ISSN: 0974-5823 Vol. 7 No. 2 February, 2022

International Journal of Mechanical Engineering

Heart Disease Prediction System using Machine Learning Techniques

Mrs.T.Sujithra

Assistant Professor, Department of Computer Science,

Mannar Thirumalai Naicker College, Madurai.

Mrs.P.Rajeswari

Assistant Professor, Department of Commerce with Computer Applications,

Mannar Thirumalai Naicker College, Madurai.

Mrs.M.Muthulakshmi

Assistant Professor, Department of Computer Science,

Mannar Thirumalai Naicker College, Madurai.

Abstract

In this paper, we carried out research on heart disease using classification techniques. Prediction of heart disease is the recent research topics in the medical and computer science fields. We used rank based feature selection for identifying the most important features of heart diseases which is main causes of heart diseases. Then we applied artificial neural network techniques on data sets of different sizes, in order to study the accuracy and stability of proposed system. Found neural networks are easier to configure and obtain much good results. The obtained results are also compared with benchmark algorithms.

I Introduction

The main cause of death worldwide is heart disease. Every year, more than 1.6 million individuals die of heart attacks. The phrase "heart disease" encompasses a wide range of cardiac conditions. Cardiovascular disease is by far the most prevalent form, which can lead to a heart attack. Other forms of cardiac illness may affect the heart's valves, or the heart can fail to pump sufficiently, leading in heart failure. Heart disease is a condition that only some people are born with. Heart disease may affect anybody, even youngsters. It occurs when a material known as plaque accumulates in your arteries. Heart disease is increased by smoking, poor dietary habits, and a lack of exercise. Heart disease can also be exacerbated by excessive cholesterol, high blood pressure, or diabetes. There are various natural techniques for preventing this condition, including as smoking cessation, weight control, eating a nutritious diet, and exercising on a regular basis. There are additional scientific techniques available, such as medications and surgery. Prediction of this disease before infection is one of the preventative approaches, and computer technologies, namely Machine Learning algorithms, are the most usually applied ways in this [ ]

The process of classifying data components using class labels is known as classification. It is the technique of guided learning. It's also the process of identifying the model that explains and differentiates data types and their values. On the basis of model training, classification is the challenge of determining which of a set of classes a new observation belongs to. In general, classification is a two-step procedure that includes learning (training) and classification (testing). The training phase entails the creation of a model using a variety of algorithms. For accurate outcomes prediction, the resulting model must be trained. The testing process involves putting the built model to the test with previously unknown data. The testing process also used to find the performances of any classification system. The rest of this paper is organized as follows; section II discusses the heart disease prediction related works, existing methods and techniques, section III discusses the overview of the proposed methods, section IV shows that the experimental results and performance analysis and section V concludes the proposed works and future enhancements.

II Related Study

Machine learning, often known as data mining, is effective for a wide range of issues. Predicting a dependent variable from the values of independent variables is one of the uses of this method. The healthcare industry is a data mining application area since it has large data resources that are difficult to manage manually. Even in wealthy countries, heart disease has been identified as one of the leading causes of mortality. The risk factors for heart disease were deemed to be sex, age, smoking, hypertension, and diabetes (Xiao et al., 2017).

The decision support system was suggested by Cherian et al., (2017) for swiftly identifying cardiac problems in a patient. As a result, both money and time are saved. The decision support system assists clinicians in diagnosing patients without

introducing unwelcome practise differences owing to the doctor's intuition or inexperience. Because the forecast is based on a historical database including a huge number of cardiac patient records, their method gives a second opinion on the patient's status, similar to that of an experienced doctor.

The approach developed by Vembandasamy et al., (2015) is based on the Navie Bayes algorithm. A good prediction of heart disease among patients was made using the Nave Bayes classification system. According to their findings, the Naive Bayes algorithm achieves 86.4198 percent accuracy in less time.

Monica et al. (2016) suggested a cardiovascular disease analysis. To forecast the sickness, they offered data mining approaches. They hoped to present an overview of existing strategies for extracting data from datasets that would be valuable to healthcare practitioners.

To predict heart disease, Mujawar et al., (2015) employed k-means and nave bayes. They created the system with the help of a historical cardiac database that provides diagnostics. The 13 qualities have been taken into account when creating the system. To extract knowledge from a dataset, data mining techniques such as clustering and classification algorithms are utilised. The classifier was built using the Cleveland Heart Database, which has 13 characteristics and 300 entries.

Heart Disease Prediction System Using Data Mining Techniques was proposed by Shahi et al., (2017). They employed Weka software to automate illness detection and assess service quality in healthcare facilities. They employed SVM, Nave Bayes, Association rule, KNN, ANN, and Decision Tree, among other techniques. They discovered that compared to other data mining algorithms, Nave bayes classification is more effective and gives higher accuracy as a result of their study.

There are a variety of data mining approaches that may be used in automated heart disease prediction systems. This (Bhatla et al., 2012) paper defines many strategies and data mining classifiers that have evolved in recent years for efficient and effective heart disease detection. According to the results, the Neural Network with 15 characteristics has the best accuracy (100%) so far. Using 15 qualities, Decision Tree, on the other hand, has also done well, with 99.62 percent accuracy. Furthermore, Decision Tree has showed 99.2 percent efficiency when combined with Genetic Algorithm and 6 characteristics.

Seema et al., (2016) focused on applying Nave Bayes, Decision Trees, Support Vector Machines (SVM), and Artificial Neural Networks to predict chronic illness by mining data from previous health records (ANN). A comparison analysis of classifiers is conducted to determine which performs better on an accuracy rate. SVM had the best accuracy rate in this trial, but Nave Bayes has the highest accuracy rate for diabetes.

With the use of machine learning, Khourdifi et al. (2019) compared the algorithms with different performance indicators. K-Nearest Neighbour, Random Forest, Nave Bayes, and Artificial Neural Network have been shown to produce the best outcomes based on their observations. The suggested models were evaluated using a heart disease dataset, and the techniques yielded a 99.65% accuracy rate.

Saini et al. (2019) suggested an Internet of Things platform for cardiac disease prediction. They used wearable devices to collect data from patients' bodies, which was then stored and processed utilising communication standards. The storage of the massive amount of data created by wearable sensor cloud storage is the health care application.

Mohan et al. (2019) proposed a hybrid random forest with linear model approach, which they called hybrid random forest with linear model. The Random Forest and Linear Model were combined. The innovative technique was able to predict cardiac illness with a high degree of accuracy. They were able to reach an accuracy of 88.7%.

Nikhar et al., (2016) aimed to present a full description of the Nave Bayes and decision tree classifiers used in research, specifically in the prediction of heart disease. A comparison of the execution of predictive data mining techniques on the same dataset was undertaken, and the results revealed that Decision Tree surpasses Bayesian classification.

Some machine learning techniques support vector machine (SVM), decision tree (DT), Naïve Bayes (NB), K-nearest neighbor (KNN) and artificial neural network (ANN) are used for the prediction of the occurrence of heart diseases which were analyzed and compared based on the obtained prediction accuracy. According to their findings, the ANN technique had the best average prediction accuracy (86.91%), while the C4.5 decision tree technique had the lowest average prediction accuracy (74.0%). (Riyaz et al., 2022).

E-Hasnony et al., (2022) have proposed the multi label active learning based model for the heart disease prediction. They have utilized five selection methods MMC, Random, Adaptive, Quire and Audi. They have used heat maps for selecting most significant features that cause the heart disease. All the selection methods were tested on the heart disease data set by 10-fold cross validation. According to their findings, the learning model could generalize based on the sample data with an accuracy and F1-score of 57.4% and 62.24%.

Verma et al., (2021) have proposed hybrid model for the prediction of heart disease. They have applied genetic algorithm based feature selection method to select the most important features from the given dataset. After that, the training of the model has been done with Ensemble deep neural network model. The result of proposed method was enhanced by Adam optimized. The obtained accuracy is 98% as it higher than the benchmark algorithms.

A hybrid deep learning model for the heart disease prediction have developed by the combination of multiple gated recurrent units (GRU), long short term memory (LSTM) and Adam optimizer which produces the highest accuracy of 98.6876% (Krishnan et al., 2021).

The proposed OANN (Optimal Artificial Neural Network) consists of two key processes: distance-based misclassified instance removal (DBMIR) and the teaching and learning based optimization (TLBO) algorithm for ANN, collectively referred to as OANN (TLBO-ANN).

A Big Data framework such as Apache Spark was used to create the proposed model. The OANN model consists of two phases: offline prediction and online prediction. The benchmark heart disease datasets will be employed to train a model and perform testing during the offline prediction stage. Similarly, at the stage of online prediction, real-time data will be streamed into an Apache Spark model, and the filtered data will be assessed using such a trained model to obtain prediction results (Thanga selvi et al., 2021).

Priyanka et al., (2020) have proposed a novel hybrid recurrent neural network (RNN)-logistic chaos-based whale optimization(LCBWO) for predicting the heart disease within 5 years of patient's records. They have applied multilayer bidirectional LSTM for the feature selection. They have obtained accuracy of 98%, specificity of 99%, precision of 96%, correlation coefficient of 97%, F1 measure of 0.9892 and ROC values of 98%.

Kavitha et al.,(2021) have proposed a novel machine learning approach for the prediction of heart disease using Cleveland heart disease dataset. They have used random forest and decision tree algorithms for obtaining the better performance in diagnosing of heart disease. From their experimental results it has been identified that the random forest and decision tree based combined approach provides 88.7% of accuracy.

Rajdhan et al., (2020) have proposed a work by exploring the four classification algorithms such as random forest, decision tree, logistic regression and naïve bayes. They have analyzed the performance of 4 classification algorithms. They have obtained the accuracies of decision tree as 81.97%, logistic regression as 85.25%, random forest as 90.16% and naïve bayes as 85.25%. they have concluded that the random forest algorithm is the most efficient algorithm with accuracy score of 90.16% for the prediction of heart disease.

Jagtap et al., (2019) have developed a web based machine learning application using UCI dataset. They have implemented three classification algorithms namely Support Vector Machine (SVM), Logistic regression and Naïve bayes. Training of classification algorithm has been done with 75% of data set and remaining 25% for testing the accuracy of the algorithm. They have observed that Naïve bayes had 60% accuracy, Logistic regression had 61.45% and SVM had 64.4%. They have suggested that the SVM is the most efficient algorithm for the web application.

Garg et al., (2021) have analyzed the performances of the supervised learning algorithms like k-Nearest Neighbor (k-NN) and random forest. They have suggested that chest pain, cholesterol level and age of the person for considering the main causes of heart disease. They have obtained accuracies of 86.885% and *7.967% for k-NN and random forest respectively.

III Proposed methods

The proposed method is based on two major parts the pre-processing phase where the system choose the most appropriate attributes and the second one train the neural networks in order to identify the risk factors of heart patients.

Pre-processing

In machine learning challenges, feature selection is one of the most significant strategies. After selecting features, the majority of learning approaches perform effectively. Over fitting, noisy data, and extensive training times may all be avoided by using feature selection. As a result, it's critical to use feature selection methods before training the model. Feature selection can improve the robustness of these algorithms by identifying the most relevant and best-representing feature subset. For example, the combination of feature selection approaches with a machine learning algorithm improved classification accuracy.

Feature selection strategies benefit the model by reducing training time, making it less complicated, simpler to read, improving accuracy, and reducing over-fitting issues. For identifying the significant features for selecting the heart illness, the researcher used the Relief feature selection approach.

Relief algorithms are effective and take into account contextual data. In issues with significant connections between characteristics, it accurately measures the quality of attributes. One of the most effective algorithms is the relief algorithm. They are feature estimators that may be used in a variety of applications.

Algorithm : Basic Relief

***Input* :** for each training instance a vector of attribute values and the class value

Output: the vector W of estimations of the qualities of attributes

Pseudo code:

1.set all weights W[A] := 0.0;
2. for i := 1 to m do begin
3. randomly select an instance R_i*;*
4. find nearest hit H and nearest miss M;
5. for A := 1 to a do

6. W[A] := W[A]-diff(A,R_i,H)/m + diff(A,R_i,M)/m;
7. end;

Assume I1, I2,... The samples are in an instance space, with each sample being a vector space of attributes Ai, i=1.....n, where n is the number of attributes in the data collection, and each sample having a class label t. All of the characteristics' weights are initially set to zero. Then, from the instance space, we select a random instance Ri and determine its nearest neighbours. One of Ri's neighbours must be from the same class as Ri, known as the nearest hit H. Another neighbour, designated as nearest miss M, must be from a different class than Ri. The weights of all the qualities are now determined based on the values of Ri, M, and H.

Neural Network

A neural network is a collection of interconnected I/O units with weighted connections that correspond to computer algorithms. It aids in the development of prediction models from massive databases. This model is based on the neurological system of humans. It aids in picture comprehension, human learning, and computer speech, among other things.

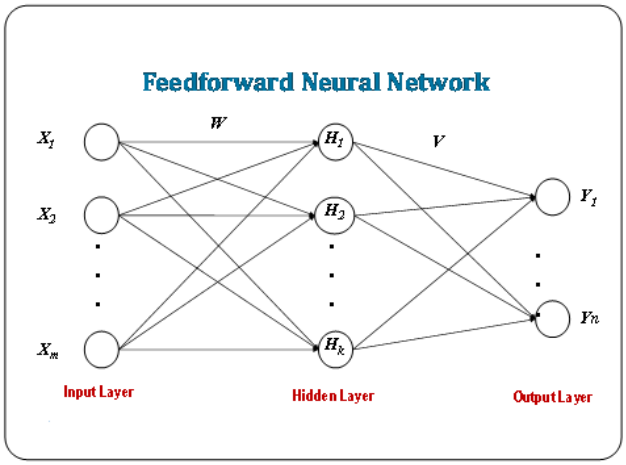

Fig.1. Architecture of feedforward neural network

Single hidden layer feedforward neural network is considered to classify the dengue infection into primary infection, DF, DHF and normal. X= [x_1 ..... x_n, 1] is the input pattern used in the input layer of the network where 1 represents bias value in the input layer. W is a matrix connecting input layer and hidden layer. V is a matrix connecting hidden layer and an output layer. H and Y are vectors that represent the output of the hidden and output layer. The sigmoidal activation function is used in the hidden layer and the linear activation function is used in the output layer to find H and Y.

$$net_h = \sum XW, \qquad net_o = \sum HV \tag{1}$$

$$H = f(net_h), \qquad Y = f(net_o) \tag{2}$$

$$f(net) = \frac{1}{1 + e^{-net}} \tag{3}$$

Backpropagation algorithm is used to train the single hidden layer feedforward network. This algorithm is learned by samples and backpropagates the error from the output layer. Weights are adjusted according to the deviation of errors. The procedure of backpropagation algorithm is shown below.

Input: set of patterns, each with input vector X and

output vector Y,

repeat

{

for each pattern p **do**

{

for each node j in the hidden layer **do**

{

$net_j = \sum_{i=1}^{ni} x_i w_{i,j}$

$h_j = f(net_j)$

$= \frac{1}{1+e^{-net_j}}$

}

for each node k in the output layer **do**

{

$net_k = \sum_{j=1}^{nh} h_j w_{j,k}$

$y_k = f(net_k)$

$= \frac{1}{1+exp^{-net_k}}$

$err_k = expec_k - y_k$

$E = \frac{1}{2} \sum_{p=1}^{np} \sum_{k=1}^{no} err_k^2$

}

$new\ V = old\ V + \lambda \frac{\partial E}{\partial V}$

$new\ W = old\ W + \lambda \frac{\partial E}{\partial W}$

}

Until (error of the network is within required accuracy)

Note: *ni, nh, no, np* represent number of input neurons, number of hidden neurons, number of output neurons and number of patterns respectively. $expec_k$ represents expected value of neuron k and λ represents learning parameter. *W* is a weight matrix connecting input layer and hidden layer and *V* is a weight matrix connecting hidden layer and output layer.

IV Results and Discussion

The proposed method has been tested with the heart disease data from machine learning repository of UCI [7]. The dataset have total 303 instances of which 164 instances belonged to the healthy and 139 instances belonged to the heart disease patients. 14 clinical features have been recorded for each instance. The dataset that was utilised to test the proposed classification approach came from the UCI Machine Learning repository. There are 14 attributes in all. There are no missing values in the dataset, which has a total of 303 occurrences. The dataset is commonly used for a variety of cardiac conditions, including conventional angina, atypical angina, non-angina discomfort, and asymptomatic angina. The age of the patient is represented by a numeric data type that varies from 29 to 65 years. The Cp is a numeric characteristic that ranges from 1 to 4 and is used to determine the pain kind. The trestbpd is the resting blood pressure, which ranges from 92 to 100, and the fbs is the fasting blood sugar level, which is either a 1 or a 0, reflecting true or false Boolean values.

The resting echocardiographic result is displayed as three instances ranging from 0 to 2 in the restecg. The thalach is the maximal heart rate achieved, which can range from 82 to 185 beats per minute. Exang is a Boolean value that represents exercise-induced angina. The illness is the dataset's target class, with yes or no answers indicating the presence of heart disease [25].

The dataset was sent into the feature selection step, which was used to identify the relevant features that cause heart disease. The Relief feature selection is used to find the dataset's most important features. All 13 attributes have their weights set to zero. To determine the ranking of all the attributes, the 10 closest hits and misses are found. Eq. 1 is used to compute the differences between the qualities (3 and 4). 0.092168, 0.86721, 0.061257, 0.044422, 0.003012, 0.00283506, 0.002632, 0.00156210, 0.002534321, 0.00283506, 0.0065, 0.0030158, 0.0002562 are the ranks given to the attributes cp, eang,chol, thal slope, trestbps, age, sex, fbs,restecg, ca, thalach and old Fig.2 Shows the summary of the computations.

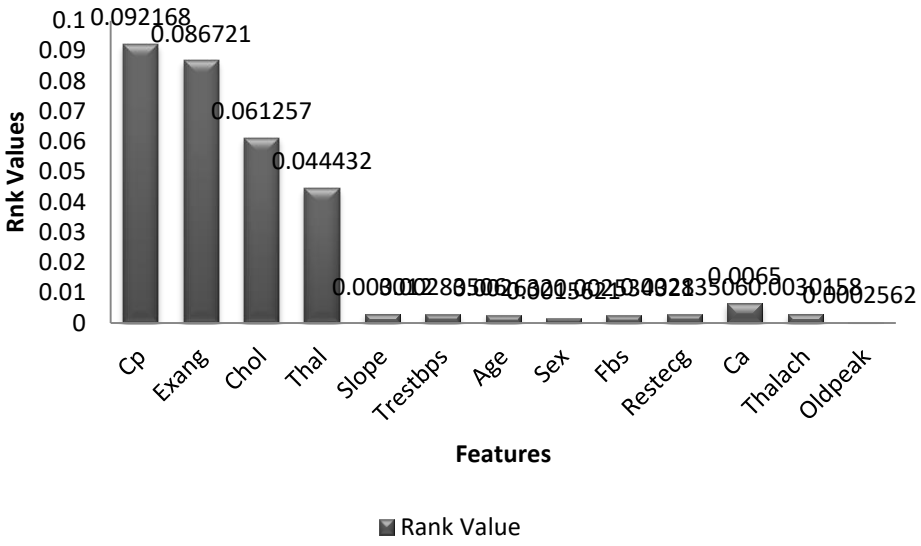

Fig.2. Results obtained from feature selection

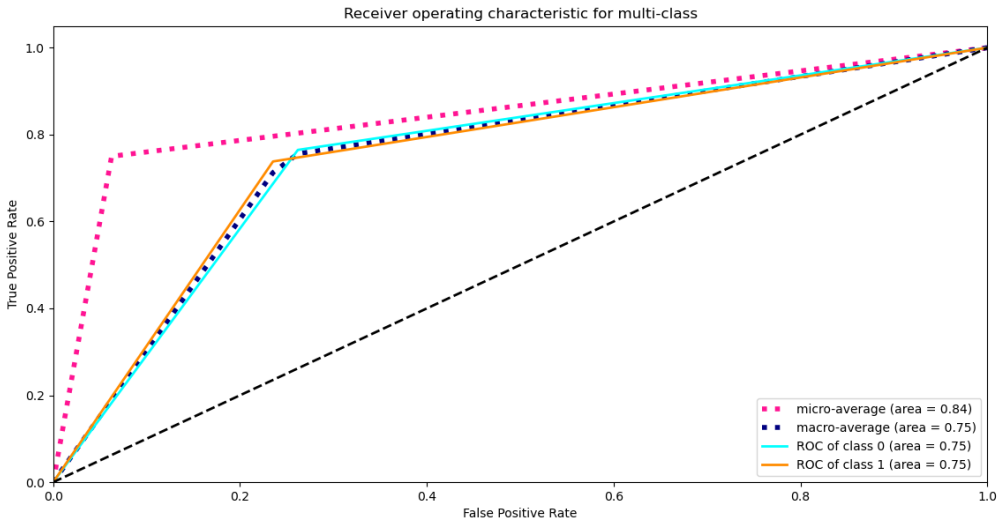

Fig. 3 Roc curve of the proposed work

ROC (Receiver Operating Characteristics) curves are commonly used to depict the relationship/trade-off between clinical sensitivity and specificity for each conceivable cut-off for a test or a set of tests in a graphical format. The suggested system's ROC curve shows that there is no difference in sensitivity and specificity. The suggested system's accuracy was measured and illustrated in Fig. 3. The model that was built in this work can be used by clinicians and health care industry to detect the heart disease in the new patients. From the patient's data, specific features (cp, exang, chol, thal and slope) were extracted from the pre-processing. The extracted features are useful because it shows the most significant attributes for predicting the heart disease. The confusion matrix represents the accuracy of proposed system on the testing phase. The elapsed time of the proposed algorithm is ***0.00078*** seconds as averaged. From the observations it has been identified that the proposed algorithm correctly classifies 61 instances.

V Conclusion

People, particularly those in our nation, are becoming more afflicted with heart disease (Algeria). As a result, being able to predict the disease before being sick lowers the chance of mortality. This is an area where a lot of study has been done. Our study is part of a larger study on heart disease detection and prediction. It is based on the use of Machine Learning methods, of which we picked the neural network algorithm on a real data set of Algerian individuals, with excellent results, with Neural Network achieving 80.37 percent accuracy.

References

1. S. N. Shivappriya, R. Navaneethakrishnan, K. S. Raj, M. Abirami and S. Chidhambaram, "Neural Network Based Heart Disease Prediction," *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, 2021, pp. 1-6, doi: 10.1109/ICAECA52838.2021.9675637.
2. Youness Khourdifi and Mohamed Bahaji (2019). "Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization", International Journal of Intelligent Engineering and Systems, Vol.12, No.1, 2019. 6. Youness
3. Nidhi Bhatla and Kiran Jyoti (2012). "An Analysis of Heart Disease Prediction using Different Data Mining Techniques", International Journal of Engineering Research & Technology, Vol. 1 Issue 8
4. Dr.S.Seema Shedole, Kumari Deepika, "Predictive analytics to prevent and control chronic disease", https://www.researchgate.net/punlication/316530782, January 2016.
5. Megha Shahi, R. Kaur Gurm, "Heart Disease Prediction System using Data Mining Techniques", Orient J. Computer Science Technology, vol.6 2017, pp.457-466..
6. Sairabi H.Mujawar, P.R.Devale, "Prediction of Heart Disease using Modified K-means and by using Naïve Bayes", International Journal of Innovative research in Computer and Communication Engineering, vol.3, October 2015, pp.10265-10273
7. Sharan Monica.L, Sathees Kumar.B, "Analysis of CardioVasular Disease Prediction using Data Mining Techniques", International Journal of Modern Computer Science, vol.4, 1 February 2016, pp.55-58.
8. K.Vembandasamy. K, R.Sasipriya and E.Deepa, "Heart Diseases Detection Using Naive Bayes Algorithm", International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 9, September 2015, pp.441-444
9. Vincy Cherian and Bindu M.S, Heart Disease Prediction Using Naïve Bayes Algorithm and Laplace Smoothing Technique, International Journal of Computer Science Trends and Technology (IJCST) – Volume 5 Issue 2, Mar – Apr 2017
10. Liu Xiao, Wang Xiaoli, Su Qiang, Zhang Mo, Zhu Yanhong, Wang Qiugen, Wang Qian. A hybrid classification system for heart disease diagnosis based on the RFRS method. Comput. Math. Methods Med. 2017;2017:1–11.
11. Sonam Nikhar and A.M. Karandikar, "Prediction of Heart Disease Using Machine Learning Algorithms", International Journal of Advanced Engineering, Management and Science, Vol-2, Issue-6, June- 2016.
12. Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivasatava (2019). "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", IEEE access volume7, 2019.
13. Komal Saini and Sandeep Sharma (2019). "Review on the Heart Disease Detection Using IoT Framework", International Journal of Computer Sciences and Engineering Vol.7(3), Mar 2019, E-ISSN: 2347-2693.
14. UCI Machine learning Repository https://archive.ics.uci.edu/ml/index.php
15. Riyaz, L., Butt, M.A., Zaman, M. and Ayob, O., 2022. Heart Disease Prediction Using Machine Learning Techniques: A Quantitative Review. In *International Conference on Innovative Computing and Communications* (pp. 81-94). Springer, Singapore.
16. El-Hasnony, I.M., Elzeki, O.M., Alshehri, A. and Salem, H., 2022. Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction. *Sensors*, *22*(3), p.1184.
17. Verma, K., Bartwal, A.S. and Thapliyal, M.P., 2021 A Genetic Algorithm based Hybrid Deep Learning Approach for Heart Disease Prediction.
18. Krishnan, S., Magalingam, P. and Ibrahim, R., 2021. Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction. *International Journal of Electrical & Computer Engineering (2088-8708)*, *11*(6).
19. Thanga Selvi, R. and Muthulakshmi, I., 2021. An optimal artificial neural network based big data application for heart disease diagnosis and classification model. *Journal of Ambient Intelligence and Humanized Computing*, *12*(6), pp.6129-6139.
20. Priyanga, P., Pattankar, V.V. and Sridevi, S., 2021. A hybrid recurrent neural network-logistic chaos-based whale optimization framework for heart disease prediction with electronic health records. *Computational Intelligence*, *37*(1), pp.315-343.
21. Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y.R. and Suraj, R.S., 2021, January. Heart disease prediction using hybrid machine learning model. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)* (pp. 1329-1333). IEEE.
22. Rajdhan, A., Agarwal, A., Sai, M., Ravi, D. and Ghuli, P., 2020. Heart disease prediction using machine learning. *International Journal of Research and Technology*, *9*(04), pp.659-662.
23. Jagtap, A., Malewadkar, P., Baswat, O. and Rambade, H., 2019. Heart disease prediction using machine learning. *International Journal of Research in Engineering, Science and Management*, *2*(2), pp.352-355.
24. Garg, A., Sharma, B. and Khan, R., 2021. Heart disease prediction using machine learning techniques. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012046). IOP Publishing.

25. Bashir, S., Khan, Z.S., Khan, F.H., Anjum, A. and Bashir, K., 2019, January. Improving heart disease prediction using feature selection approaches. In *2019 16th international bhurban conference on applied sciences and technology (IBCAST)* (pp. 619-623). IEEE.