2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC) | 978-1-6654-2867-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICESC51422.2021.9532705

Heart Disease Detection using Machine Learning Technique

Likitha KN
UG Student,
Dept. of ECE,
CMR Institute of Technology,
Bengaluru, India
Liki17ec@cmrit.ac.in

Nethravathi R
UG Student,
Dept. of ECE,
CMR Institute of Technology,
Bengaluru, India
netr17ec@cmrit.ac.in

Nithyashree K
UG Student,
Dept. of ECE,
CMR Institute of Technology,
Bengaluru, India
nith17ec@cmrit.ac.in

Ritika Kumari
UG Student,
Dept. of ECE,
CMR Institute of Technology,
Bengaluru, India
riti17ec@cmrit.ac.in

Sridhar N
Asst. Prof,
Dept. of ECE,
CMR Institute of Technology,
Bengaluru, India,
nsridhar83@gmail.com

Venkateswaran K
Assoc.Prof,
Dept. of ECE,
CMR Institute of Technology,
Bengaluru, India
venkateswaran.k@cmrit.ac.in

Abstract: **Heart disease (HD) is a serious health problem which may affect large numbers of people across the world. As a result, detecting a heart condition at an early stage will be beneficial to treatment. The number of persons with heart disease is rapidly increasing, necessitating the development of a system that can detect heart disease more easily. The presence or absence of disorder is determined by the patient's smoking status. The cardiac disease system can identify patients who are high-risk and define the most important variables in cardiovascular patients but also build a model so that they can distinguish between them easily and understandably. The Machine Learning algorithms are applied and compared based on the characteristics like age, Chest ache, Blood Pressure (BP), sex, cholesterol and heartbeat. The main focus of this paper is to develop a basic machine learning model to enhance the diagnosis of the heart condition in the right manner. The different techniques such as Logistic Regression, K-Nearest Neighbor (K-NN), Decision Tree, Naive Bayes, Random Forest and Support Vector Machine are applied for machine learning and achieved the better results in this work.**

Keywords - Heart Disease, Diagnosis, Dataset, Machine Learning Algorithm, High-Risk Patient, Decision Tree.

I. INTRODUCTION

Heart related issues are critical in human beings and which can even lead to demise [1]. Every year, far too many people die as a result of a heart problem. When modern technology and specialists are unavailable, diagnosing and treating cardiac conditions is extremely challenging. A cardiac issue is frequently recognized by symptoms such as a high BP, discomfort, hypertension, asystole, and so on. Chest ache, difficulty in breath, and discomfort are the most prevalent signs. Heart disease is caused by a variety of factors, including birth abnormalities, increased vital signs, diabetes, smoking, medications, and alcohol. Nowadays, there are several automated methods for diagnosing intestinal disease, including data processing, machine learning, deep learning, and so forth [2]. As a result, in this paper, we employ machine learning repositories to train the datasets. In this system, a heart condition data set is utilized. The primary purpose of this research is to predict the chance of a cardiovascular disease arising in affected persons in terms of a percentage. This is typically performed through the use of data processing categorization techniques. A classification technique is being used to categories the whole data into two groups: *Yes* and *No* [4]. The dataset will be trained using machine learning classification techniques such as Decision Tree (DT) and Naive Bayes (NB) Classification techniques. They may enhance the level of accuracy of classification techniques and helps in performing both classification and prediction models, which makes use of python programming language to perform these models. Heart predictor system may utilize the data mining knowledge to provide an approach to hidden patterns in the data and can be utilized by the medical care specialists to improve quality of service.

In this paper, the required parameters for the early detection of heart condition is included. The comparison of various results shows the accuracy of the model with respective parameter. When it involves comparing two or more machine learning algorithm, it's most difficult because each algorithm differs in ways. There are just one option to realizing the efficiency of the algorithm for the actual dataset is implementing them. The consequences of the work are mentioned on the idea of assessment metric of the algorithms. At the end conclusion and future directions are discussed.

II. RELATED WORK

Various machine learning-based diagnostics strategies are offered in the research by scientists to diagnose HD. Many studies have been conducted in relation to disease prediction systems employing various machine learning algorithms. Detrano et al. [8] created an HD model utilizing machine learning classification techniques, and the model's accuracy was found to be 77 percent. Cleveland was used as the dataset, and the approach worldwide evolutionary and feature selection technique was utilized. Gudadhe et al. [10] developed an HD classification diagnostic approach based on multi-layer perception and SVMs approaches, with an efficiency of more percent of accuracy. Humar et al. [11] used a neural net and symbolic logic to construct an HD classification system. The correctness of the arrangement was assessed to be 87.4 percent. Resul et al. [12] used the statistical measurement tool to construct an ANN ensemble-based diagnosis tool for HD, with 89.01 percent accuracy, 80.09 percent sensitivity, and 95.91 percent specificity. In [13], an HD diagnostics system based on learning have developed. In combination with the Feature Selection (FS) approach the ANN-DBP approach proved successful. Palaniappan et al. In [14], a technique of diagnosis by an HD detection specialist has given. For systems development purposes, the prediction for learning algorithms, such as NB, DT and Recurrent Neural Network (RNN), has been determined to be 86.12 percent NB precision, 88.12 and 80.4 per cent DT classification precision. Olaniyi et al. [15] created a three-phase methodology for HD prediction in angina that supported the synthetic neural network methodology and achieved 88.89 percent accuracy. For the diagnosis of HD, the integrated medical decision network, supported by an artificial neural system and a Fuzzy AHP was constructed by Samuel et al. [16]. The suggested development in terms of accuracy was determined to be 91.10%. In [17], the high definition arrangement based on relief and raw techniques are given. The classification accuracy of the approach suggested was 92.32%. The strategy used by feature selection and classification algorithms for HD identification has been described [18]. Feature selection SBS FS (Sequential Backward Selection Algorithm) The K-NN classification performance was examined in both the whole and selected sets of features. This method outperformed the others in terms of accuracy. Mohan et al. [19] developed an HD prediction system employing hybrid machine learning approaches in another investigation, also provided a substitute approach for significant feature selection for effective machine learning classifier training and testing, and an accuracy of 88.07 percent was discovered. Geweid et al. [20] created HD detection approaches by modifying an SVM-based duality optimization methodology. Early prediction of coronary artery disease is an important research now [21]. The sensitivity, specificity, accuracy of the training and testing set are used in Genetic Algorithm based approach evaluation [22]. Table.1 summarizes the limits and advantages of existing machine learning algorithms. In order to provide effective and accurate identification in initial periods for superior treatment to get well, the predictive accuracy of the HD detection technique should be further enhanced. Therefore, low precision and long time for calculations are important problems with these older methodologies, which may be linked to the usage of irrelevant characteristics in the data set. New ways to HD detection are needed to solve these problems. Improving prediction accuracy may be a major problem and a research gap.

TABLE I. COMPARATIVE STUDY OF VARIOUS ALGORITHMS

S.No	Author, year	Purpose	Technique used	Accuracy	Remarks
1)	Sonam Nikhar, 2016[3]	Heart disease prediction utilizing algorithms for machine learning.	NB classifier and DT	In comparison to the naive Bayes classification, the decision-tab is more accurate.	Improved data storage for legal and practical objectives. No Random Forest &KNN concentration.
2)	V.V Ramalingam, 2018 [4]	Prediction of heart disease Use of techniques of machine learning.	NB, SVM, KNN, DT, RF, Ensemble Model.	SVM is more precise than other technologies	Used for large and complex data.
3)	Avinash Golande, 2019 [5]	Effective Machine Learning Techniques Heart Disease Prediction.	Decision tree, KNN, K-mean clustering, Ad boost.	Decision tree (86.60%)	Applicable for various ML algorithms. No much focus on Random Forest and Naive Bayes
4)	Mr Santhana Krishnan. J,2019, [6]	Heart disease prediction utilizing the algorithm of machine learning.	Naive Bayes, Decision tree.	Decision Tree (91%) Accuracy	Increased accuracy for effective heart disease diagnosis. No focus on Random Forest, KNN&SVM

The health care businesses are frequently keen on isolating individuals based on their health, particularly their cardiac condition. Before making any decisions in this subject, the Machine Learning (ML) methodology has proven to be highly useful.

Although many authors have proposed and created several ways for early detection of HD, there is still need for research on machine learning algorithms to improve their prediction accuracy.

III. MACHINE LEARNING ALGORITHM FOR DETECTING HEART DISEASE

The dataset for heart condition prediction was obtained from the Kaggle website, a knowledge repository for data analysts, and a small amount of data was obtained from a scan centre. The variables in the dataset are as follows: age, gender, pain, cholesterol, blood glucose level, and resting blood glucose [7]. The acquired dataset was pre-processed, with records with missing or irrelevant values deleted and replaced with the right values, before being passed to the classifier to assess and calculate the estimated accuracy. From among the achieved accuracies, the acceptable method was picked and evaluated with the test data.

A. Heart Disease Predictions

The primary notion of gut illnesses is to reinforce the forecast with a more accurate value, which will be accomplished through the use of various algorithms. A tremendous quantity of knowledge is generated in the medical industry, and it is critical to mine that data to assist practitioners in early disease identification.

B. Techniques for treating Heart Disease

The attributes listed, such as age, pain, sex, cholesterol, and heartbeat, are classified using RF, DT, LR and NB approach [12]. The input dataset is split into two parts: a training dataset and a test dataset with 70% & 30% of the data respectively. A training dataset is a collection of data that is used to train a model. The trained model performance is evaluated using the testing dataset.

(i) Decision Tree (DT):

This is a kind of controlled ML approach that discovers many means of separating information which is supported continuously by a certain parameter. The tree has two entities: a node of choice and a node of the leaf.

(ii) Naive Bayes (NB):

This classification is a statistical method in order to handle binary or multi-class classification issues using Bayes theorem [2]. A classification of Naive Bayes is straightforward to design, as forecasts are supposedly independent. This notion is extremely unbelievable in actual world facts.

(iii) Random forest (RF):

A random forest tree is a kind of monitored learning process that generates many decision-making trees for regression and classification. The random wood method produces decision-making trees based on data samples, predicting their outcome and then selecting the simplest response or output.

(iv) Support Vector Machine (SVM):

An SVM machine may be a discriminative classifier that finds a hyperplane that maximizes the margin between two classes. SVM has its advantage and disadvantage for the set of data. Since in the medical data field, the data set can be non-linear, so SVM would be one of the good options. SVM is a supervised learning approach that may be used for both classification and regression.

(v) K- Nearest Neighbor (KNN):

This is a slow classification technique, it takes more time to train an algorithm's classification than others; it is divided into two steps: training from data and testing it on a fresh instance. The KNN working rule is based on each data point that is labelled as a neighbor.

The already processed dataset is used to support the prototype, and the previously mentioned techniques are investigated and used. The confusion matrix is used to generate the metrics and describes the model's performance for the various methods used in this prototype.

Within the proposed system, the temporal trajectories of the checkup measure among various sorts of algorithms, we've used RF, DT, LR, KNN and NB algorithms.

IV. Experimental Set-up and Flow Diagram

The Cardiac System for heart condition prediction followed with flow diagram is discussed.

A.simulation setup

Given schematic in Fig.1 shows the parameters that are required for the prediction of heart condition using the various algorithms of machine learning. The parameters considered are same for all the models, since a standard data set has been used.

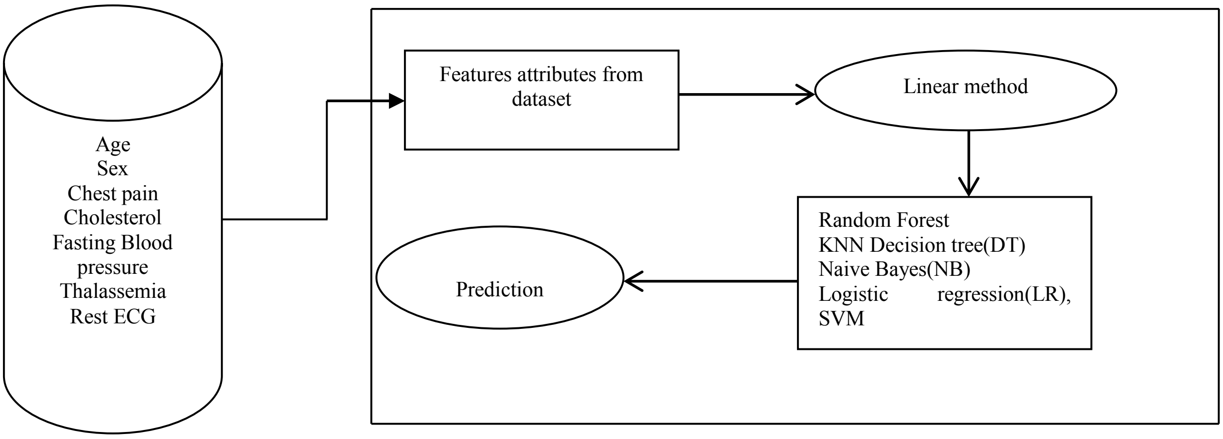

Fig.1. The cardiac disease system is depicted as a block diagram

Attributes in the Standard dataset has been selected by using the *Head()* function.

A goal prediction value based on separate factors is modelled linearly. The link between variables and forecasts is mostly determined by this method. Different linear models change accordingly: the type of link between dependent and separate variable is taken into account and the quantity of separate variable is utilized. A linear regression is the linear model.The linear technique is chosen because it predicts a dependency of the variable (y) based on a certain independent variable (x). This technique of regression thereby establishes a linear link between an input and output.

B. Flow diagram

The flow of the prediction of the disease is given in Fig.2. The Data collected from the patient about the characteristics of their behaviour such as age, gender, cholesterol, blood sugar level, fasting blood pressure, thalassemia , rest ECG is pre-processed and classified according to their characteristics. Pre-processing is the initial step conducted in the analysis.Data pre-processing is a technique where the raw datum is transformed into a proper format and unnecessary values are removed.

The disease can be predicted accurately after training the data set, extracting the feature and testing the data.

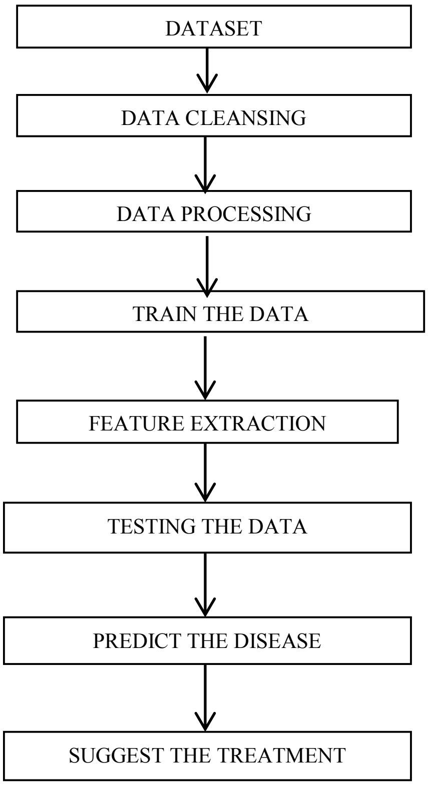

Fig.2.Flow Diagram of the Proposed System.

Once the disease is predicted, the treatment can be suggested.

V. Results & Discussions

A.Results

Metrics such as precision, sensitivities, and specificity are employed to evaluate the performance of the classification approach. These metrics are generated from the confusion matrix. The model's accuracy reflects the overall performance and is calculated using the formula given in equation (1).

On the confusion matrix, a True Negative (TN) indicates the patient does not have a cardiac condition.

- A true positive (TP) implies that the patient has a cardiac condition.
- False Positive (FP) shows that the patient does not have the condition but that the model predicted that he or she did, meaning that the model misclassified a healthy individual.
- False Negative (FN) shows that the patient has disease although the model predicted that he or she did not, meaning that the model incorrectly identified a person with a cardiac condition.

$$Accuracy = \left(\frac{TP+TN}{TP+TN+FP+FN}\right) * 100 \qquad (1)$$

Table II. Performance of the Proposed System for Machine Learning Techniques

Sl. No	Technique Used	Accuracy
1	Random Forest	96.6%
2	KNN	88%
3	Decision tree	94%
4	Naive Bayes	85.23%
5	SVM	92%
6	Logistic Regression	87.33%

With the various algorithm and dataset, observed the system performing better and produces the specified range of values.

B. Discussions

As table III indicates, the novelty of our platform is that the heart disease prediction model is often considered because the better predicting model for the first detection. The readability of the code is straightforward because it is programing language and algorithm is employed, then Hybrid technology makes it easy to be used of the system in HD Identification which helps in the future development of the work.

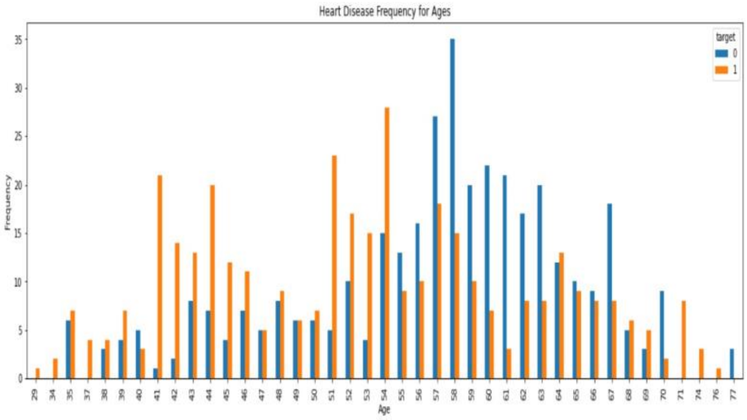

Fig.3. Heart Disease Frequency For Ages

Fig.3 represents the presence of heart disease based on the age of the person in the form of histograms.

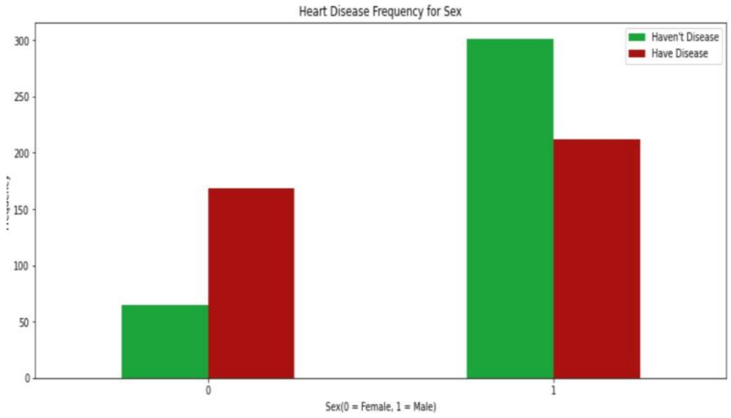

Fig.4. Heart disease frequency for sex

Fig.4 characterizes the existence of heart condition supported the sex of an individual indicating '0' for female and '1' for male within the sort of histogram where the green color is that the count of the healthy person and red color is that the count of the sick person.

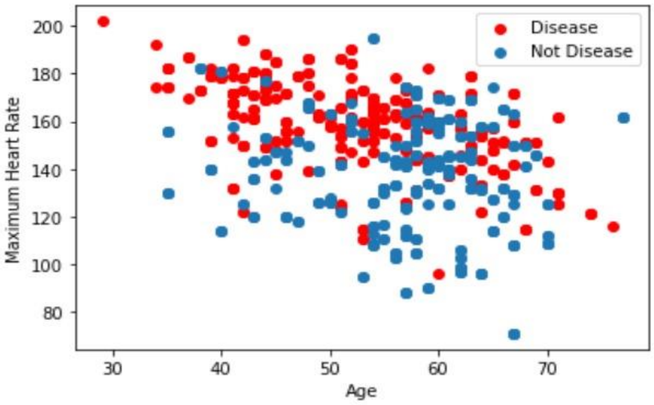

Fig.5.Scatter Plot

Fig.5 gives the utmost number of individuals having heart condition supported the age from the taken dataset

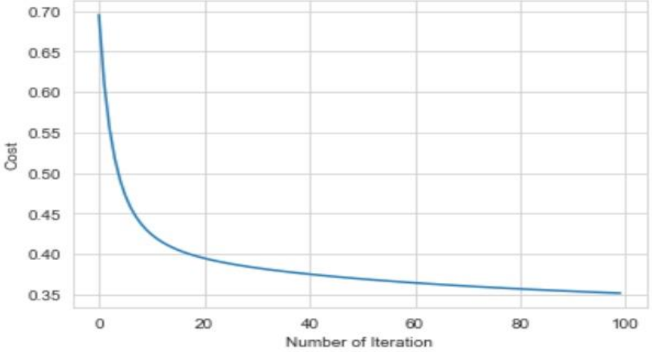

Fig.6.Logistic regression plot

Fig.6 is the plot between cost and number of iterations

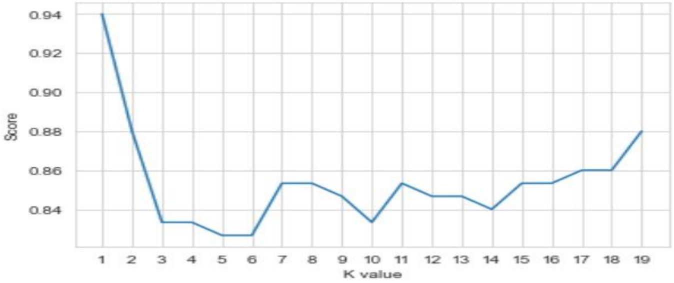

Fig.7.K-nearest neighbors

Fig.7 shows the plot of scores for various values of K

The model's performance is described by the Confusion matrix in Table.III, which gives the summary of the results predicted based on classification techniques. A Confusion matrix consists of rows that contain the values predicted by the classifier and columns which contains the actual value. The results of the confusion matrix indicates whether the patient has the cardiac condition or not, and therefore whether the examination is favorable or unfavorable.

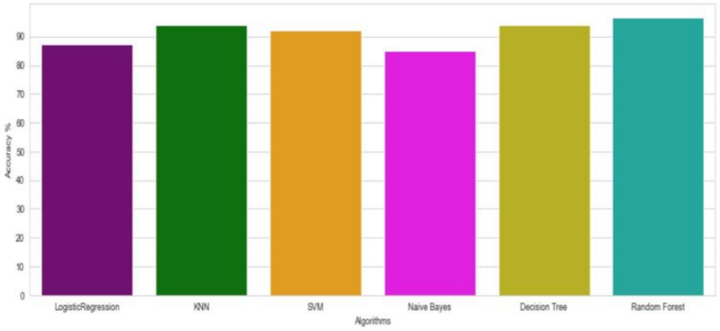

Fig.8.Accuracy Plot

Fig.8 shows the accuracies of the varied algorithms utilized in the paper, Random Forest has the highest accuracy followed.

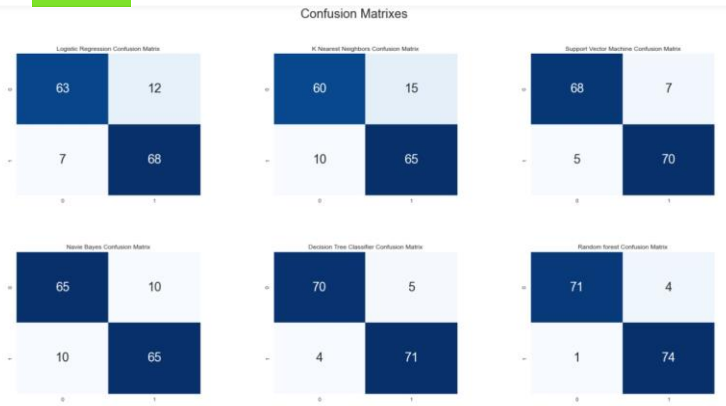

Fig.9.Confusion Matrices

Fig.9. represents a table that allows visualization of the performance of all the algorithms used.

Table III. Confusion Matrix Values Using Various Algorithms

Algorithms	True Negative	False Negative	False Positive	True Positive
Logistic Regression(LR)	63	7	12	68
K-NN	60	10	15	65
SVM	68	5	7	70
Naive Bayes	65	10	10	65
Decision Tree	70	4	5	71
Random Forest	71	1	4	74

The Confusion matrix is used to categorize the right prediction of a model for various classes and errors.

VI.CONCLUSION AND FUTURE SCOPE

Various machine learning strategies for predicting cardiac status are presented in this research. There often exist missing parts in checkup data. As in case of human, is not taking their annual check-up of body may be difficult to examine once a while all the problems. This approach reduces medical failure, eliminate unnecessary variations in the practice and improve patient safety and outcomes through the integration of clinical assessment aid with processer based patient records. Data mining may provide a wise environment that may assist in enhancing the quality of healthcare judgments greatly. By examining their attributes, we arrived at a simple algorithm and this helps in the prediction of missing data. In different instances, each algorithm has produced a different result and these are helped in the detection of early heart condition. This machine learning technique using a basic algorithm will help in the early identification of heart illness. But still there is some limitation since medical diagnoses are seen as an important but complex work, which must be performed accurately and swiftly. Clinical choices are usually based not on knowledge-rich facts included in the database, but on medical intuition and experience. This approach leads to undesirable preconditions, blunders, and unreasonable medical costs that undermine patient's quality of care. Data exploitation can produce a rich environment in which the quality of therapeutic judgments can be considerably improved. In future, it is necessary to automate the system and this would be really useful to enhance the accuracy of forecasting the early heart disease.

REFERENCES

[1] A. L. Bui, T. B. Horwich, and G. C. Fonarow, "Epidemiology and risk profile of heart failure", *Nature Rev. Cardiol.*, vol. 8, no. 1, p. 30, 2011.

[2]M. Durairaj and N. Ramasamy, "A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate," *Int. J. Control Theory Appl.*, vol. 9, no. 27, pp. 255-260, 2016.

[3] Sonam Nikhar, A.M. Karandikar, "Prediction of Heart Disease Using Machine Learning Algorithms", *International Journal of Advanced Engineering, Management and Science (IJAEMS)* Infogain Publication, Vol-2, Issue-6, June- 2016.

[4] V.V. Ramalingam, Ayantan Dandapath, M Karthik Raja, "Heart disease prediction using machine learning techniques: a survey", *International Journal of Engineering & Technology*, 7 (2.8) (2018) 684-687.

[5]Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", *International Journal of Recent Technology and Engineering (IJRTE),* ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019.

[6] Mr Santhana Krishnan. J, Dr Geetha.S, "Prediction of Heart Disease Using Machine Learning Algorithms", 1st International Conference on Innovations in Information and Communication Technology (ICIICT), doi:10.1109/ICIICT1.2019.8741465, 2019.

[7] S. Nazir, S. Shahzad, S. Mahfooz, and M. Nazir, "Fuzzy logic based decision support system for component security evaluation," *Int. Arab J.Inf. Technol.*, vol. 15, no. 2, pp. 224-231, 2018.

[8] R. Detrano, A. Janosi, W. Steinbrunn, M. P_sterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease,"' *Amer. J. Cardiol.*, vol. 64, no. 5, pp. 304-310, Aug. 1989.

[9] J. H. Gennari, P. Langley, and D. Fisher, "Models of incremental concept formation," *Artif. Intell.*, vol. 40, nos. 1-3, pp. 11-61, Sep. 1989.

[10] M. Gudadhe, K. Wankhade, and S. Dongre, "Decision support system for heart disease based on support vector machine and artifcial neural network,"' in *Proc. Int. Conf. Comput. Commun. Technol. (ICCCT)*, Sep. 2010, pp. 741_745.

[11] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for diabetes and heart diseases," *Expert Syst. Appl.*, vol. 35, nos. 1-2, pp. 82_89, Jul. 2008.

[12] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7675-7680, May 2009.

[13] M. A. Jabbar, B. Deekshatulu, and P. Chandra, "Classifcation of heart disease using arti_cial neural network and feature subset selection," *GlobalJ. Comput. Sci. Technol. Neural Artif. Intell.*, vol. 13, no. 3, pp. 4-8, 2013.

[14] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *Proc. IEEE/ACS Int. Conf. Comput. Syst.Appl.*, Mar. 2008, pp. 108_115.

[15] E. O. Olaniyi, O. K. Oyedotun, and K. Adnan, ``Heart diseases diagnosis using neural networks arbitration," *Int. J. Intell. Syst. Appl.*, vol. 7, no. 12, p. 72, 2015.

[16] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy AHPfor heart failure risk prediction," *Expert Syst. Appl.*, vol. 68, pp. 163-172,
Feb. 2017.

[17] X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang, and Q. Wang, "A hybrid classi_cation system for heart disease diagnosis based on the RFRS method," *Comput. Math. Methods Med.*, vol. 2017, pp. 1-11, Jan. 2017.

[18] A. U. Haq, J. Li, M. H. Memon, M. H. Memon, J. Khan, and S. M. Marium, "Heart disease prediction system using a model of machine learning and sequential backward selection algorithm for features selection," in *Proc.IEEE 5th Int. Conf. Converg. Technol. (ICT)*, Mar. 2019, pp. 1-4.

[19] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542_81554, 2019.

[20] Jian Ping, Amin Ul Haq "heat Disease Identification Method Using Machine Learning Classification In E-Healthcare, 2020.

[21] Chen, Joy long Zong and P. Hengjinda, "Early prediction of coronary Artery Disease by Machine Learning Method-A Comparative Study", *Journal of Artifical intelligence and Capsule Networks*, Vol. 3, No 01, 2021.

[22] B Kaan Uyar , Ahmet llhan, "Diagnosis of heart disease using trained genetic algorithm based trained recurrent fuzzy neural networks", *procedia computer science*, Vol. 120, 2017.