Mathematical Statistician and Engineering Applications
ISSN: 2326-9865
https://doi.org/10.17762/msea.v71i1.40

Predictive Hybrid Approach Method to Detect Heart Disease

Dr. Surendra Kumar Yadav
Professor
Poornima College of Engineering
Jaipur,India
surendra.yadav@poornima.org

Yati Chouhan
Research Scholar
Poornima College of Engineering
Jaipur,India
yatichouhan@gmail.com

Manish Choubisa
Assistant Professor
Poornima College of Engineering
Jaipur,India
manish.choubisa@poornima.org

Article Info
Page Number: 36 – 47
Publication Issue:
Vol 71 No. 1 (2022)

Abstract

Data Mining has indicated a capable result in prediction and detection of Heart illness, the data removal methods are widely applied which are used for predictions, identification in different type of heart diseases. The discipline of this process is to extract the data and information from the large dataset by combining methods of artificial intelligence and database management. Here a Proposed method is used to predict the better approach followed by others. According to the comparative learning of different data removal method are applied which are used as predictions of heart diseases and similar medical problems. Diverse types of algorithms which are tied with data removal which has helped in educating the performance in medical domain this paper comprises of proposed approach of hybrid method to determine the Heart Syndrome based on artificial neural network. As the first approach followed by normal ANN (artificial neural network) which consists of vast arrangement of layers and is applied on the dataset. This gives a view to use neural networks. Here the accuracy is calculated by applying hidden layer of auto encoder surrounded by the layers of neural network which is high comparable with others .Considering and comparing both the cases without auto encoder and with auto encoder is to be calculated. The performance is more robust as compared with the other methods and literature papers. The results calculated in this hybrid approach gives better and optimized solution.

Keywords: - data mining, Machine Learning, auto encoder, Neural Network.

Article History
Article Received*: 18 November 2021*
Revised*: 01 December 2021*
Accepted*: 15 December 2021*
Publication*: 27 January 2022*

I. INTRODUCTION

Data mining has shown a capable result in forecast and discovery of disease, the data removal technique widely applies for predictions, identification and for different types of heart diseases. Data mining can be suitable method to support medical professionals in detecting the disease by obtaining the information and knowledge regarding the disease and symptoms from patient's data set [1].Information removal techniques comprises of the hidden methods to produce awareness in organization surroundings. This can support too broadly to get recovered from the functioning along with the excellence of medical decision. According to proportional study different data mining techniques were applied as predictions of heart diseases and similar medical problems [2].Hence there are different data mining algorithms which are to be used and are compared in terms of their higher efficiency. Some areas in which it had been used vastly are:

A. Data modeling for physical condition care applications.
B. Administrative Information System for health care.
C. Prediction management costs and requirement of resources.
D. Public fitness information processing.[3].

There are different approaches which are to be used with higher accuracy which are used on numerical and continuous data set. To search out the high quality out- come with higher presentation. Various approaches such as Naïve Byes algorithm, Decision tree classifier model, Random forest, Logistic regression, Voting classifier ,SVM, K neighbors classifier and ANN are to be used [4].ANN is used in order to improve the performance learning and to provide high accuracy and lower error rate [5].

II. RELATED LITERTAURE

In this related literature, there are different studies that deals with the different approaches used in order to detect the empathy illness symptoms with techniques of information removal which helps in illustrating the different approaches used to detect the Heart illness symptoms using information extracting algorithms. Hence the best way to analyze the literature review is in the form of table. As elaborated in Table1 which contains different papers summary in appropriate year and the algorithms used with accuracy mentioned with them. First literature view is of the paper published in 2016 whose title is given as "Long term Kidney illness study is using Data removal sorting technique" in this paper the main objective of this article work is to predict Long term Kidney illness (CKD) using sorting procedure method. Similar to Naive Byes as well as Artificial Neural Network (ANN).The investigational outcome implemented by using quick miner tool which shows to facilitate Naive Bayes generates more precise outcome than Artificial Neural system [6].

Second paper is of the year 2017, whose title is named as "Calculation of Heart illness with k-means and Artificial Neural system as fusion approach" to get better accuracy. The center of the article is based whose principle in order to serve well-organized prediction method to resolve and remove the unfamiliar facts of heart illness using fusion arrangement of K-means clustering algorithm and artificial neural system. Here, UCI data set is to be used which makes grouping of a variety of attributes, it also uses k-means algorithm in addition to predict its uses Back propagation method in neural networks. The key purpose of this paper is to expand a sample for predicting of heart illness with high precision rate [7].

Next paper is of 2019, year whose title is "Calculation and analysis of Heart illness Patients are using Data removal method" which aim on diverse heart illness and to make all probable safety measures to stop at

ISSN: 2326-9865
https://doi.org/10.17762/msea.v71i1.40

early phase itself with reasonable time. Here 'Data removal' method has the feature which are fed into SVM, forest algorithm, KNN, and artificial neural system taxonomy Algorithms for the calculation of heart illness. In the beginning context and methods there are different techniques used in order to know the outcome to perceive heart illness at starting stage and can be totally cured by using precautions [8]. Another paper was published in 2019 whose title is "Enhanced sparse auto encoder based artificial neural system" this method follows for prediction of heart illness also this paper comprises of two phase manner which is planned to successfully compute heart disease. The primary stage involves teaching a better sparse auto encoder (SAE), an unverified neural system, to study the finest depiction of the preparing data. The next stage involves with a use of artificial neural system (ANN) to forecast the health position based on the academic report. The SAE was optimized so as to teach well-organized model. The tentative result shows that the planned technique improves the presentation of the ANN classifier, and is further tough as compared to additional methods [9].

III. Different Data Mining Algorithms

A. Naïve Bayes

It illustrates the view of all taking part feature for the expected place. Raw Byes or Byes parameter is the first peak for numerous numbers of devices to learn and statistics removal procedure. The rule is to construct duplication by means of uncertain capability. It gives descriptive performance to find out within productive figure. Naive byes algorithm is considered as the fastest and accurate one among all ML algorithms, it performs and gives good results in multiclass predictions it used the concept of bayes theorem which is used for calculating conditional probabilities [11].

B. Decision Tree

Conclusion trees are measured to be the central Classification algorithm that increases the expansion of data mining. Accepted conclusion hierarchy algorithms secure Quinlan's ID3, C4.5, C5 and Breiman CART. Considering as the forename suggest recursively split explanation in branches to build a tree for the principle of refining to calculate exactness. This is also used to predict the decision about some conditions. This model consists of the hierarchical structure containing one root node and leaf node in this the dataset consist of labels which is defined in structure [12].

C. Random Forest

Random forest as the name suggests, it is used generate a forest in a random way surrounded by the forest. RF is preferred because of the following properties such as

- It has precision comparable or improved than Ad boost.
- It is comparatively vigorous to outliers and noise,
- It is faster than boosting,
- As more trees are extra to the forest, the inclinations to over fit decrease

Random forest is the fast boosting method it is supervised machine learning algorithm in which the input is given and the output is produced according to the behavior of the input given [13].

D. *Artificial Neural Network*

Artificial Neural network works similarly as near to human brain. Theoretically artificial neural method is stimulated by neural system our brain contains millions of nerves like in neural system also there are different layer on which it works with the collaboration of machine learning. ANN takes in numerous inputs and produces a distinct output [14].The neurons and nodes has weights attached to them each node have a weight which is calculated in last. There are many diverse algorithms used to instruct neural networks with too many option.ANN also comprises of biological structure in which there is biological representation consists of dendrites, cell body [15].ANN is the vast topic in which contains millions of layers functioning between then with a hidden layer also there is bias function included into it. Here is the depiction an artificial neural system (ANN) to get a number of fair thought on how neural system function works. Basically there are three layers in the neural network [16].

- Input layer
- Hidden layer
- Output layer

IV. MATERIALS AND METHODS

A. *Illustrations of dataset*

As there are diverse data sets intended for heart illness here in this article the dataset which is used is Heart Statlog Cleveland Hungary dataset. This dataset consists of 1190 records of patient from different parts of nation like US, UK, Switzerland and Hungary. It has 11 attributes and 1 goal variable. Target variable is set to 0 or 1

The attributes used in the paper are 12 which are described as:

Era –Victim era in days (digits)

Gender- Victim femininity boy as 1 feminine like 0 (supposed)

chest pain category -Type of upper body pain classify into 1 distinctive, 2 distinctive angina, 3 non-angina pain, 4 asymptomatic

hidden BP - stage of blood anxiety at hidden form in mm/HG (Numerical)

Cholesterol -Serum cholesterol in mg/dl (digits)

Fasting blood sugar -Blood sugar levels on fasting > 120 mg/dl symbolize as 1 inside holder of accurate and 0 as fake

Latent ECG - result of electrocardiogram while at rest are represented in 3 distinct values 0

Max heart velocity -high heart rate attain (digits)

Train angina - Angina induced by apply 0 show No 1 and show fair enough (supposed)

Elder crest- work out encourage ST-anxiety in contrast with the condition of relax (digits)

B. *Illumination of Algorithm*

i. Import Libraries
ii. Load Dataset of heart disease (CSV file)
iii. Preprocessing on dataset is applied
iv. Importing the class called SimpleImputer from impute model in sklearn (Simple Imputer).

Mean is calculated (np.nan, strategy =mean)
This step allows the handling of missing data

v. Apply hot encoder and label encoder, helps for handling categorical data
{
One hot encoder=One Hot Encoder ()
Fit transform (value of column to be fitted)
Add back to original frame apply loop from arrange (data. shape [1])
Dropping the column name and then print (column)}

vi. Label encoding
Preprocessing of labeling the encoder
Fit the label in which want to transform (x).

vii. Divide the set into instruct and trying data
{
X (train, test),Y (train, test) }

viii. Feature scaling transform (train ,test) datasets

End preprocessing

ix. Apply Auto encoder layer in NN(neural network)

{
Create the R layers in which input layer of neural network
Encode the dense layer by using activation _ rectified linear unit function
Apply for loop up to n to encode and decode the layers.}
Return input and output layer
Calculate the parameters used in encode and decode of layers

x. Fit the auto encoder (epoch ,batch size)

xi. Creation of hidden layer of auto encoder in neural networks
(Adding the layers of auto encoder from 1 to n)

xii. Build neural network
Train the model
Add layer by using dense.

Loop is applied in order to repeat the layers, RELU (rectified linear activation unit),epoch and sigmoid functions are added into model. After this the accuracy is achieved is 91% after the validation and applying auto encoder into it. Here the dataset is trained in order to use for deploying the algorithms on tool used as Anaconda and Google collaborative.

V. PREPROCESSING ON DATASET

Here the set of preprocessing on data is to be applied which makes the raw data to clean dataset. Now our contained dataset is not certain so through this the method of preprocessing is applied in order to make certain dataset. For data preprocessing script Anaconda Navigator is used.

- Importing the libraries

Firstly, all the libraries are to be imported by using pip install. Then the data is to be read from our dataset i.e. Heart Statlog Cleveland Hungary as in comma-separated values file which means that the dataset has been imported.

- Checking the Missing Data

This method involves all the removal of missing data contained in heart disease dataset. Now this method might cause some issues because our dataset contains large records so our data can get lost the alternative and best method is to take the mean of the columns of heart disease dataset. Here np.nan is used which signifies that we are targeting missing values. Here the missing data in our dataset is taken the mean.

- Handling of Categorical Data

In this method our data is divided into categories. In this hot encoder and label encoder are to be imported and applied on 1st column of dataset. The data in our data set is categorized as 0 and 1.Like in sex there are male and female to be categorized.

- Splitting the dataset into teaching and trying sets

In the direction to check any information mining algorithm, accuracy needed to be checked. In order to do that the information set is separated into teaching and trying dataset as the name suggests our specified data set is to be trained. So our information set is separated into X training set plus Y training set similarly for testing both X and Y. Here the function train_test_split is used on our Heart Statlog Cleveland Hungary dataset. In this 20% of the data is test and 80% is train.

- Characteristic Marking

This technique is use to standardize the collection of autonomous uneven and featured information. In the used dataset Standard Scale function is used applied on X train and X test. In our data set if the range is going out it makes into put into scale.

VI. Results and Performances

In order to display the efficiency and presentation of the features educated by our proposed method used is auto encoder, first we trained an ANN using the raw data then ANN layer is added with hidden layer of auto encoder between them this auto encoder is used to compressed the representation of raw data and are also known as data de noising. This idea of auto encoder comes from the base paper named as “Improved sparse auto encoder based artificial neural network approach for prediction of heart disease” in which the accuracy achieved is of 91% which works on SAE i.e. sparse auto encoder technique .The SAE model is used to optimized as well as to train the data set, the dataset used in this is Framingham dataset Here the performance of ANN classifier is more vigorous as compared to others. Likely considered in this article firstly the dataset is preprocessed in order to make suitable for our model after preprocessing the data is split into testing and training dataset the dataset selected is named as Heart Statlog Cleveland Hungary dataset .In this method the model is to trained and dense layer is to be created applying rectified linear activation function and activation function into it. Now a new function of auto encoder comes the next step and the second most steps to apply is of hidden layer of auto encoder into it in which the preprocessing of data is to be done and auto encoder is applied which makes the compression of Raw data. Here after preprocessing and applying auto encoder max Accuracy during validation achieved is 92.50% here there is graphical view shown in Figure 1 in which the auto encoder loss curve is shown between loss and value loss. Another Figure 2 shows the graph between plotting of train and testing accuracy. Whereas Red lines in the graph

shows train loss and blue lines shows value loss. Here the table2 shows all the algorithms compared with their accuracy in which the highest and optimized we get is of ANN. Ann gives a better accuracy when applied to the dataset while comparing with others. Also ROC curve for each algorithm is shown in figure 3, 4,5,6,7 of different mining algorithms used.

Here, the table 2 comprises of different data mining schemes such as Naïve bayes, SVM classifier, decision tree, random forest, logistic regression and voting classifier with the accuracy mentioned with them. Table 3 comprises of the approach used in the article i.e. proposed approach in which the accuracy gained by ANN is 83% and by applying the proposed approach it is extended as 92.50% .As shown in figure 9, the bar graph between the hybrid method and ANN with the SAE calculated and shown in figure. Here the base paper in which there is the concept of SAE named sparse auto encoder in which the concept of auto encoder is used the concept is applied as hidden layer of auto encoder in our approach. Also figure 8 shows the graph between the comparisons of three of them.

A. Figures and Tables

a) Here table 1 comprises of comparing the algorithms.

TABLE I. COMPARISON OF DIFFERENT LIERTAURE REVIEW

Author name	**Literature Review**		
	Dataset used	***Performance Measurement***	***Year***
Veenita Kunwar, KhushbooChandel , A. Sai Sabitha, Abhay Bansal [6]	UCI Machine Learning Repository	Naïve bayes - 100% , ANN -72%	2016
AmitaMalav, Kalyani Kadam, Pooja Kamet [7]	UCI Heart stroke	88%,93% hybrid - 97%	2017
Mamatha Alex P and Shaicy P Shaji [8]	Jubilee Mission Medical College and study foundation Thrissur	Support Vector Machine(SVM)- 85.88%, Random forest- 85.88%, KNN- 83.21% , ANN- 92.21%	2019
Ibomoiye Domor Mienye, Yanxia Sun , Zenghui Wang [9]	Framingham, Massachusetts.	ANN - 91%	2020
Dr. Mohammed Ismail, Dr .K.	Movie Lens Dataset	K and clustered	2018

Author name	Literature Review		
	Dataset used	***Performance Measurement***	***Year***
Bhanu Prakash, Dr. M. Naga bhushana Rao [10]		algorithms - 90%	

b) Accuracy Results Calculated by different data mining algorithms

TABLE II. ACCURACY RESULTS OF MINING ALGORITHMS

Mining Algorithm Used	Model Accuracy
SVM Classifier Model	86.88%
Naive Bayes Classifier Model	87.89%
Decision Tree Classifier Model	78.68%
Random Forest Classifier Model	81.96%
K Neighbors Classifier Model	81.96%
Logistic Regression Model	85.24%
Voting Classifier Model	86.88%

c) Accuracy Compared with Proposed Approach

TABLE III. RESULT OF PROPOSED APPROACH

Algorithm name	**Accuracy**
Artificial neural network(ANN)	83%
SAE (sparse auto encoder)[9]	91%
Proposed Approach	92.50%

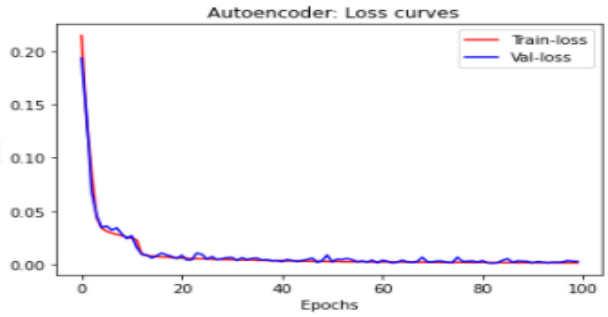

Fig1: Auto encoder loss curves between train and value loss

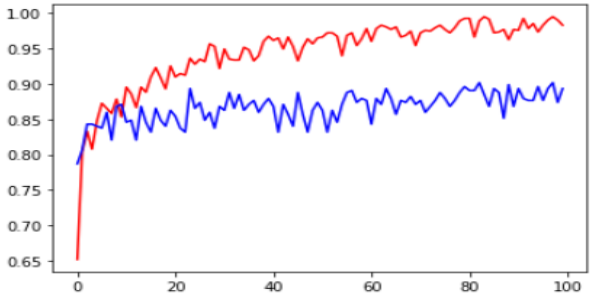

Figure2: Plotting of train and testing accuracy

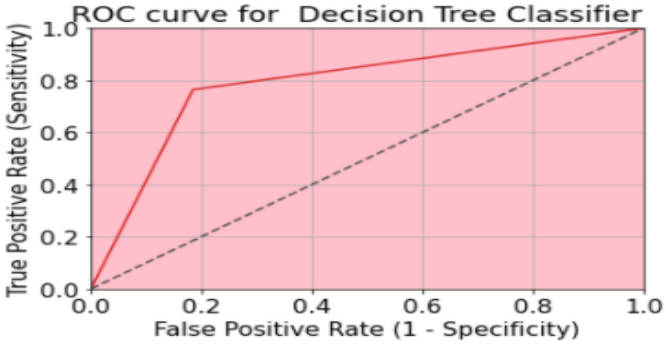

Figure 3: ROC curve of Decision tree algorithm between sensitivity and specificity

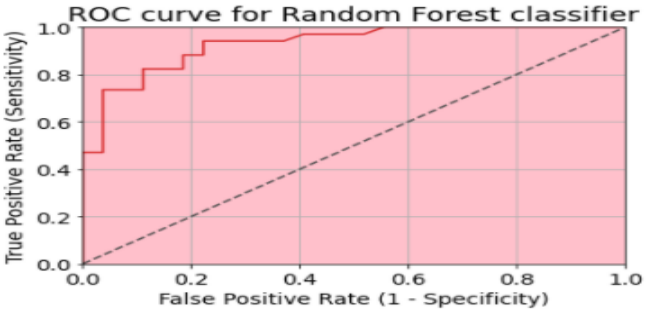

Figure 4: ROC curve of Random forest algorithm between sensitivity and specificity

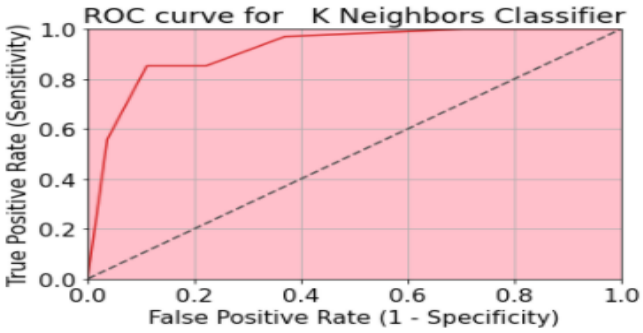

Figure 5: ROC curve of K Neighbors classifier algorithm between sensitivity and specificity

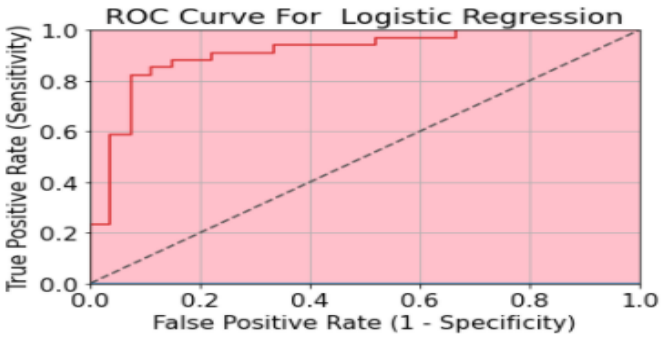

Figure 6: ROC curve of Logistic regression algorithm between sensitivity and specificity

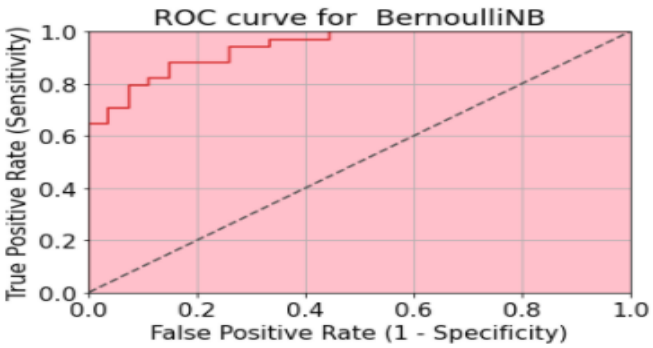

Figure 7: ROC curve of Bernoulli NB algorithm between sensitivity and specificity

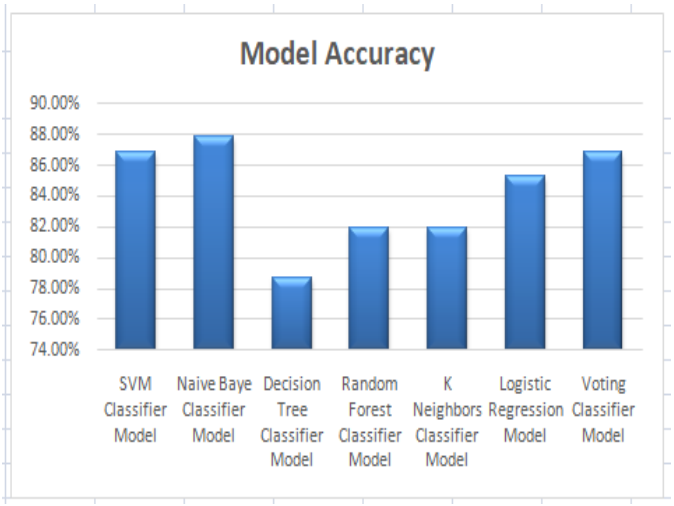

Figure 8: Model Accuracy between algorithms of Data mining

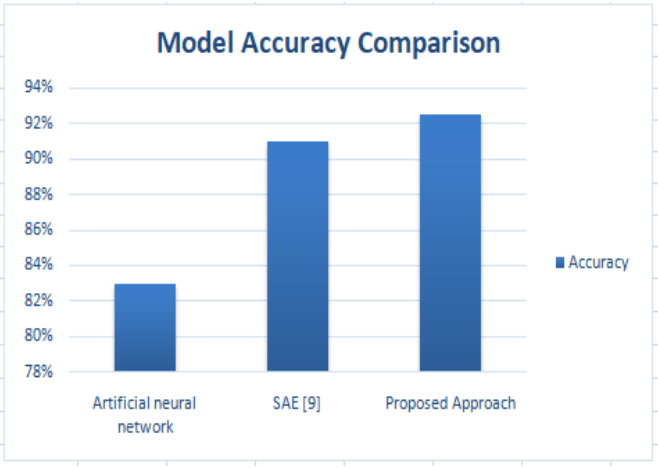

Figure 9: Model accuracy comparison of our proposed approach used

CONCLUSIONS AND OUTLOOK WORK

In this paper there is computation of different algorithms such as Naive Bayes Classifier, Decision Tree Classifier, Random Forest Classifier, K Neighbors Classifier, Logistic Regression, voting model classifier based on the accuracies we found highest we got is of 88%which is of voting classifier. By all these accuracies the conclusion we get is Machine learning performs better in these situations. Many researchers suggested and they work on different layers of neural network and the summary is to achieve higher accuracy with best and efficient method. Here in this paper, we applied a hidden layer of auto encoder on 12 attributes of heart illness dataset too which can be used by applying other efficient and optimized method in order to gain more accuracy of system. Also, by changing the dataset our accuracy differs and which method is used. Also, add boost function can be used in order to enhance the performance which includes different methods and libraries.

Declarations

Conflict of interest: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

[1].Sharma, Sumit, and Mahesh Parmar. "Heart diseases prediction using deep learning neural network model." International Journal of Innovative Technology and Exploring Engineering (IJITEE) 9.3 (2020).

[2].Wu CS, Badshah M, Bhagwat V. Heart disease prediction using data mining techniques. In Proceedings of the 2019 2nd international conference on data science and information technology 2019 Jul 19 (pp. 7-11).

[3].Ashwini Shetty A, Chandra Naik, "Different Data Mining Approaches for Predicting Heart Disease", International Journal of Innovative in Science Engineering and Technology, Vol.5, May 2016, pp.277-281.

[4].]. DeFreitas, Kyle, and Margaret Bernard. "Comparative performance analysis of clustering techniques in educational data mining." IADIS International journal on computer science & Information systems 10.2 (2015).

[5]. Musleh, Musleh M., et al. "Predicting Liver Patients using Artificial Neural Network." International Journal of Academic Information Systems Research (IJAISR) 3.10 (2019).

[6]. Kunwar, Veenita, et al. "Chronic Kidney Disease analysis using data mining classification techniques." 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence). IEEE, 2016.

[7].Malav, Amita, Kalyani Kadam, and Pooja Kamat. "Prediction of heart disease using k-means and artificial neural network as Hybrid Approach to Improve Accuracy." International Journal of Engineering and Technology 9.4 (2017): 3081-3085.

[8]. Shaji, Shaicy P. "Predictionand Diagnosis of Heart Disease Patients using Data Mining Technique." 2019 international conference on communication and signal processing (ICCSP) IEEE, 2019.

[9].Mienye, Ibomoiye Domor, Yanxia Sun, and Zenghui Wang. "Improved sparse auto encoder based artificial neural network approach for prediction of heart disease." Informatics in Medicine Unlocked 18 (2020): 100307.

[10]. Lakshmi, K. Naga, et al. "Design and Implementation of Student Chat Bot using AIML and LSA." International Journal of Innovative Technology and Exploring Engineering 8.6 (2019): 1742-1746.

[11]. Krishnan, Santhana, and S. Geetha. "Prediction of Heart Disease Using Machine Learning Algorithms." 2019 1st international conference on innovations in information and communication technology (ICIICT). IEEE, 2019.

[12].Shekar, K. Chandra, Priti Chandra, and K. Venugopala Rao. "An ensemble classifier characterized by genetic algorithm with decision tree for the prophecy of heart disease." Innovations in computer science and engineering. Springer, Singapore 2019. 9-15.

[13].Li, Liam, and Ameet Talwalkar. "Random search and reproducibility for neural architecture search." Uncertainty in artificial intelligence. PMLR,2020.

[14].Mienye, Ibomoiye Domor, and Yanxia Sun. "Performance analysis of cost-sensitive learning methods with application to imbalanced medical data." Informatics in Medicine Unlocked 25 (2021): 100690.

[15].Hossain, Adiba Ibnat, et al. "Applying Machine Learning Classifiers on ECG Dataset for Predicting Heart Disease." 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI). IEEE, 2021.

[16].Tarawneh, Monther, and Ossama Embarak. "Hybrid approach for heart disease prediction using data mining techniques." International Conference on Emerging Internetworking, Data & Web Technologies. Springer, Cham, 2019.