IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 19, 2019, accepted December 20, 2019, date of publication December 27, 2019,
date of current version January 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962755

MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis

ANKUR GUPTA[1,*], (Student Member, IEEE), RAHUL KUMAR[1,*], HARKIRAT SINGH ARORA[2], AND BALASUBRAMANIAN RAMAN[1], (Member, IEEE)

[1]Machine Vision Laboratory, Department of Computer Science and Engineering, IIT Roorkee, Roorkee 247667, India
[2]Department of Chemical Engineering, IIT Roorkee, Roorkee 247667, India

Corresponding authors: Ankur Gupta (agupta4@cs.iitr.ac.in) and Rahul Kumar (rkumar9@cs.iitr.ac.in)

*Ankur Gupta and Rahul Kumar contributed equally to this work.

ABSTRACT Cardiovascular disease tops the list among all major causes of deaths worldwide. Though, prognostication and in-time diagnosis can help in reducing the mortality rate as well as increases the survival rate of patients. Unavailability or scarcity of radiologists and doctors in different countries due to several reasons is a significant factor for hindrance in early diagnosis. Among various efforts of developing the decision support systems, computational intelligence is an emerging trend in the field of medical imaging to detect, prognosticate and diagnose the disease. It helps radiologists and doctors to get relief from being over-burdened and minimizes the induced delays for in-time diagnosis of patients. In this work, a machine intelligence framework for heart disease diagnosis *MIFH* has been proposed. *MIFH* utilizes the factor analysis of mixed data (FAMD) to extract as well as derive features from the UCI heart disease Cleveland dataset and train the machine learning predictive models. The framework *MIFH* is validated using the holdout validation scheme. Experimentation results show that *MIFH* performed well over several baseline methods of recent times in terms of accuracy and comparable in terms of sensitivity and specificity. *MIFH* returns best possible solution among all input predictive models considering performance criteria and improves the efficacy of the system, hence can assist doctors and radiologists in a better way to diagnose heart patients.

INDEX TERMS Cleveland, UCI repository, FAMD, random forest, feature selection, cardiovascular.

I. INTRODUCTION

According to World Health Organization (WHO), cardiovascular disease (CVD) is one of the lethal diseases leads to the most number of deaths worldwide and prevalent in United States [1], [2]. CVD is a condition when heart is not functioning properly to pump the required amount of blood to other parts of the body that causes heart failure [3]. Blockage of coronary arteries is the prime reason for heart failure. The early signs of CVD is irregular heartbeat, chest pain or discomfort, shortness of breath, swollen feet or ankles, fatigue and fainting. Early detection and prognosis is a way to increase the life-span of a patient. A serious issue which is bottleneck in this regard is the lack of resources and unavailability of doctors and radiologists in developing or low-income countries in a fair proportion of the population which leads to diagnosis in advance stage of disease. This is one of the prime causes that the survival rate of approximately 50% of CVD patients is $1-2$ years [4]. The risk factors that lead to CVD are patient's medical background, age, sex, life-style *etc.* Alteration in life-style such as physical activity and non-smoking can reduce the risk factors by controlling the cholesterol level and blood pressure. These risk factors are under observation of medical expert. Analysis of early signs, alteration in life-style and medical examination report by medical experts help in diagnosing the disease. Though, the expertise used to examine patient records differs depending on the level of knowledge and due to human errors, therefore precision and prognostication cannot be guaranteed [5], [6].

Angiography is one of the promising methods to diagnosticate the severity of CVD. Though, side effects of angiography and high level of expertise required are one of the prime reasons for researchers inclination towards an automated solution that can help in simplifying the diagnosis process. Researchers and academicians are looking forward to develop such automated machine intelligent expert systems during last few decades to reduce the associated risk of medical examination. The decision support systems can exploit the machine learning (ML) predictive models or their ensembled versions such as logistic regression (*LR*),

The associate editor coordinating the review of this manuscript and approving it for publication was Sara Dadras.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

k-nearest neighbour (*kNN*), support vector machine (*SVM*), naive bayes (*NB*), adaboost (*AB*), K-means clustering, linear regression, decision trees (*DT*s) and random forest (*RF*) *etc.*

Main Contribution: The research contribution in the proposed work includes designing of a machine intelligence framework based on predictive models for the diagnosis of CVD. To validate the efficacy of the proposed framework hold-out validation scheme is applied on the publicly available Cleveland heart disease dataset available on University of California Irvine (UCI) repository. Since the Cleveland dataset is mixed type (*i.e.*, comprises of both numeric and categorical features), the factor analysis of mixed data (FAMD) mechanism is used to extract or derive the features from the dataset. The proposed design exploits the machine learning models *LR*, *kNN*, *SVM*, *DT* and *RF* to classify the subjects either into the normal ones or heart patients. In order to evaluate the performance of the proposed framework, the performance metric $\langle Acc, Sens, Spec, Score, MCC, AUC\rangle$ is calculated, where *Acc*, *Sens*, *Spec*, *Score*, *MCC* and *AUC* represent the accuracy, sensitivity, specificity, F1-score, Matthew's correlation coefficient and area under the curve, respectively. The performance metric is explained in detail in Section VII-B. The major research contributions can be listed as:

- The performance of well-known classifiers *LR*, *kNN*, *SVM*, *DT* and *RF* is measured on extracted as well as derived features from Cleveland heart disease dataset. The performance is measured by tuning the hyper-parameters and evaluating the metric $\langle Acc, Sens, Spec, Score, MCC, AUC\rangle$.
- The classifiers' performance is measured on the extracted as well as derived features' range $2-28$ with hold-out validation scheme.
- The proposed framework, *i.e. MIFH*, eventually, returns the best combination of feature set and classifier along with its tuned hyper-parameters for accurate classification of the subjects in the Cleveland dataset.

The rest of the paper is organized as follows. Section II represents the state-of-the-art techniques and research contribution towards developing the decision support system to classify the subjects into normal one or heart patient considering the Cleveland heart disease dataset. A motivation and the objective for the proposed work is formulated in Section III. The UCI heart disease Cleveland dataset along with its features and their associated datatype are explained in Section IV. The dataset pre-processing steps are explained in Section V. The proposed framework, *MIFH*, is presented in Section VI. The experimental results and analysis are discussed in detail in Section VII. Finally, the work is concluded along with future scope in Section VIII.

II. STATE-OF-THE-ART

The state-of-the-art or more accurately say baseline methods which came into existence in recent times are presented to review the various relevant, impactful and effective research contributions based on several machine learning and fuzzy logic based classification methods for heart disease diagnosis. The existing traditional invasive methods used to diagnose heart disease are based upon medical history of a patient as well as family genealogical history, physical examination report which includes, but not limited to these elements only, high blood cholesterol, high blood pressure, obesity, smoking, and medical experts assessment for the symptoms involved. Most of these methods cause inaccurate diagnosis because of human intervention and mistakes and often delayed in the diagnosis outcomes. The human-oriented process also incurs high cost and is complicated in terms of computation and requires significant time for assessment [7]. In order to overcome the effect of these traditional factors, a non-invasive medical decision-making diagnostic support system based on predictive models of machine learning such as *NB*, *LR*, *kNN*, *SVM*, *DT*s, *RF*, fuzzy logic (*FL*), rough sets (*RS*) and many more have been developed and used by academicians and researchers from industry and academia and is being commonly used for the diagnosis of heart disease. With the support of these expert medical decision-making support systems based on machine learning, the cardiovascular mortality ratio has been reduced [8]–[22]. In literature, the heart disease diagnosis through machine-learning predictive models is widely used, and significant performance metrics have been reported over UCI heart disease Cleveland dataset.

Long *et al.* [14] proposed a diagnostic system for heart disease which is based on attribute reduction of rough sets using chaos firefly algorithm and interval type-2 fuzzy logic by re-defining the dimensions and performed validation over the heart disease dataset. Krishnaiah *et al.* [15] have used the minimum distance-based fuzzy-kNN classifier to diagnose heart disease and the reported classification accuracy is 91%. Iftikhar *et al.* [17] used SVM and particle swarm optimization (PSO) to create an analytical model for health care. The proposed approach is used to define cardiovascular risk factors. The approach is validated on UCI heart disease Cleveland dataset to measure the efficiency of proposed analytical model for health care. Vijayashree and Sultana [19] proposed heart disease classification method which collaboratively utilizes the PSO with SVM and reported the classification accuracy 84.36%. Ismaeel *et al.* [24] used extreme learning-based algorithms to diagnose heart disease and by validating their method on the Cleveland dataset, which gives precision of 80%. Purushottam *et al.* [25] proposed a rule-based classifier for heart disease prediction and achieved an accuracy of 86.7%. Esfahani and Ghazanfari [26] have applied ensembling of predictive models for detection of CVD and validated the performance over heart disease dataset and reported the classification accuracy 89%. Shah *et al.* [27] created a novel selection strategy based on the probabilistic principal component analysis (PPCA) of probabilistic electronic medical record. PPCA's essential function is used to obtain the most significant predictive characteristics for heart disease prediction. Tomar and Agarwal [28] suggested an early diagnosis heart disease model based on the least square twin support vector machine (LSTSVM) which is formulated in [29].

The proposed model is used to utilize the statistics from F1-score to identify weight of each feature. In order to predict the heart disease, rule-based fuzzy logic (RBFL) and opposition firefly with BAT method is used by Reddy and Khare [30]. The hybrid technique for diagnosing heart disease, namely OFBAT-RBFL, is also suggested based on rule-based fuzzy logic (RBFL) and oppositional BAT firefly (OFBAT) method. Furthermore, Arabasadi *et al.* [31] suggested a neural network based hybrid neural network with genetic algorithm method to classify heart disease datasets effectively.

III. MOTIVATION AND PROBLEM FORMULATION

According to WHO, non-communicable diseases top the list for being the prime reason of deaths worldwide. Approximately 9.6 million deaths encountered by Ischemic Heart Disease (IHD) only [2]. The early detection and prognosis of disease help in increasing the survival rate and increases the life-span of the patient. The availability of doctors and radiologists is uneven worldwide depending upon the economical structure and GDP of the country [32]. This leads to the motivation for researchers and academicians to come-up with a framework or decision support system that can help in predicting the disease in an early stage. The objective of developing such system is to extend the reach of healthcare in low-income countries at affordable rate.

The Cleveland heart disease dataset for coronary heart disease (CHD) is publicly available in UCI repository and used in the proposed work for developing the machine intelligent design. The Cleveland dataset $\mathcal{D}$, along with output class $\mathcal{C}$, comprises of feature set $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_m\}$ with m features and instances $\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_n\}$ corresponding to n subjects.

Definition 1: A dataset $\mathcal{D}$ is composition of feature set $\mathcal{F} = \{\mathcal{F}_i \mid 1 \leq i \leq m\}$, where m is the number of features and instances $\mathcal{I} = \{\mathcal{I}_j \mid 1 \leq j \leq n\}$, where n is the total number of subjects.

Definition 2: An instance $\mathcal{I}_j$ is represented by feature values $\mathcal{F}_i$, such as $\mathcal{I}_j = \{\mathcal{F}_i \mid 1 \leq i < m\}$ and m is the number of features in $\mathcal{D}$. The value $\mathcal{F}_i$ is either numeric or categorical.

The goal behind designing a machine intelligent framework for accurate predictions ($\mathcal{P}$) using a learning algorithm ($\mathcal{L}$) is to make the predictive model learn to fit by analyzing the behaviour of data and converges by reducing the error (**E**) present in all instances ($\mathcal{I}$) collectively as depicted in Equation 1.

$$\mathbf{P} \xrightarrow{\mathcal{C}} min\left[\mathbf{E}\left\{\mathcal{C} - \mathcal{P}\left[\mathcal{L}\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\mathcal{D}(\mathcal{F}_i, \mathcal{I}_j)\right)\right]\right\}\right] \quad (1)$$

The predictions $\mathcal{P}$ from $\mathcal{L}$ is evaluated on the basis of performance (**P**) which is measured using metrics $\langle Acc, Sens, Spec, Score, MCC, AUC\rangle$. The metrics is explained in detail in Section VII-B.

IV. DATASET SPECIFICATIONS

Most of the existing research considers publicly available University of California (UCI) heart disease Cleveland dataset as benchmark for CHD prediction [23]. The Cleveland dataset contains 76 features but only a maximum of 14 features are popularly used for research purposes. These 14 features along with their data types and values are: age (in years; numeric), sex (male/female = 1/0; binary), chest pain type (typical angina/atypical angina/non-anginal pain/asymptomatic = 1/2/3/4; nominal), trestbps (resting blood pressure in mm Hg on admission to the hospital; numeric), chol (serum cholesterol in mg/dl; numeric), fbs (fasting blood sugar > 120 mg/dl *i.e.*, 1 = *true*, 0 = *false*; binary), restecg (resting electrocardiographic results, 0 = normal, 1 = ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria; nominal), thalach (maximum heart rate achieved; numeric), exang (exercise induced angina, 1 = *yes*, 0 = *no*; binary), oldpeak (ST depression induced by exercise relative to rest; numeric), slope (the slope of the peak exercise ST segment, upsloping/flat/downsloping = 1/2/3; nominal), ca (number of major vessels (0 − 3) colored by flourosopy; numeric), thal (normal/fixed defect/reversible defect = 3/6/7; nominal), num (diagnosis of heart disease, angiographic/disease = 0/1, 2, 3, 4; binary). The pictorial representation of the used attributes of UCI Cleveland CHD dataset is represented in Table 1. The range of values for all 14 features of Cleveland dataset along with the unique values, mean and median of each feature is depicted in Table 2.

The features of Cleveland CHD dataset contains subjects medical background and clinical features as well. Apparently, the features are not independent. The correlation between the features of the Cleveland CHD dataset is depicted in the feature heatmap in Figure 1 which is derived using Pearson correlation coefficient. The heatmap scale represents the degree of correlation between features, where −0.40 and 1.00 represent the negatively correlated (*red* in color) and positively correlated (*green* in color) features, respectively.

V. DATASET PRE-PROCESSING

The performance of machine learning systems heavily depends upon the form of data. The statistics shows that well-organized data generates impactful results. Several data pre-processing methods exist to pre-process the data or normalize in case of real-time data. The following methods have been used in our proposed work; one hot encoding for categorical features, data standardization using z-score normalization to normalize the features and data stratification to divide the dataset into training and validation sets to eliminate the unbalancing effect of disease classes.

A. DATA IMPUTATION

In real-life scenario, it is difficult or almost impossible in some situations to collect the complete information from the

TABLE 1. Used features description of Cleveland dataset from UCI heart disease repository [23].

Feat. #	Feature	Feature Description	Feature Type	Feature Data Range	Missing Values	Data Imputation
1	age	Number of years	Numeric	[29, 77]	No	—
2	sex	Sex	Binary	0 = female 1 = male	No	—
3	cp	Chest pain	Nominal	1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 = asymptomatic	No	—
4	trestbps	Resting blood pressure in mm Hg on admission to the hospital	Numeric	[94, 200]	No	—
5	chol	Serum cholesterol in mg/dl	Numeric	[126, 564]	No	—
6	fbs	Fasting blood sugar > 120 mg/dl	Binary	0 = false 1 = true	No	—
7	restecg	Resting electrocardiographic results	Nominal	0 = normal 1 = ST-T wave abnormality 2 = left ventricular hypertrophy	No	—
8	thalach	Maximum heart rate achieved	Numeric	[71, 202]	No	—
9	exang	Exercise induced angina	Binary	0 = no 1 = yes	No	—
10	oldpeak	ST depression induced by exercise relative to rest	Numeric	[0, 6.2]	No	—
11	slope	The slope of the peak exercise ST segment	Nominal	1 = upsloping 2 = flat 3 = downsloping	No	—
12	ca	Number of major vessels colored by flourosopy	Nominal	0 – 3	Yes	Majority
13	thal	Defect type	Nominal	3 = normal 6 = fixed defect 7 = reversible defect	Yes	Majority
14	num	Diagnosis of heart disease	Binary	0 = less than 50% diameter narrowing (Normal) 1 = greater than 50% diameter narrowing (Patient)	No	—

TABLE 2. Statistical overview of features of Cleveland dataset.

Statistical Indices	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Num
Range	[29, 77]	[0, 1]	[1, 4]	[94, 200]	[126, 564]	[0, 1]	[0, 2]	[71, 202]	[0, 1]	[0, 6.2]	[1, 3]	[0,3]	[3,7]	[0, 1]
Unique Values	41	2	4	50	152	2	3	91	2	40	3	4	3	2
Mean	54.4	0.68	3.16	131.69	246.69	0.15	0.69	149.61	0.33	1.04	1.6	0.66	—	0.33
Median	56	1	3	130	241	0	1	153	0	0.8	2	0	—	0

subject such as interruptions in the data flow, privacy concerns, inability of the patient to co-operate *etc*. The Cleveland dataset has missing information of the features as well. To make the dataset complete and reasonable for processing, data imputation is done to fill the missing values of the features with the new labels. Since the missing values in the imputed Cleveland dataset are filled, the number of instances are same as in the original Cleveland dataset. The numerical attributes of Cleveland dataset are *age*, *tresbps*, *chol*, *thalach* and *oldpeak*. The categorical attributes are *sex*, *cp*, *fbs*, *restecg*, *exang*, *slope*, *ca* and *thal* along with one categorical output class *num*. The Cleveland dataset has two categorical attributes, namely *ca* and *thal* with missing values which is imputed with *majority label* as depicted in Table 1. The research inclination shows data impuation is more effective than removing instances from dataset. The feature *ca* and *thal* is missing for 4 and 2 instances, respectively, and has value 0 as majority label in 176 out of 299 and value 3 as majority label in 166 out of 301 instances, respectively, therefore the missing instances of *ca* and *thal* are represented

TABLE 3. Imputed Cleveland dataset and its attributes.

Dataset	# of instances	# of normal patients	# of heart subjects	# of train samples§	# of validation samples$	# of used attributes	# of attributes after formalization
Cleveland	303	164	139	(131, 111)	(33, 28)	14	29

§ : $(\alpha, \beta) :=$ (# of normal patients, # of heart subjects) in train samples

$: $(\gamma, \delta) :=$ (# of normal patients, # of heart subjects) in validation samples

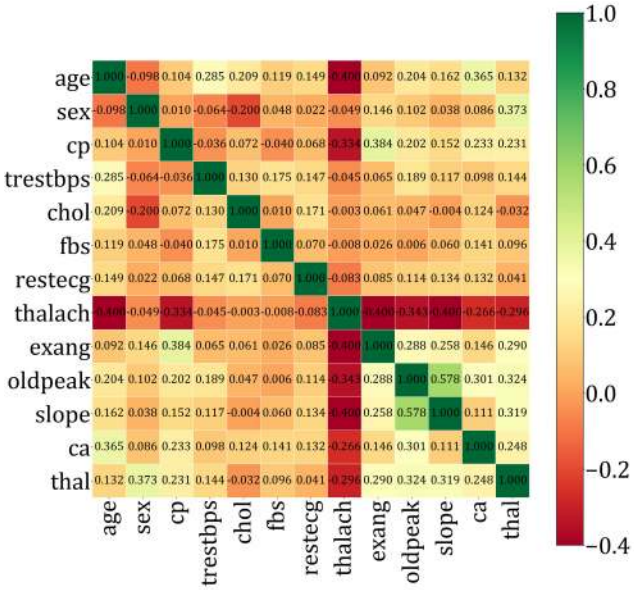


FIGURE 1. Heatmap of features of the Cleveland CHD dataset derived using Pearson correlation coefficient. The red and green color in heatmap scale is used to represent the negatively correlated and positively correlated features, respectively.

with 0 for those 4 and 2 instances, respectively, well. The statistics for total number of instances (*i.e.*, normal subject and heart patient count), normal subjects, heart patients and the number of used attributes for imputed Cleveland dataset is represented pictorially in Table 3.

B. DATA STANDARDIZATION

The medical data is mostly discrete, therefore data standardization is essential to converge the characteristics of data. Z-score normalization is one of the popular method for data standardization which exploits mean and standard deviation of the attribute to normalize the data of the attribute. *Data standardization* is a transformation process of diverse data into a normalized and consistent form. Considering μ_i and σ_i as the mean and standard deviation of the i^{th} attribute $\mathcal{F}_i$ of $\mathcal{D}$, then the z-score ($\mathcal{Z}_{ij}$) is calculated for j^{th} instance $\mathcal{I}_j$ as depicted in Equation 2.

$$\mathcal{Z}_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \tag{2}$$

where μ_i is calculated for attribute i as given in Equation 3.

$$\mu_i = \frac{1}{n} \sum_{\delta=1}^{n} x_\delta \tag{3}$$

C. DATA STRATIFICATION

Data stratification is a technique to divide the dataset into smaller ones with well-defined strata based upon predetermined set of criteria. The Cleveland dataset is divided into training and validation sets. The proposed framework, *MIFH*, is trained and validated the learning of model using hold-out validation scheme considering the train and validation ratio of 80 : 20 for imputed Cleveland dataset. The division of dataset using hold-out validation scheme leads to 242 and 61 training and validation instances. The heart subjects and normal patients are 111 and 131, and 28 and 33 for training and validation samples, respectively, as depicted in Table 3.

VI. MIFH: MACHINE INTELLIGENCE FRAMEWORK FOR HEART DISEASE DIAGNOSIS

The objective of developing a machine intelligence framework for heart disease diagnosis is to potentialize the system in predicting the heart disease in order to increase the survival rate of patients by the accurate, precise and early detection of disease. The intended goal leads to provide an automated solution to assist doctors and radiologists in prognostication and making decisions with precision and higher confidence along with saving the analysis time. The nature of medical data is of high variance, therefore dataset pre-processing is a viable step. The framework comprises of extraction of features in association with analyzing the impact of machine learning predictive models. Extraction of features includes selection and reduction of features, and derivation of new features as well. The aim is to design a machine intelligence framework for heart disease detection which focuses on predicting labels of UCI Cleveland heart disease dataset. Since Cleveland dataset is of mixed type and contains both numeric and categorical features, factor analysis of mixed data (FAMD) is suitable for feature extraction. The extracted as well as derived features are used to train the models for classification of normal subjects and heart patients [14], [33]–[39]. To facilitate understanding of the proposed framework *MIFH*, the schematic illustration of workflow is explained in Figure 2. The pseudo-code of *MIFH* is shown as **Algorithm 1**.

The framework *MIFH* has number of steps which includes Data Imputation and Partitioning, Feature Extraction using FAMD, Features Normalization, Machine Learning Approach and Performance Metric Evaluation.

Algorithm 1 *MIFH*($\mathcal{D}$)

begin
Performance metric $\mathbf{P} = \{\}$;
Assign weight vector w for performance metric $\mathbf{P}$;
$\mathcal{D}^I \longleftarrow Data\_Imputation(\mathcal{D})$;
$(\mathcal{D}^T, \mathcal{D}^V) \longleftarrow Dataset\_Stratification\_HoldOut(\mathcal{D}^I, val\_ratio)$;
for *feature_count* : $fc \leftarrow 2$ *to* 28 **do**
  $\mathcal{F}^{fc} \longleftarrow FAMD(\mathcal{D}^T)$;
end
$\mathcal{F}' \longleftarrow \langle \mathcal{F}^2, \mathcal{F}^3, \cdots, \mathcal{F}^{28} \rangle$;
$\mathcal{D}^T \longleftarrow \langle \mathcal{D}^2, \mathcal{D}^3, \cdots, \mathcal{D}^{28} \rangle$;
$\mathcal{D}^T \longleftarrow Dataset\_Normalization(\mathcal{D}^T)$;
for *ML_approach* : $\langle LR, kNN, SVM, DT, \cdots, RF \rangle$ **do**
  $algo \longleftarrow ML\_approach(i)$;
  for *feature_count* : $fc \leftarrow 2$ *to* 28 **do**
    $\langle Acc, Sens, Spec, Score, MCC, AUC \rangle^{fc} \longleftarrow FAMD\_MLBox(\mathcal{D}^T, \mathcal{F}^{fc}, algo)$;
    $\mathbf{P}^{fc} \longleftarrow w(Acc) \times Acc + w(Sens) \times Sens + w(Spec) \times Spec + w(Score) \times Score + w(MCC) \times MCC + w(AUC) \times AUC$;
  end
  $\mathbf{P}^i \longleftarrow max(\mathbf{P}^2, \mathbf{P}^3, \cdots, \mathbf{P}^{28})$;
end
$algo \longleftarrow ML\_approach$ whose performance metric is highest according to w;
$(\mathbf{P}, method) \longleftarrow \langle \mathbf{P}^i, algo \rangle$;
$\mathcal{D}^V \longleftarrow Dataset\_Normalization(\mathcal{D}^V)$;
for *Selected feature set* $\mathcal{F}^{fc}$ *and ML approach method* **do**
  Validate the dataset $\mathcal{D}^V$ using $\mathcal{F}^{fc}$ in *method*;
  Evaluate performance metric $\mathbf{P} \longleftarrow \langle Acc, Sens, Spec, Score, MCC, AUC \rangle$;
end
return $(\mathbf{P}, \mathcal{F}^{fc}, method)$;
end

Algorithm 2 *Data_Imputation*($\mathcal{D}$)

begin
$\langle \mathcal{F}, \mathcal{I} \rangle \longleftarrow \mathcal{D}$;
$\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_m\}$;
$\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_n\}$;
foreach $\mathcal{F}_i \in \mathcal{F}$ **do**
  for $\mathcal{I}_j \in \mathcal{I}, where 1 \leq j \leq n$ **do**
    $\mathcal{I}_j$ have values missing in majority; Drop the column $\mathcal{F}_i$;
    $\mathcal{I}_j$ has missing values less than 40% and are categorical in nature; Fill the missing values with the majority label of $\mathcal{F}_i$;
    $\mathcal{I}_j$ has missing values less than 40% and are numeric in nature; Fill the missing values with the median value of $\mathcal{F}_i$;
  end
end
return $\mathcal{D}^I$;
end

1) DATA IMPUTATION AND PARTITIONING

The Cleveland CHD dataset $\mathcal{D}$ has missing values for features *ca* and *thal* which is filled by *majority label* and the pseudo-code to deal with missing values for data imputation is shown as **Algorithm 2**. Since, the validation of framework is imperative in terms of real-time performance because the new patient data would be completely unknown to the system, the imputed Cleveland dataset $\mathcal{D}^I$, returned from **Algorithm 2**, is partitioned into two sets, training and validation sets, using holdout validation scheme with holdout 0.2.

Algorithm 3 *Dataset_Stratification_HoldOut*($\mathcal{D}^I$, *val_ratio*)

begin
- $\langle \mathcal{F}, \mathcal{I} \rangle \longleftarrow \mathcal{D}^I$;
- Based on $\mathcal{F}^{\mathcal{C}}$, $\langle \mathcal{I}_H, \mathcal{I}_N \rangle \longleftarrow \mathcal{I}$;
- **foreach** *set in* $\langle \mathcal{I}_H, \mathcal{I}_N \rangle$ **do**
  - $|\mathcal{I}_H^V| \longleftarrow val\_ratio \times |\mathcal{I}_H|$; $|\mathcal{I}_H^T| \longleftarrow |\mathcal{I}_H| - |\mathcal{I}_H^V|$;
  - $|\mathcal{I}_N^V| \longleftarrow val\_ratio \times |\mathcal{I}_N|$; $|\mathcal{I}_N^T| \longleftarrow |\mathcal{I}_N| - |\mathcal{I}_N^V|$;
  - Select random $\lceil |\mathcal{I}_H^V| \rceil$ and $\lceil |\mathcal{I}_N^V| \rceil$ instances from $\mathcal{I}_H$ and $\mathcal{I}_N$, respectively;
  - Remaining instances are $\lfloor |\mathcal{I}_H^T| \rfloor$ and $\lfloor |\mathcal{I}_N^T| \rfloor$ of $\mathcal{I}_H$ and $\mathcal{I}_N$, respectively;
  - $\mathcal{D}^V \longleftarrow \langle \mathcal{I}_H^V, \mathcal{I}_N^V \rangle$;
  - $\mathcal{D}^T \longleftarrow \langle \mathcal{I}_H^T, \mathcal{I}_N^T \rangle$;
- **end**
- **return** partitioned output class balanced dataset ($\mathcal{D}^T$, $\mathcal{D}^V$);

end

Algorithm 4 *FAMD*($\mathcal{D}^T$)

begin
- $\langle \mathcal{F}, \mathcal{I} \rangle \longleftarrow \mathcal{D}^T$;
- $\langle \mathcal{F}^{ql}, \mathcal{F}^{qn} \rangle \longleftarrow \mathcal{F}$;
- **foreach** *quantitative variable q* **do**
  - Calculate $r(q, \ell)$, which is the correlation coefficient between variables q and ℓ, and $\ell \leftarrow 1, 2, \cdots, ql$;
  - Calculate $\eta^2(q, p)$, which is the squared correlation ratio between variables q and p, and $p \leftarrow 1, 2, \cdots, qn$;
- **end**
- Maximize $\mathcal{F}^{fc} \longleftarrow \sum_{ql} r^2(q, \ell) + \sum_{qn} \eta^2(q, p)$, where $\ell \leftarrow 1, 2, \cdots, ql$ and $p \leftarrow 1, 2, \cdots, qn$;
- **return** $\mathcal{F}^{fc}$

end

To prohibit class imbalanced partitioning, data stratification is done to ensure the equal proportion of normal subjects and heart patients in both the sets. The holdout along with data stratification is depicted in **Algorithm 3**. The number of distinct instances of heart subjects and normal patients in training set is 111 and 131 and in validation set is 28 and 33, respectively, as shown in Table 3.

2) FEATURE EXTRACTION USING FACTOR ANALYSIS OF MIXED DATA

Cleveland dataset $\mathcal{D}$ is of mixed type *i.e.*, it contains both numeric and categorical features. In statistics, numeric features are quantitative and categorical features are qualitative. Factor analysis of mixed data (FAMD) is a factorial method dedicated to both types of features and have working nature of principal components analysis (PCA) and multiple correspondence analysis (MCA) for corresponding categorical and numeric features, respectively. Mathematically, the numeric features are scaled down to unit variance and categorical features are transformed into a disjunctive representation of crisp coding and scaled using the selective format of MCA. FAMD also helps in visualizing the graphical representation of the objects, correlation between numeric to numeric and categorical features and association between features. The pseudo-code of FAMD is represented in **Algorithm 4**. The heatmap of 28 derived features using Pearson correlation coefficient after applying FAMD is shown in Figure 3. It is observed from the heatmap that FAMD is capable of deriving the features from the dataset $\mathcal{D}$ which are orthogonal. The information carried by negatively correlated features is distinct in nature and help in increasing the prediction capability of the model. The contribution of features $\mathcal{F}_i$, where $1 \leq i \leq 13$, of $\mathcal{D}$ in the principal dimensions are shown in Figure 4a and 4b. It can be observed that numeric features *oldpeak* and *thalach* has contributed most to the first principal dimension and categorical features *sex* and *age* contributed most to the second principal dimension. The correlation of numeric and categorical features with respect to the principal dimensions whose contribution is 16.92% and 8.44%, respectively, in the derived features of newly formed dataset along with their coordinates is shown in Figure 5. In a broader aspect, the contribution of numeric and categorical features with respect to explained variances in their correlation in terms of distinct features is depicted in Figure 6. The correlation circle represents the relationship between the categorical features in $\mathcal{D}$ and the correlation between features and dimensions as shown in Figure 7. It can be observed that categorical features *age* is most contributing features

Algorithm 5 *Dataset_Normalization*($\mathcal{D}^X$)

begin
1. $\langle \mathcal{F}, \mathcal{I} \rangle \longleftarrow \mathcal{D}^X$;
2. $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_m\}$;
3. $\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_n\}$;
4. An element in $\mathcal{D}^X$ is represented by x_{ij} where i and j represents the entry corresponding to j^{th} instance $\mathcal{I}_j$ of feature $\mathcal{F}_i$.;
5. **if** *dataset is training dataset* $\mathcal{D}^T$ **then**
6. **foreach** $\mathcal{F}_i \in \mathcal{F}$, *where* $1 \leq i \leq m$ **do**
7. Calculate mean $\mu_i \longleftarrow \frac{1}{n}\sum_{\delta=1}^{n} x_{i\delta}$;
8. Calculate standard deviation $\sigma_i \longleftarrow \sqrt{\frac{1}{n}\sum_{\delta=1}^{n}(x_{i\delta} - \mu_i)^2}$;
   end
   end
9. **else**
10. **if** *dataset is validation dataset* $\mathcal{D}^V$ **then**
11. Use μ_i and σ_i, calculated for $\mathcal{F}_i$ in $\mathcal{D}^T$, for normalizing the $\mathcal{F}_i$ in $\mathcal{D}^V$;
    end
    end
12. **foreach** $x_{ij} \in \mathcal{D}^X$, *where* $1 \leq i \leq m$ *and* $1 \leq j \leq n$ **do**
13. $x_{ij} \longleftarrow \frac{x_{ij} - \mu_i}{\delta_i}$;
    end
14. **return** updated dataset $\mathcal{D}^X$;

end

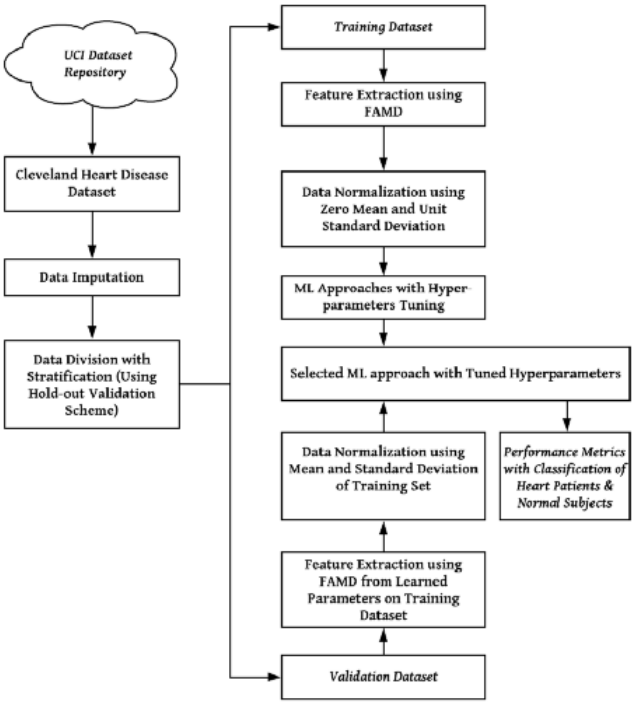


FIGURE 2. MIFH: A machine intelligence framework for Cleveland heart disease dataset.

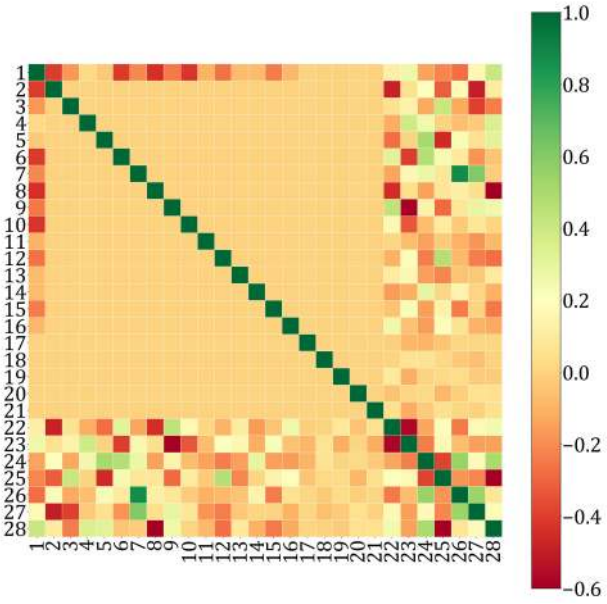
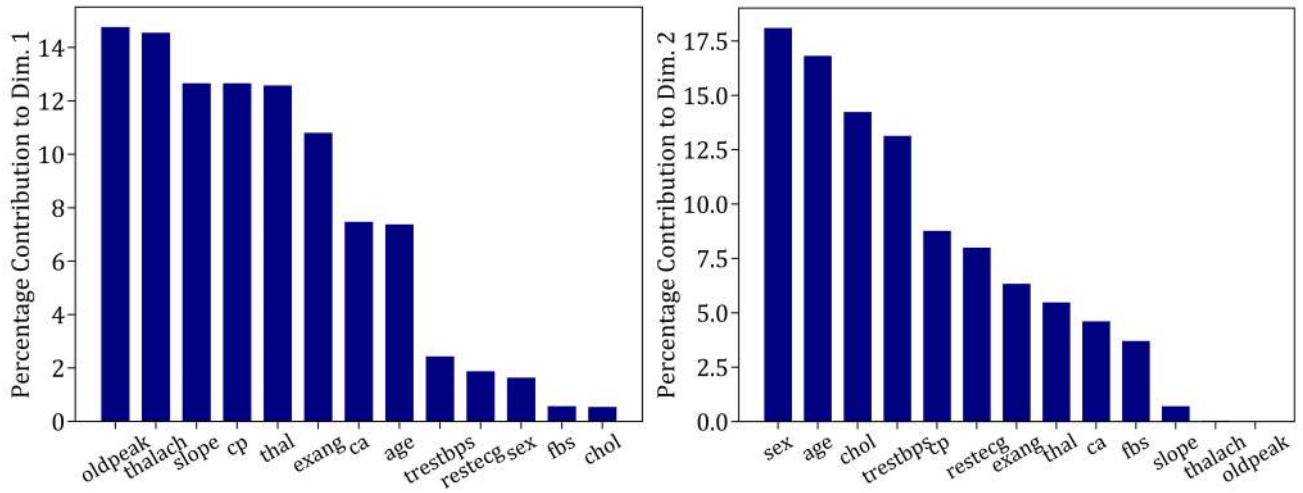


FIGURE 3. Heatmap of derived features (28 in total) of the Cleveland heart disease dataset derived using Pearson correlation coefficient after applying FAMD. The red and green color in heatmap scale is used to represent the negatively correlated and positively correlated features, respectively.

and numeric features *oldpeak* and *thalach* are second most contributing features with respect to the principal dimensions, Dimension 1 and Dimension 2.

3) FEATURES NORMALIZATION

The training dataset is created by extracting features using FAMD. The features are extracted as well as derived from 2 to 28 for Cleveland training dataset ($\mathcal{D}^T$). In this way, total of 27 new training datasets are created with varying feature



(a) (b)

FIGURE 4. Contribution of derived features from Cleveland heart disease dataset $\mathcal{D}$ concluded as (a) Dimension 1 (b) Dimension 2.

Algorithm 6 *FAMD_MLBox*($\mathcal{D}^T$, $\mathcal{F}^{fc}$, *algo*)

begin

On the basis of derived feature set $\mathcal{F}^{fc}$, the machine learning model *algo* is trained by tuning the corresponding hyper-parameters;

The performance metric $\langle Acc, Sens, Spec, Score, MCC, AUC\rangle$ is calculated by using Equations 4-8;

return $\langle Acc, Sens, Spec, Score, MCC, AUC\rangle$;

end

set $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_m\}$, where $2 \leq m \leq 28$. These features $\mathcal{F}_i$, where $1 \leq i \leq m$, are normalized with unit mean and zero standard deviation. In a similar way, the features are extracted from the validation set ($\mathcal{D}^V$) from 2 to 28 using the learned parameters of respective training dataset, $\mathcal{D}^T = \{\mathcal{D}_2^T, \mathcal{D}_2^T, \cdots, \mathcal{D}_{28}^T\}$, for all 27 new validation datasets $\mathcal{D}^V = \{\mathcal{D}_2^V, \mathcal{D}_2^V, \cdots, \mathcal{D}_{28}^V\}$, where subscript i and j, $2 \leq i \leq 28$, $2 \leq j \leq 28$, denotes the number of features in the training and validation datasets $\mathcal{D}_i^T$ and $\mathcal{D}_i^V$, respectively. The validation datasets $\mathcal{D}_i^V$, where $2 \leq i \leq 28$ are normalized with the mean and standard deviation of the respective training datasets $\mathcal{D}_j^T$, where $2 \leq j \leq 28$ and $i = j$. The pseudo-code for features normalization is depicted in **Algorithm 5**.

4) MACHINE LEARNING APPROACH AND PERFORMANCE METRIC EVALUATION

The training datasets, $\mathcal{D}_i^T$, where $2 \leq i \leq 28$, are fed in to the popular machine learning methods (*e.g.*, *LR*, *kNN*, *SVM*, *DT*s, *RF*, *etc.*). Though, top five machine learning approaches are selected for training and validation, however it can be extended as desired. The methods are trained and their respective hyper-parameters are tuned in the training phase for all training datasets $\mathcal{D}_i^T$, where $2 \leq i \leq 28$. The performance metric $\langle Acc, Sens, Spec, Score, MCC, AUC\rangle$ is evaluated and returned for all 27 training datasets as shown in **Algorithm 6**. On the basis of weight matrix w, the cumulative performance ($\mathbf{P}^i$) for each training dataset $\mathcal{D}_i^T$, where $2 \leq i \leq 28$, is calculated and performance $\mathbf{P} \leftarrow max(\mathbf{P}^i)$ and its associated machine learning approach *algo* is selected.

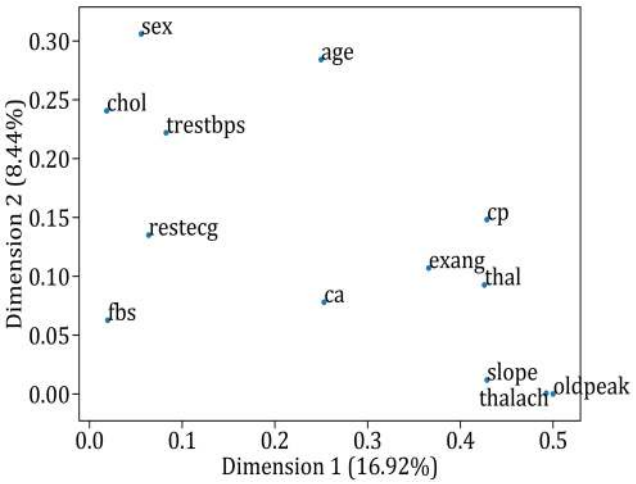


FIGURE 5. Coordinates of the various features representing their contributions to principal dimensions, Dimension 1 and Dimension 2.

The validation datasets, $\mathcal{D}_j^T$, where $2 \leq j \leq 28$, are validated using the machine learning approach *algo* with tuned hyper-parameters and performance metric $\mathbf{P}$ is evaluated as shown in **Algorithm 1**.

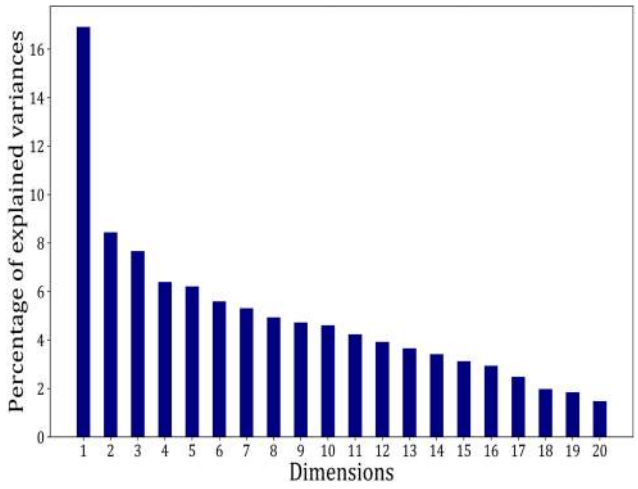


FIGURE 6. Initial 20 derived features contribution to the proposed model, *MIFH*.

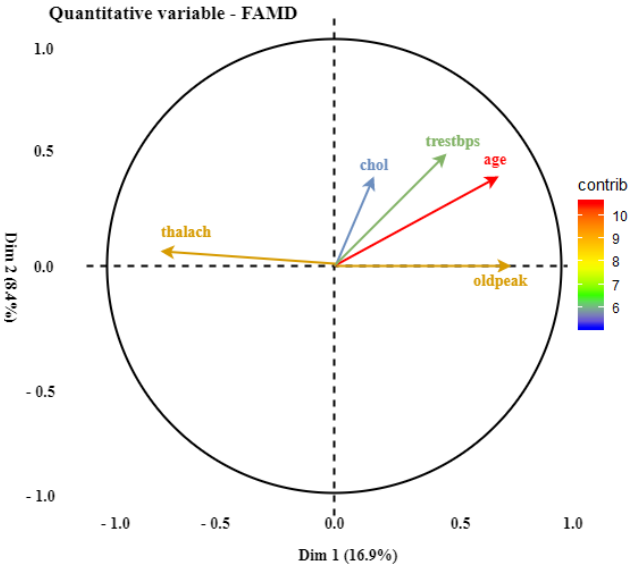


FIGURE 7. Representation of correlation between categorical features and principal dimensions.

VII. RESULTS ANALYSIS AND DISCUSSION

A. EXPERIMENTAL SETUP

We have implemented the proposed machine intelligence framework, *i.e., MIFH*, for heart disease diagnosis using python programming language of version 3.6 and executed the programs in a Linux machine with Intel i7 3.40GHz CPU and 8GB memory.

B. EVALUATION CRITERIA

In order to evaluate the performance of the proposed framework, *MIFH*, the confusion matrix is calculated which gives an idea about the learning extent of the machine learning approach and its ability for accurate classification. The prime components of confusion matrix are true positive (TP), true negative (TN), false positive (FP) and false negative (FN). In case of Cleveland heart disease dataset $\mathcal{D}$, TP and TN represents the proportion of instances where heart patients are identified as heart patients and normal subjects are identified as normal subjects, respectively. Depending upon the learned statistics, the framework may not be able to classify all instances accurately. To report those instances, FP and FN represents the inaccurate classification identifying normal subjects as heart patients and heart patients as normal subjects, respectively. This confusion matrix $\langle TP, FP, FN, TN\rangle$ is utilized to evaluate various performance measures such as accuracy (Acc), sensitivity ($Sens$) / recall (Rec) / true positive rate (TPR), specificity ($Spec$) / true negative rate (TNR), F1-Score ($Score$), false negative rate (FNR) / Miss-rate (MR), false positive rate (FPR) / fall-out ($Fout$), prevalence ($Prev$), positive predictive value (PPV) / precision ($Prec$), false omission rate (FOR), positive likelihood ratio ($LR+$), negative likelihood ratio ($LR-$), false discovery rate (FDR), negative predictive value (NPV), diagnostic odds ratio (DOR), Mathew's correlation coefficient (MCC), and area under the curve (AUC), *etc*. The total instances in $\mathcal{D}$, instances of heart patients ($TP+FN$) and normal subjects ($FP+TN$), predicted instances of heart patients and normal subjects are denoted by S, P and N, P' and N', respectively. The categorization of confusion matrix components and their collective interpretation is depicted in Table 4.

The state-of-the-art methods for heart disease diagnosis follow the metric $\langle Acc, Sens, Spec, Score, MCC, AUC\rangle$ [40]. The accuracy (Acc) represents the accurate classification of normal subjects and heart patients collectively and is depicted mathematically in terms of confusion matrix components as given in Equation 4. Though, only Acc could not determine the accurate discrimination between normal subjects and heart patients separately. The classification of heart patients and normal subjects is termed as $Sens$ and $Spec$, and is depicted mathematically in terms of confusion matrix components as given in Equations 5-6, respectively. The incorrect measures or mis-classifications are intuitive from $Sens$ and $Spec$, and are termed as $Fout$ and MR, where normal subjects are identified as heart patients and heart patients are identified as normal subjects, respectively. $Score$ is weighted average of recall and precision and can be represented mathematically in terms of confusion matrix components as given in Equation 7.

$$Acc = \frac{TN + TP}{TP + FP + FN + TN} \times 100\% \tag{4}$$

$$Sens = \frac{TP}{TP + FN} \times 100\% \tag{5}$$

$$Spec = \frac{TN}{FP + TN} \times 100\% \tag{6}$$

$$Score = \frac{2 \times TP}{2 \times TP + FP + FN} \times 100\% \tag{7}$$

Mathew's correlation coefficient (MCC) metric is used to predict the classification score ranging between $[-1, +1]$. The values $+1, -1$ and near to zero indicate the ideal, completely wrong and random predictions, respectively. The MCC is

TABLE 4. Confusion matrix.

		Actual Values				
	S	P	N	$Prev = \frac{\sum P}{\sum S}$	$Acc = \frac{\sum TP + \sum TN}{\sum S}$	
Predicted Values	P'	TP	FP	$PPV,\ Prec = \frac{\sum TP}{\sum P'}$	$FDR = \frac{\sum FP}{\sum P'}$	
	N'	FN	TN	$FOR = \frac{\sum FN}{\sum N'}$	$NPV = \frac{\sum TN}{\sum N'}$	
		$TPR, Rec, Sens = \frac{\sum TP}{\sum P}$	$FPR, Fout = \frac{\sum FP}{\sum N}$	$LR+ = \frac{TPR}{FPR}$	$DOR = \frac{LR+}{LR-}$	$Score = 2 \times \frac{Prec \times Rec}{Prec + Rec}$
		$FNR,\ MR = \frac{\sum FN}{\sum P}$	$TNR, Spec = \frac{\sum TN}{\sum N}$	$LR- = \frac{FNR}{TNR}$		

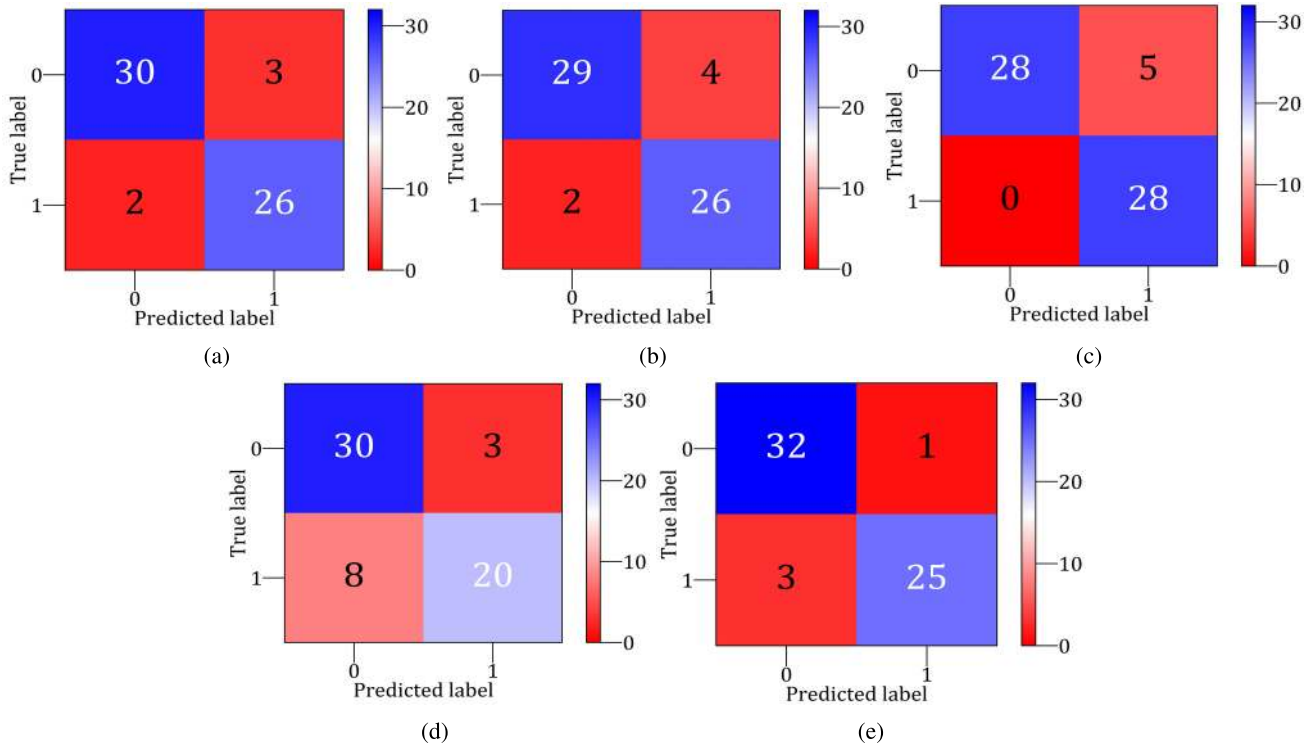


FIGURE 8. Confusion matrix derived from *MIFH* over Cleveland CHD validation dataset: (a) *LR* (b) *kNN* (c) *SVM* (d) *DT* and (e) *RF*.

evaluated mathematically in terms of confusion matrix as given in Equation 8.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \times 100\% \quad (8)$$

ROC and AUC: The performance metric AUC - ROC curve is used to represent the classification using the curve area. Receiver operating characteristics (ROC) represents a probability curve and area under the curve (AUC) represents the degree of separability of classifier to accurately classify instances between classes. Higher value of AUC implies the capability of model in distinguishing heart patients as heart patients and normal subjects as normal subjects.

C. PERFORMANCE EVALUATION OF MIFH

The proposed framework, *i.e., MIFH*, inputs the UCI Cleveland CHD dataset $\mathcal{D}$, imputed the dataset for missing values *ca* and *thal* using majority labels as presented in Section V-A.

TABLE 5. Performance evaluation metrics using FAMD and ML approaches *LR*, *kNN*, *SVM*, *DT* and *RF* over UCI heart disease Cleveland dataset.

Predictive model	# of Features	Hyper-parameter	TP	FP	TN	FN	Accuracy (in %)	Sensitivity (in %)	Specificity (in %)	F1- Score (in %)	MCC (in %)	AUC (in %)
FAMD + LR	17	C = 1, Regularization = L1	26	3	30	2	91.80	92.85	90.90	91.22	83.58	91.88
	18 - 22, 24 - 27	C = 1, Regularization = L1	26	4	29	2	90.16	92.85	87.78	89.65	80.47	90.36
	11	C = 1, Regularization = L2	25	4	29	3	88.52	89.28	87.78	87.71	76.99	88.52
	23	C = 1, Regularization = L1	26	5	28	2	88.52	92.85	84.84	88.13	77.49	88.85
	28	C = 1, Regularization = L1	23	2	31	5	88.52	82.14	93.93	86.79	77.05	88.04
FAMD + kNN	28	k = 11	26	3	30	2	90.16	92.85	87.87	89.65	80.47	90.36
	27, 24	k = 66, 60	26	6	27	2	86.88	92.85	81.81	86.66	74.51	87.33
	26, 11	k = 80, 6	26	7	27	2	85.24	92.85	78.78	85.24	71.64	85.82
	2	k = 3	27	9	24	1	83.60	96.42	72.72	84.37	70.07	84.57
	3	k =2	22	4	29	6	83.60	78.57	87.87	81.48	66.95	83.22
FAMD + SVM	28	C = 10, Kernel = Linear	28	5	28	0	91.80	100	84.84	91.80	84.84	92.42
	9	C = 10, Kernel = Linear	25	4	29	3	88.52	89.28	87.78	87.71	76.99	88.58
	8	C = 10, Kernel = Linear	25	5	28	3	86.88	89.28	84.84	86.20	73.89	87.06
	11	C = 1, Kernel = Linear	24	5	28	4	85.24	85.71	84.84	84.21	70.41	85.28
	24	C = 0.1, Kernel = RBF	21	3	30	7	83.60	0.75	90.90	80.76	67.23	82.95
FAMD + DT	27	Gini Index	20	3	30	8	81.96	71.42	90.90	78.43	64.09	81.16
	28	Entropy	20	4	29	8	80.32	71.42	87.78	76.92	60.49	79.65
	5	Gini Index	15	0	33	13	78.68	53.57	100	69.76	61.99	76.78
	17	Gini Index	23	9	24	5	77.04	82.14	72.72	76.66	54.75	77.43
	8	Gini Index	19	5	28	9	77.04	67.85	84.84	73.07	53.76	76.35
FAMD + RF	28	Gini Index	25	1	32	3	93.44	89.28	96.96	92.59	86.91	93.12
	10, 18, 20, 21, 26	Entropy	25	2	31	3	91.80	89.28	93.93	90.90	83.49	91.61
	22	Entropy	27	4	29	1	91.80	96.42	87.78	91.52	84.03	92.15
	5	Entropy	24	1	32	4	91.80	85.71	96.96	90.56	83.77	91.34
	27	Gini Index	26	3	30	2	91.80	92.85	90.90	91.22	83.58	91.88

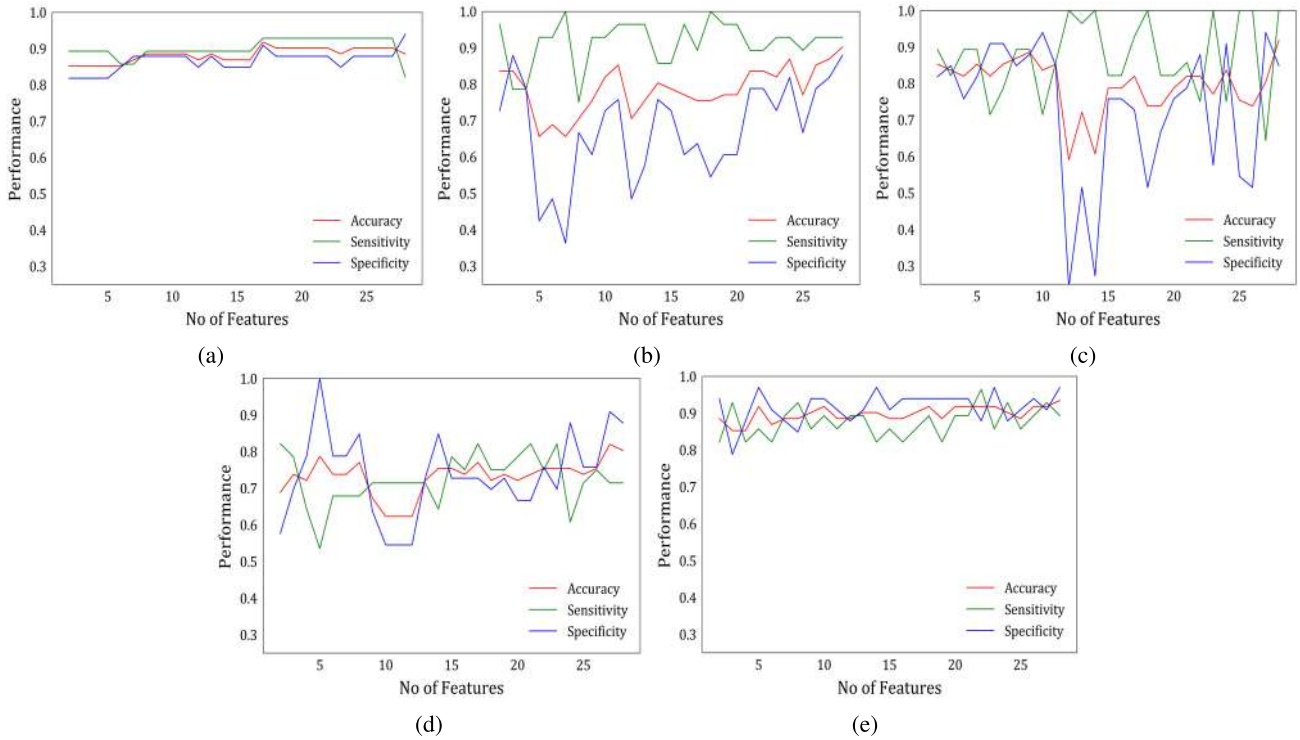


FIGURE 9. Performance of respective classifiers in terms of *Acc*, *Sens*, *Spec* (a) *LR* (b) *kNN* (c) *SVM* (d) *DT* and (e) *RF*, over Cleveland heart disease dataset with respect to the number of selected features.

The imputed Cleveland dataset is partitioned into training and validation datasets, *i.e.*, $\mathcal{D}^T$ and $\mathcal{D}^V$, respectively using the hold-out validation scheme with validation ratio 0.2. Stratification is performed to keep the partitioning balanced for heart patient and normal subject instances in both datasets, $\mathcal{D}^T$ and $\mathcal{D}^V$. The proposed framework *MIFH* is trained on $\mathcal{D}^T$ and the performance is validated on $\mathcal{D}^V$ for the purpose of verifying the robustness of framework. From $\mathcal{D}^T$, features are extracted as well as derived starting from 2 to 28, hence total of 27 new datasets have been prepared with new feature sets $\mathcal{F}' = \langle \mathcal{F}^2, \mathcal{F}^3, \cdots, \mathcal{F}^{28} \rangle$. These new training datasets are $\mathcal{D}^T = \langle \mathcal{D}^2, \mathcal{D}^3, \cdots, \mathcal{D}^{28} \rangle$. The dataset $\mathcal{D}^i$ is corresponding

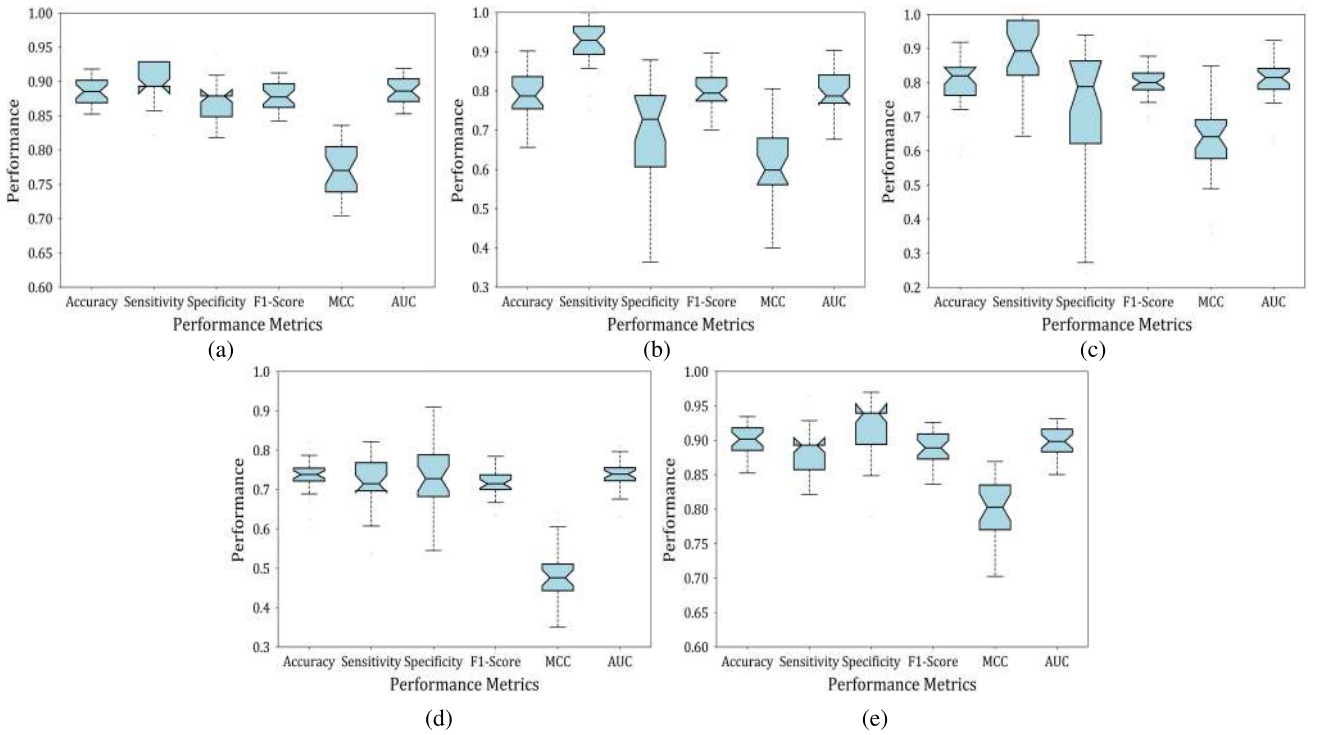


FIGURE 10. Whisker boxplot depicting the distribution of respective classifiers (a) *LR* (b) *kNN* (c) *SVM* (d) *DT* and (e) *RF*, in terms of *Acc*, *Sens*, *Spec*, *Score*, *MCC*, and *AUC* where Whisker boxplot extreme ends show the maximum and minimum value for the Cleveland heart disease dataset.

to feature set $\mathcal{F}^i$, where $2 \leq i \leq 28$. Each $\mathcal{D}^i$ in $\mathcal{D}^T$ is normalized using zero mean and unit standard deviation. The framework *MIFH* is learned from these datasets using different machine learning approaches and returns the best possible outcome based upon the mentioned performance metrics according to the weight assigned to each element of metric. In the proposed experimentation, considered machine learning approaches are $\langle LR, kNN, SVM, DT, RF\rangle$, the performance metrics is $\langle Acc, Sens, Spec, Score, MCC, AUC\rangle$ and the most concerned element of metrics is *Acc* along with *Sens* and *Spec*. The weight score is equal for these performance metric elements in our *MIFH* model. The selected algorithm *algo i.e.*, *Random Forest* (*RF*) in our case, is validated using the performance metrics $\langle Acc, Sens, Spec, Score, MCC, AUC\rangle$ for the validation dataset $\mathcal{D}^V$ after normalizing with the mean and standard deviation of $\mathcal{D}^T$ and learned parameters from FAMD during training phase. The selective experimentation statistics for datasets $\mathcal{D}^2, \mathcal{D}^3, \cdots, \mathcal{D}^{28}$ (in terms of number of features) using FAMD and ML approaches considering performance metrics $\langle Acc, Sens, Spec, Score, MCC, AUC\rangle$ along with the hyper-parameters used during their tuning is depicted in Table 5. Moreover, the confusion matrix elements *i.e.*, TP, FP, TN, FN, specified in the Table 5 is representing the classified and mis-classified count of normal subject and heart patient from the validation dataset, $\mathcal{D}^V$.

The performance of ML approaches with varying features and metrics is shown in Figure 8, 9 and 10, respectively. It can be observed from confusion matrix of classifiers that *RF* is comparatively better to classify normal subject as normal subject and heart patient as heart patient. There are 1 and 3 cases where *RF* predicts normal subjects as heart disease and heart disease as normal subjects. Though, *SVM* performed comparatively well as *SVM* did not classify any heart patient as normal subject but comparatively higher false positives. It can be observed that *RF* is best classifier among all in terms of accuracy exploiting all 28 derived features with Gini Index as criteria for splitting nodes. It has achieved *Acc* 93.44%, with *Sens* 89.28% and *Spec* 96.96%. The performance of *RF* can also be observed from Figure 9e and 10e, which depicts the consistency with minute deviation while varying features from 2 to 28. The classifier *LR* shows consistent results with varying number of features and performance as shown in Figure 9a and 10a. It achieves the performance $\langle 91.80\%, 90.90\%, 92.85\%\rangle$ for $\langle Acc, Sens, Spec\rangle$, along with $C = 1$, penalty as $L1$-regularization and the number of features taken as 17. The classifiers *SVM* and *kNN* show high variance and very inconsistent in performance while varying the number of features. It can be observed from Figure 9c and 10c for *SVM* and Figure 9b and 10b for *kNN* that due to its sharp deviations in performance, it cannot be accounted for final deployment, even when *SVM* has achieved *unit sensitivity*. The best results for *SVM* are $\langle 91.80\%, 100\%, 84.84\%\rangle$, for $\langle Acc, Sens, Spec\rangle$ with $C = 10$ and linear kernel and for *kNN*, the performance is $\langle 90.16\%, 92.85\%, 87.78\%\rangle$ for $\langle Acc, Sens, Spec\rangle$ with number of features and nearest neighbors for classification are 28 and 11, respectively. Although, the classifier *DT* is significantly consistent in performance as depicted in Figure 9d and 10d, but not performed well as

TABLE 6. Performance metrics of the proposed method *MIFH* along with baseline methods statistics on UCI heart disease Cleveland dataset.

Research Contribution	Method	Accuracy (in %)	Sensitivity (in %)	Specificity (in %)
Purushottam *et al.*, 2016, [25]	Rules based classifier	86.7	—	—
Shah *et al.*, 2017, [27]	PPCA	82.18	75	90.57
Vijayashree *et al.*, 2018, [19]	PSO with SVM	84.36	—	—
Haq *et al.*, 2018, [7]	Relief + LR	89	77	98
Haq *et al.*, 2018, [7]	mRMR + NB	84	77	90
Haq *et al.*, 2018, [7]	LASSO + SVM	88	75	96
Saqline *et al.*, 2019, [20]	RBF kernel-based SVM	81.19	72.92	88.68
Mohan *et al.*, 2019, [21]	HRFLM	88:7	92.8	82.6
Ali *et al.*, 2019, [22]	L1 Linear SVM + L2 Linear & RBF SVM	92.22	82.92	100
MIFH (Proposed)	**FAMD + RF**	**93.44**	**89.28**	**96.96**

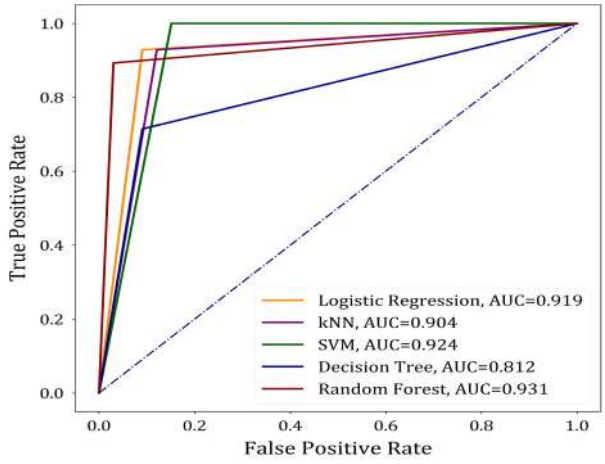


FIGURE 11. ROC-AUC curve obtained by classifiers *LR*, *kNN*, *SVM*, *DT* and *RF* with respect to the number of selected features.

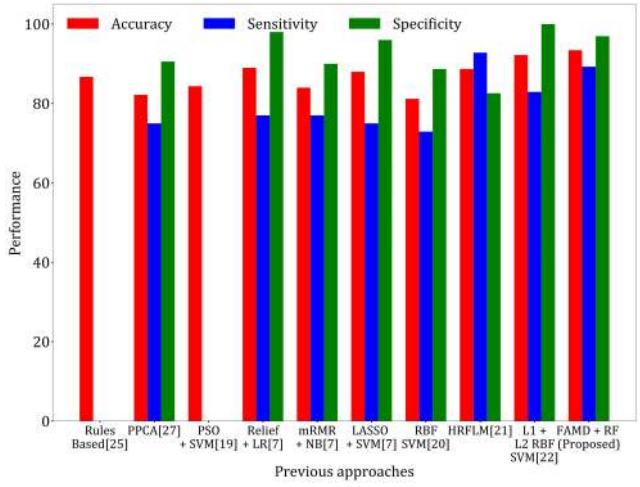


FIGURE 12. Comparison of proposed machine intelligence framework *MIFH* for Cleveland heart disease dataset with existing state-of-the-art methods.

compared to its ensembles *RF* and *LR*. It achieves the best performance ⟨81.96%, 71.42%, 90.90% for ⟨*Acc*, *Sens*, *Spec*⟩ along with 27 derived features. $AUC - ROC$ score represents the capability of model to distinguish among classes. From Figure 11, it can be clearly observed that *RF* ($AUC = 0.931$) is best classifier followed by *SVM* ($AUC = 0.924$) and *LR* ($AUC = 0.919$).

The performance of proposed framework for heart disease diagnosis, *MIFH*, is compared with several baseline methods recently proposed and came into existence by the academicians and researchers to contribute in developing the decision support system for heart disease diagnosis [7], [19]–[22], [25], [27]. It can be observed that *MIFH* improves the system in terms of overall accuracy in the prediction of heart disease as presented statistically as well as pictorially in Table 6 and Figure 12, respectively. The sensitivity and specificity of *MIFH* is comparable to the existing recent methods such as Mohan *et al.* has achieved sensitivity 92.8% and Ali *et al.* has achieved the specificity 100% while *MIFH* gives 89.28% and 96.96% using *RF* predictive model.

VIII. CONCLUSIONS AND FUTURE WORK

In the proposed work, a machine intelligence framework *MIFH* is presented for heart disease diagnosis. The proposed framework *MIFH* can be used to predict the instances either as normal subjects or heart patients. *MIFH* utilizes the characteristics of FAMD to extract as well as derive features from the UCI heart disease Cleveland dataset and train the machine learning predictive models for classification of instances as well as prediction of heart disease and normal subjects. MIFH returns the best classifier based upon the weight matrix corresponding to performance metrics.

As a future perspective, the multi-class classification of heart disease datasets can be considered. In addition, the data

collected at a medical hospital and institution is usually class imbalanced. The study says normal predictive models are neither efficient nor specifically designed to handle the class imbalanced data. In addition, class imbalanced datasets can also be explored to deal with real-life scenarios in hospitals and medical institutions.

ACKNOWLEDGMENT

The author R. Kumar would like to thank Ministry of Electronics and Information Technology (MeitY), Govt. of India, for their support through ''Visvesvaraya Ph.D. Scheme for Electronics and IT'' to carry out the research work. *(Ankur Gupta and Rahul Kumar contributed equally to this work.)*

REFERENCES

[1] P. A. Heidenreich *et al.*, ''Forecasting the future of cardiovascular disease in the United States,'' *Circulation*, vol. 123, no. 8, pp. 933–944, 2011.

[2] *World health organization: Who*. Accessed: Jul. 25, 2019. [Online]. Available: https://www.who.int

[3] A. L. Bui, T. B. Horwich, and G. C. Fonarow, ''Epidemiology and risk profile of heart failure,'' *Nature Rev. Cardiol.*, vol. 8, no. 1, pp. 30–41, Jan. 2011.

[4] J. López-Sendón, ''The heart failure epidemic,'' *Medicographia*, vol. 33, no. 4, pp. 363–369, 2011.

[5] L. A. Allen, ''Decision-making in advanced heart failure,'' *Circulation*, vol. 125, no. 15, pp. 1928–1952, 2012.

[6] T. Tirkes, M. A. Hollar, M. Tann, M. D. Kohli, F. Akisik, and K. Sandrasegaran, ''Response criteria in oncologic imaging: Review of traditional and new criteria,'' *Radio Graph.*, vol. 33, no. 5, pp. 1323–1341, Sep. 2013.

[7] 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.

[8] K. Polat, S. ahan, and S. Güne , ''Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and *k*-NN (nearest neighbour) based weighting preprocessing,'' *Expert Syst. Appl.*, vol. 32, no. 2, pp. 625–631, Feb. 2007.

[9] I. Babaoglu, O. Findik, and E. Ülker, ''A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine,'' *Expert Syst. Appl.*, vol. 37, no. 4, pp. 3177–3183, Apr. 2010.

[10] M. A. Karaolis, J. A. Moutiris, D. Hadjipanayi, and C. S. Pattichis, ''Assessment of the risk factors of coronary heart events based on data mining with decision trees,'' *IEEE Trans. Inform. Technol. Biomed.*, vol. 14, no. 3, pp. 559–566, May 2010.

[11] P. Anooj, ''Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules and decision tree rules,'' *Open Comput. Sci.*, vol. 1, no. 4, pp. 482–498, 2011.

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

[13] P. Anooj, ''Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules,'' *J. King Saud Univ. Comput. Inf. Sci.*, vol. 24, no. 1, pp. 27–40, Jan. 2012.

[14] N. C. Long, P. Meesad, and H. Unger, ''A highly accurate firefly based algorithm for heart disease prediction,'' *Expert Syst. Appl.*, vol. 42, no. 21, pp. 8221–8231, Nov. 2015.

[15] V. Krishnaiah, G. Narsimha, and N. S. Chandra, ''Heart disease prediction system using data mining technique by fuzzy *k*-NN approach,'' in *Proc. 49th Annu. Conv. Comput. Soc. India (CSI)*, vol. 1, 2015, pp. 371–384.

[16] I. Paschalidis. (2017). *How Machine Learning is Helping Us Predict Heart Disease and Diabetes*. [Online]. Available: https://hbr.org/2017/05/how-machine-learning-is-helping-us-predict-heart-disease-and-diabetes

[17] S. Iftikhar, K. Fatima, A. Rehman, A. S. Almazyad, and T. Saba, ''An evolution based hybrid approach for heart diseases classification and associated risk factors identification,'' *Biomed. Res.*, vol. 28, no. 8, pp. 3451–3455, 2017.

[18] A. K. Paul, P. C. Shill, M. R. I. Rabin, and K. Murase, ''Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease,'' *Appl. Intell.*, vol. 48, no. 7, pp. 1739–1756, Jul. 2018.

[19] J. Vijayashree and H. P. Sultana, ''A machine learning framework for feature selection in heart disease classification using improved particle swarm optimization with support vector machine classifier,'' *Program. Comput. Soft.*, vol. 44, no. 6, pp. 388–397, Nov. 2018.

[20] S. M. Saqlain, M. Sher, F. A. Shah, I. Khan, M. U. Ashraf, M. Awais, and A. Ghani, ''Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines,'' *Knowl. Inf. Syst.*, vol. 58, no. 1, pp. 139–167, Jan. 2019.

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

[22] L. Ali, A. Niamat, J. A. Khan, N. A. Golilarz, X. Xingzhong, A. Noor, R. Nour, and S. A. C. Bukhari, ''An optimized stacked support vector machines based expert system for the effective prediction of heart failure,'' *IEEE Access*, vol. 7, pp. 54007–54014, 2019.

[23] *UCI Machine Learning Repository: Heart Disease Data Set*. Accessed: Jul. 25, 2019. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[24] S. Ismaeel, A. Miri, and D. Chourishi, ''Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis,'' in *Proc. IEEE Canada Int. Humanitarian Technol. Conf. (IHTC)*, May 2015, pp. 1–3.

[25] Purushottam, K. Saxena, and R. Sharma, ''Efficient heart disease prediction system,'' *Procedia Comput. Sci.*, vol. 85, pp. 962–969, Jan. 2016.

[26] H. A. Esfahani and M. Ghazanfari, ''Cardiovascular disease detection using a new ensemble classifier,'' in *Proc. IEEE 4th Int. Conf. Knowl.-Based Eng. Innov. (KBEI)*, Dec. 2017, pp. 1011–1014.

[27] S. Shah, S. Batool, I. Khan, M. Ashraf, S. Abbas, and S. Hussain, ''Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis,'' *Phys. A, Stat. Mech. Appl.*, vol. 482, pp. 796–807, Sep. 2017.

[28] D. Tomar and S. Agarwal, ''Feature selection based least square twin support vector machine for diagnosis of heart disease,'' *Int. J. Bio-Sci. Bio-Technol.*, vol. 6, no. 2, pp. 69–82, Apr. 2014.

[29] M. Arun Kumar and M. Gopal, ''Least squares twin support vector machines for pattern classification,'' *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7535–7543, May 2009.

[30] G. T. Reddy and N. Khare, ''An efficient system for heart disease prediction using hybrid OFBAT with rule–based fuzzy logic model,'' *J. Circuits, Syst. Comput.*, vol. 26, no. 4, Apr. 2017, Art. no. 1750061.

[31] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, ''Computer aided decision making for heart disease detection using hybrid neural network-genetic algorithm,'' *Comput. Methods Programs Biomed.*, vol. 141, pp. 19–26, Apr. 2017.

[32] *European Coordination Committee of the Radiological, Electromedical, and Healthcare it Industry: Medical Imaging Equipment Age Profile & Density*. Accessed: Jul. 25, 2019. [Online]. Available: http://www.cocir.org/uploads/media/16052_COC_AGE_PROFILE_web_01.pdf

[33] H. Liu and L. Yu, ''Toward integrating feature selection algorithms for classification and clustering,'' *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005.

[34] E. Avci, ''A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier,'' *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10618–10626, Sep. 2009.

[35] W. Guan, A. Gray, and S. Leyffer, ''Mixed-integer support vector machine,'' in *Proc. NIPS Workshop Optim. Mach. Learn.*, Dec. 2009.

[36] A. Khemphila and V. Boonjing, ''Heart disease classification using neural network and feature selection,'' in *Proc. 21st Int. Conf. Syst. Eng.*, Aug. 2011.

[37] J. Nahar, T. Imam, K. S. Tickle, and Y.-P.-P. Chen, ''Computational intelligence for heart disease diagnosis: A medical knowledge driven approach,'' *Expert Syst. Appl.*, vol. 40, no. 1, pp. 96–104, Jan. 2013.

[38] J. A. Sanz, M. Galar, A. Jurio, A. Brugos, M. Pagola, and H. Bustince, ''Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system,'' *Appl. Soft Comput.*, vol. 20, pp. 103–111, Jul. 2014.

[39] S. Shilaskar and A. Ghatol, ''Feature selection for medical diagnosis: Evaluation for cardiovascular diseases,'' *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4146–4153, Aug. 2013.

[40] M. H. Zweig and G. Campbell, ''Receiver-Operating Characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine,'' *Clin. Chem.*, vol. 39, no. 4, pp. 561–577, 1993.

ANKUR GUPTA (Student Member, IEEE) received the B.Tech. degree in computer science and engineering from the Maharana Pratap Engineering College, Kanpur, affiliated to Gautam Buddh Technical University, Lucknow, India, in 2010, and the M.Tech. degree in computer science and engineering from the National Institute of Technology (NIT) Hamirpur, Hamirpur, India, in 2015. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, IIT Roorkee, Roorkee, India. From 2010 to 2013, he served the society as a faculty in computer science and engineering and fulfills several academic and social responsibilities, including the Center Superintendent for the State University Examination. He was a Visiting Scientist with the Next Generation Computing (NGC) Lab, College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan, from November 2017 to December 2017. He is also a Research Scholar with the Department of Computer Science and Engineering, IIT Roorkee. He has published two ACM transactions, two international conference papers, and one international workshop paper. His research interests include machine (deep) learning and its applications in the emerging trends, including the healthcare sector using medical imaging techniques, algorithm design for optimization problems, graph theory, and algorithmic microfluidics along with electronic design automation for embedded solutions.

RAHUL KUMAR received the B.Tech. degree in information technology from the Muzaffarpur Institute of Technology Muzaffarpur, affiliated to Aryabhatta Knowledge University, Patna, India, in 2012, and the M.Tech. degree in computer science and engineering from the Dr. B. R. Ambedkar National Institute of Technology Jalandhar, Jalandhar, India, in 2015. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, IIT Roorkee, Roorkee, India. His research interests include image processing, and application of machine and deep learning in biomedical imaging.

HARKIRAT SINGH ARORA is currently pursuing the bachelor's degree with the Department of Chemical Engineering, IIT Roorkee, Roorkee, India. His research interests include using machine and deep learning techniques in the field of medical imaging, and focused on improving healthcare with the help of these techniques.

BALASUBRAMANIAN RAMAN (Member, IEEE) received the Ph.D. degree from IIT Madras, in 2001. He was a Postdoctoral Fellow with the University of Missouri, Columbia, MO, USA, from 2001 to 2002, and a Postdoctoral Associate with Rutgers, The State University of New Jersey, USA, from 2002 to 2003. He was a Visiting Professor and a member with the Computer Vision and Sensing Systems Laboratory, Department of Electrical and Computer Engineering, University of Windsor, Canada, in 2009. He is currently a Professor with the Department of Computer Science and Engineering, IIT Roorkee, Roorkee, India. His research interests include fractional transform theory, wavelet analysis, biometrics, content-based video retrieval, video skimming and summarization, medical imaging, long-range imaging, and hyperspectral imaging. He has more than 150 research publications in reputed journals and conference proceedings. He is a member of the IEEE Society, Uttar Pradesh Section, and acted as a Joint Secretary of the Executive Committee, IEEE Uttarakhand Sub-Section, from 2011 to 2013. He was a recipient of the BOYSCAST Fellowship from DST, India.