JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Risk Prediction of Heart Disease using Gradient Boosting based Machine Learning Algorithm

Arpita Sharma, Prof. Rupali Chaure, Dr. Ritu Shrivastava

M. Tech. Scholar, Professor, Head of Department

Department of Computer Science & Engineering

SIRT, Bhopal

Abstract: Cardiac disease is a major global health problem in modern medicine. The twenty-first century adage consummate proliferation in life expectancy and a significant transference in the causes of heart disease bereavement throughout the world. The criticality of cardiac diseases is more crucial and can even lead to vulnerable consequences if it is not detected at an earlier stage. The techniques such as electronic health records, body area networks are emerged to continuously monitor and diagnose patient's health conditions through the projection of medical sensors and wearable devices across human bodies. Since the data generated from the body area networks are continuous and tremendous in volume, the machine learning techniques are used for efficient health data classification processes. However, health data classification is the most challenging process as it needs to be executed accurately with an earlier prediction of heart diseases. Machine learning (ML) provides a reliable and excellent support for prediction of a heart diesease with correct case of training and testing. Diagnosis of heart mellitus desires great support of machine learning classifiers to detect diabetes disease in early stage, since it cannot be cured which brings great complication to our health system.

Index Terms **–** Cardiac Disease, Heart Disease, Gradient Boosting, ML

I. INTRODUCTION

Presently, Heart Disease (HD) is considered as a major reason for the increased mortality rate. Based on the study reported by World Heart Federation Report, it is stated that one-third of the death rate can be reduced in the case of earlier identification of Heart Disease. Basic symptoms of Heart Disease are chest pain, breathing issues, neck pain, jaw, esophagus, upper stomach, or back. Also, few limiting factors that assist in minimizing the risks of Heart Disease are controlled Blood Pressure (BP), low cholesterol, avoiding smoking, and habitual exercises. In most of the cases, the Heart Disease could not be identified until a heart attack, or stroke occurs. So, it is needed to observe the cardiovascular parameters and discuss with doctors [1, 2]. The technological enhancements in data and computing have enabled the medical domain to gather and save continuous medical data, which aids in crucial medical decisions. The data which is stored might be examined to create essential medical decisions that might include diagnosis, line of treatments, prediction, and image analysis [3, 4].

The data available in the healthcare system are rich. DM techniques act as a major role in resolving highly nonlinear prediction and classification as well as complex problems over the recent decades. Therefore, it is probable to build a model that might predict the absence or presence of Heart Disease depending on different symptoms of heart-related features. It is a vital necessity of any task of disease prediction to segment the unhealthy and healthy patient precisely. Else, a healthy patient might under needless treatment with a result of misclassification. It is highly significant to predict any occurrence of disease accurately [5, 6].

Cardiac disease is a major global health problem in modern medicine. The twenty first-century adage consummate proliferation in life expectancy and a significant transference in the causes of heart disease bereavement throughout the world. Today it is interpreted for approximately thirty percent decrease across the globe including approximately 40 percent in the high-income country and twenty-eight percent in low and middle-income countries. Compelled by economic development, suburbanization and associated with circadian life changes this constant transition is arising around the world among all races, ethnic groups, and nations at an even faster rate than the last century. A recent development of modern life style exponentially increases the heart failure rates [7].

II. TYPES OF CARCIAC DISEASE

There are a few classifications of heart sicknesses. Figure 2 shows the different kinds of coronary illness dependent on clinical conditions. These classifications are comprehensively delegated myocardial dead tissue, cardiovascular breakdown, heart arrhythmia, angina pectoris, cardiomyopathy, atrial fibrillation dependent on their clinical proof. Coronary illness has numerous highlights, which influence the capacity or structure of the heart [8].

Coronary Artery Disease

The coronary conduit infection is inconvenience prompt by drained course of blood .The consumption supply in corridors will harm the vein and produce the uneasiness to the standard systolic and diastolic capacity of the heart [9].

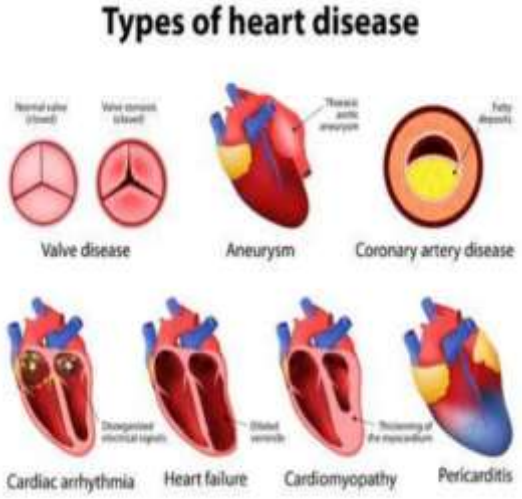

Fig. 1: Types of Cardiac Disease

Acute myocardial infarction

Clinical name for a heart failure is intense myocardial localized necrosis. A heart failure is a condition that greasy substances present in the blood esteem influence the pace of stream which results tissue harm on corridors. The blockage corridors will most likely be unable to supply the oxygenated blood supply to the body which will bring about the brokenness to different organs. Figure 2 clarifies a kind of heart capture brought about by extreme weight [10].

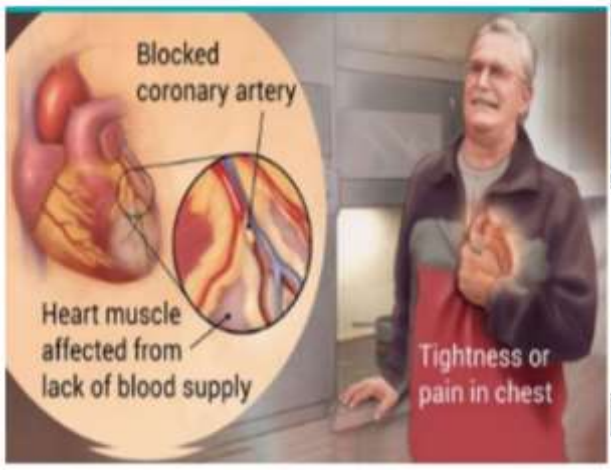

Fig. 2: Acute Myocardial Infarction

Chest Pain (Angina)

Clinical name of chest pressure is Angina. It is overwhelming clinical consideration need crisis treatment for the patients. Patients needs to treated with ventilators promptly on the off chance that we experience this sort of distress. Because of the helpless stock of blood stream will cause the tension on the blood dividers and influence the veins. Which will makes tension on the blood vessals results chest torment. Figure 4 shows common angina caused in the coronary vessel. Stable angina is the condition causes in peritoriam. Sporadic blood stream between the peritoris dividers. The fundamental reasons of flimsy angina are way of life adjustment, social propensities. Figure 3 shows run of the mill unstableangina caused in the coronary vessel [11].

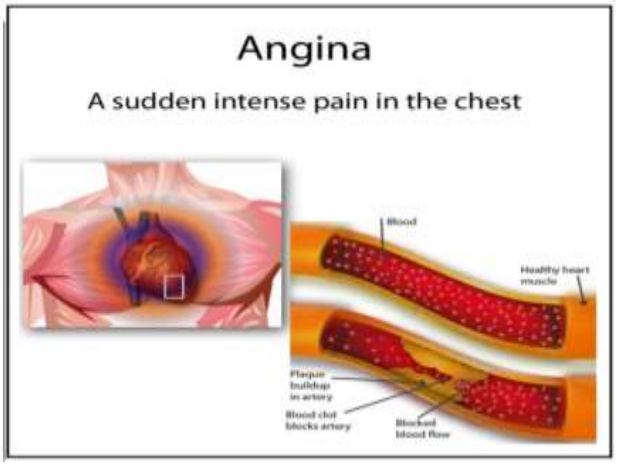

Fig. 3: Angina

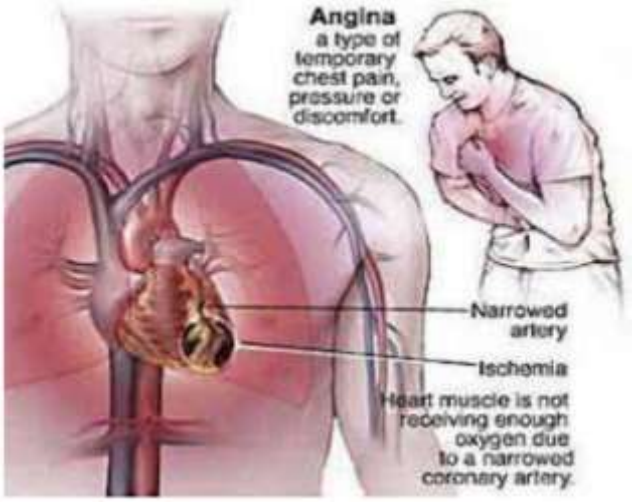

Fig. 4: Unstable Angina

III. PROPOSED METOD

Supervised machine learning classifiers can be categorized into multiple types. These types include naïve Bayes, linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), generalized linear models, stochastic gradient descent, support vector machine (SVM), linear support vector classifier (Linear SVC) decision trees, neural network models, nearest neighbours and ensemble methods. The ensemble methods combine weak learners to create strong learners. The objective of these predictive models is to improve the overall accuracy rate. This can be achieved using two strategies. One of the strategies is the use of feature engineering, and the other strategy is the use of boosting algorithms. Boosting algorithms concentrate on those training observations which end up having misclassifications. There are five vastly used boosting methods, which include AdaBoost, CatBoost, LightGBM, XGBoost and gradient boosting.

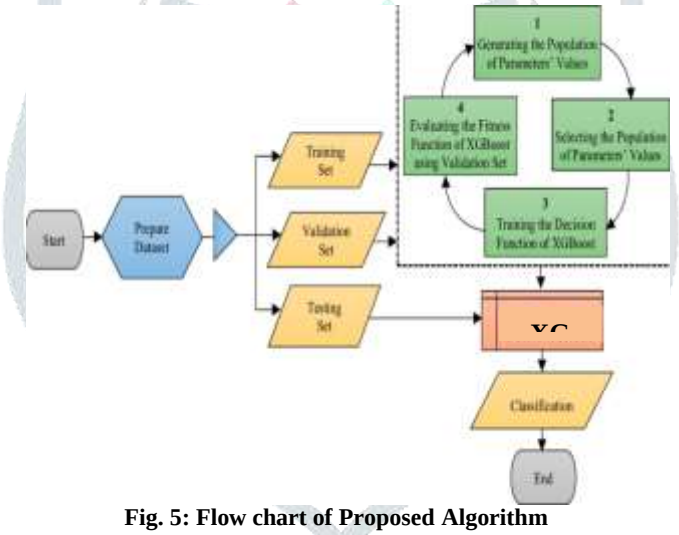

Fig. 5: Flow chart of Proposed Algorithm

Dataset was divided into two datasets (70%/30%, training/testing) to avoid any bias in training and testing. Of the data, 70% was used to train the ML model, and the remaining 30% was used for testing the performance of the proposed activity classification system. The expressions to calculate precision and recall are provided in Equations (1) and (2).

$$\text{Precision} = TP\ TP + FP \times 100 \quad (1)$$

$$\text{Recall} = TP\ TP + FN \times 100 \quad (2)$$

Precision provides a measure of how accurate your model is in predicting the actual positives out of the total positives predicted by your system. Recall provides the number of actual positives captured by our model by classifying these as true positive. F-measure can provide a balance between precision and recall, and it is preferred over accuracy where data is unbalanced.

Algorithm steps:

Input: D= {($\boldsymbol{x1}$,), ($\boldsymbol{x2}$,) ,…, ($\boldsymbol{xN}$,$\boldsymbol{yN}$)}, $\boldsymbol{L}(\boldsymbol{y},\boldsymbol{O}(\boldsymbol{x}))$

Where: ($\boldsymbol{y}$,($\boldsymbol{x}$)) is the approximate loss function.

Begin

Initialize: ($\boldsymbol{x}$) = $\frac{\boldsymbol{argmin}}{\boldsymbol{w}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{L}(\boldsymbol{yi},\boldsymbol{w})$

$\boldsymbol{for\ m}$=$\boldsymbol{1}$:$\boldsymbol{M}$

$$\boldsymbol{r_{im}} = -\frac{\boldsymbol{\partial L(yi, O(xi))}}{\boldsymbol{\partial O(xi)}}$$

Train weak learner $C_m(x)$ on training data
Calculate w: $w_m = \arg\min \sum_{i=1}^{N} L(yi, O_{m-1}(xi) + wC_m(xi))$
Update : $O_m(x) = O_{m-1}(x) + wC_m(x)$
End for
End
Output: $O_m(x)$

Table 1: Heart Disease Attributes Datasets

Sr. No.	Attribute	Representative icon	Details
1	Age	AGE	Patient age (In years)
2	Sex	SEX	Gender of patient (male-0 female-1)
3	Chest Pain	CP	Chest pain type
4	Rest blood pressure	TRESTBPS	Resting blood pressure (in mm Hg on admission to hospital ,values from 94 to 200)
5	Serum cholesterol	CHOL	Serum cholesterol in mg/dl, values from 126 to 564)
6	Fasting blood sugar	FBS	Fasting blood sugar>120 mg/dl, true-1 false-0)
7	Rest electrocardiograph	RESTECG	Resting electrocardiographics result (0 to 1)
8	Max Heart rate	THALCH	Maximum heart rate achieved (71 to 202)
9	Exercise-induced angina	EXANG	Exercise included agina(1-yes 0-no)
10	ST depression	OLDPEAK	ST depression introduced by exercise relative to rest (0 to .2)
11	Slope	SLOPE	The slop of the peak exercise ST segment (0 to 1)
12	No. Of vessels	CA	Number of major vessels (0-3)
13	Thalassemia	THAL	Defect types; 3—normal; 2—fxed defect; 1—reversible defect
14	Target	TARGET	0 or 1

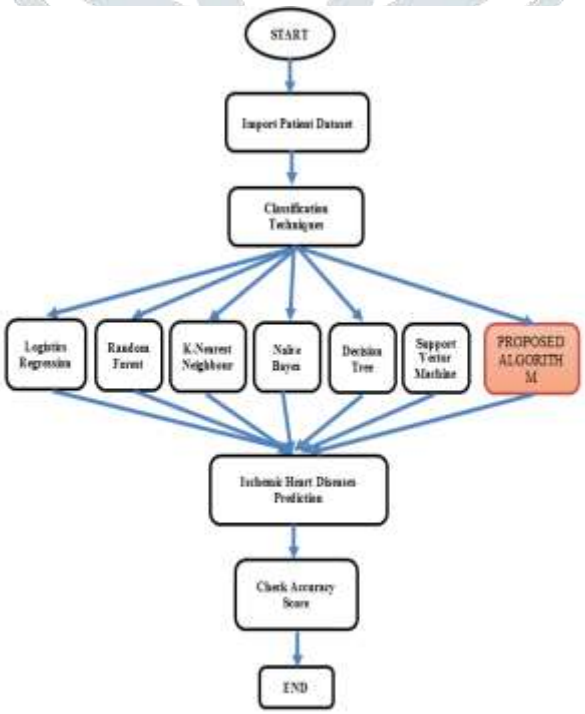

Fig. 6: Flow chart of Proposed Algorithm

IV. SIMULATION RESULTS

There are two classes found in Scikit-learn machine learning library called LabelEncoder and OneHotEncoder. LabelEncoder basically transforms the categorical values into numbers which are ordinal in nature. In data set used for this study, there are categorical variables such as Cp, chest pain type which is represented as 1,2,3 and 4. 1,2,3 and 4 does not have ordinal relationship with each other therefore it gives wrong results when applied directly to machine learning algorithms. Thus, OneHotEncder is used to encode chest pain type values into binary values, this resolves the issue of ordinality. In this data set the dependent variable or the value to be predicted is multi class. It ranges from 0 to 4.

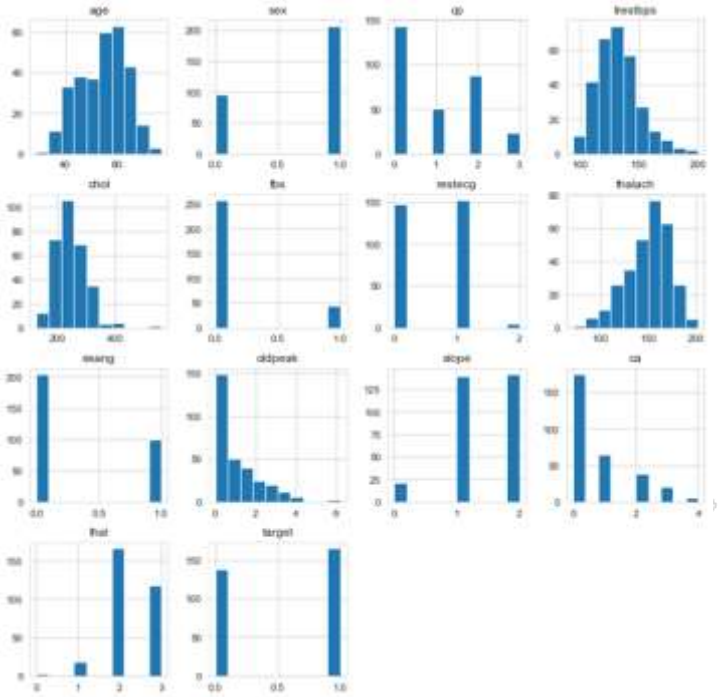

Fig. 7: Histogram of Dataset

Training set is the portion of data in which the model is trained. In this study, 70 percent of data was used for training. In general, in machine learning communities, it is a norm to used 60 to 70 percent of data for training but it varies diversely according to the need 31 and purpose of the experiment. In data training, often the accuracy of training is high, meaning the model shows high level of accuracy performance in the training set but when tested against the test set, the performance is poor. So to avoid performance error, k-fold cross validation was used. In k-fold cross validation, for example 10-fold cross validation, training set is split in 10 parts and and from each 10 part, training and test set is defined and model is employed and the result of all the 10 parts are averaged, this helps to minimize the over fitting and under fitting of the data.

Figure 6 shows the histogram of attributes shows the range of dataset attributes and code which is used to create it.

In proposed algorithm we used an ensemble of SVM, KNN and navies bayes to achieve an accuracy of 92.615%. The Majority vote-based model as demonstrated which comprises of random foreest, Decision Tree and Support Vector Machine classifiers, gave an accuracy of 83.51%, sensitivity of 72.52% and specificity of 82.41% for UCI heart disease dataset.

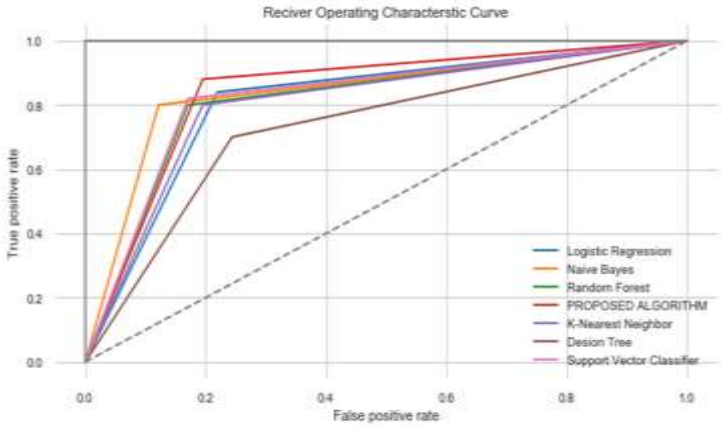

Fig. 8: Roc Curve for Accuracy

After performing the machine learning approach for testing and training we find that accuracy of the knn is much efficient as compare to other algorithms.

Table 2: Accuracy Comparison

Algorithm	Accuracy
Logistic Regression	81.31868131868131
Naive Bayes	83.51648351648352
Random Forest	79.31868131868131
K-Nearest Neighbor Classifier	80.21978021978022
Decision Tree Classifier	72.52747252747253
Support Vector Machine	82.41758241758241
PROPOSED ALGORITHM	**92.61538461538461**

Accuracy should be calculated with the support of confusion matrix of each algorithms as shown in Figures here number of count of TP, TN, FP, FN are given and using the equation (2) of accuracy, value has been calculated and it is conclude that proposed algorithm is best among them with 92.615% accuracy and the comparison is shown in Table 2.

V. CONCLUSION

This research work achieved the highest accuracy of 83.51% with navies bayes without PCA, 82.61% with navies bayes with PCA. Though ensemble classifiers using Boosted Tree, Bagged Tree, Subspace DA, and Subspace navies bayes without and with PCA are trained, it was observed that the ensemble classifiers did not perform better than the single classifiers in term of accuracy.

In the future, intend to perform hyper parameter optimization and conduct more experiments by using feature selection algorithms on a dataset with more observations to improve the classifier's performance.

REFERENCES

[1] Karna Vishnu Vardhana Reddy, Irraivan Elamvazuthi, Azrina Abd Aziz, Sivajothi Paramasivam and Hui Na Chua, “Heart Disease Risk Prediction using Machine Learning with Principal Component Analysis”, International Conference on Intelligent and Advanced Systems (ICIAS), IEEE 2021.

[2] Archana Singh and Rakesh Kumar, “Heart Disease Prediction Using Machine Learning Algorithms”, International Conference on Electrical and Electronics Engineering (ICE3-2020), IEEE 2020.

[3] M.Ganesan and Dr. N. Sivakumar, “IoT based heart disease prediction and diagnosis model for healthcare using machine learning models”, International Conference on System, Computation, Automation and Networking (ICSCAN), IEEE 2019.

[4] Priyan Malarvizhi Kumar, Usha Devi Gandhi, "A novel threetier Internet of Thingsnarchitecture with machine learning algorithm for early detection of heart diseases", Computers and Electrical Engineering, Vol.65, pp. 222–235, 2018.

[5] Prabal Verma, Sandeep K. Sood, "Cloud-centric IoT based disease diagnosis healthcare framework", J, Parrallel Distrib. Comput., 2018.

[6] M.Ganesan, Dr.N.Sivakumar, “A Survey on IoTrelated Patterns”,International Journal of Pure and Applied Mathematics,Volume 117 No. 19, 365-369, 2017.

[7] R.Rajaduari, M.Ganesan,Ms.Nithya “A Survey on Structural Health Monitoring based on Internet of Things”International Journal of Pure and Applied Mathematics,Volume 117 No. 18, 389-393, 2017.

[8] Amin Khatami, AbbasKhosravi, C. L. (2017), ‘Medical image analysis using wavelet transform and deep belief networks’, Journal of Expert Systems With Applications 3(4), 190–198.

[9] Zhang, Shuai, Y.-L. S. A. (2017), ‘Deep learning based recommender system: a survey and new perspectives’, Journal of ACM Computing Surveys 1(1), 1–35.

[10] Zhiyong Wang, Xinfeng Liu, J. G. (2016), ‘Identification of metabolic biomarkers in patients with type-2 diabetic coronary heart diseases based on metabolomic approach’, 6(30), 435–439.

[11] Ashwini Shetty, Naik, C. (2016), ‘Different data mining approaches for predicting heart disease’, International journal of innovative research in science, engineering and technology 3(2), 277–281.

[12] Berikol, B. and Yildiz (2016), ‘Diagnosis of acute coronary syndrome with a support vector machine’, Journal of Medical System 40(4), 11–18.

[13] Chebbi, A. (2016), ‘Heart disease prediction using data mining techniques’, International journal of research in advent technology 25(3), 781–794.

[14] Cheng-Hsiung Wenga, Tony Cheng-Kui Huang, R.-P. H. (2016), ‘Disease prediction with different types of neural network classifiers’, Journal of Telematics and Informatics (4), 277–292.

[15] Ghadge, Prajakta, K. (2016), ‘Intelligent heart attack prediction system using big data’, International journal of recent research in mathematics computer science and information technology 2(2), 73–77.

[16] Lafta, R., Zhang, J. and Tao (2016), ‘An intelligent recommender system based on short-term risk prediction for heart disease patients’, Journal of web intelligence and intelligent agent technology (12), 102–105.