Implementation of Machine Learning Model to Predict Heart Failure Disease

Fahd Saleh Alotaibi[1]
Information Systems Department
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

Abstract—**In the current era, Heart Failure (HF) is one of the common diseases that can lead to dangerous situation. Every year almost 26 million of patients are affecting with this kind of disease. From the heart consultant and surgeon's point of view, it is complex to predict the heart failure on right time. Fortunately, classification and predicting models are there, which can aid the medical field and can illustrates how to use the medical data in an efficient way. This paper aims to improve the HF prediction accuracy using UCI heart disease dataset. For this, multiple machine learning approaches used to understand the data and predict the HF chances in a medical database. Furthermore, the results and comparative study showed that, the current work improved the previous accuracy score in predicting heart disease. The integration of the machine learning model presented in this study with medical information systems would be useful to predict the HF or any other disease using the live data collected from patients.**

Keywords—Machine learning model; medical data; heart failure diagnoses

I. INTRODUCTION

The main cause of heart stroke is due to blockage in arteries. It has many other names such as cardiovascular disease and arterial hypertension [1]. Approximately, there are almost 26 million people around the world affecting with heart disease [2]. The worry point is, this ratio is expected to increase rapidly in coming years, if precautions are not taken efficiently [3]. Apart from making life style healthy and diet control, the right time diagnosing and comprehensive analysis are other essential factors, which can ultimately save the lives [4]. Therefore, this paper has taken a small step towards saving the lives of HF patients and describes a way to improve the performance of diagnosing the patients on the bases of their medical history.

Most of the time patients goes for several tests, which can overburden them with extra physical activities, time, and for sure additional financial charges [5]. As previous studies suggested the common reasons behind heart disease can be unhealthy food, tobacco, excessive sugar, overweight or extra body fat [3], [6]. Whereas the common symptoms can be pain in arms and chest [7]. Noticeably, these reasons are independent from each other; proper analysis on this kind of dataset can improve the process of diagnosing and can assist the heart surgeons as well. Previously, different researches used number of techniques to improve the HF diagnosis process such as Extreme Learning Machine [8], heart disease classification [9], and machine learning classifiers [1]. Therefore, this research attempts to improve the performance of the classifiers by doing experiments using multiple machine-learning models to make better use of the dataset collected from different medical databases.

The paper is further divided into the following sub-sections: the next section describes a comprehensive overview on the use of machine learning models for predicting the heart disease. Section III explains the data overview, number of attributes and description of each attribute. Section IV shows the data preprocessing steps applied in this study. Furthermore, Sections V, VI and VII present the experiment design, implementation, and performances of the classifiers respectively. Finally, in Section VIII the study has been concluded.

II. MACHINE LEARNING CLASSIFIERS FOR HEART DISEASE PREDICTION

The identification of heart disease in a patient is complex and requires various details, laboratory tests, and equipment [10]. This research is not for replacing the traditional approach use for diagnosing and predicting the chances of heart failure, rather the study attempted to support this process using advanced technologies such as Machine Learning (ML). The ML is not a new technique and has been used several times for different applications.

A cloud based decision support system proposed by [11] in order to helps the heart consultants during diagnosis process. This system used machine-learning methods for predicting the heart disease. The system was proposed to provide the assistance in affordable way, where the system have the capacity to integrate with existing system. In that research clustering method used for categorizing the dataset based on particular groups in unsupervised manner. The author in [12] used an approach by implementing multiple clustering algorithms on heart disease dataset to understand the optimal solution, which can maximize the prediction accuracy ratio. ML approaches proved to be an effective in predicting the heart disease using historical data is further proved in a research conducted using Naïve Bayes, Decision Tree, support vector model and other models [13]. The results indicated that the support vector machine provided the optimal results between other implemented approaches.

Bashir *et al.*, (2019) attempted to improve the performance of heart disease prediction using feature selection approach. Different models such as Naïve Bayes, Random Forest and other used in the experiment implemented using Rapid Miner tool. The output indicated the high accuracy measured due to feature selection approach [7]. Furthermore, the Extreme Learning Machine techniques using feedforward neural network applied on Cleveland data based on 300 patients, suggested 80% accuracy in forecasting the heart disease in a patient [8]. In another research, the neural network applied using multi-layer perceptron, which also known as supervised learning. The system was proposed to determine the potential heart disease risk in a patient, using patient's historical data [5]. HF ratio using preserved ejection fraction is another work presented using multiple factors like strain rate, hypertensive situation, and velocity, where overall accuracy computed was more than 80% [14].

The main idea behind this discussion is to put stress on how helpful machine learning approaches are to predict the heart disease using medical data. This research is also emphasizing onto overcome the vulnerable situation and proposed a computerized system, so that heart consultants cannot miss any information due to improper reading and understanding of the data. Such a situation described in a research, that most of the heart diseases would not always detect just by doing ECG (a kind of test for diagnosing the working capability of a heart) [15]. Therefore, this kind of research overwhelmed those situations where doctors are puzzled and left behind some evidences. To support them, computerized medical system [16]–[19] with full of functionalities are there, which specially built to assist healthcare industry for the sake of patient's time, money, and most importantly to help the surgeons to save the patient's life. A kind of system proposed by [1] using ML approach for predicting the heart failure using heart sound reports. The study shown another proof that machine learning methods can applied on life saving system such as heart failure detection.

Moreover, a review report presented in a study [20] that described the importance of classification models and further explained the details of the models already implemented in the healthcare industry. The paper highlighted that there are many researches attempted data mining techniques successfully on medical cases. In the same way, another comparative study shown the performances of the multiple classifiers applied on two different tools; Matlab and Weka. Overall, the accuracy of the decision tree, Linear SVM and other models was recorded between 52% to 67.7%, although the accuracy were considerably low [9]. As per the researches discussed above, still different kind of models are providing variation in the prediction score. Thus, the dimensionality reduction and feature engineering can improve the process of data selection, which ultimately can improve the accuracy estimation [21].

TABLE. I. THE ACCURACY MEASURED IN PREVIOUS WORK

Techniques	[UCI, Rapid Miner, 2019] [7]	[UCI, Matlab, 2017] [9]	[UCI, Weka, 2017] [9]
Decision Tree	82.22%	60.9%	67.7%
Logistic Regression	82.56%	65.3%	67.3%
Random Forest	84.17%	X	X
Naïve Bayes	84.24%	X	X
SVM	84.85%	67%	63.9%

In conclusion, the clear research gap found in the previous researches is that, the measured accuracy is not up to the mark. Somewhere, the common machine learning approaches has not used as shown in Table I. Therefore, this section described the comprehensive overview on the previous work accompanied to predict the heart disease in a patient using ML approaches. The idea of this study is to improve the previous work using the selected dataset and ML models, as described in the next section. The performance of each model is discussed in the result section. Although, the models and dataset selected in this research are based on the previous work. The most commonly ML approaches found and used in this study are; Decision Tree, Naïve Bayes, Random Forest, Support Vector Machine, and Logistic Regression. This study used the dataset collected from Kaggle, the data set originally published on UCI data repository for machine learning. Previously, the experiments attempted for predicting the heart disease, the details and accuracy measured is shown in Table I. Finally, in the result section the comparative study is presented to understand the performance of the classifiers in this study and in the previous work.

III. DATA OVERVIEW

The dataset used in this research is collected from Kaggle platform, the dataset is also known as Heart Disease Dataset [22]. Altogether, the data was the combination of four different database, but only Cleveland data used in this experiment. It is an open dataset, having number of attributes, but for this experiment only fourteen attributes selected as described and suggested by different scholars that selected 14 attributes are most useful to predict the heart disease in a patient [7], [23]. In addition, the database file contains the record of 303 patients. The complete description of each attribute and the number of values for each attribute is shown in the Table II below:

TABLE. II. DATA OVERVIEW AND ATTRIBUTES DESCRIPTION

S.No.	Attribute Description	Distinct Values
1	***Age*** - The first attribute is defining the age of the person. [Minimum Age: 29, Maximum Age: 77]	Multiple values between 29 and 77
2	***Sex*** - The attribute number two describes the gender of a person. ["0" means Female and "1" means Male]	0, 1
3	***CP*** - The third attribute is defining the level of chest pain (CP) a patient suffering from, when reached to the hospital. There are four kind of distinct values defined for this attribute, where each value is describing a level of chest pain.	0, 1, 2, 3
4	***RestBP*** - The next attribute describes about the blood pressure (BP) figure for the patient while admitted to the hospital. [Minimum BP: 94, Maximum BP: 200]	Multiple values between 94 and 200
5	***Chol*** - This column is showing the cholesterol level recorded while admitting the patient in the hospital. [Minimum Chol: 126, Maximum Chol: 564]	Multiple values between 126 and 564
6	***FBS*** - The next attribute is describing the fasting blood sugar level in the patient. It has binary classified values. The values are depending on, if the patient has more than 120mg/dl sugar = 1, if not = 0.	0,1
7	***RestECG* -** This parameter is showing the result of ECG from 0 to 2. Where each value is showing the severity of the pain.	0, 1, 2
8	***HeartBeat*** - The maximum value of heartbeat counted at the time of admission [Minimum: 71, Maximum: 202]	Multiple values between 71 and 202
9	***Exang* -** This parameter was used to understand about, does exercise induce angina or not. If yes, the value will be "1", and "0" for not.	0, 1
10	***OldPeak* -** The next attribute is defining the patient's depression status. It is assigned as different real number values falls between 0 and 6.2.	Multiple real number values between 0 and 6.2.
11	***Slope* -** The condition of the patient during peak exercise. This value defined into three segments [Upsloping, Flat, Down sloping]	1, 2, 3
12	***CA***: This attribute is showing status of fluoroscopy. It is showing that how many vessels are colored.	0, 1, 2, 3
13	***Thal* -** This parameter is another kind of test required for the patient having chest pain or breathing difficulty. Four kind of values showing the result of Thallium test.	0, 1, 2, 3
14	***Target*** – This is the last column in the dataset. This Target column is also known as Class column or Label column. As this column describes the number of categories, (classes) defined in the data file. As per the dataset taken in this experiment. There are two different types of classes (0,1), where "0" means there is no chances of Heart Failure, whereas "1" imply that there are strong chances of heart failure in a patient. The value "0" and "1" is based on the other 13 parameters described in this dataset above.	0, 1

IV. DATA PREPROCESSING

Data preprocessing is an essential step use to clean the data and make it useful for any experiment associated with machine learning or data mining [24]. In this study, multiple preprocessing steps applied on the selected dataset. Firstly, the size of the dataset was found not enough for the implementation of machine learning approaches. As described by [25] the size of the dataset for machine learning implementation may create biasness and would also effect on the results generated through machine learning models. Therefore, for each attribute using minimum and maximum values, the random number generation technique applied to generate random values for each column [26]. This helped us to enhance the capacity of the data, which has created the positive impact on the performance of the classifier as can be seen in the results section. In conclusion, the data have increased the volume by three times.

Secondly, using rapid miner, data cleaning step applied to find out missing values and noisy data values. The data has some missing values which has been imputed using K Nearest Neighbor (KNN) method. As KNN method is proved to be a useful method for missing data imputation [27]. In addition, the outlier detection methods used to estimate the noise in the data. The data has not found noisy values and no outlier detected in the dataset. The outlier detection applied using rapid miner's operator with distances method [28]. In order to check the other discrepancies in the dataset, data discretization, transformation and binning techniques were applied as well.

The next step was to transform the data values into appropriate data type. In this study, multiple models were applied to check the performance of the prediction accuracy. Therefore, it was essential to convert the data type of some attributes as per the required format based on the model specification. Mainly, the experiment design built using binary classification, which is the process of categorizing the dataset according to predefined classes, which has been widely used in applying machine learning algorithms [29]. Hence, the same binary classification was used in the given dataset, where the binary classification provided the better way to show the performance accuracy of the selected classifier in this study.

Most of the attributes were nominal in the selected dataset i.e. Slope, CA, Thal, and CP. For example, Thal attribute is describing the value of the Thallium test based on the four predefined values (0, 1, 2, 3). In the same way, CP was another independent attribute in the dataset, which highlighting the condition of the chest pain using (0, 1, 2, 3) in the patient at the time of admitting in the hospital, where the "0" means normal and "3" means the worst condition. The Target column in the dataset that also known as class attribute has two types of predefined classes known as "0" and "1". This attribute represent the overall condition of the patients using other independent variables. Whereas the value "0" means that patient does not have chances of heart failure, and "1" means that patient has high probability of heart failure. For example, using a value of all independent variables, if a patient has high blood pressure, sugar and contain high values in Thallium test can have chances of heart failure and vice versa.

V. The Experiment Preparation

This research used five different models to predict the heart disease using collected dataset. The performance of each classifier and comparison with previous work is presented in the next section. After successful implementation of the data preprocessing step, in this section discussed about the selected models, their descriptions and overall methodology used for the experiment. The work presented in this study was a sequel of the research presented by [7], which used the same dataset and implementation tool. That research reduced the number of attributes to 14. The number of records they used for the experiment were 300. The 5-fold cross-validation technique was applied to improve the accuracy and reduce the chances of duplication in record selection. Overall, the experiment computed the accuracy from 82% to 85%.

In this work, the target was to improve the accuracy of the model; therefore different amendment was done in the experiment design. For example, data expansion, 10-fold cross validation, execution of all model at the same time to understand the actual differences in the accuracy measurement. Following are the procedural steps of designed methodology applied in this research.

Algorithm: Predicting Heart Failure Disease

```
Step 1: Selection of dataset/Data Preprocessing
        {
                Data overview
                Detect and remove outliers
                Detect and impute missing data
                Data enhancement using random number
                generators
                Applying suitable normalization techniques
        }
Step 2: Model Selection
        {
                Understanding data value (classes)
                Machine learning model selection
        }
Step 3: Model Implementation using Rapid Miner
        {
        Import Data
        Implementing all models together using Rapid Miner
        }
Step 4: Performance Measurement
        {
        Calculate Accuracy using "Performance" operator
        Analyzing the result through Confusion Matrix
        }
Step 5: Result Comparison
        {
        Comparing the accuracy among all models
        Comparing the result with previous work
        Calculate final output
        }
```

As discussed above that five machine learning models used for predicting the heart disease in a patient and to analyze up to optimal performance among all. The short description of each model explained in this section.

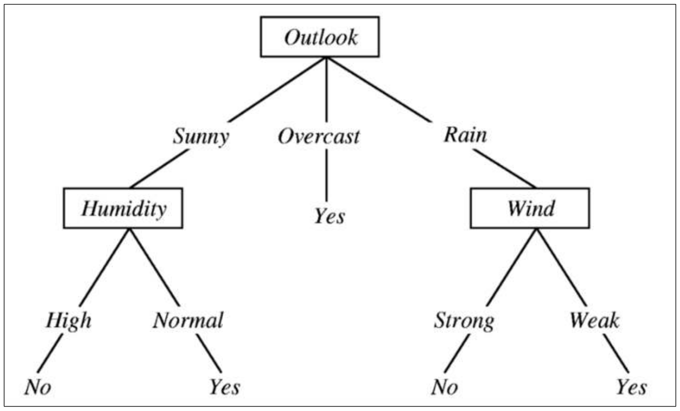

Fig. 1. A Decision Tree Example [31].

A. Decision Tree

It's a tree like classification model, which built a structure consisting of branches and nodes on the bases of evidence collected for each attributes during model learning phase [30]. The decision tree's branches and nodes connect according to the number of entities described in the dataset. The forwarding process uses the number of values dedicated for each attribute. Furthermore, following the rules describe on each branch and node it reached to the decision for each transaction. Finally, according to the decision node the class label will be assigned to the record. This procedure is iterative and repeat till each transaction got a class category. Therefore, this algorithm converts the attributes into a branches and nodes, and select one of the attributes as decision node, which also known as class label. The class label in rapid miner can select while importing the dataset. A decision tree example is shown in Fig. 1.

B. Naïve Bayes

The next classifier used in this study is known as Naïve Bayes. It is also a supervised learning classification model, which classify the data by computing the probability of independent variables. After calculating the probability of each class, the high probability class do assign for the complete transaction [7]. Naïve Bayes is a common approach used to predict classes for different types of dataset such as educational data mining [32] and medical data mining [18]. This model also useful for classifying different kind of dataset like sentiment analysis [33] and virus detection [34]. It works by using the values for independent variables and predict a pre-defined class for each record. It measures the probability of A given that B as shown in following equation. Then working on finding out the distinct class for each attributes, in this scenario all other variables are not dependent on each other [18]. Naïve Bayes uses the following equation for measuring the probability:

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)} \qquad [34]$$

C. Random Forest

Random forest is the next model selected and implemented in this research. As this model is from classification family, therefore it is also known as supervised learning algorithm. During the learning phase, this model first generates multiple random trees called a forest [35]. For example, a dataset contains "x" number of attributes, it first selects some feature randomly known as "y". Using all features; (i.e. "y"), it produces nodes using best rift method. Furthermore, the algorithm will work for creating a complete forest by repeating the previous steps. Then during the prediction process, the algorithm tries to combine the trees using estimated outcome and voting procedure [36]. The purpose of merging the random trees through voting in a forest is to opt out the highest forecasted tree, which can enhance the prediction accuracy for future data.

D. Logistic Regression

Logistic regression is another kind of classification model, which learn and predict the parameters in the given dataset using regression analysis [7]. The learning and prediction processes are based on measuring the probability of binary classification. Logistic regression model requires class variable that should be binary classified. Likewise, in this dataset the "target" column has the two type of binary numbers, "0" for the patient who has no chances of heart failure, and "1" for the patients who has predicted as heart failure patients. On the other side, the independent variables can be of binary classified, nominal or polynomial types [37]. The equation of logistic regression is as follows:

$$Logit(p) = ln\left(\frac{p}{1-p)}\right) = \frac{prob.\ of\ presence\ of\ characteristics}{prob.\ of\ absence\ of\ characteristics} \quad [37]$$

E. SVM

The final machine learning algorithm used in this research is known as support vector machine. This is also called a supervised machine learning model where the classes in the dataset should be pre-defined [7]. It works by categorizing the objects in the given dataset according to the predefined classes. It classify the transactions by assigning one or more classes to maximize the performance in accuracy [38]. Previously, SVM has implemented on medical data application to predict the accurate class for the heart disease patient [39]. Another model proposed by [40] to predict a class using attribute extraction method.

VI. Implementation

This research used five machine learning models, using predictive approach to forecast the chances of heart failure in a patient admitted in the hospitals. Therefore, as described above the dataset taken from Kaggle having patients' records, which have been collected from multiple locations. The dataset has a list of 14 attributes, which collectively used for diagnosing the heart disease in a patient. For experiment execution, the Rapid Miner tool used in this study. Rapid Miner, is an open source software, which provides a wide range of pre-programmed operators for numerous tasks related to machine learning, data mining, statistical and others [41].

The proposed algorithm as presented in previous section, this study used five ML models; Naïve Bayes, Decision Tree, Random Forest, Logistic Regression, and SVM. At the first step of implementation, the training dataset used to learn the ML model. For this, the dataset was imported using "*Read_CSV*" operator in Rapid Miner. Furthermore, to connect the dataset with ML models, it was copied five times. To avoid similar values selection during model learning and testing phase, 10-fold Cross Validation operator was used. It helps to divide the data into k equal subsets and to give a chance for each subset to be a part of training and testing phase. The working of cross validation operator considers as an efficient, as it repeats the learning phase k times, where every time the testing data selection is different from previous. Finally, it repeats the experiment k times and uses the average results. The cross validation is widely used operator for learning and testing purpose. It provides the data selection in four different ways; liner sampling, shuffled sampling, stratified sampling, and automatic [42]. Whereas, the stratified sampling is used in this study. The experiment execution in rapid miner is shown in Fig. 2.

As shown in Fig. 2, the model name is representing on each cross-validation operator. This operator computes the statistical analysis and model performance of a learning and testing phase. The cross-validation operator is also called a nested operator, which have two types of sub-processes; training sub-process and testing sub-process. The training sub-process is used to handle the training session by learning the model through given dataset and model, while testing sub-process is used to validate the model and estimate the performance of the model, which also known as model accuracy. For clear understanding about sub-processes, the validation operator's for Naïve Bayes is shown in Fig. 3, while the remaining operators used the same strategy. To maintain the experiment quality and to know the exact accuracy, all models connected with the same dataset and executed at the same time.

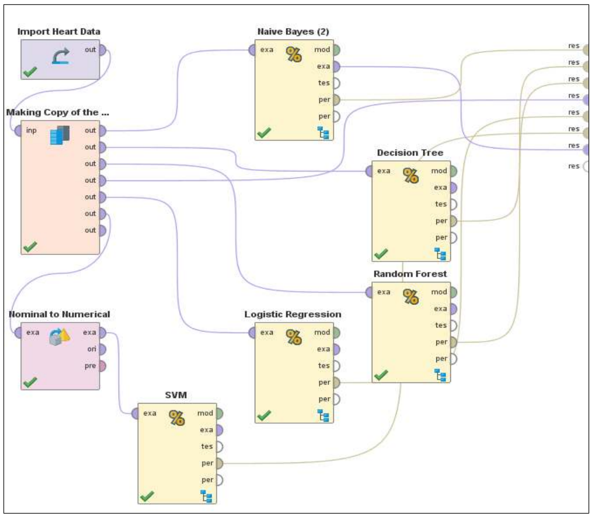

Fig. 2. The Process of Model Implementation in Rapid Miner.

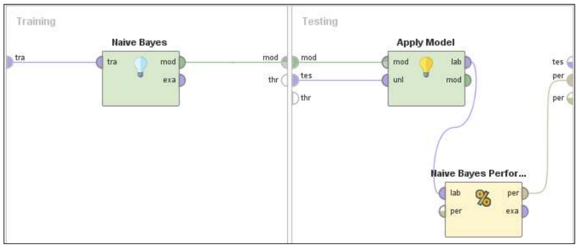

Fig. 3. Training and Testing Sub-Processes.

VII. Discussion on Model Performances and Comparisons

The model performance in the form of confusion matrix is displayed in Table III. A confusion matrix is a table used for describing the performance of a classifier that executed on given test data where the "True" values are considered known data values. In this table, the True Class (1) means the known values for the class category (1); the patients having chances of heart failure. On the other side, True Class (0) denotes the known values for class category (0); the patients showing healthy sign. In the same way, the rows values illustrating the prediction computed for both classes. Accordingly, based on the True values and predicted values, the class precision and class recall values computed and presented in the table. The class recall and class precision values are helpful to identify the overall accuracy of the classifier. As per the displayed values in the table, the precision and recall values for decision tree classifier are maximum, while Naïve Bayes computed minimum among all.

There were total 1013 number of patient's record in the given dataset. For example, the first classifier Naïve Bayes is showing that 434 patients were known in the dataset as heart failure patient and predicted correctly under the class category (1). However, 57 records were initially belonging to heart failure patients but predicted wrongly under the category (0); non-heart failure patient. In the same way, originally total 522 patients were non-heart failure patients and 450 were predicted correctly and 72 estimations recorded as wrong. Overall, the Naïve Bayes classifier performance was the lowest and the performance of Decision Tree classifier computed highest, among all classifier.

Table IV presented the comparison of the experiment's results conducted in this study with the previous work. Overall, every classifier has shown good performance in this study as compare to the previous work. According to the table, the accuracy of the decision tree model was the highest between all models, while the performance of the Naïve Bayes has shown the lowest accuracy in this study. The best two model in our experiment are known as SVM and decision tree. Every model significantly enhanced the performances in previous work and shown the satisfactory enhancement, which is greater than 85%.

In comparison with the previous work illustrated in the third column of Table IV, the research applied the experiment using five algorithms [7]. That research used the same dataset (UCI) with feature selection approach. The accuracy of our model has improved the performance of the classifiers. For example, Naïve Bayes accuracy increased 3%, Logistic Regression and Random Forest enhanced 5%, Decision Tree improved the accuracy ratio 11%, and lastly the SVM machine learning classifier increased 8%. However, in previous work, the study also used the same platform, which is Rapid Miner. But accuracy performance augmented in our work might be because of using 10-fold cross validation, while the previous work used 5-fold cross validation approach. More iteration during the learning phase can help to generate more accurate results. Another possible reason behind the positive enrichment in the accuracy is the size of the dataset, which has been amplified in this study as discussed in data overview section. It highlights that the large size of the dataset can create positive impact on classifier accuracy as it enhances the learning process.

TABLE. III. Model Performances through Confusion Matrix

Naïve Bayes	**True (1)**	**True (0)**	**Class Precision**
Prediction (1)	434	72	85.77%
Prediction (0)	57	450	88.76%
Class Recall	88.39%	86.21%	
Decision Tree	**True (1)**	**True (0)**	**Class Precision**
Prediction (1)	458	36	92.71%
Prediction (0)	33	486	93.64%
Class Recall	93.28%	93.10%	
Random Forest	**True (1)**	**True (0)**	**Class Precision**
Prediction (1)	436	55	88.80%
Prediction (0)	55	467	89.46%
Class Recall	88.80%	89.46%	
Logistic Regression	**True (1)**	**True (0)**	**Class Precision**
Prediction (1)	435	72	85.80%
Prediction (0)	56	450	88.93%
Class Recall	88.59%	86.21%	
SVM	**True (1)**	**True (0)**	**Class Precision**
Prediction (1)	475	62	88.45%
Prediction (0)	16	460	96.64%
Class Recall	96.74%	88.12%	

TABLE. IV. Performance Comparison with Previous Studies

Technique	**This Study**	**[UCI, Rapid Miner, 2019]** [7]	**[UCI, Matlab, 2017] [9]**	**[UCI, Weka, 2017] [9]**
Decision Tree	93.19%	82.22%	60.9%	67.7%
Logistic Regression	87.36%	82.56%	65.3%	67.3%
Random Forest	89.14%	84.17%	X	X
Naïve Bayes	87.27%	84.24%	X	X
SVM	92.30%	84.85%	67%	63.9%

Table IV, column number four and five are related to the another research conducted for predicting and classifying the heart disease patient's records [9]. They used the same dataset "UCI", whereas the main idea of that research was to do the experiment in two different platform; i.e. Matlab and Weka, and then compare the results from both perspectives. Decision tree, logistic regression and SVM were the similar models used in both researches. Altogether, it is evident from the above table that in our study, the decision tree classifier has increased the accuracy more than 30% from the previous work. In the same way, logistic regression and SVM also outperform and were computed the better score than previous work. It illustrates that the performance of the Rapid miner has shown better accuracy and performance of the classifiers.

VIII. Conclusion and Future Work

The ratio of heart failure patients has been increasing every day. To overcome this dangerous situation and deteriorate the chances of heart failure disease, there is a need of a system that can generate rules or classify the data using machine learning approaches. Therefore, this research discussed, proposed and implemented a machine learning model by combining five different algorithms. Rapid miner is the tool used in this research, which computed the high accuracy than Matlab and Weka tool. In comparison with the previous researches, this study has shown significant improvement and high accuracy than previous work. As far as UCI dataset concerns, the dataset needs to be amplified. As the main limitation in this work is the small size of the dataset. The dataset has limited number of patient's records; therefore, the dataset was augmented using appropriate techniques. In future, the results indicated that the system can be useful and helpful for the doctors and heart surgeons for timely diagnoses the chances of heart attack in a patient.

References

[1] M. Gjoreski, A. Gradišek, M. Gams, M. Simjanoska, A. Peterlin, and G. Poglajen, "Chronic heart failure detection from heart sounds using a stack of machine-learning classifiers," in Proceedings - 2017 13th International Conference on Intelligent Environments, IE 2017, 2017, pp. 14–19.

[2] G. Savarese and L. Lund, "Global Public Health Burden of Heart Failure," Card. Fail. Rev., vol. 3, no. 1, 2017.

[3] E. J. Benjamin, P. Muntner, and et al. Alonso, Alvaro, "Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association," Circulation, vol. 139, no. 10, 2019.

[4] M. Ramaraj and T. A. Selvadoss, "A Comparative Study of CN2 Rule and SVM Algorithm and Prediction of Heart Disease Datasets Using Clustering Algorithms," Netw. Complex Syst., vol. 3, no. 10, pp. 1–6, 2013.

[5] A. Gavhane, G. Kokkula, I. Pandya, and P. K. Devadkar, "Prediction of Heart Disease Using Machine Learning," in Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology, ICECA 2018, 2018, pp. 1275–1278.

[6] H. Murthy and M. Meenakshi, "Dimensionality reduction using neuro-genetic approach for early prediction of coronary heart disease," in International Conference on Circuits, Communication, Control and Computing (I4C), 2014, pp. 329–332.

[7] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, "Improving Heart Disease Prediction Using Feature Selection Approaches," in 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), 2019, pp. 619–623.

[8] S. Ismaeel, A. Miri, and D. Chourishi, "Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis," in 2015 IEEE Canada International Humanitarian Technology Conference, IHTC 2015, 2015, pp. 1–3.

[9] S. Ekiz and P. Erdogmus, "Comparative study of heart disease classification," in 2017 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, EBBT 2017, 2017, pp. 1–4.

[10] K. Chen, A. Mudvari, F. G. G. Barrera, L. Cheng, and T. Ning, "Heart Murmurs Clustering Using Machine Learning," in 2018 14th IEEE International Conference on Signal Processing (ICSP), 2018, pp. 94–98.

[11] E. Maini, B. Venkateswarlu, and A. Gupta, "Applying Machine Learning Algorithms to Develop a Universal Cardiovascular Disease Prediction System," in International Conference on Intelligent Data Communication Technologies and Internet of Things, 2018, pp. 627–632.

[12] S. Kodati, R. Vivekanandam, and G. Ravi, "Comparative Analysis of Clustering Algorithms with Heart Disease Datasets Using Data Mining Weka Tool," in Soft Computing and Signal Processing, Singapore: Springer, 2019, pp. 111–117.

[13] K. Deepika and S. Seema, "Predictive analytics to prevent and control chronic diseases," in 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2016.

[14] M. Tabassian et al., "Diagnosis of Heart Failure With Preserved Ejection Fraction: Machine Learning of Spatiotemporal Variations in Left Ventricular Deformation," J. Am. Soc. Echocardiogr., vol. 31, no. 12, pp. 1272–1284, 2018.

[15] T. R. Reed, N. E. Reed, and P. Fritzson, "Heart sound analysis for symptom detection and computer-aided diagnosis," Simul. Model. Pract. Theory, vol. 12, no. 2, pp. 129–146, 2004.

[16] P. Amnarayan et al., "Measuring the Impact of Diagnostic Decision Support on the Quality of Clinical Decision Making: Development of a Reliable and Valid Composite Score," J. Am. Med. Informatics Assoc., vol. 10, no. 6, pp. 563–572, 2003.

[17] C.-S. Lee and M.-H. Wang, "A fuzzy expert system for diabetes decision support application.," IEEE Trans. Syst. MAN, Cybern. B Cybern., vol. 41, no. 1, pp. 139–153, 2011.

[18] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, "Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field," in Machine Learning Paradigms, 2019, pp. 71–99.

[19] K. Shameer, K. W. Johnson, B. S. Glicksberg, J. T. Dudley, and P. P. Sengupta, "Machine learning in cardiovascular medicine: are we there yet?," Heart, vol. 104, no. 14, pp. 1156–1164, 2018.

[20] D. Tomar and S. Agarwal, "A survey on Data Mining approaches for Healthcare," Int. J. Bio-Science Bio-Technology, vol. 5, no. 5, pp. 241–266, 2013.

[21] V. V. Ramalingam, A. Dandapath, and M. Karthik Raja, "Heart disease prediction using machine learning techniques : a survey," Int. J. Eng. Technol., vol. 7, no. 2.8, pp. 684–687, 2018.

[22] UCI, "Heart Disease Data Set." [Online]. Available: https://www.kaggle.com/ronitf/heart-disease-uci. [Accessed: 20-Apr-2019].

[23] S. A. Tiwaskar, R. Gosavi, R. Dubey, S. Jadhav, and K. Iyer, "Comparison of Prediction Models for Heart Failure Risk: A Clinical Perspective," in Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2019, pp. 1–6.

[24] F. Al-Mudimigh, A. S., Ullah, Z., & Saleem, "A framework of an automated data mining systems using ERP model.," Int. J. Comput. Electr. Eng., vol. 1, no. 5, 2009.

[25] A. L'Heureux, K. Grolinger, H. El Yamany, and M. Capretz, "Machine Learning with Big Data: Challenges and Approaches," IEEE Access, 2017.

[26] D. DiCarlo, "Random Number Generation: Types and Techniques," 2012.

[27] R. Pan, T. Yang, J. Cao, K. Lu, and Z. Zhang, "Missing data imputation by K nearest neighbours based on grey relational structure and mutual information," Appl. Intell., vol. 43, no. 3, 2015.

[28] R. Miner, "Outlier Detection Using Rapid Miner." [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/cleansing/outliers/detect_outlier_distances.html. [Accessed: 20-Apr-2019].

[29] R. Kumari and S. K. Srivastava, “Machine Learning: A Review on Binary Classification,” Int. J. Comput. Appl., vol. 160, no. 7, 2017.

[30] R. S, P. Raj H. R., A. C, and V. K, “Medical data mining and analysis for heart disease dataset using classification techniques,” in National Conference on Challenges in Research & Technology in the Coming Decades National Conference on Challenges in Research & Technology in the Coming Decades (CRT 2013), 2013, pp. 1.09-1.09.

[31] J. Shubham, “Decision Trees in Machine Learning,” Medium, 2018.

[32] F. Razaque, N. Soomro, S. Ahmed, S. Soomro, J. A. Samo, and H. Dharejo, “Using Naïve Bayes Algorithm to Students ’ bachelor Academic Performances Analysis,” in Engineering Technologies and Applied Sciences (ICETAS), 2017 4th IEEE International Conference, 2017, pp. 1–5.

[33] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, “Sentiment analysis of review datasets using naive bayes and k-nn classifier,” arXiv Prepr. arXiv, vol. 16, no. 10, 2016.

[34] O. Qasim and K. Al-Saedi, “Malware Detection using Data Mining Naïve Bayesian Classification Technique with Worm Dataset,” Int. J. Adv. Res. Comput. Commun. Eng., vol. 6, no. 11, pp. 211–213, 2017.

[35] N. Donges, “The Random Forest Algorithm,” Towards Data Science, 2018.

[36] S. S. Bashar, M. S. Miah, A. H. M. Z. Karim, A. Al Mahmud, and Z. Hasan, “A Machine Learning Approach for Heart Rate Estimation from PPG Signal using Random Forest Regression Algorithm,” in 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1–5.

[37] H. DW Jr, L. S, and S. RX, Applied Logistic Regression, 3rd ed. New Jersey: John Wiley & Sons, 2013.

[38] V. Vapnik, Statistical Learning Theory. New York: Wiley, 1998.

[39] P. Tabesh, G. Lim, S. Khator, and C. Dacso, “A support vector machine approach for predicting heart conditions,” in Proceedings of the 2010 Industrial Engineering Research Conference, 2010, p. 5.

[40] K. M, S. F, Z. Z, and P. N, “Predicting MOOC dropout over weeks using machine learning methods,” in Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, aclweb.org, 2014, pp. 60–65.

[41] R. M. Team, “Rapid Miner.” [Online]. Available: https://rapidminer.com/.

[42] Rapid Miner, “Cross Validation Operator.” [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/validation/cross_validation.html. [Accessed: 01-Apr-2019].