IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 6, June 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106097|

An Analytical Study on the use of Machine Learning Techniques in Heart Attack Diagnosis and its Risk Prediction

Ms. Harsh Arora

Assistant Professor, Institute of Innovation in Technology and Management Janakpuri, Delhi, India

ABSTRACT: One of the most common diseases nowadays is heart disease. This disease is sort of common nowadays, we used different attributes which may be related to the present heart diseases well to seek out the higher method to predict, and that we are also used algorithms for prediction. Naive Bayes, the algorithm is analyzed on the dataset is based on risk factors.[2] Decision trees and a combination of algorithms for the prediction of the heart condition are used based on some attributes. The results showed that when the dataset is a little naive Bayes algorithm gives accurate results and when the dataset is large decision trees give accurate results. The automatic prediction of heart disease isone of the most vital health problems in the real world.[1] Heart disease affects the functionality of blood vessels and causes arterial coronary infections that weaken the body of the patient, especially adults and old people. The main matter is prognosis using machine learning techniques. Machine learning is widely used nowadays in manybusiness applications like e-commerce and many more. One of the areas where this machine learning used is prediction, our topic is about the prediction of heart disease by processing patient datasets and data of patients to whom we'd like to predict the prospect of occurrence of heart disease. Cardiovascular diseases are often identified by conducting medical tests and using wearable sensors.However, extracting valuable risk factors for heart conditionfrom electronic medical tests is difficult as physicians attempt to quickly and accurately diagnose patients.[3] The integration of the machine learning model presented during this study with medical information systems would be useful to predict the HF or the other disease using the live data collected from patients.

KEYWORDS*:* Machine learning model, medical data, heart failure diagnoses.

I. INTRODUCTION

The right prediction of heart condition is important to efficiently treat cardiac patients before an attack occurs. This ambition is often achieved using an ideal machine learning model with rich healthcare data on heart diseases. Various systems supported by machine learning are presented recently to predict and diagnose a heart condition. However, these systems cannot handle high-dimensional datasets. Thanks to the shortage of a sensible framework which will use different sources of knowledge for heart condition prediction[7]. Heart failure may be a critical condition with a high prevalence (about 2% within the adultpopulation in developed countries, and quite 8% in patients older than 75 years). The costs are very high, reaching up to3% of the total health costs in the developed countries.

Building an efficient disease management strategy requires analysis of an outsized amount of knowledge, early detection of the disease, assessment of the severity, and early prediction of adverse events [11]. Every year almost 26 million people are affected by this kind of disease. From theheart surgeon's point of view, it is complex to predict heart failure at the correct time. Fortunately, classification and predicting models can aid the medical field and can illustrate how to use medical data efficiently. This paper aims to enhance the HF prediction accuracy using the dataset. For this, multiple machine learning approaches will not understand the info and predict the HF chances during a medical database. Furthermore, the results and comparative study showed that the present work improved the previous accuracy score in predicting heart condition. The integration of the machine learning model presented during this study with medical information systems would be useful to predictHF or other diseases using the live data collected from patients. The main aim of this paper is to provide a tool for doctors to detect heart disease at an early stage [5]. This successively will help to supply an effective treatment to patients and avoid severe consequences. ML plays a really vital role to detect the hidden discrete patterns, and thereby analyze the given data.[8] After the analysis of knowledge, ML techniques help in heart condition prediction, and early diagnosis. This paper presents the performance analysis of varied ML techniques like Naive Bayes, Decision Trees, Logistic Regression, and Random Forest for predicting heart condition at an early stage.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106097|

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 isn't for replacing the normal approach use for diagnosing and predicting the probabilities of coronary failure, rather the study attempted to support this process using advanced technologies such as Machine Learning (ML). The ML isn't a replacement technique and has been used several times for various applications. A cloud-based decision support system advised by [11] to helps the heart counselor during the diagnosis process.

This system used machine-learning methods for predicting the guts disease. The system was proposed to supply help in an affordable way, where the system has the capacity to integrate with an existing system. In that research clustering method used for categorizing the dataset based on particular groups in an unsupervised manner. The authors in [12] used an approach by implementing multiple clustering algorithms on the heart condition dataset to know the optimal solution, which may maximize the prediction accuracy ratio.

ML approaches proved to be well organized in foretelling heart disease using past data is further proved in research conducted using Naïve Bayes, Decision Tree, support vector model, and other models [13]. The results indicated that the support vector machines provided the optimal results between other implemented approaches. Bashar et al., (2019) attempted to enhance the performance of heart condition prediction using a feature selections approach. Different models such as Naïve Bayes, Random Forest and blew for the Mode most frequent class for classification problem. The value of k is user-specified other used in the experiment implemented using the Rapid Miner tool. The output indicated the high accuracy measured due to the feature selection approach [7]. Furthermore, the acute Learning Machine techniques using a feed-forward neural network applied on Cleveland data supported 300 patients, suggested 80% accuracy in forecasting the guts disease during a patient [8].In another research, the neural network applied using multi- layer perception, which also referred to as supervised learning. The system was proposed to work out the potential heart condition risk during a patient, using the patient'shistorical data [5].

HF ratio using preserved ejection fractionis alternative work presented using numerous factors like to strain rate, hypertensive situation, and velocity, where overall accuracy computed was more than 80% [14]. The main concept behind this talk is to put stress on how helpful machine learning approaches are to foretelling heart disease using medical data. This research is additionally emphasizing onto overcome the vulnerable situation and proposed a computerized system, in order that heart consultants cannot miss any information thanks to improper reading and understanding of the data. Such a situation described during research, that the majority of the gut'sdiseases wouldn't always detect just by doing ECG [15].Therefore, this type of research overwhelmed those situations where doctors are puzzled and left behind some evidence. To support them, computerized medical system[16] - [19] filled with functionalities are there, which specially built to help the healthcare industry for the sake of patient's time, money, and most importantly to assist the surgeons to save lots of patient's life. A kind of system proposed by [1] using the ML approach for predicting the guts failure using heart sound reports. The study showed another proof that machine learning methods can be applied to life-saving systems like coronary failure detection.

Moreover, a review report presented during a study [20] described the importance of classification models and further explained the small print of the models already implemented within the healthcare industry. The paper highlighted there are many types of research attempted data mining techniques successfully on medical cases. In the same way, another qualified study showed the achievement of the multiple classifiers applied on two different tools; MATLAB and Weka. Overall, the accuracy of the choice tree, Linear SVM, and other models was recorded between 52% to 67.7%, although the accuracy was considerably low [9]. As per the researches discussed above, still different quite models are providing variation within the prediction score. Thus, dimensional reduction and have engineering can improve the method of knowledge selection, which ultimately can improve accuracy estimation [21].

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%

TABLE. I. THE ACCURACY MEASURED IN PREVIOUS WORK

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106097|

In conclusion, the clear research gap found in the previous researches is that the measurement accuracy is not up to the mark. Somewhere, the common machine learning approaches have not used as shown in Table I. Therefore, this section described the great overview of the previous work accompanied to predict the guts disease during a patient using ML approaches. The idea of this study is to enhance the previous work using the chosen dataset and ML models, as described in the next section. The performance of every model is discussed in the result section. Although, the models and dataset selected during this research are supported the previous work. The most commonly ML approaches found and utilized 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 info set originally published on the UCI data repository for machine learning. Previously, the experiments attempted for predicting the guts disease, the small print and accuracy measured is shown in Table I. Finally, within the result section the comparative study is presented to know the performance of the classifiers during this study, and within the previous work.

III. DATA OVERVIEW

The dataset used in this research is collected from the Kaggle platform, the dataset is additionally referred to as heart condition Dataset [22]. Altogether, the info was the combination of 4 a different database, but only Cleveland data utilized in this experiment. it's 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 guts disease during a patient [7], [23]. In addition, the database file contains a record of 304 patients. The complete description of each attribute and therefore, the number of values for every attribute is shown in the table II below:

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

TABLE. II. DATA OVERVIEW AND ATTRIBUTES DESCRIPTION

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106097|

IV. DATA PREPROCESSING

Data pre-processing 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 pre-processing steps were applied to the selected dataset. Firstly, the proportions of the dataset were found insufficient for the implementation of machine learning approaches. As described by [25] the dimensions of thedataset for machine learning implementation may create biases and would also affect the results generated through machine learning models. Therefore, for each attribute using minimum and maximum values, the random number generation technique was applied to generate random values for each column [26]. This helped us to enhance the capacity of the info, which has created a positive impact on the performance of the classifier as can be seen in the resultssection. In conclusion, the data have increased the volume by three times.

Secondly, using the rapid miner, the data cleaning stepapplied to find out missing values and noisy data values.The data has some missing values which have been imputed using K The Nearest Neighbour (KNN) method. As KNN method is proved to be a useful method for missing data imputation [27]. Besides, the outlier detection methods are used to estimate the noise in the data. the information has not found noisy values and no outlier was detected in the dataset. The outlier detection was applied using the rapid miner's operator with the distances method [28]. 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 the 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 is 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 utilized in the given dataset, where the binary classification provided the higher thanks to showing the performance accuracy of the chosen classifier in this study. Most of the attributes were nominal in the selected dataset i.e. Slope, CA, That, and CP.For example, the That attribute is describing the value of theThallium test based on the four predefined values (0, 1, 2,3).In the same way, CP was another independent attribute in the dataset, which highlights the condition of the pain using (0, 1, 2, 3) in the patient at the time of admitting to the hospital, where the ―0‖ means normal and ―3‖ means the worst condition. The Target column in the dataset that is also known as a class attribute has two types of predefined classes known as ―0‖ and ―1‖. This attribute represents 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 a high probability of heart failure. For example, employing 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.[4]

V. THE EXPERIMENT PREPARATION

This research used five different models to predict heart disease using the collect dataset. The performance of each classifier and comparison with previous work is presented inthe next section. After the successful implementation of the data preprocessing steps, this section discussed the chosen models, their descriptions, and the overall methodology used for the experiment. The work presented in this study was a sequel to the research presented by [7], which used an equivalent dataset and implementation tool. That research reduced the number of attributes to 14. The number of records they used for the experiment was 300. The 5-fold cross-validation approach was an appeal to develop 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 enhance the accuracy of the model; therefore, the different amendment was done in the experiment design. For example, data expansion, 10-foldcross-validation, execution of all models at the same time to understand the actual differences in the accuracy measurement. Following are the procedural steps of the 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 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}

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106097|

- Step 5: Result Comparison

{Comparing the accuracy among all models Comparingthe 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. A short description of each model explained in this section.

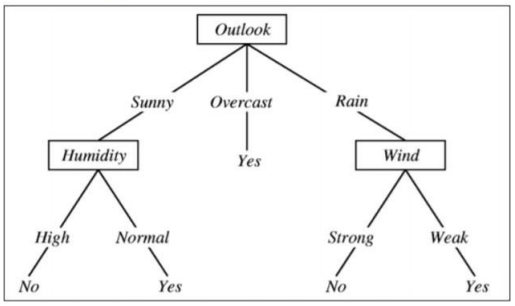

Fig. 1. A Decision Tree Example [31].

A. *Decision Tree*

It's a tree-like classification model, which built a structure consisting of branches and nodes Based on evidence collected for every attribute during the model learning phase[30]. The decision tree's branches and nodes connect according to the number of entities described in the dataset. The advancement procedure uses the number of utility dedicated for each attribute. Furthermore, following the principles described on each branch and node it reached the choice for every 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 branches and nodes and chooses one of the attributes as a decision node, which is also referred to as a 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 analogous used in this course is known as Naïve Bayes. it's also a supervised learning classification model, which classifies the info by computing the probability of independent variables. After calculating the probability of each class, the high probability class does assign for the complete transaction [7]. Naïve Bayes is a common approach used to predict classes for various sorts of datasets like educational data processing [32] and medical data processing [18]. This model also useful for classifying different quiet datasets like sentiment analysis [33] and virus detection [34]. It works by using the values for independent variables and predict a predefined class for each record. It measures the probability of A as long as B as shown in the following equation. Then working on finding out the distinctclass for each attribute, in this scenario all other variablesare not dependent on each other [18]. Naïve Bayes uses the following equation for measuring the probability[34]:

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)}$$

C. *Random Forest*

Random forest is the next model selected and implemented in this research. As this model is from classification family, therefore it's also referred to as a supervised learning algorithm. During the training 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 the best rift method. Furthermore, the algorithm will work for creating an entire forest by repeating the previous steps. Then during the prediction process, the algorithm tries to mix the trees using estimated outcome and voting procedure [36]. The purpose of merging the random trees through voting during a forest is to opt out the very best forecast tree, which may enhance the prediction accuracy for future data.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106097|

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 a class variable that should be binary classified. Likewise, in this dataset, the target column has two type of binary numbers, —0‖ for the patient who has no chances of heart failure, and — 1 ‖ for the patients who have 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[37]:

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

E. *SVM*

The final machine learning algorithm used in this research isknown as a support vector machine. This is also called a supervised machine learning model where the classes in the dataset should be predefined [7]. It works by categorizingthe objects within the given dataset consistent with the predefined classes. It classifies 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 the predictive approach to forecast the probabilities of coronary failure during a patient admitted to 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 heart disease in a patient. For experiment execution, the Rapid Miner tool used in this study. Rapid Miner is open-source software, which provides a good range of reprogrammed operators for numerous tasks related to machine learning, data processing, statistical et al. [41].The proffered algorithm as give out in the previous section, this study used five ML models; Naïve Bayes, Decision Tree, Random Forest, Logistic Regression, and SVM. At thefirst step of implementation, the training dataset want 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 had been copied five times. To avoid similar values selection during the model learning and testing phase, a 10-fold Cross Validationoperator 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 the training and testing phase. The working of cross- validation operator considers as 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. Cross-validation is a widely used operator for learning and testing purposes. It provides the data selection in four, different ways; liner sampling shuffled sampling, stratified sampling, and automatic [42]. Whereas, stratified sampling is used in this study. The experiment execution in a rapid miner is shown in Fig. 2. As shown in Fig. 2, the model's 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 whichhas two types of sub-processes; training sub-process and testing sub-process. The training subprocess is used to handle the training session by learning the model throughthe given dataset and model, while the testing subprocess is used to validate the model and estimate the performance of the model, which also referred to as model accuracy. For a clear understanding of sub-processes, the validation operators for Naïve Bayes are shown in Fig. 3, while the remaining operators used the same strategy. To maintain theexperiment quality and to know the exact accuracy, all models connected with an equivalent dataset and executed atthe same time.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106097|

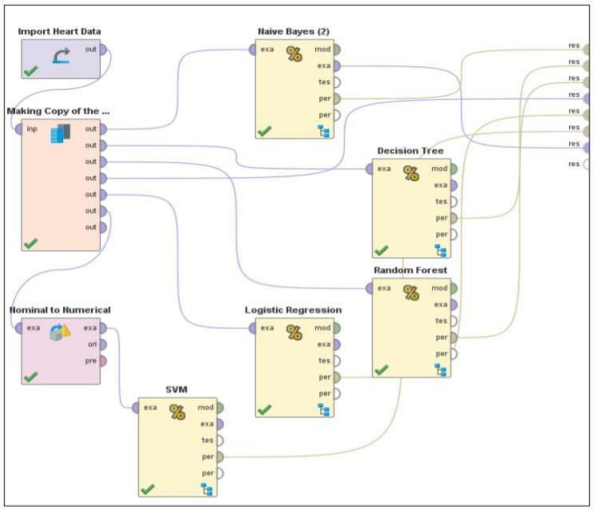

Fig. 2. The Process of Model Implementation in Rapid Miner

VII. DISCUSSION ON MODEL PERFORMANCES AND COMPARISONS

The model performance in the form of a confusion matrix is displayed in Table III. A confusion matrix is a table used fordescribing the performance of a classifier executed on given test data where the true values are considered known data value. In this table, truth Class (1) means the known values for the class category (1); the patients having chances of heart failure. On the opposite side, True Class (0) denotesthe known values for class category (0); the patients showing healthy sign. In the same way, the values of therow 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 within the table. The class recall and sophistication precision values are helpful to spot the overall accuracy of the classifier. As per the displayed values within the table, the precision and recall values for the decision tree classifier are maximum, while Naïve Bayes computed minimum among all. There was a total 1013 number of patient's record within the given dataset. For example, the first classifier Naïve Bayes is showing that 434 patientswere known within the dataset as heart failure patient and predicted correctly under the 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 523 patients were non-heart failure patients and 451 were predicted correctly and 73 estimations recorded as wrong. Overall, the Naïve Bayes classifier performance was the lowest and the performance of the Decision Tree classifier computed highest, among all classifier.

The comparison of the experiment's results conducted during this study with the previous work has been done. Overall, every classifier has shown good performance during this study as compared to the previous work. According to the table, the accuracy of the decision tree model was the highest among all models, while the performance of the Naïve Bayes has shown the lowest accuracy in this study[7]. The best two models in our experiment are referred to as SVM and decision tree. Every model significantly enhanced the performances in previous work and shown satisfactory enhancement, which is greater than 85%. That research used thesame dataset (UCI) with a 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 5.1%, Decision Tree improved the accuracy ratio 11%, and lastly, the SVM machine learning classifier increased 8%.However, in previous work, the study also used an equivalent platform which is Rapid Miner. But accuracy performance augmented in our work could be because of using 10-fold cross-validation, while the previous work used a 5-fold cross-validation approach. More iteration during the training phase can help to get more accurate results. Another possible reason behind the positive enrichment in the accuracy is the size of the dataset, which has been amplified during this study as discussed in the data overview section. It highlights that the massive size of the dataset can create a positive impact on classifier accuracy because it enhances the learning process work and shown satisfactory enhancement, which is greater than 85%. Incomparison to the previous work illustrated within the third column of Table IV, the research applied the experiment using five algorithms [7]. That research used the same dataset (UCI) with a 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 5.1%, Decision Tree improved the accuracy ratio 11%, and lastly, the SVM machine learning classifier increased 8%.However, in previous work, the study also used anequivalent platform which is Rapid

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106097|

Miner. But accuracy performance augmented in our work could be because of using 10-fold cross-validation, while the previous work used a 5-fold cross-validation approach. More iteration during the training phase can help to get more accurate results. Anotherpossible reason behind the positive enrichment in the accuracy is the size of the dataset, which has been amplified during this study as discussed in the data overview section.It highlights that the massive size of the dataset can create a positive impact on classifier accuracy because it enhances the learning process.

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. III. MODEL PERFORMANCES THROUGH CONFUSION MATRIX

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. PERFORMANCE COMPARISON WITH PREVIOUS STUDIES

Table IV, column number four and five are associated with the other research conducted for predicting and classifying the heart disease patient's records [9]. They used the same dataset ―UCI‖, whereas the most idea of that research was to try to the experiment in two, different platforms; i.e. Matlab and Weka, and then compare the results from both perspectives. Decision tree, logistic regression and SVMwere similar models used in both types of research. Altogether, it is evident from the table above that in our study, the decision tree classifier has increased the accuracy by more than 30% from the previous work. In the same way, logistic regression and SVM also outperform and were computed the

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106097|

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

The ratio of coronary failure patients has been increasing every day. To overcome this dangerous situation and deteriorate the chances of heart failure disease, there is a need for a system which will generate rule or classify the info 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 to the previous researches, this study has shown significant improvement and high accuracy than previous work. As far as UCI dataset concerns, the dataset must be amplified. As the main limitation during this work is the small size of the dataset. The dataset has a limited number of patient's records; therefore, the dataset was augmented using appropriate techniques. In the future, the results indicated that the system can be useful and helpful for the doctors and heart surgeons for timely diagnoses of the chances of a 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 neurogenetic 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.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106097|

[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.UCI, ―Heart Disease Data Set.‖ [Online]. Available: https://www.kaggle.com/ronitf/heart-disease-uci. [Accessed: 20-Apr2019].
[22] 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.
[23] 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.
[24] A. L'Heureux, K. Grolinger, H. El Yamany, and M. Capretz,―Machine Learning with Big Data: Challenges and Approaches,‖ IEEE Access, 2017.
[25] D. DiCarlo, ―Random Number Generation: Types and Techniques,‖ 2012.
[26] 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.
[27] R. Miner, ―Outlier Detection Using Rapid Miner.‖ [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/cleansing/outliers/det ect_outlier_distances.html. [Accessed: 20-Apr-2019]. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019 268 | P a g e www.ijacsa.thesai.org [29] R. Kumari and S. K. Srivastava, ―Machine Learning: A Review on BinaryClassification,‖ 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_vali dation.html.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details