International Journal of Advanced Science and Technology
Vol. 29, No. 8s, (2020), pp. 1665-1674

Effective Ensemble Strategies for Predicting the Cardiac Diseases

P.Vasanthi[1], Prathiba Jonnala[2], Ummadi Janardhan Reddy[3]
[1,2,3]*Assistant Professor*
[1]*Department of CSE,* [2]*Department of ECE,* [3]*Department of IT*
[1]*GATES Institute of Technology, Gooty*
[2,3]*Vignan's Foundation for Science, Technology & Research (Deemed to be University), Guntur*

Abstract

Number of people losing their lives due to heart disease is growing day by day, revealing the need of a model which predicts beforehand. An initiative has to be taken to aid the people by giving them a cautionary advice about the disease at the correct time. It is not easy for everyone to afford expensive treatments and medications so there is urgency of a structure which can quickly go through the information of the patient and inform them at an earlier stage if they test positive. We need a logical process that analyzes and finds unrevealed data and figures in the medical data. Thus, we propose to perform the analysis of given dataset by performing the data validation and preprocessing techniques, exploration data analysis visualization and training a model, build a classification model and then performance measurements of supervised machine learning algorithms with evaluation classification report, identify the confusion matrix and categorizing data from priority. The main objective is to make a predictive analytics model to diagnose the various stages of heart patients by ensemble learning methods like Bagging, Boosting and Voting which aims to enhance the accuracy of the deficient algorithms. Outcome of these ensemble techniques are analyzed and the one that proves to enhance the precision is considered and showcased using a GUI.

Keywords: *Heart Disease Prediction, Machine Learning, Classification Algorithms, Ensemble Learning Classifier, Cleveland Database.*

Introduction

Human life is completely dependent on the efficient working of the brain and heart. Heart disease can be caused under different circumstances that lead to a abnormality in the pumping of blood. [1] poor diet, obesity, smoking, alcohol, cholesterol, age, high blood pressure, sex, physical inactivity, family history and diabetes are the determinant conditions that could make a person test positive for heart related disease. [1],[14] cardiomyopathy, arrhythmias, congestive and congenital failure, coronary heart disease, angina pectoris are some of the possible heart disease and it is troublesome job to see the possibilities of obtaining a heart disease for a person based on risk factors single handedly.

Every hospital, healthcare center, diagnostic center generates a huge amount of data such as patient particulars, test reports, etc. The patient detail contains many features which help to predict the heart diseases. The historical medical data needs analytical methods to analyze the data and to extract the potential information from it for which the data mining concepts turn helpful to discover the unrevealed figures, relationships from database and machine learning approaches to further diagnose a patient. The problem domain contains to predict the heart disease presence and provides treatment in early stage. Diagnosis is complicated and sometimes may give incorrect results like the prediction may go wrong ending up with wastage of money for the expensive treatments of the patients.

In this paper we aim to resolve the issue of on-time and accurate prediction of heart disease for a patient by firstly gathering appropriate data and elements related to our study, identifying necessary attributes that could aid us in the system and [18] see the usual patterns that could be helpful in selection of the model and parameters. To achieve using machine learning classification methods like Naive Bayes, Decision tree, Support Vector Machines, Random Forest, K-Nearest Neighbors and Logistic Regression that could give precise results for every new insert. Evaluate and analyze statistical and visualized results, which find the standard pattern for all regiments. The main objective is to make a predictive analytics model to diagnose the various stages of heart patients by ensemble learning methods like Bagging, Boosting and Voting which aims to enhance the accuracy of the

ISSN: 2005-4238 IJAST

deficient algorithms. We further predict the outcome of a patient whether he is likely to suffer from a heart disease or not using a GUI application. In this system, Cleveland dataset is used.

Literature Survey

In this paper they propose a model to identify the features used in the [1] prediction with a hybrid method of ML algorithm using the R Studio rattle. Combinations of attributes keep on repeating during the feature selection and modeling. To achieve better results they compare [5] the precision, error of classification, accuracy support, F-measure, sensitivity and specificity. The algorithm has resulted in a accuracy of 88.7%. They also mentioned that a mixture of machine learning algorithms could further be used along with new feature selection methods to increase the performance.

In this paper intention is to predict the presence of heart disease with the basic symptoms [6] using neural networks which proves to be more accurate and reliable comparatively. Using multi-layer perceptron , an algorithm of neural networks to further train and test the dataset. It consists of three layers for the input, output and one or more for hidden layers between these two layers. The nodes of the input layer each are connected to output nodes via the hidden layers. PyCharm IDE using python code is used to develop this system. For future work they have suggested to use technologies like big data, data mining strategies, fuzzy, cloud and machine learning.

The study tell us that there [3] needs a advance system to give a better treatment so they have proposed a system which utilizes back propagation algorithm in artificial neural network in MATLAB of R2015a. The system was trained to tell us the absence or presence of heart stroke with precision of 95%. They concluded [2] that the model can be developed as a model which is hybrid with other classification techniques to diagnosis in a better way.

In this paper proposes to use the different [4] machine learning algorithms namely Naive Bayes, Gradient Boosting, Random forest, Support Vector Machine, logistic regression and classifier. The dataset used in this paper is UCI machine learning repository containing attributes such as cholesterol, age, sex, etc. A total of 14 attributes are chosen out of a total of 76 attributes to train the model. The model is then trained using R language. The accuracies obtained is as follows: Logistic regression = 0.8651685, Random forest = 0.8089888, Naïve Bayes = 0.8426966, Gradient boosting = 0.8426875, SVM = 0.7977528

This research proposes a [5] model that uses python programming language to predict the probability of the occurrence of heart disease. The paper makes use of two machine learning algorithm: Naive Bayes Algorithm and Decision Tree Algorithm. The accuracy obtained from decision tree algorithm is 91 % and Naive Bayes Algorithm results in a accuracy of 87%. The dataset used is Cleveland's dataset containing 76 attributes out of which 14 appropriate attributes are chosen such as age, cholesterol, sex etc.

The research study proposes using [6] Ensemble Machine Learning method for classifying and predicting Gene Expression Data. The various ensemble ML techniques used are Bagging Boosting and Arcing. The consecutive accuracy obtained from these methods are 94.4%, 91.7% and 88.9%. The data used to build the model is Gene Expression dataset (Golubet aI.).

The paper proposes to build a model that uses four different machine learning algorithms to predict heart disease using [7] additional weighted fuzzy rules, which contributes in increasing the accuracy of a model. The resulting accuracy for the model is: Logistic Regression (86.1%), SVM (83.87%), Decision Tree (70.97%), and Random Forest (77.42%).

The paper instructs on building a model that [8] predicts a coronary heart disease in a patient and uses the algorithms, Decision tree and Naive Bayes Classifier for the prediction process. The model after training obtains an accuracy of 91% for Decision Tree and 87% for Naive Bayes Classifier Model. The dataset used here is Cleveland's Machine Learning data.

ISSN: 2005-4238 IJAST

This research proposes a model that makes use of [9] data mining techniques and Machine Learning methods for classifying and predicting for Heart Disease. The paper makes use of Support Vector Machine and Artificial Neural Networks for the prediction and classifying process. The resulting accuracy for SVM is 85.6% while that of Artificial Neural Network is 83.3% .The data set used in the model is Cleveland's Repository.

This r This research proposes a model which uses Data Mining and Hybrid Intelligent Techniques in order to predict Heart Disease in a patient. The data set used here is Cleveland's UCI repository. [10] Hybrid intelligent techniques are used in order to improve the accuracy and hence the usefulness of the model in Prediction processes. The accuracy obtained with the model is 89.3%.

Proposed Model

A. Data Source

Every newly entered patient detail filled during hospital appointments are used as the dataset further. These training datasets are used to predict the presence of a particular disease after testing.

Heart Disease Dataset used here is the Cleveland database available at UCI which is used by various testers. [21].

The Cleveland dataset is widely used for the prediction of heart disease. 303 instances and 76 attributes are available in the dataset, but only 14 of them are used. All the related work in this study is done using only 14 attributes of the dataset [3]. These 14 attributes mentioned in Table I give insightful information about the patient.

B. Software Requirements

Operating system used here is Windows, having installed the tool Anaconda navigator and launching the Jupyter Notebook platform and using Python language to work on this dataset. Libraries like sk-learn, pandas, numpy, matplotlib and seaborn have been imported to help do each task.

Table 1: Attributes used

Attributes	Description
age	age in years
sex	(1 = male; 0 = female)
cp	chest pain type (4 values)
trestbps	resting blood pressure (in mm Hg on admission to the hospital)
cholserum	cholesterol in mg/dl
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecgresting	electrocardiographic results (values 0,1,2)
thalach	maximum heart rate achieved
exang	exercise induced angina (1 = yes; 0 = no)
oldpeakST	depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment
ca	number of major vessels (0-3) colored by fluoroscopy
thal	3 = normal; 6 = fixed defect; 7 = reversible defect
target	1 or 0

Table 1 shows the attributes and their description used in the dataset.

C. Architecture Diagram

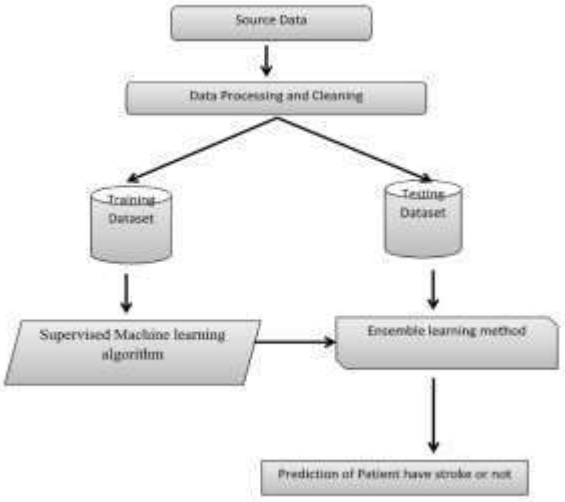

Figure 1: Architecture Diagram

The above figure 1 explains the workflow of this system that is to take the Cleveland dataset and perform the following

1. Data validation process and pre-processing techniques.
2. Exploration data analysis visualization process and train a model.
3. Performance measurements of Supervised Machine learning algorithms.
4. Performance measurements of Ensemble method.
5. Showing the output with GUI application.

D. Description

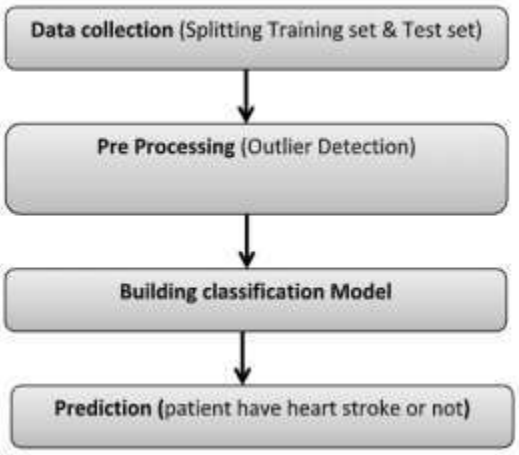

Figure 2: Data Flow Diagram for Machine Learning Model.

Figure 2 explains how each machine learning model is created in an orderly manner. Each of the process included in building a model are explained further.

Splitting the dataset- [26] dataset is split into training data and second into testing data. The train data outcome is known and the models study's it so that it can be applied onto another data. Here the dataset is split in the ratio 67:33 / 70:30. This test dataset is then used as a template to verify with various mentioned algorithms and compare multiple algorithms.

Data Wrangling- It will check whether the dataset is clean and then trim and clean the data for further analysis.

Data analysis for variable exploration- The data from the dataset is checked for missing and null values. For this we import various library packages to perform data cleaning. We perform certain DQM checks to ensure the data consistency.

Single variate analysis of data- the data now obtained is saved in the proper data format to avoid inconsistencies. Exploratory data analysis of bi-variate and multivariate- The cleaned and pre-processed data is plotted in the form of bar chart, histogram, etc. The data is then split into training and testing datasets.The training dataset is then used to train the model using machine learning algorithms. After the model has been trained, the testing dataset is fed to the model.

Data validation techniques are used to get the error rate of the machine learning model, which can be considered as close to the true error rate of the dataset. It helps you know your dataset better and to choose an appropriate algorithm to build the model. The library packages are imported to clean the data such as removing the missing values, and to access other useful functionalities and tune models accordingly.

Data Visualization is the graphical way of representing data in order to get insights into the data such as patterns, outliers etc. The advantage of data visualization is that it is easily understood by naive users.

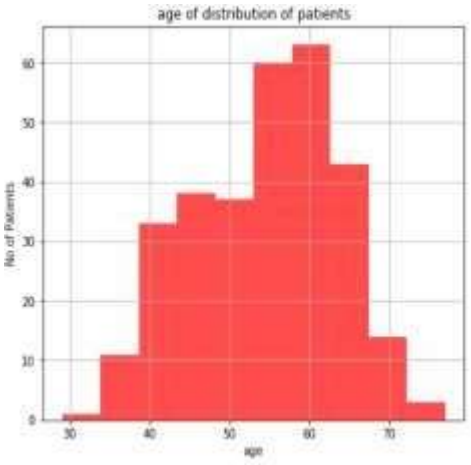

Figure 3: Age of distribution of patient

Fig 3 Shows the distribution of number of patients according to their age for gaining qualitative understanding about the dataset .

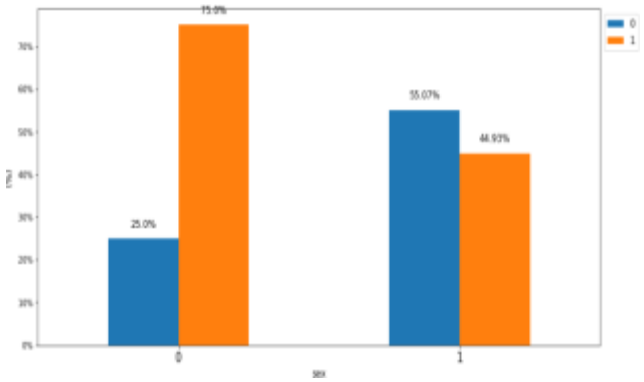

Figure 4: Percentage value of gender having heart disease or not.

Figure 4 gives an insight about percentage probability of a gender to have heart disease or not. 0 describes for not having heart disease and 1 for having heart disease. The colour orange depicts of a patient having heart disease whereas blue depicts that of not having.

Six various supervised machine learning algorithms like Support Vector Machine, Random Forest, Naïve Bayes, Logistic Regression, Decision Tree, K-Nearest Neighbour are used for prediction of heart disease. Each of these models will give different results and the accuracy may not always be the same always. Hence we use ensemble learning methods like Bagging, Boosting and Voting further to improve the currently obtained accuracy and give a reliable system and predict the same with the highest accuracy algorithm in a GUI.

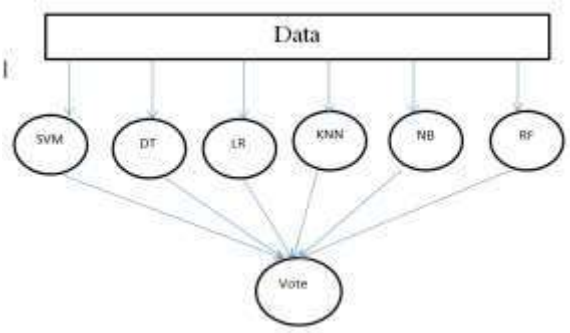

Figure 5: Shows the ensemble structure

Figure 5 shows how ensemble learning is implemented where it first considers all the 6 supervised machine learning algorithms and then one by one implement ensemble techniques on the precise model to further test on all samples of the dataset and in the end take a vote of the majority.

Algorithms Used

In this study we make use of the [9] various supervised machine learning algorithms such as: Support Vector Machine, Random Forest, Naïve Bayes, Logistic Regression, Decision Tree, K-Nearest Neighbour for prediction of heart disease. First we take the dataset and split it into test and train dataset in the ratio 67:33 / 70:30. This test dataset is then used as a template to verify with various mentioned algorithms and compare multiple algorithms. Each of these models will give different results and the accuracy may not be the same. Hence we use ensemble learning methods like Bagging, Boosting and stacking further to improve the currently obtained accuracy and to make a system which is more reliable using Python with scikit-learn.

E. Logistic Regression

Logistic Regression is an algorithm of machine learning that tells whether and event will take place or no (0 or 1) i.e a binary classification algorithm. An easier method of prediction is provided by this model which gives the base scores of accuracy and compares with other models not involving any assumptions to the form [16]. The outcome of a variable is categorical or binary and it requires large sized samples. [17] P(Y=1) predicts as a function of X where Y is linearly related to the log odds in the mathematically interpreted Logit function.

F. Decision Tree

Decision tree [16] builds a model in the form of a flowchart that takes the whole training set as the root node and then breaks down into smaller subsets by the application of simple decision rules .[10] These smaller subsets are called branches that are the outcome against each test node and leaf nodes that represents the final decision. They deal with categorical and numerical data.

G. Random Forest

Random forest is a classifier that considers N random records from the dataset and builds many decision trees based on these records and can be used for classification, regression and other tasks. It then outputs the class by individual trees that is the mode of the class's output. It is an ensemble learning method that is believed to give accurate results among the algorithms [8].

H. Support Vector Machine

Support Vector Machine classifies the data using a hyper plane. Therefore it takes the labeled data and provides a hyper plane that classifies the data as the output.[18].It is an effective way to solve the problem of high-dimensional space [15].

I. Naïve Bayes

Naïve Bayes [14] is a statistical classifier which considers all the attributes to be independent known as conditional independence. It makes calculations simpler, easier and faster for large datasets comparatively and [15] is based on Bayes theorem.

J. K-Nearest Neighbour

K- Nearest Neighbor is an instance-based learning and non-parametric method. It is widely used for classification and regression containing the input of k closest training neighbors. K-NN classification is done by voting by its neighbors. Every new instance is to be checked to using some distance metric like Euclidean distance. K can have any value 1, 2 an so on [15].

K. Ensemble learning method

Ensemble learning methods [14] are used to increase the results of machine learning algorithms. It does so by using multiple learning algorithms together for the very same task. This technique provides a higher predictive performance than that of a single model. In this a combination of various diverse model sets are used in order to improve the predictiveness and stability of the model.[11].

- Reducing Overfitting
- Higher accuracy (lower errors)
- Single model overfits.
- Reduction in variance and bias
- Results worth the extra training
- It can be used for classification as well as regression

For eg when a single model overfits like a Decision tree usually does, we can then use a Random forest or we can use esemble of multiple similar models to give us a better fit. The difference in accuracy between the single model and the constructed ensemble model is worth the extra training done.

Firstly you construct base classifiers and base models and then perform voting on them like averaging and weighted average for the ensemble learning.

Bagging- The technique called Bagging [7] is an ensemble machine learning algorithm. The purpose of Bagging technique is to improve and increase the accuracy of the various machine learning algorithms for Regression and Classification purposes. Bagging also known as Bootstrap Aggregation reduces fluctuation of data and helps in overcoming over fitting. It uses multiple random training sets with the concept of sampling with replacement. Each new set contains the original set with few changes. These sets are then taken from the data and trained with the classifier and with voting result is selected [11].

Boosting- The technique called Boosting [7] is a meta-algorithm that is used to reduce the bias and variance from supervised machine learning technique, it aims to convert weak learners to string ones. [6] Boosting works by re-weighting the previous training sets based on the error rates the base classifier used previously.

Voting- Voting is a technique where in many models are used to make predictions for each point of data in the dataset. Each prediction from every model is seen as a vote. Then finally this ensemble learning model concludes with the model that majority of the prediction were proved right. That model is considered is the final model with highest accuracy for future predictions. It is a very direct method which immediately gives output .It takes into consideration different ML models and then wraps the models to take the aggregate of the predictions of these ML models. It can later be used to make a prediction on some other data.

Performance Evaluation

On comparing the resulting performance of various supervised machine learning techniques from given healthcare department dataset with evaluation classification report we have achieved the following accuracies with Logistic Regression giving the highest of 86.81 % .

Table 2: Comparison Of Accuracies

Sr.No	Algorithm Used	Accuracy
1	Logistic Regression	86.81
2	Decision Tree	74.73
3	Random Forest	79.12
4	Support Vector Machine	54.95
5	Naïve Bayes	81.32
6	K-Nearest Neighbour	65.93

Table 2 compares all the obtained accuracies of the Supervised Machine Learning Algorithms. And checks for the precision and their classification reports.

Figure 6: Accuracy score for supervised machine learning algorithms

Figure 6 shows With Logistic Regression giving the highest accuracy , on applying ensemble learning techniques like Bagging, Boosting and Voting individually has resulted in the following accuracy in below table.

Table 3: Comparison Of ensemble Accuracies

Ensemble Methods	Enhanced Accuracy
Bagging	89.655
Boosting	92.155
Voting	96.551

Table III concludes with voting ensemble technique showing a drastic improvement of nearly 10% over supervised Logistic Regression we have achieved a more reliable and accurate system for the

healthcare department to rely on. Further using GUI application the model predicts with the most accurate ensemble method and concludes whether a person is affected with heart disease or not. GUI application uses Voting Classifier to check the presence of heart disease as we put in the attributes of a significant person and check for the result.

Figure 7: Sample GUI Page

Figure 7 shows a sample GUI page where in we need to insert the patient specific attributes to predict the result with Voting Classifier Method.

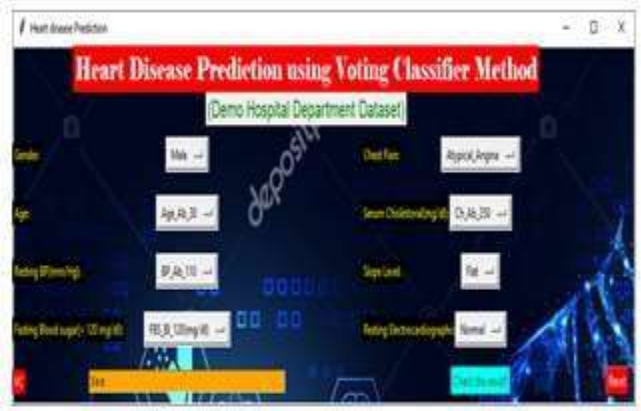

Figure 8: Sample GUI Page with test output predicting a person has heart disease.

Figure 8 shows a output run on GUI for a specific patient details.On checking the result for these entered attributes the GUI predictes YES implying that the person has Heart Disease.

Conclusion

In this research paper a process of analyzing started from cleaning of data and then processing, finding missing values and finally performing an evaluation of the supervised machine learning classifiers has been formed to further perform ensemble learning method to improve the accuracy of the learning algorithms using Python with scikit learn. Algorithms like Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Naïve Bayes, K-Nearest Neighbour are assessed for prediction of heart disease where Logistic Regression outperformed by giving a better accuracy of 86.81%. Ensemble techniques Bagging, Boosting and Voting have been analyzed. The best accuracy on test set is given by Voting Classifier technique with 96.551% hence by giving early diagnosis many lives could be saved.

References

[1]. Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava, *"Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,"* IEEE published ,vol 3, July 3 2019.

[2]. Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Prof. Kailas Devadkar, *" Prediction of Heart Disease Using Machine Learning ",* IEEE Conference Record # 42487; IEEE Xplore ISBN:978-1-5386-0965-1 (ICECA 2018)
[3]. Tülay Karayılan and Özkan Kılıç *" Prediction of Heart Disease Using Neural Network ",* IEEE-Published 2017.
[4]. Dinesh Kumar G, Arumugaraj K, Santhosh Kumar D, Mareeswari V , *" Prediction of Cardiovascular Disease Using Machine Learning Algorithms ",* Publisher: IEEE, Proceeding of 2018 IEEE International Conference.
[5]. Mr Santhana Krishnan. J , Dr Geetha. S, *" Prediction of Heart Disease Using Machine Learning Algorithms ",* 26 April, 2019.
[6]. Ching Wei Wang, *"New Ensemble Machine Learning Method for Classification and Prediction on Gene Expression Data",* Proceeding of 2016 IEEE International Conference.
[7]. P. K. 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. doi: 10.1016/j.jksuci.2011.09.002.
[8]. A. S. Abdullah and R. R. Rajalaxmi, *"A data mining model for predicting the coronary heart disease using random forest classifier"* in Proc. *Int. Conf. Recent Trends Comput. Methods,Commun.Controls,*Apr.2012,pp.22–25.
[9]. Rairikar, A., Kulkarni, V., Sabale, V., Kale, H., & Lamgunde, A. (2017, June). " Heart disease prediction using data mining techniques. " In 2017 *International Conference on Intelligent Computing and Control (I2C2)* (pp. 1-8). IEEE.
[10].J.Vijayashree and N.Ch.SrimanNarayanaIyengar, *"Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques : A Review " ,* Vol.8, No.4 (2016), pp. 139-148
[11].C. Beulah Christalin Latha, S. Carolin Jeeva *," Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques ",Informatics in Medicine Unlocked* 16 (2019)
[12].Burak Kolukisa, Hilal Hacilar, Gokhan Goy, Mustafa Kus, Burcu Bakir-Gungor, Atilla Aral, Vehbi Cagri Gungor," Evaluation of Classification Algorithms, Linear Discriminant Analysis and a New Hybrid Feature Selection Methodology for the Diagnosis of Coronary Artery Disease ",2018 *IEEE International Conference on Big Data (Big Data)*
[13].M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang*,"Disease prediction by machine learning over big data from healthcare communities",* IEEE Access, vol. 5, 2017.
[14].Aditya Methaila, Prince Kansal, Himanshu Arya, Pankaj Kumar, *" Early Heart Disease Prediction Using Data Mining Techniques ",* 2014.
[15].Seyedamin Pouriyeh, Sara Vahid, Hamid Reza Arabnia and Giovanna Sannino, *"A Comprehensive Investigation on Comparison of Machine Learning Techniques on Heart Disease Domain ",* July, 2017
[16].An Dinh, Stacey Miertschin, Amber Young and Somya D.Mohanty, *"A data-driven approach to predicting diabetes and cardiovascular disease with machine learning ",* 2019.
[17].Pahulpreet Singh Kohli, Shriya Arora, "Application of Machine Learning in Disease Prediction",*Proceeding of 2018 IEEE International Conference.*
[18].S. U. Ghumbre and A. A. Ghatol, "Heart Disease Diagnosis Using Machine Learning Algorithm," Advances in Intelligent and Soft Computing Proceedings of the *International Conference on Information Systems Design and Intelligent Applications, January 2012.*.
[19].Dhiraj Dahiwade, Prof. Gajanan Patle, Prof. Ektaa Meshram*, "Designing Disease Prediction Model Using Machine Learning Approach ",* IEEE, ICCMC 2019
[20].M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang*,"Disease prediction by machine learning over big data from healthcare communities",* IEEE Access, vol. 5, 2017.
[21].Heart disease Dataset-WWW.UCI Repository. com