with the confusion matrix. Totally, four outcomes are generated by confusion matrix, namely TP (**True Positive**), TN (**True Negative**), FP (**False Positive**) and FN (**False Negative**). The following measures are used for the calculation of the accuracy, sensitivity, specificity.

Accuracy = (TN+TP) / (TN+TP+FN+FP) =105+155/295= 0.8847

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2923707, IEEE Access

TABLE 5
RESULT OF VARIOUS MODELS WITH PROPOSED MODEL

	Overall error rate			Best Model	Overall classification error rate			Best Model
Data Split	**DT**	**RF**	**LM**		**DT**	**RF**	**LM**	
1	14.9	4	6.7	RF	14.9	14.9	16.2	DT /RF
2	34.9	12.2	22.6	RF	39.6	37.7	38.7	RF
3	50	11.1	16.6	RF	50	27.8	50	RF
4	62.5	20.9	29.2	RF	62.5	54.1	54.2	RF
5	60	13.3	13.3	RF/ LM	60	53.4	53.3	LM
6	54.6	12	18.1	RF	60.6	57.6	54.6	LM
7	57.1	0	28.5	RF	57.1	28.5	42.8	RF
8	36.4	18.2	9.1	LM	36.4	27.3	27.3	RF/ LM

Sensitivity = (TP/TP+FN) =155/155+12 = 92.8
Specificity = (TN/TN+FP) =105/105+22 = 82.6
Precision = TP /TP+FP = 155/155+22==87.5
F-Measure= 2TP/ 2TP+FP+FN =310 /310+22+12 = 0.90

VI. EVALUATION RESULTS

The prediction models are developed using 13 features and the accuracy is calculated for modeling techniques. The best classification methods are given below in Table 3. This table compares the accuracy, classification error, precision, F-measure, sensitivity and specificity. The highest accuracy is achieved by HRFLM classification method in comparison with existing methods.

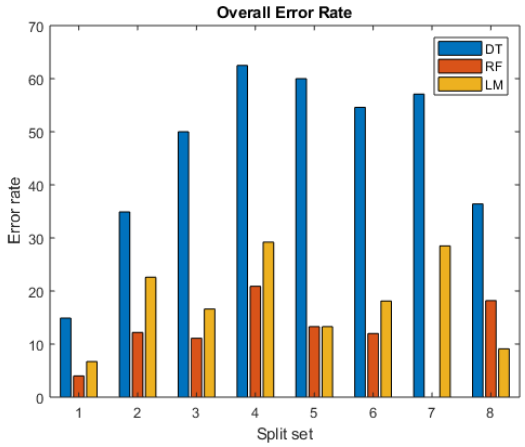

FIGURE 3. Overall error rate of the dataset

VII. DISCUSSION OF HRFLM TO IMPROVE THE RESULTS

The UCI dataset is further classified into 8 types of datasets based on classification rules. The classification rules are listed in Table 4. Each dataset is further classified and processed by **R Studio Rattle**. The results are generated by applying the classification rule for the dataset.

The classification rules generated based on the rule after data pre-processing is done. After pre-processing, the data's three best ML techniques are chosen and the results are generated. The various datasets with DT, RF, LM are applied to find out the best classification method. Table 5 shows that results of existing and proposed methods.

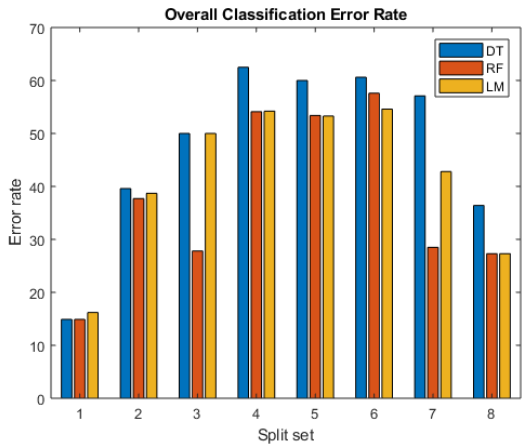

FIGURE 4. Overall classification error rate of the dataset

The results show that RF and LM are the best. The RF error rate for dataset 4 is high (20.9%) compared to the other datasets. The LM method for the dataset is the best (9.1%) compared to DT and RF methods. We combine the RF method with LM and propose HRFLM method to improve the results. Table 6 show the results of the proposed method. Figure 3 shows the overall error rate of the dataset. Figure 4 shows the overall classification error rate of the dataset.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2923707, IEEE Access

TABLE 7
COMPARISON OF VARIOUS MODELS WITH THE PROPOSED MODEL

Source	Sex	cp	Fbs	restecg	exang	Oldpeak	Slope	Ca	thal
Bhatla & Jyoti (2012	0	1	0	0	1	1	0	1	1
Nahar, Imam, Tickle & Chen (2013)	0	1	1	1	1	0	0	0	0
Sen, Patel & Shukla (2013)	1	1	1	1	1	0	0	0	0
Chaurasia & Pal (2013)	1	1	1	1	1	0	1	0	0
Tomar & Agarwal (2014)	0	1	1	1	0	1	1	1	0
Nahato, Harichandran & Arputharaj (2015)	0	1		1	0	1	0	1	1
Paul, Shill, Rabin & Akhand (2016)	1	1	0	1	1	1	1	1	1
Dey, Singh & Singh (2016)	1	1	1	1	1	0	1	0	0
Wiharto et.al (2017)	1	1	0	0	1	1	1	0	0
Liu, Wang, et.al (2017)	0	1	0	1	0	0	1	1	1
Total	5	10	5	8	7	5	6	5	4

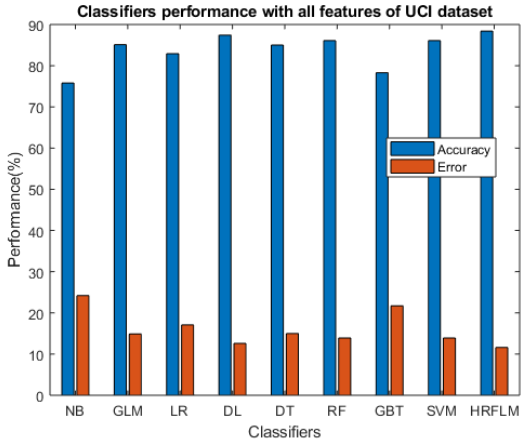

FIGURE 5. Performance comparison with various models

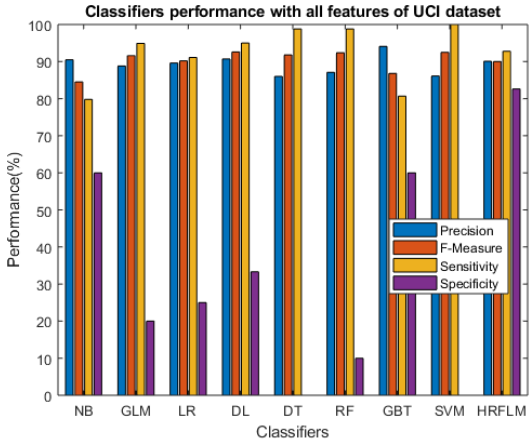

FIGURE 6. Performance comparison with various models

A. BENCHMARKING OF THE PROPOSED MODEL

Benchmarking is needed to compare the performance of the existing models compared with the proposed model. This method is used to identify whether the proposed method is the best and improves accuracy or not. The accuracy is calculated with the number of feature selection and the model generated results. HRFLM has no restriction in selecting of features to use. All the features selected in this model accomplish the best results. Table 7 shows that comparison of various models with our proposed method. Figure 5 and Figure 6 shows the performance comparison of the various model with respect to proposed method respectively.

Table 5 depicts the details of features selected by various models from the UCI dataset for heart disease. The proposed method is used on all 13 attributes and classified, based on the error rate. This result clearly proves that all the features selected and ML techniques used, prove effective in accurately predicting heart disease of patients compared with known existing models.

VIII. CONCLUSION

Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible. Further extension of this study is highly desirable to direct the investigations to real-world datasets instead of just theoretical approaches and simulations. The proposed hybrid HRFLM approach is used combining the characteristics of Random Forest (RF) and Linear Method (LM). HRFLM proved to be quite accurate in the prediction of heart disease. The future course of this research can be performed with diverse mixtures of machine learning techniques to better prediction techniques. Furthermore, new feature-selection methods can be developed to get a broader perception of the significant features to increase the performance of heart disease prediction.

REFERENCES

[1] Abdullah, A.S., 2012. A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier. , (Icon3c), pp.22 −25.

[2] Alkeshuosh, A.H., Moghadam, M.Z. Mansoori, I. Al, 2017. Diagnosis of Heart Disease. , pp.306−311.

[3] Al-milli, N., 2013. Backpropogation Neural Network for Prediction of Heart Disease., 56(1), pp.131−135.

[4] A. Devi,S. Rajamhoana, C, K. Umamaheswari, R. Kiruba, K. Karunya and R. Deepika, Analysis of Neural Networks Based Heart Disease Prediction System, 2018 11th International Conference on Human System Interaction (HSI), Gdansk, 2018, pp. 233-239.

[5] Anooj, P.K., 2012. Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. Journal of King Saud University - Computer and Information Sciences, 24(1), pp.27−40. Available at: http://dx.doi.org/10.1016/j.jksuci.2011.09.002.

[6] Baccour, L., 2018. Amende d fuse d TOPSIS-VIKOR for classification (ATOVIC) applied to some UCI data sets R. Expert Systems With Applications, 99, pp.115−125. Available at: https://doi.org/10.1016/j.eswa.2018.01.025.

[7] Cheng, C. Chiu, H., 2017. An Artificial Neural Network Model for the Evaluation of Carotid Artery Stenting Prognosis Using a National-Wide Database. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp.2566−2569.

[8] H. A. Esfahani and M. Ghazanfari,"Cardiovascular disease detection using a new ensemble classifier," 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), Tehran, 2017, pp. 1011-1014.

[9] Dammak, F., Baccour, L. Alimi, A.M., The Impact of Criterion Weights Techniques in TOPSIS Method of Multi-Criteria Decision Making in Crisp and Intuitionistic Fuzzy Domains. 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), (9), pp.1−8.

[10] Das, R., Turkoglu, I. Sengur, A., 2009. Expert Systems with Applications Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications, 36(4), pp.7675−7680. Available at: http://dx.doi.org/10.1016/j.eswa.2008.09.013.

[11] Durairaj, M. Revathi, V., 2015. Prediction Of Heart Disease Using Back Propagation MLP Algorithm. , 4(08), pp.235−239.

[12] Gandhi, M., 2015. Predictions in Heart Disease Using Techniques of Data Mining. 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), pp.520−525.

[13] Gavhane, A., 2018. Prediction of Heart Disease Using Machine Learning. 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), (Iceca), pp.1275−1278.

[14] Jpdlo, V. et al., 2018. Heart diseases prediction with Data Mining and Neural Network Techniques. , 6(7 2), pp.1−6.

[15] K, S.B.N., 2016. Prediction of Heart Disease at early stage using Data Mining and Big Data Analytics: A Survey. , pp.256−261.

[16] Kelwade, J.P., 2016. Radial basis function Neural Network for Prediction of Cardiac Arrhythmias based on Heart rate time series. , (Cmi), pp.454−458.

[17] Krishnaiah, V. Chandra, N.S., 2016. Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review. , 136(2), pp.43−51.

[18] Kumar, P.S. et al., 2016. A Computational Intelligence Method for Effective Diagnosis of Heart Disease using Genetic Algorithm. , 8(2), pp.363−372.

[19] Liberatore, M.J. Nydick, R.L., 2008. The analytic hierarchy process in medical and health care decision making: A literature review. , 189, pp.194−207.

[20] Mahboob, T., Irfan, R. Ghaffar, B., 2017. Evaluating Ensemble Prediction of Coronary Heart Disease using Receiver Operating Characteristics. , pp.110−115.

[21] Nahar, J., Imam, T., Tickle, K.S., et al., 2013. Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Systems with Applications, 40(1), pp.96−104. Available at: http://dx.doi.org/10.1016/j.eswa.2012.07.032.

[22] Nahar, J., Imam, T., Tickle, K.S., et al., 2013. Expert Systems with Applications Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems With Applications, 40(4), pp.1086−1093. Available at: http://dx.doi.org/10.1016/j.eswa.2012.08.028.

[23] Nayak, S. et al., 2018. Applicability of the Cleveland clinic scoring system for the risk prediction of acute kidney injury after cardiac surgery in a South Asian cohort. Indian Heart Journal, 70(4), pp.533−537. Available at: https://doi.org/10.1016/j.ihj.2017.11.022.

[24] Tulay Karaylan and Ozkan Kilic, "Prediction of heart disease using neural network," 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, 2017, pp. 719-723.

[25] R, T.P. Thomas, J., 2016. Human Heart Disease Prediction System using Data Mining Techniques.

[26] Raju, C. et al., 2018. Mining Techniques. 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), (March), pp.253−255.

[27] Ravish, D.K. Shenoy, N.R., 2014. Heart Function Monitoring , Prediction and Prevention of Heart Attacks: Using Artificial Neural Networks. , pp.1−6.

[28] Sabahi, F., 2018. Bimodal fuzzy analytic hierarchy process (BFAHP) for coronary heart disease risk assessment. Journal of Biomedical Informatics, 83(April), pp.204−216. Available at: https://doi.org/10.1016/j.jbi.2018.03.016.

[29] Shafenoor Amin, M., Kia Chiam, Y. Dewi Varathan, K., 2018. Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0736585318308876.

[30] Shah, S.M.S. et al., 2017. Feature extraction through parallel Probabilistic Principal Component Analysis for heart disease diagnosis. Physica A: Statistical Mechanics and its Applications, 482, pp.796−807. Available at: http://dx.doi.org/10.1016/j.physa.2017.04.113.

[31] Shao, Y.E., Hou, C. Chiu, C., 2014. Hybrid intelligent modeling schemes for heart disease classification. Applied Soft Computing Journal, 14, pp.47−52. Available at: http://dx.doi.org/10.1016/j.asoc.2013.09.020.

[32] Sonawane, J.S. Student, P.G., 2014. Prediction of Heart Disease Using Multilayer Perceptron Neural Network. , (978).

[33] Sowmiya, C., 2017. Analytical Study of Heart Disease Diagnosis Using Classification Techniques.

[34] Tarle, B., 2017. An Artificial Neural Network Based Pattern Classification Algorithm for Diagnosis of Heart Disease. 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), pp.1−4.

[35] Tran, V.P. Al-jumaily, A.A., 2017. Non-Contact Doppler Radar Based Prediction of Nocturnal Body Orientations Using Deep Neural Network for Chronic Heart Failure Patients. , pp.3−7.

[36] Uyar, K. Ilhan, A., 2017. Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. Procedia Computer Science, 120, pp.588−593.

[37] Vivekanandan, T. Sriman Narayana Iyengar, N.C., 2017. Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease. Computers in Biology and Medicine, 90(April), pp.125−136.

[38] S. Radhimeenakshi, "Classification and prediction of heart disease risk using data mining techniques of Support Vector Machine and Artificial Neural Network," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 3107-3111.

[39] Wagh, R. Paygude, S.S., 2016. CDSS for Heart Disease Prediction Using Risk. , pp.12082−12089.

[40] Williams, O. et al., 2017. An integrated decision support system based on ANN and Fuzzy AHP for heart failure risk prediction. Expert Systems With Applications, 68, pp.163−172.

[41] Zaman, S., 2017. Codon Based Back Propagation Neural Network Approach to Classify Hypertension Gene Sequences. , pp.443−446.

[42] Zhang, W. Han, J., 2017. Towards Heart Sound Classification Without Segmentation Using Convolutional Neural Network. , 44, pp.1-4.

[43] Meidan, Yair, Michael Bohadana, Asaf Shabtai, Juan David Guarnizo, Martin Ochoa, Nils Ole Tippenhauer, and Yuval Elovici. "ProfilIoT: a machine learning approach for IoT device identification based on network traffic analysis." In Proceedings of the symposium on applied computing, pp. 506-509. ACM, 2017.

[44] J. Wu, S. Luo, S. Wang and H. Wang, "NLES: A Novel Lifetime Extension Scheme for Safety-Critical Cyber-Physical Systems Using SDN and NFV," IEEE Internet of Things Journal, no. 6, no. 2, pp. 2463-2475, 2018.

[45] J. Wu, M. Dong, K. Ota, J. Li, Z. Guan, Big Data Analysis based Security Cluster Management for Optimized Control Plane in Software-Defined Networks, IEEE Transactions on Network and Service Management, vol. 15, no. 1, pp. 27-38, 2018.

[46] J. Wu, M. Dong, K. Ota, J. Li, Z. Guan, FCSS: Fog-Computing-based Content-Aware Filtering for Security Services in Information-Centric So-

cial Networks, IEEE Transactions on Emerging Topics in Computing, DOI: 10.1109/TETC.2017.2747158, pp. 1-12, 2017.

[47] G. Li, J. Wu, J. Li, K. Wang, T. Ye, Service Popularity-based Smart Resources Partitioning for Fog Computing-enabled Industrial Internet of Things, IEEE Transactions on Industrial Informatics, vol. 14, no. 10, pp. 4702-4711, Oct. 2018.

[48] J. Wu, M. Dong, K. Ota, C. Li, A Hierarchical Security Framework for Defending against Sophisticated Attacks on Wireless Sensor Networks in Smart Cities, IEEE Access, vol.4, pp. 416-424, Jan. 2016.

[49] Li, He, Kaoru Ota, and Mianxiong Dong. "Learning IoT in edge: deep learning for the internet of things with edge computing." IEEE Network 32, no. 1 (2018): 96-101.

.

APPENDIX A DATA SPLIT BASED ON DT

Dataset	**DT**						**RF**						**LM**					
	Confusion matrix					**Error**	**Confusion matrix**					Error	**Confusion matrix**					**Error**
Original	150	8	1	1	0	6.2	159	1	0	0	0	0.6	152	5	2	0	1	5
	24	13	7	10	0	75.9	10	39	4	1	0	27.8	23	20	6	5	0	63
	8	5	15	7	0	57.1	5	1	28	1	0	20	8	8	11	6	2	68.6
	4	9	5	17	0	51.4	5	4	2	24	0	31.4	3	10	11	11	0	68.6
	2	2	3	6	0	100	1	0	1	2	9	30.8	2	3	1	5	2	84.6
1	63	0	0	0	0	0	63	0	0	0	0	0	61	0	1	0	1	3.2
	3	0	0	0	0	100	2	1	0	0	0	66.7	2	1	0	0	0	66.7
	2	0	0	0	0	100	0	0	2	0	0	0	0	0	2	0	0	0
	4	0	0	0	0	100	1	0	0	3	0	25	0	1	0	3	0	25
	2	0	0	0	0	100	0	0	0	0	2	0	0	0	0	0	2	0
2	26	0	0	0	0	0	26	0	0	0	0	0	22	2	2	0	0	15.4
	5	0	0	0	0	100	3	2	0	0	0	60	3	2	0	0	0	60
	2	0	0	0	0	100	0	0	2	0	0	0	0	0	2	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	3	0	0	0	100	2	1	0	0	0	33.3	3	0	0	0	0	0
	0	6	0	0	0	0	0	6	0	0	0	0	0	6	0	0	0	0
	0	4	0	0	0	100	0	2	2	0	0	50	0	1	3	0	0	25
	0	5	0	0	0	100	0	2	0	3	0	40	0	1	0	4	0	20
	0	2	0	0	0	100	0	0	0	0	2	0	0	0	0	0	2	0
4	16	0	0	0	0	0	16	0	0	0	0	0	13	1	2	0	0	18.8
	6	0	0	0	0	100	1	5	0	0	0	16.7	1	5	0	0	0	16.7
	4	0	0	0	0	100	1	0	3	0	0	25	1	0	3	0	0	25
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	5	0	0	0	100	2	3	0	0	0	60	2	2	0	1	0	60
	0	7	0	0	0	0	0	7	0	0	0	0	0	6	0	1	0	14.3
	0	1	0	0	0	100	0	0	1	0	0	0	0	0	1	0	0	0
	0	4	0	0	0	100	0	2	0	2	0	50	0	0	2	2	0	50
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	10	0	0	0	0	0	10	0	0	0	0	0	8	0	2	0	0	20
	2	0	0	0	0	100	1	1	0	0	0	50	1	1	0	0	0	50
	1	0	0	0	0	100	0	0	1	0	0	0	0	0	1	0	0	0
	3	0	0	0	0	100	1	0	0	2	0	33.3	1	0	0	2	0	33.3
	2	0	0	0	0	100	1	0	0	0	1	50	0	0	1	0	1	50
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	7	0	0	100	0	4	3	0	0	42.9	0	4	3	0	0	42.1
	0	0	15	0	0	0	0	0	15	0	0	0	0	1	13	0	1	13.3
	0	0	5	0	0	100	0	0	2	2	1	60	0	0	2	3	0	40
	0	0	3	0	0	100	0	0	0	0	3	0	0	0	0	0	3	0
8	0	1	0	0	0	100	1	0	0	0	0	0	1	0	0	0	0	0
	0	6	0	3	1	40	0	6	0	3	1	40	1	4	0	3	2	60
	0	1	0	3	3	100	0	2	5	0	0	28.6	0	0	6	1	0	14.3
	0	5	0	11	1	35.3	0	1	0	14	2	17.6	0	4	0	12	1	29.4
	0	0	0	2	4	33.3	0	0	0	0	6	0	0	0	0	0	6	0

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2923707, IEEE Access

APPENDIX B FEATURE EXTRACTION FROM LM

	DT						RF						LM					
data	**Confusion matrix**					**Error**	**Confusion matrix**					Error	**Confusion matrix**					**Error**
Original	150	8	1	1	0	6.2	158	2	0	0	0	1.2	152	5	2	0	1	5
	24	13	7	10	0	75.9	11	40	2	1	0	25.9	22	21	6	5	0	61.1
	8	5	15	7	0	57.1	5	1	29	0	0	17.1	8	8	12	5	2	65.7
	4	9	5	17	0	51.4	5	5	2	23	0	34.3	3	10	11	11	0	68.6
	2	2	3	6	0	100	0	1	2	1	9	30.8	2	3	1	5	2	84.6
1	98	0	0	0	0	0	98	0	0	0	0	0	96	1	0	1	0	2
	11	0	0	0	0	100	5	6	0	0	0	45.5	8	3	0	0	0	72.7
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	100	0	0	0	1	0	0	0	0	0	1	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	26	0	0	0	0	0	26	0	0	0	0	0	23	1	2	0	0	11.5
	5	0	0	0	0	100	3	2	0	0	0	60	1	3	1	0	0	40
	2	0	0	0	0	100	0	0	2	0	0	0	0	0	2	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	3	0	0	0	100	2	1	0	0	0	33.3	2	0	1	0	0	33.3
	0	6	0	0	0	0	0	6	0	0	0	0	1	5	0	0	0	16.7
	0	4	0	0	0	100	0	2	2	0	0	50	0	1	3	0	0	25
	0	5	0	0	0	100	0	2	0	3	0	40	0	1	0	4	0	20
	0	2	0	0	0	100	0	0	0	0	2	0	0	0	0	2	0	0
4	16	0	0	0	0	0	15	0	1	0	0	6.2	13	3	0	0	0	18.8
	6	0	0	0	0	100	1	5	0	0	0	16.7	1	5	0	0	0	16.7
	4	0	0	0	0	100	1	0	3	0	0	25	1	0	3	0	0	25
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	5	0	0	0	100	2	2	0	1	0	60	2	2	0	1	0	60
	0	7	0	0	0	0	0	7	0	0	0	0	0	6	1	0	0	14.3
	0	1	0	0	0	100	0	0	1	0	0	0	0	0	1	0	0	0
	0	4	0	0	0	100	0	2	0	2	0	50	0	0	1	3	0	25
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	10	0	0	0	0	0	8	1	0	1	0	20	8	0	2	0	0	20
	2	0	0	0	0	100	1	1	0	0	0	50	1	1	0	0	0	50
	1	0	0	0	0	100	0	0	1	0	0	0	0	0	1	0	0	0
	3	0	0	0	0	100	1	0	0	2	0	33.3	0	0	0	2	1	33.3
	2	0	0	0	0	100	1	0	0	0	1	50	0	0	1	0	1	50
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	7	0	0	100	0	4	3	0	0	42.9	0	4	3	0	0	42.9
	0	0	15	0	0	0	0	0	15	0	0	0	0	0	14	0	1	6.7
	0	0	5	0	0	100	0	0	3	2	0	60	0	0	3	2	0	60
	0	0	3	0	0	100	0	0	0	0	3	0	0	0	0	0	3	0
8	0	1	0	0	0	100	1	0	0	0	0	0	1	0	0	0	0	0
	0	6	0	3	1	40	0	6	0	3	1	40	0	6	0	3	1	40
	0	1	0	3	3	100	0	1	5	1	0	28.6	0	1	5	1	0	28.6
	0	5	0	11	1	35.3	0	1	1	14	1	17.6	0	1	1	13	2	23.5
	0	0	0	2	4	33.3	0	0	0	0	6	0	0	0	0	0	6	0

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2923707, IEEE Access

APPENDIX C HYBRID MODEL WITH HRFLM

Data	Confusion matrix					Error	MODEL
1	98	0	0	0	0	0	RF
	5	6	0	0	0	45.5	
	0	0	0	0	0	0	
	0	0	0	1	0	0	
	0	0	0	0	0	0	
2	26	0	0	0	0	0	RF
	3	2	0	0	0	60	
	0	0	2	0	0	0	
	0	0	0	0	0	0	
	0	0	0	0	0	0	
3	3	0	0	0	0	0	LM
	0	6	0	0	0	0	
	0	1	3	0	0	25	
	0	1	0	4	0	20	
	0	0	0	0	2	0	
4	16	0	0	0	0	0	RF
	1	5	0	0	0	16.7	
	1	0	3	0	0	25	
	0	0	0	0	0	0	
	0	0	0	0	0	0	
5	2	3	0	0	0	60	RF
	0	7	0	0	0	0	
	0	0	1	0	0	0	
	0	2	0	2	0	50	
	0	0	0	0	0	0	
6	8	1	0	1	0	20	RF
	1	1	0	0	0	50	
	0	0	1	0	0	0	
	1	0	0	2	0	33.3	
	1	0	0	0	1	50	
7	0	0	0	0	0	0	RF
	0	4	3	0	0	42.9	
	0	0	15	0	0	0	
	0	0	3	2	0	60	
	0	0	0	0	3	0	
8	1	0	0	0	0	0	RF
	0	7	0	1	2	30	
	0	1	5	1	0	28.6	
	0	1	0	14	2	17.6	
	0	0	0	0	6	0	

ACKNOWLEDGMENT

The authors would like to thank IEEE Access journal and their respective Universities for their support.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2923707, IEEE Access

AUTHORS

DR.SENTHILKUMAR MOHAN is working as an Assistant professor(senior) in School of Information Technology and Engineering at Vellore Institute of Technology, Vellore. He worked as a Project Associate in IIT Madras (2009-2010). He received Ph.D. in VIT University in 2017. He completed his master's degree M.Tech -IT networking (2013) and M.S Software Enginnering (2007). He has 10 years of experience in teaching and research. His areas of interest include Big Data and Machine / Deep learning.

PROF.CHANDRASEGAR THIRUMALAI currently pursuing his Ph.D at VIT University, India. About his education, completed his Master of Technology in Computer Science and Engineering at Pondicherry Central University and Bachelor of Engineering in Computer Science and Engineering at Dr.Pauls Engineering College affiliated to Anna University, India. His area of specialization includes Linear Cryptanalysis, Public Key Cryptosystems, Fuzzy Expert Systems, Automata and Networking. Published several International journals and International conferences

DR. GAUTAM SRIVASTAVA was awarded his B.Sc. degree from Briar Cliff University in U.S.A. in the year 2004, followed by his M.Sc. and Ph.D. degrees from the University of Victoria in Victoria, British Columbia, Canada in the years 2006 and 2012, respectively. He then taught for 3 years at the University of Victoria in the Department of Computer Science, where he was regarded as one of the top undergraduate professors in the Computer Science Course Instruction at the University. From there in the year 2014, he joined a tenure-track position at Brandon University in Brandon, Manitoba, Canada, where he currently is active in various professional and scholarly activities. He was promoted to the rank Associate Professor in January 2018. Dr. G, as he is popularly known, is active in research in the field of Data Mining and Big Data. In his 8-year academic career, he has published a total of 50 papers in high-impact conferences in many countries and in high-status journals (SCI, SCIE) and has also delivered invited guest lectures on Big Data, Cloud Computing, Internet of Things, and Cryptography at many Taiwanese and Czech universities. He is an Editor of several international scientific research journals. He currently has active research projects with other academics in Taiwan, Singapore, Canada, Czech Republic, Poland and U.S.A. He is constantly looking for collaboration opportunities with foreign professors and students. Assoc. Prof. Gautam Srivastava received *Best Oral Presenter Award* in FSDM 2017 which was held at the National Dong Hwa University (NDHU) in Shoufeng (Hualien County) in Taiwan (Republic of China) on November 24-27, 2017.