International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST

Machine Learning Based Heart Disease Prediction System

Poonam Jadhav[1], Prof. Prajakta Khairnar[2]

[1]Student, Dr. D. Y. Patil School of Engineering, Lohegaon, Pune.

[2]Prof. Department of Electronics and Tele Communication, DYPSOE, Pune, India

ARTICLEINFO

Article History:

Accepted: 05 May 2023
Published: 22 May 2023

Publication Issue
Volume 10, Issue 3
May-June-2023

Page Number
330-340

ABSTRACT

Heart disease is the main cause of the enormous deaths in the world over the past few decades and it turned out to be the deadliest not only in India but all over the world. It is therefore necessary to have a reliable, precise and functional one. To diagnose these diseases in time for appropriate treatment we used machine learning algorithms techniques. Heart is larger organs than the brain, which have a higher status in the human body to pump blood and delivers it to all organs throughout the body. Data analysis is useful for making predictions based on more information. Huge amounts of patient data are stored per month. Stored data can be useful as a source for predicting future events diseases. Machine learning techniques are used to predict heart disease, such as Artificial Neural Network, Random Forest and Support Vector Machine (SVM). The main reason of demise is because of Heart assault in world. Deaths due to heart disease have become one of the major problems with roughly one person losing their life every minute due to heart disease. Machine learning, when implemented in healthcare, is capable of early and accurate disease detection for heart disease. The datasets used have attributes of medical parameters. The datasets are processed in python using the ML Algorithm. This technique uses past old patient records to get prediction of new ones in the early stages of preventing loss of life. In this work, a heart disease prediction system is implemented using a powerful Random Forest algorithm, SVM, Naive Bayes, Decision Tree, Logistic Regression. It loads the patient data record in the form of a CSV file. After accessing the data set, the operation is performed and the effective heart attack level is create. The advantage is high performance, level of accuracy. it is very flexible and has high rates. WHO surveyed 10 million people affected by heart disease. The problem facing the healthcare industry in today's life is the timely prediction of disease after a person is affected. History records are very extensive and real-world data can be incomplete, inconsistent.

In the past, it was not possible to predict the
treatment of the disease for every patient in the early stages. Come up with the idea of predicting heart disease with 90% Accuracy is achieved in testing now. Practical use of data collect from previous records is time-consuming. To overcome this, we implement Random Forest algorithm to get accurate results in less time. The dataset Pre-processing, we use contains NaN.

Real-time application in a variety of contexts.

Keywords- ML :Machine Learning, Vector Quantization, Random Forest algorithm, Decision Trees. Neural Network, Support vector machine.

I. INTRODUCTION

The work in article focuses primarily on various data mining practices used to predict heart disease. Any heart abnormality can cause pain in other parts of the body. Any disruption to the normal functioning of the heart can be referred to as heart disease. In the modern world, heart disease is one of the leading causes of most deaths. Heart disease can result from an unhealthy lifestyle, smoking, alcohol consumption, and eating large amounts of fat, which can lead to high blood pressure. According to the World Health Organization, heart disease kills 17.7 million people each year, accounting for 31% of all deaths in the world. Heart disease has also become the leading cause of death in India. It killed 1.7 million Indians in 2016. Heart disease increases healthcare costs and decreases a person's productivity. WHO estimates show that India lost up to $237 billion between 2005 to 2015, for heart or circulatory diseases.

Thus, it is very important to possible and accurate to predict heart disease. Medical organizations around the world collect data on various health problems. This data can be used using various machine learning techniques to obtain useful information. But the data collected is very extensive and numerous times this data can be very strong. Those datasets that are too overwhelming for the human mind can be easily explored using various machine learning techniques. Therefore, these algorithms have become very useful recently times to accurately predict the presence or absence of heart disease. The use of information technology in healthcare is increasing day by day to support physicians in decision-making. Help doctors heal diseases, administer medicines and discover patterns and dependencies between diagnostic data.

Current methods of predicting cardiovascular risk fail to recognize many people who are affected would benefit from preventive treatment, while others would have unnecessary surgery. Machine Learning Offerings the ability to improve accuracy by exploiting the complex interactions between risk factors. We evaluated whether Machine learning could improve cardiovascular risk prediction. The greatest challenge of modern healthcare is to provide the highest quality services and efficient and accurate diagnoses. All accuracy in the treatment of the disease lies in its rapid detection. Chen came up with the idea of predicting heart disease. He used IS AI vector quantization technique for prediction purposes with 85 accuracy is achieved in testing. Practical use of data collected from previous records is time-consuming. To overcome this, we implement Random Forest algorithm to get accurate results in less

time.The proposed work is an attempt to detect these heart diseases early to avoid catastrophic consequences. Big Medical Data records created by medical experts are available for analysis and extraction of valuable knowledge.

Data mining techniques are ways to extract valuable and hidden information from a large amount of available data. The medical database consists mainly of discrete information. Therefore, making decisions using discrete data becomes a complex and difficult task. Machine learning , a subfield of data mining, efficiently processes a well-formed large data set. In medicine, machine learning can be used to diagnose, detect and predict various diseases. The main goal of this article is to provide physicians with a tool for early detection of heart disease . This, in turn, will help ensure that patients are treated effectively, and serious consequences are avoided. Machine learning plays a very important role in detecting hidden discrete patterns and then analyzing the data. After analyzing the data, ML techniques help to predict heart disease and diagnose it early. This article presents a performance analysis of various ML techniques such as Naive Bayes, Decision Tree, Logistic Regression and Random Forest for predicting heart disease early stage.

II. RELATED WORK

Chalay Beyonce et al. [1] recommended predicting and analyzing the incidence of heart disease using data mining techniques. The main goal is to predict the occurrence of heart disease for automatic early detection of the disease in no time. The proposed methodology is also fundamental for organizing healthcare with experts has no more knowledge and skills. It uses various medical attributes like blood sugar and heart rate, age, gender Certain attributes are included to determine whether a person has heart disease. Data set analysis is calculated with WEKA software.

Senthil Kumar Mohan et al. [2] implemented hybrid machine learning to predict heart disease. The dataset used is Record Cleveland. The first step is data preprocessing. In doing so, the tuples are removed from the existing data set omitted values. The age and gender attributes of the dataset are also not used as the authors consider them to be personal data information and does not affect the prediction. The other 11 attributes are considered important because they contain relevant medical records. They proposed their Linear Hybrid Random Forest (HRFLM) method. Combination of random forest (RF) and linear method (LM). In the HRFLM algorithm, the authors used four algorithms. The first algorithm deals with the partitioning of the input data set. It is based on the decision tree for which it is performed of each record sample.Once the feature space is identified, the data set is divided into leaf nodes. First outing Algorithm is a split of the data set. So in the second algorithm, they apply the rules to the data set and the output is here data classification using these rules. In the third algorithm, features are extracted using the Less Error classifier. The algorithm takes care of finding the minimum and maximum error rates of the classifier. The output of this algorithm is objects with classified attributes. The fourth algorithm uses the classifier, which is a hybrid error-based method bids for extracted features. Finally, they compared the results obtained after using HRFLM with others classification algorithms such as decision tree and support vector machine. Consequently, RF and LM perform better results on top of each other, the two algorithms are combined and a new single HRFLM algorithm is created. authors propose to further improve accuracy by using a combination of different machine learning algorithms.

Nagaraj M Lute Math, et al. [3] has performed the heart disease prediction using Naive bayes classification and SVM (Support Vector Machine). The performance measures used in analysis are Mean Absolute Error, Sum of Squared Error and Root Mean Squared Error, it is established that SVM was emerged

as superior algorithm in terms of accuracy over Naive Bayes.

Singh, Yesh Vendra K. et al. [4], deal with various supervised machine learning algorithms such as Random Forest, Support Vector Machine, Logistic Regression, Linear Regression, Decision Tree with 3-fold, 5 fold and 10 fold cross-validation techniques. They used a Cleveland dataset with 303 tuples, with some tuples missing attributes. During data preprocessing, they simply removed a tuple of the missing value from a six-inch data set. , then separated from the remaining 297 tuples linear regressions were used. In it, they defined the dependency of an attribute on others that can be linearly separated from each other. Basically, the classification is done for binary classification. They performed best 10 times, which is 83.82%. Logistic regression classification is performed using a sigmoid function. This heart disease prediction algorithm shows the maximum and 5-fold cross-validation and is 83.83%. Support Vector Machine is a classification algorithm for supervised modes machine learning. In this case, the classification is done using a hyperplane. Maximum accuracy achieved by SVM 3 times authors used different numerical subdivisions and different numbers of leaf nodes to find best accuracy. From 37 The highest accuracy of digital divisions and 6 leaf nodes, equal to 79.12%, was achieved. accuracy from decision tree 79.54% with 5 times. The random forest algorithm applied to a non-linear dataset delivers better results than the decision tree. A random forest is a set of decision trees created by knots. From this set of decision trees one can vote first and then classify from there reach the maximum number of votes.

Fahd Saleh Alotaibi designed a comparison of ML models five different algorithms [5]. Rapid miner tool used, resulting in higher accuracy than MATLAB Instrument Weka. In this study, the accuracy of the decision tree, Logistic Regression, Random Forest, Naive Bayes and SVM classification algorithms were compared. decision tree the algorithm had the highest accuracy.

Anjan Nikhil Repaka, possibly Tel., in [6] proposed a system that the uses NB (Naive Bayesian) techniques for classification and the AES algorithm (Advanced Encryption Standard). for secure data transmission for disease prediction.

Prince Therese. R et al. [7] conducted a survey among various classification algorithm for predicting heart disease.The classification techniques used are Naive Bayes, KNN (K Nearest Neighbor), Decision Tree, Neural Network, and The accuracy of classifiers for different numbers was analyzed attributes.

The main idea behind the proposed system after reviewing the above papers was to create a heart disease prediction system based on the inputs. We analysed the classification algorithms namely Decision Tree, Random Forest, Logistic Regression and Naive Bayes based on their Accuracy, Precision, Recall and f-measure scores and identified the best classification algorithm which can be used in the heart disease prediction. The dataset used was the Heart disease Dataset which is a combination of 4 different database, but only the UCI Cleveland dataset was used. This database consists of a total of 76 attributes but all publish experiments refer to using a subset of only 14 features .The complete description of the 14 attributes used in the proposed work is mentioned in Table shown below.

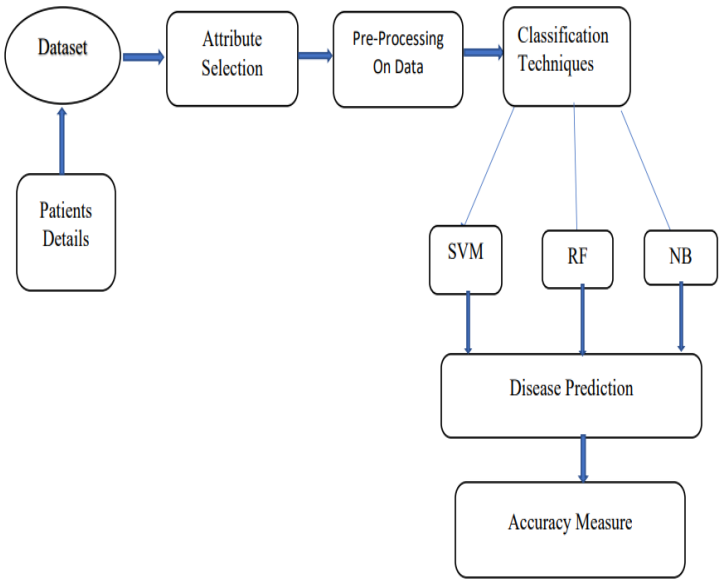
TABLE:

FEATURES SELECTED FROM DATASE

No	Attribute Description	Distinct Values of Attribute
1.	Age- represent the age of a person	Multiple values between 29 & 71
2.	Sex- describe the gender of person (0-Feamle, 1-Male)	0,1
3.	CP- represents the severity of chest	0,1,2,3

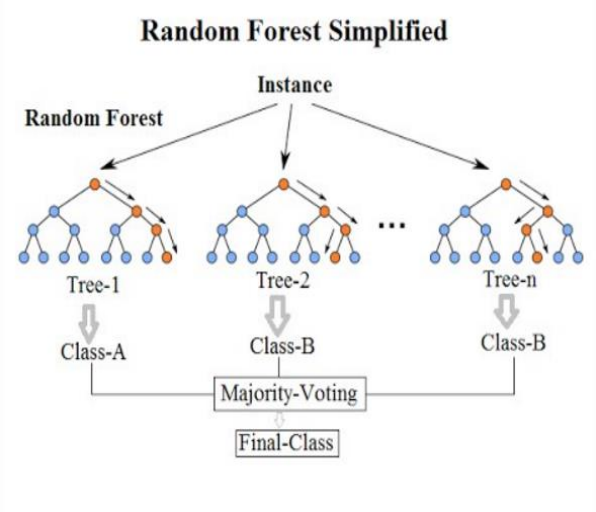
	pain patient is suffering.	
4.	Rest-It represents the patient's BP.	Multiple values between 94& 200
5.	Chol-It shows the cholesterol level of the patient.	Multiple values between 126 & 564
6.	FBS-It represent the fasting blood sugar in the patient.	0,1
7.	Resting ECG-It shows the result of ECG	0,1,2
8.	Heartbeat- shows the max heart beat of patient	Multiple values from 71 to 202
9.	Exang- used to identify if there is an exercise induced angina. If yes=1 or else no=0	0,1
10.	OldPeak- describes patient's depression level.	Multiple values between 0 to 6.2.
11.	Slope- describes patient condition during peak exercise. It is divided into three segments (Unsloping, Flat, Down sloping)	1,2,3.
12.	CA- Result of fluoroscopy.	0,1,2,3
13.	Thal- test required for patient suffering from pain in chest or difficulty in breathing. There are 4 kinds of values which represent	0,1,2,3
	Thallium test.	
14.	Target-It is the final column of the dataset. It is class or label Colum. It represents the number of classes in dataset. This dataset has binary classification i.e. two classes (0,1).In class "0" represent there is less possibility of heart disease whereas "1" represent high chances of heart disease. The value "0" Or "1" depends on other 13 attribute.	0,1

PROPOSED METHOD-



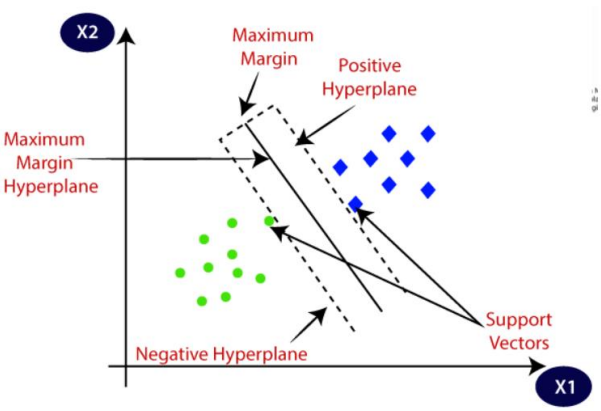
1.Random Forest:

Random Forest is a supervised machine learning algorithm. Classification tasks, but generally performs better on classification tasks. As the name suggests, Random Forest The technique considers multiple decision trees before reporting a result. So it's basically a lot. This technique is based on the belief that more trees will converge towards the right decision. For classification, uses a grading system and then decides the grade, while in regression it takes the average of all the scores or decision trees. Works well with large datasets with high multidimensionality.



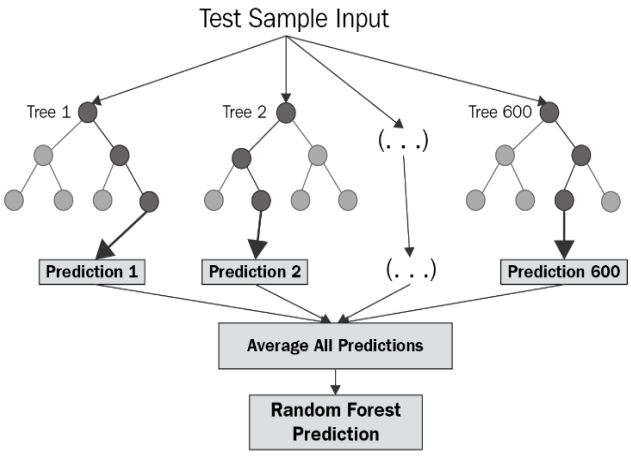
2.Support Vector Machines (SVMs):

Support vector machines exist in different forms, linear and non-classifier. What is usual in this context, two different datasets are involved with SVM, training and a test set. In the ideal situation the classes are linearly separable. In such situation a line can be found, which splits the two classes perfectly. However not only one line splits the dataset perfectly, but a whole bunch of lines do. From these lines the best is selected as the "separating line". A SVM can make some errors to avoid over Support vector machines classifiers are applied in many applications. They are very popular in recent research. This popularity is due to the good overall empirical performance. Comparing the naive Bayes and the SVM classifier, the SVM has been applied the most



3.Decision Tree:

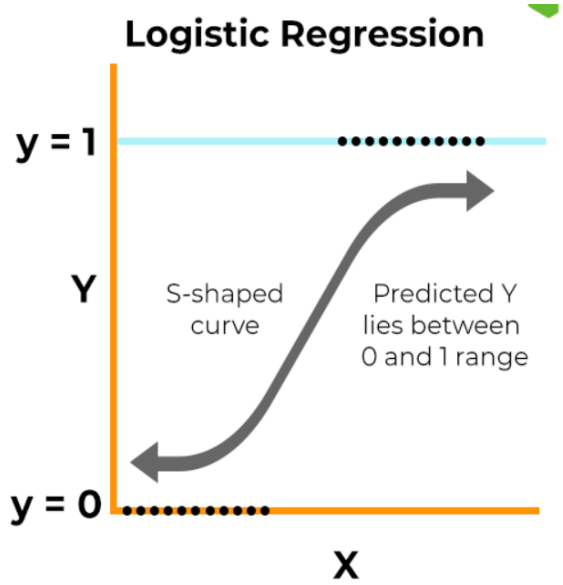
Decision Tree algorithm is in the form of a flowchart where the inner node represents the dataset attributes and the outer branches are the outcome. Decision Tree is chosen because they are fast, reliable, easy to interpret, and very little data preparation is required. In Decision Tree, the prediction of class label originates from root of the tree. The value of the root attribute is compared to record's attribute. On the result of comparison, the corresponding branch is followed to that value and jump is made to the next node.



4.Logistic Regression:

Logistic Regression is a classification algorithm mostly used for binary classification problems. In logistic

regression instead of fitting a straight line or hyper plane, the logistic regression algorithm uses the logistic function to squeeze the output of a linear equation between 0 and 1. There are 13 independent variables which makes logistic regression good for classification.

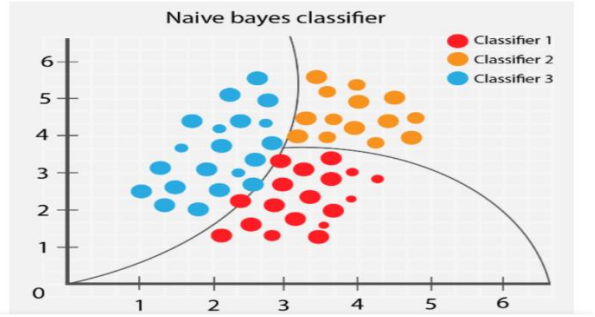


Logistic Regression – Sigmoid Function

5.Naive Bayes:

Naïve Bayes algorithm is based on the Bayes rule. The independence between the attributes of the dataset is the main assumption and the most important in making a classification. It is easy and fast to predict and holds best when the assumption of independence holds. Bayes' theorem calculates the posterior probability of an event (A) given some prior probability of event B represented by P(A/B)[10] as shown in equation 1 :

$$P(A|B) = (P(B|A)P(A)) / P(B) \quad (1)$$



SOFTWARE:

Python:

Python web scraper programmed in Python is used to collect the data. According to Wikipedia, Python's syntax allows developers to express concepts in fewer lines of code. Guido van Rossum of CWI in the Netherlands started implementing Python in December 1989. Python 2.0 was released on October 16, 2000 and Python 3.0 was released on December 3, 2008.Why use Python for web scraping and not something else? Python offers a module called "urllib2" that has the right functions to open web pages and retrieve information easily. Python is used to program the web scraper responsible for collecting weather data for the model.

MS Excel:

Microsoft Excel is a spreadsheet application developed by Microsoft for Windows and Mac OS X. It features calculation, graphing tools, pivot tables and a macro-programming language. The first version was released in 1987. Why choose MS Excel versus another similar type of software? MS Excel is a very complete spreadsheet application tool, which supports almost any kind of file extension, and it has a lot of features. Its user-friendly interface helps you most of the time. However, if this doesn't seem enough, I will say that, apart from the typical things a normal user would do in Excel (Charts, Calculation…), it enables you to use the VBA language to create functions to use on the spreadsheets you've created. Excel can also be used as if it were an SQL database as was explained in a previous chapter. Having said this, for me it is the

perfect program.MS Excel is used a lot throughout the project, to visualize the data and perform cleaning tasks on it.

III. RESULTS AND DISCUSSION

This project aims to know whether the patient has heart disease or not. The records in the dataset are divided into the training set and test sets. After preprocessing the data. The data classification technique namely Support vector machine, Decision Tree, Logistic Regression, Naive Bayes Artificial neural network, Random Forest method were applied. The project involved analysis of the heart disease patients dataset with proper data processing. Then, models are trained and tested with maximum scores as follows:

1. Random Forest Classifier: 90.0 %
2. Support vector machine:85.0%
3. Decision Tree:82.0%
4. Logistic Regression: 86.0%
5. Naïve Bayes: 87.0%

Table-1: Sample data set

Age	63	37	41	56
Cp	3	2	1	1
Trestbps	145	130	130	120
Chol	233	250	204	236
Fbs	1	0	0	0
Thalach	150	187	172	178
Exang	0	0	0	0
Old Peak	2.3	3.5	1.4	0.8
Thal	1	2	2	2
Target	1	1	1	1
ECG	3	2	4	3

The above mentioned attributes in table-1 are enough to predict if a person is affected with heart disease or not. Each attribute in the data set result of functionality of heart. For example,

- Cp- The type of chest pain categorized into 4 values.
  (1. Typical angina 2.Atypical angina 3.Non-angical pain 4. Asymptomatic)
- Trest bps- Level of blood pressure at resting mode.
- Chol-Serum cholesterol in mg/dl.
- Fbs- Blood sugar levels on fasting (if>120mg/dl represented as 1 otherwise 0)
- Resting ecg- Results of electrocardiogram while at rest.
- Exang- Angina induced by exercise (0-No, 1-Yes)
- Old peak- Exercise induced ST depression in comparison with state of rest.
- ECG: Electrocardiogram- Electrical signal from the heart

Table-2: Sample data set with results.

AGE	63	37	17	56
CP	1	1	0	1
TRESTBPS	3	2	0	0
CHOL	233	250	170	200
FBS	1	0	1	1
RESTECG	0	1	0	1
THALCH	150	187	77	79
EXANG	0	0	1	1
OLDPEAK	2.3	3.5	0	1
SLOPE	0	0	2	2
THAL	1	2	3	3
TARGET	1	1	1	1
HEART DISEASE	YES	YES	NO	NO

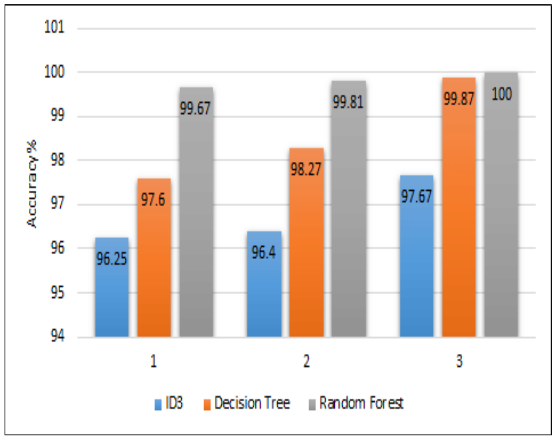


Fig: Graph comparing accuracy of various algorithms.

we can say that the application when implemented using random forest algorithm has more accuracy rate when compared to other algorithms.

IV. CONCLUSION

This project provides insight into machine learning techniques for heart disease classification role. The classifier is essential in the medical industry so that the results can be used to predict potential treatment donated to patients. Existing techniques are examined and compared to find efficient and precise systems. Machine learning techniques significantly improve the accuracy of patient cardiovascular risk prediction can be identified in the early stages of the disease and may benefit from preventive treatment can be deducted that machine learning algorithms have great potential to predict cardiovascular or heart-related diseases. Each of the above algorithms performed exceptionally well in some cases, but poorly in others. Cases.

V. FUTURE SCOPE

In future this application can extended by updating some features like, if the user will affected with heart disease all his family members will be notified with a message in early and also the information should be passed to the nearest hospital. Another feature is there should be online doctor consultation with the nearest doctor available.

In this regard, ML applications using various efficient algorithms are utilized not only in disease prediction and diagnosis but also in the field of radiology, bioinformatics and medical imaging diagnosis. We will study to explores the potential application of Fano factor constrained TQWT for the automated heart sound signals classification. we intend to improve the performance of classification by using more data and also explore other transforms like empirical wavelet transform. We will develop machine learning system can be employed in polyclinics and hospitals as triage to assess the cardiac health of patients.

VI. ACKNOWLEDGEMENTS

I would like to express special thanks of gratitude to my guide Prof. Prajakta Khairnar who gave me the golden opportunity to do this wonderful project on the topic Heart Disease Prediction Using Machine Learning, which also helped me in doing a lot of research and I came to know about so many new things. who helped me a lot in finalizing this project within the limited time frame. I would also like to thank our H.O.D. of ENTC-Dr.S.C. Inamdar, for providing opportunity.

VII.REFERENCES

[1]. Kaan Uyar and Ahmet İlhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks" in B.V ICTASC, Elsevier, pp

[2]. Ashish Chhabbi, Lakhan Ahuja, Sahil Ahir and Y. K. Sharma, "Heart Disease Prediction Using Data Mining Techniques", © IJRAT Special Issue National Conference "NCPC-2016", pp. 104-106, 19 March 2016.

[3]. Berry JD, Lloyd-Jones DM, Garside DB, et al. Framingham risk score and prediction of coronary heart disease death in International journal of computer applications, published in 2006.

[4]. Theresa Princy and R, J. Thomas, "Human Heart Disease Prediction System using Data Mining Techniques", © IEEE ICCPCT, 2016.

[5]. Kaur h Beant and William jeet Singh, "Review on Heart Disease Prediction System using Data Mining Techniques", © IJRITCC, vol. 2, no. 10, pp. 3003-08, 2014.

[6]. Kirmani, M.M., Ansarullah, S.I.: Prediction of heart disease using decision tree a data mining technique. IJCSN Int. J. Comput. Sci. Netw. 5(6), 885–892 (2016)

[7]. Salam Ismaeel, Ali Miri et al., "Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis", IEEE Canada International Humanitarian Technology

[8]. Tahira Mahboob, Rida Irfan and Bazelah Ghaffar et al." Evaluating ensemble prediction of coronary heart disease using receiver operating characteristics" ©2017 IEEE

[9]. Ammar Asjad Raja, Irfan-ul-Haq , Madiha Guftar Tamim Ahmed Khan "Intelligence syncope Disease Prediction Framework using DM-techniques" FTC 2016 –Future Technologies Conference 2016.

[10]. M.A. Jabbar, B.L. Deekshatulu, and Priti Chandra, " Intelligent heart disease prediction system using random forest and evolutionary approach", Journal of Network and Innovative Computing, Vol. 4, pp.174-184, 2016.

[11]. N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," Machine Learning,(1997).

[12]. Ayon Dey, Jyoti Singh, N. Singh "Analysis of supervised machine learning algorithms for heart disease prediction".

[13]. Q. K. Al-Shayea, "Artificial neural networks in medical diagnosis," International Journal of Computer Science Issues, published in 2011. young men. Am Heart J. 2007;154(1):80–6.

[14]. Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Technique", International Journal of Pure and Applied Mathematics, 2018. [15]. Mohan, Senthil kumar, Chand rasegar Thirumalai, and Gautam Srivastava, "Effective heart disease prediction using hybrid machine learning techniques" IEEE Access 7 (2019): 81542-81554. [3]. Ali, Liaqat, et al, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure" IEEE Access 7 (2019): 54007-54014.

[15]. Singh Yeshvendra K., Nikhil Sinha, and Sanjay K. Singh, "Heart Disease Prediction System Using Random Forest", International Conference on Advances in Computing and Data Sciences. Springer, Singapore, 2016. [5]. Prerana T H M1, Shivaprakash N C2 , Swetha N3 "Prediction of Heart Disease Using Machine Learning ,Algorithms- Naïve Bayes, Introduction to PAC Algorithm, Comparison of Algorithms and HDPS" International Journal of Science and Engineering Volume 3, Number 2 – 2015 PP: 90-99

[16]. B.L Deekshatulua Priti Chandra "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm" International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013.

[17]. Michael W.Berryet.al, Lecture notes in data mining, World Scientific(2006) [8]. S. Shilaskar and A. Ghatol, "Feature selection for medical diagnosis :Evaluation for cardiovascular diseases," Expert Syst. Appl., vol. 40, no. 10, pp. 4146–4153, Aug. 2013.

[18]. C.-L. Chang and C.-H. Chen, "Applying decision tree and neural network to increase quality of dermatologic diagnosis," Expert Syst.

Appl., vol. 36, no. 2, Part 2, pp. 4035–4041, Mar. 2009.

[19]. T. Azar and S. M. El-Metwally, “Decision tree classifiers for automated medical diagnosis,” Neural Comput. Appl., vol. 23, no. 7–8, pp. 2387–2403, Dec. 2013.

Cite this Article

Poonam Jadhav, Prof. Prajakta Khairnar, "Machine Learning Based Heart Disease Prediction System", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 10 Issue 3, pp. 330-340, May-June 2023.

Journal URL : https://ijsrst.com/IJSRST52310359