Methodologies and Techniques for Heart Disease Classification and Prediction

Priyanka S.Sangle[1], R. M. Goudar[2],
A.N.Bhute[3]
[1]priyasangle90@gmail.com,
[2]rmgoudar@comp.maepune.ac.in
[3]anbhute@mitaoe.ac.in
[1,2,3] School of Computer
Engineering and Technology,
MIT Academy of Engineering,
Pune, Maharashtra, India

Abstract— Heart disease is one of the most common reasons of increased mortality rate in the world. The term heart disease refers to disease of heart & blood vessel system within it. Heart disease prediction at a premature phase is a critical challenge in the area of clinical data analysis. Nowadays appropriate decision support systems are being used by many hospitals that results in minimizing the cost of clinical tests ultimately. The objective of this paper is to explore and analyze the different techniques and methodologies for classifying the heart diseases which gives better results in terms of evaluation parameters viz. accuracy and performance.

Keywords—Heart Disease, Prediction, classification, Machine Learning

I. INTRODUCTION

Heart is an important central muscular organ of all living individual, which plays an essential role of blood pumping to the rest of the organs through the blood vessels of the circulatory system. The proper working of the heart is directly related to life such a way that, improper functioning of heart directly influences the other body parts such as brain, kidney, etc.

According to the World Health Organization (WHO), more than 12 million deaths occur every year worldwide because of the different types of heart diseases known as cardiovascular diseases. Heart disease consists of many diseases like Cardiomyopathies, Angina Pectoris, Coronary diseases, Congenital etc. that are very heterogeneous and clearly affect the heart and arteries. Even young people are also getting influenced by heart diseases. The reason behind the increased possibility may be because of the risk factors like family history, age, smoking, high blood pressure, high blood cholesterol, poor diet, hyper tension, obesity, physical inactivity among which some of the risk factors are controllable.

Apart from the above factors, major risk factors are considered to be common risk factors which are caused due to some lifestyle habits such as physical inactivity, eating habits, and obesity. Chest pain, strong compressing or flaming in the chest, discomfort in chest area, sweating, light headedness, dizziness, shortness of breath, pain spanning from the chest to arm and neck, fluid retention are the symptoms of heart disease.

It is not easily possible to manually find out the heart disease based on risk factors. However, the disease can be predicted by using machine learning classification techniques according to the existing data. An attempt is made in this paper to discuss not only related work of the researchers but also describes dataset used for heart disease with some of the machine learning classification techniques to predict heart disease from the risk factors and dataset.

II. RELATED WORKS

In the last few years, researchers have done several studies to evaluate classification accuracies by employing different machine learning classification techniques to Cleveland heart disease dataset for heart disease risk prediction.

S. Mohan et.al [1], proposed machine learning techniques to treat baseline data and states unique perception towards heart disease. They proposed a novel technique by combining the features of Random Forest (RF) and Linear Method (LM) called hybrid random forest with linear method (HRFLM) approach and they have proved that HRFLM is an absolutely accurate in the heart disease prediction. In evaluation, results shows that HRFLM achieved highest accuracy of 88.7% as related to other classification methods.

K Mathan et.al [2], tested augury frameworks for heart disease by using more number of information attributes. In this research, they proposed an altered solution with decision trees for classification that provides accurate results when differed strikingly with other calculations. The work explains that neural networks and Gini index prediction models provides most conspicuous and accurate results for prediction. They used voting technique of the discretization strategies to get more exact decision trees and calculated accuracy and sensitivity. In the execution, research work inspected the results by applying set of instructions to the different kinds of decision trees and attained the accuracy and sensitivity by the executing the selected methods of decision tree.

Min Chen [3] proposed a multimodal disease risk prediction based on convolutional neural network (CNN-MDRP) algorithm which is recently came into existence and achieved 94.8% prediction accuracy with the greater convergence speed when compared to other similar enhanced algorithms like CNN-based unimodal disease risk prediction (CNN- UDRP) algorithm.

T. Vivekanandan et.al [4], developed a model which can handle the enormous amount of e-healthcare data efficiently. In this paper, they selected critical features from the large set of available features to diagnose the heart disease and performed some rigorous tasks. They have used an algorithm named as modified differential evolution (DE) to select features and to optimize them. The DE/rand/2/exp strategy has been considered for comparative study. Fuzzy AHP and a feed-forward neural network been used with selected features to carry out heart disease prediction. Authors integrated modified DE algorithm with fuzzy AHP and a feed-forward neural network to evaluate various performance measures. They proposed hybrid model with accuracy of 83%, which is greater than the other existing models and also evaluated prediction time for the same. Authors observed that the error minimization using effective back propagation model will be the future work to improve performance measures without great deviation.

Zeinab A et.al [5], proposed a novel and hybrid methodology to analyze the cause and to recognize coronary artery disease and augmented the performance measures of neural network around 10%. To do so, they have used hybrid of two algorithms such as genetic algorithm to enhance its initial weights and neural network and achieved accuracy of 93.85% by working on Z-Alizadeh Sani dataset. Authors have suggested to use many evolutionary strategies like particle swarm optimization (PSO) strategy to have better performance instead of using proposed genetic algorithm.

Seyedamin et.al [6] applied different machine learning techniques and measured the term accuracy as an output while comparing them. Researcher used various techniques on Cleveland dataset and then results were compared with each other. SVM is applied on the heart disease dataset that results in a classifier. The mentioned techniques like Bagging, Stacking, Boosting are applied to improve accuracy. Using Stacking technique, MLP and SVM achieved best accuracy of 84.15% which is higher than other techniques.

In [7], authors proposed a technique named as ridge expectation maximization imputation (REMI) to determine the missing values in the databases. They considered two powerful classification algorithms, SVM and extreme learning machine (ELM) and compared their performance. Authors used two databases such as STULONG and UCI databases to experiment on proposed algorithm. They have improved the performance speed and removed irrelevant attributes that results in reduction of size of features. For this they used conditional likelihood maximization method. In experimental outputs they improved accuracy of risk prediction as compared to other works. They have shown that the performance of proposed REMI approach/technique was significantly better than using conventional techniques.

M. A Jabbar et.al [8], used various feature selection measures. They have calculated the accuracy and performance of Naïve Bayes classifier for the prediction of heart disease and obtained greatest accuracy of 86.29%. As per the experiments conducted and results achieved it is seen that the Hidden Naïve Bayes (HNB) shows optimal accuracy and superior to naïve bayes. But they have also found the limitation that the dependencies among the attributes can't be modeled in naïve bayes classifier.

Jagdeep Singh et.al [9], implemented various association and classification methods on the heart datasets to predict the heart disease. They modeled a hybrid technique by using a classification and association and developed decision system for heart disease prediction for classification associative rules (CARs) on weka environment. They compared different classifiers using aprior association on heart dataset and result shows that IBk (k Nearest Neighbor) achieved highest accuracy of 99.19% for prediction.

Theresa Princy R et.al [10], investigated different classification techniques like Naïve Bayes, ID3, Neural Network and KNN for investigation of heart disease by selecting different attributes like age, gender, pulse rate, cholesterol. Author analyzes that the accuracy of finding the risk is depends on the number of attributes such that it is increased as the attributes are increased. Using different methods, it is also viable to increase the accuracy with less number of attributes. Authors achieved the increased accuracy level to 80.6%.

In [11], authors used three classifiers such as Decision Trees, Naïve Bayes and K Nearest Neighbor to focus on the heart disease prediction. They showed that predictors performed better when in practical use. Analogous to the model establishment, the Result shown that KNN provides highest accuracy which is conventional since KNN reminds all the factors. But Decision Tree performed well as analyzed with other two techniques for the given dataset when used for prediction.

The objective of the author in [12] is to analyze and anticipate the heart disease more accurately. Researchers incorporated a *Fuzzy K-NN* classifier assuming minimum distance to categorize the data amongst multifarious sets and to eliminate the uncertainty of the data. After experimentation, result shows that the method is able to eliminate the excess of data and obtained a system with

better accuracy. In performance analysis authors shows that the *fuzzy K-NN* classifier acquired the more accuracy than *K-NN* classifier.

Table 2 describes the accuracy comparison chart in which particular methodologies with accuracy obtained by researchers in their research works are shown. This table provides researchers a way to easily find out the technique to get more precise solution for risk prediction in future work.

III. DATASET DESCRIPTION

In this study, the dataset used is the Cleveland Heart Disease data set taken from the University of California, Irvine (UCI) learning data set repository [15]. The dataset contains a total of 303 patient records with some missing values. Dataset consists of total number of 76 attributes. However, the many of the researchers uses at most of 14 attributes which are nearly related to the heart disease, out of which, 13 are input diagnostic attributes with their range of values and one is predictable attribute "Num" for the inspection of heart disease that gives the heart related disease state as value 0 represents "Absent" and value 1 represents "Present". Table 1 shows the detailed information of UCI dataset with data type and range of values the attributes used.

Table 1. Description of UCI dataset 13 input attributes

S. No	Attribute Description	Range or Values	Data type
1	Age(Age in years)	<33= Young 34–40= Medium 41–52= Old >52= Very Old	Numeric
2	Sex(Patient's Gender)	1=male,0=female	Nominal
3	Cp(Chest Pain Type)	1 = typical type 1 2 = typical type angina 3 = non-angina pain 4 = asymptomatic	Nominal
4	trestbps(Resting Blood Pressure)	<128= Low 128–142=Medium 143–154= High >154=Very High	Numeric
5	chol(Serum cholesterol)	(126 to 564)Continuous value in mg/dl	Numeric
6	fbs(Fasting blood sugar)	1 ≥ 120 mg/dl 0 ≤ 120 mg/dl	Nominal
7	Restecg(Resting electrographic results)	0 = normal 1 = having_ST_T wave abnormal 2 = left ventricular hypertrophy	Nominal
8	thalach(Maximum heart rate achieved)	<112= Low 112–152= Medium >152= High	Numeric
9	exang(Exercise induced angina)	1= True (YES) 0= False (NO)	Nominal
10	Oldpeak(ST depression induced by exercise relative to rest)	<1.5= Low 1.5–2.55= Risk >2.55= Terrible	Numeric
11	Slope(Slope of the peak exercise ST segment)	1 = up sloping 2 = flat 3 = down sloping	Nominal
12	Ca(Number of major vessels colored by fluoroscopy)	0=Fluoroscopy-0 1=Fluoroscopy-1 2=Fluoroscopy-2 3=Fluoroscopy-3	Numeric and Categorical
13	Thal (Defect type)	3 = normal 6 = fixed 7 = reversible defect	Nominal and Categorical

IV. MACHINE LEARNING ALGORITHMS

Various researchers used different existing machine learning classifiers, namely Decision Tree (DT), Naïve Bayes (NB), Artificial Neural Network (ANN), k- Nearest Neighbor (k-NN), Random Forest (RF) and Support Vector Machine (SVM), and new ensemble learning techniques such as bagging, boosting and stacking which are discussed below. In each outline, the performance measures like accuracy, precision, recall and F-measure has been calculated by them.

1. Decision Tree

A Decision tree is a notable supervised machine learning algorithm mostly used to solve classification problems. This technique works on continuous and categorical data and consists of a tree-like graph or a model of decisions (nodes) in which trees are generated by determining the best split at each node. The data (information) gain, Gini Index and Gain Ratio are computed in this technique. Here, the entropy of each attribute is calculated initially and then the dataset is divided with the help of variables with max data gain or min entropy. Following two steps [14] performs in recursion manner with rest of the factors.

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

In [2], K. Mathan et.al used decision tree based classifier and accuracy with voting technique and also without voting technique of discretization strategy as 79.8% and 78.3% respectively. In [6], decision tree has worst accuracy of 77.55% but after using boosting technique with decision tree, authors achieved better accuracy of 82.17%.

2. K-Nearest Neighbor (K-NN) Algorithm

K Nearest neighbor is a nonparametric classification technique which is based on instances as it makes use of all the instances and used in classification and regression problems to obtain instance values by employing user defined value as "k". It categorizes an object by the

majority voting of its closest neighbors. Putting it differently, the class of a new instance will be anticipated based on some distance metrics where the distance metric used can be a simple Euclidean distance.
In [10], authors detected heart disease by using hybrid of KNN and ID3 and achieved the accuracy of 80.6%. S. Joshi and et.al compared 3 classification algorithms such as DT, NB, and KNN and found that KNN has highest accuracy as compared to other techniques [11]. In [12], authors used Fuzzy KNN algorithm to predict heart disease and got 90% accuracy.

3. Support Vector Machine(SVM)
It is a supervised machine learning technique which categorizes data into two classes over a hyper plane. In two dimensional spaces, hyperplane is a line that divides a plane into two parts as shown in figure 1. SVM technique tries to enhance the margin which is the distance between the hyper plane and the two nearest data points from each class. A SVM is a discriminatory and effective classifier mostly used for classification, sentiment analysis and high dimensionality space problems than the regression challenges. It finds out an optimal hyper plane which classifies new examples on given labeled training data.
In [6], authors applied SVM algorithm using boosting technique and achieved 85% accuracy which is greater than using only SVM. S. Nikan in [7], proposed two classification methods such as SVM and extreme learning (ELM) machine and obtained 86.95% accuracy from SVM which is less than accuracy from ELM.

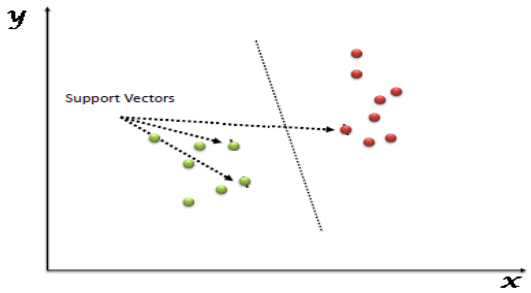
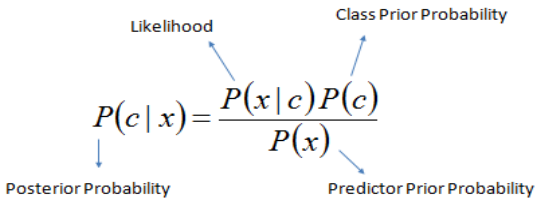


Fig. 1: Support Vector Machine

4. Naive Bayes(NB)
It is a simple and effective probabilistic classifier, is based on Bayes theorem used to construct classifiers. This classifier assumes the input attributes whose values are independent of the values of other attributes, therefore it is called "naïve". It is very simple to carry out the NB algorithm and mostly used for huge datasets. Even with this it is powerful algorithm used for real time Prediction, text classification/ spam Filtering, recommendation System. Bayes theorem is stated as follows:



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

where X is the data tuple and C is the class such that P(X) is constant for all classes [14].
M. A Jabbar et.al [8], considered Naïve Bayes classifier for diagnosis of heart disease and achieved greatest 86.29% accuracy.

5. Artificial Neural Network (ANN)
An artificial neural network (ANN) is a nonlinear statistical model which is also known as "neural system" or "neural framework" used to find out patterns. It consists of input variables, output variables and weights. It is comprised of three interconnected layers namely, input layer, hidden layer and output layer and it also contains association between each layer with particular weights.
The first input layer receives the raw data into the neural network. A neural network may have several hidden layers whose performance is depends on input and weights and relations with other hidden layers. The performance of output layer is determined by activity by hidden layers and weights. In [6], Multilayer perceptron neural network (MLPNN) has been used and achieved accuracy of 82.83%.
Feed forward neural network (FFNN) is one of the simple type of ANN in which connections does not form cycle because of single directed information. In [4], authors integrated fuzzy analytical hierarchical process (Fuzzy AHP) and modified DE algorithm with feed forward neural network and achieved 83% accuracy.

6. Random Forest
Random Forest is an ensemble classifier based on decision tree used for both classification and regression techniques. It constructs several decision trees and integrates them to obtain the best output which is the mean prediction for classification. It usually refers bagging or bootstrap aggregation. In [1], authors proposed hybrid of random forest and linear method (HRFLM) and achieved an accuracy of 88.7%.

7. Ensemble Learning Methods
It is an accumulation of distinct classifiers in which weak classifiers combined with strong classifiers to improve efficiency of weak classifier. These techniques can be used to enhance the accuracy of a classifier. The purpose of combining the multiple classifiers is to obtain better performance as compared with an individual classification method. There are three different models of ensemble technique such as stacking, bagging and boosting[13]. Bagging is also called bootstrap aggregation in which the prophecies of absolutely the similar type are integrated by

voting. In Boosting original dataset is partitioned into various subsets which is trained with classifier and results into series of models. Stacking is a technique of combining multiple classification patterns of distinct types via meta classifier.

Table 2. Prediction accuracy comparison of existing research work

Sr. No	**Contributers**	**Machine Learning Algorithms**	**Accuracy [In %]**
1	S. Joshi et al.,Springer,2015 [11]	DT,NB,KNN	100
2	V. Krishnaiah et al.,Springer,2015 [12]	Fuzzy KNN	90
3	T. Princey et al.,IEEE,2016 [10]	KNN,ID3	40.3,80.62
4	J. Singh et al.,IEEE,2016 [9]	Hybrid of Classification & Association	99.19
5	S. Pouriyeh et al.,ISCC,2017 [6]	MLP & SVM (Stacking T)	84.15
6	T. Vivekanandan et al.,Elsevier,2017[4]	Fuzzy AHP,FFNN	83
7	K. Mathan et al.,Springer,2018[2]	DT+NN	79.8
8	S. Mohan et al., IEEE Access,2019[1]	HRFLM (Voting T.)	88.7

V. DISCUSSION

This paper analyzed several research works carried out in the field of machine learning and deep learning to classify and predict the heart disease.

For the prediction, many of the researchers have used very few factors rather than considering some affecting factors like alcohol, obesity, chest pain (CP), resting electrocardiographic (RECG). One cannot find the precise solution for the prediction without using these affecting factors. All the influencing factors should be considered for accurate prediction. A performance of classifier depends upon the feature selection method which is a very challenging task.

In some research works, researchers have used large datasets but for classification they did not applied proper error handling mechanism such as use of global values at empty spaces, taking mean value of factors and elimination of raw data.

Real-time medical datasets should be used with each individual classification technique or with hybrid technique to enhance the prediction model's performance and accuracy.

VI. CONCLUSION

Heart disease prediction is a very challenging job in the healthcare field. However, the mortality rate can be reduced if the disease is identified at an early phase. This paper investigates various research works done and describes the different classification algorithms and techniques which have been used by the researchers in their work for accurate prediction. New ensemble techniques like hybrid models or multiple learning models that involves combination of different classification methods provides better results with high accuracy. The usages of such hybrid techniques on real-time medical dataset and more influencing parameters of heart disease are expected in future work. Improving the prediction accuracy and performance by removing the existing drawbacks will results in increasing the survival rate.

REFERENCES

[1] S. Mohan, C. Thirumalai & G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", IEEE Access,vol.7, 2019.

[2] K. Mathan , P. M. Kumar, P. Panchatcharam , G. Manogaran, R. Varadharajan, " A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease", Sprinjer, April 2018.

[3] Min Chen, YixueHao, Kai Hwang, Fellow,Lu Wang, and Lin Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", IEEE, vol. 15, pp. 215-227, 2017.

[4] T. Vivekanandan, N.C. Sriman Narayana Iyengar, "Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease", Computers in Biology and Medicine , vol.90, pp. 125–136, 2017.

[5] Z. Arabasadi, R. Alizadehsani , M. Roshanzamir , H. Moosaei , A.Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm", Computer Methods and Programs in Biomedicine, vol. 14 , pp.19–26, 2017.

[6] S. Pouriyeh, S. Vahid, G. Sannino, G. D. Pietro and H. Arabnia, J.Gutierrez, "A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease," 22nd IEEE Symposium on Computers and Communication (ISCC 2017): Workshops-ICTS4eHealth, 2017.

[7] S. Nikan, F. Gwadry-Sridhar, and M. Bauer "Machine Learning Application to Predict the Risk of Coronary Artery Atherosclerosis", IEEE, August 2016.

[8] M. A. Jabbar, S. Samreen, "Heart disease prediction system based on hidden naive bayes classifier", IEEE, 2016.

[9] J. Singh, A. Kamra, H. Singh, "Prediction of Heart Diseases Using Associative Classification", IEEE, 2016.

[10] Therasa Princy R, J. Thomas, " Human Heart Disease Prediction System Using Data Mining Techniques", International Conference on circuit, Power and Computing Technologies [ICCPCT], IEEE, 2016.

[11] Sujata Joshi and Mydhili K. Nair, "Prediction of Heart Disease Using Classification Based Data Mining Techniques", Computational Intelligence in Data Mining – Vol. 2,Smart Innovation, Systems and Technologies 32,Springer, vol.2, 2015.

[12] V. Krishnaiah, G. Narsimha, and N. Subhash Chandra ,"Heart Disease Prediction System Using Data Mining Technique by Fuzzy K-NN Approach", *Emerging ICT for Bridging the Future* ,Advances in Intelligent Systems and Computing 337,Springer, Vol. 1,pp.371– 384, 2015.

[13] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques, 2016.

[14] V. Ramalingam, A. Dandapath and M. Raja, "Heart Disease Prediction using Machine Learning Techniques: a survey", International Journal of Engineering and Technology [IJET], vol.7, pp.684–687, 2018.

[15] "Cleveland Heart Disease Dataset," http://archive.ics.uci.edu/ml/datasets/Heart+Disease