High Accurancy and Low Risk Prediction and Diagnosis Heart Disease using Gradient Boosting Algorithm

Sachin Sahu
M. Tech. Scholar
Department of Computer Science and Engineering
All Saints' College of Technology, Bhopal

Zuber Farooqui
Professor
Department of Computer Science and Engineering
All Saints' College of Technology, Bhopal

ABSTRACT
This paper gives an endeavor to productively arrange and foresee heart illnesses at a beginning phase with high exactness and execution measures. The huge commitment of this exposition is isolated into two sections. Initial, a powerful way to deal with prior location and grouping of coronary illness is portrayed. Next, a fourier change based clinical proposal model is introduced for the previous conclusion of heart diesease. 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. In this paper the implemented result with the help of gradient boosting algorithms.

Keywords
Gradient Boosing, Support Vector Machine, Neural Network, Classification, Heart Disease

1. INTRODUCTION
Heart illness is a significant worldwide medical issue in current medication. The twenty-first-century is aphorism perfect expansion in future and a critical transaction in the reasons for coronary illness mourning all through the world. Today it is deciphered for roughly 30% diminishing over the globe remembering around 40% for the big league salary nation and 28 percent in low and center pay nations. Constrained by financial turn of events, suburbanization and related with circadian life changes this consistent progress is emerging far and wide among all races, ethnic gatherings, and countries at a significantly quicker rate than the only remaining century [1]. An ongoing improvement of current way of life dramatically expands the cardiovascular breakdown rates.

Ongoing examination indicated that the proof of cardiovascular breakdown is significantly increased in the last a quarter century. Ongoing examination expresses that Chronic noninfectious sickness like cardiovascular infection is one of the unmistakable reasons of downfall around the globe. Worldwide ascent in heart sickness impacts from an emotional assignment in the wellbeing status of people far and wide [2].

The heart infection turned into unquestionably the regular schedule of death around the world. The worldwide ascent in heart infection impacts from an emotional concession in the wellbeing status of people far and wide. Heart sicknesses are inauspiciously expanding step by step in the course of recent many years, and it has gotten one of the chief purposes behind mourning in the vast majority of the nations over the globe. Late cardiovascular wellbeing focused overview persuaded that practically 1.2 billion individuals die each year as a result of heart illnesses. There is no single answer for the rising weight of coronary illness, given the monstrous changes in cultural, ethnic, and financial environs. Generally cardiovascular breakdown anticipation is exceptionally an intriguing errand in the night before significant expense proportions [3].

The enormous and complex nature of the clinical consideration information put away across electronic-wellbeing information bases catches the eye of analysts towards clinical applications. This amend parts of clinical administrations with cutting edge electronic-wellbeing methods. Clinical applications, when all is said in done, contain six huge exercises, for example, screening, analysis, treatment, visualization, observing, and the executives [4]. This paper centers principally around the order cycle with wellbeing hazard expectation. The basic target of the section is to expand the forecast system characterized in past parts this proposition points to present scanty standard part investigation strategy for include decrease, and for order, the fluffy min-max neural organization (FMMN) with double cuckoo search is introduced for improvement. The hybridization strategies improve forecast exactness with information pre-preparing and highlight decrease procedures [5,6].

This part accepts the current information mining philosophies to expand upon, and the proposed calculation aids cardiovascular danger expectation and suggestion measures. The coupling of quick fourier changes with AI models is accepted to chip away at the premise of half breed order draws near. It is accepted that the utilization of quick fourier change aids time arrangement investigation of the patient's information and the troupe model backings compelling forecast and clinical suggestion measure. Further, the information dataset related with the proposed framework is thought to be liberated from clamor and missing qualities [7].

2. TYPES OF CARDIAC 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].

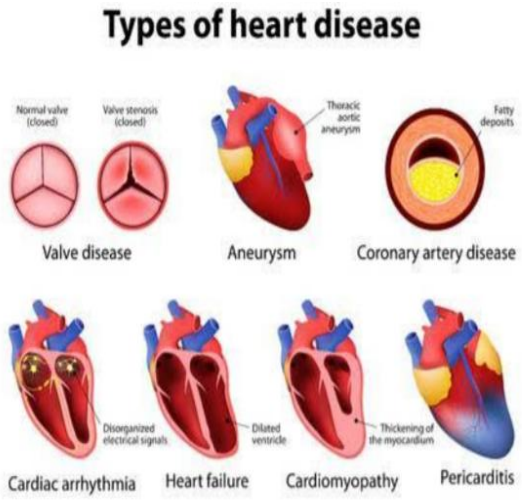

Fig. 1: Types of Cardiac Disease

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].

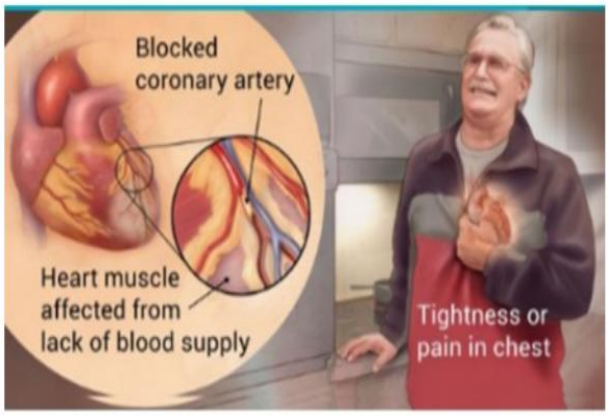

Fig. 2: Acute Myocardial Infarction

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].

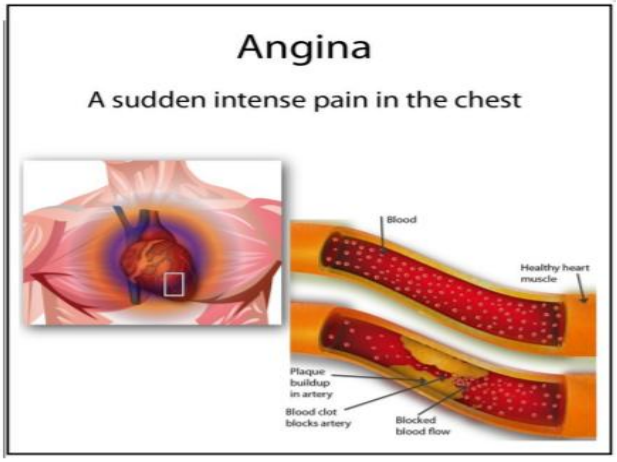

Fig. 3: Angina

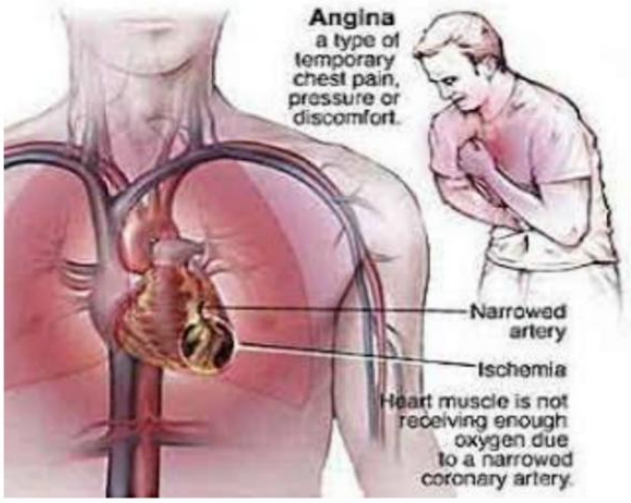

Fig. 4: Unstable Angina

3. PROPOSED METHODS

Gradient boosting (GB) sequentially creates new models from an ensemble of weak models with the idea that each new model can minimize the loss function. This loss function is measured by gradient descent method. With the use of the loss function, each new model fits more accurately with the observations, and thus the overall accuracy is improved. However, boosting needs to be eventually stopped; otherwise, the model will tend to overfit. The stopping criteria can be a threshold on the accuracy of predictions or a maximum number of models created.

The data is collected from UCI machine learning repository. The data set is named Heart Disease Data Set and can be found in UCI machine learning repository. The UCI machine learning repository contains vast and varied amount of datasets which include datasets from various domains. These data are widely used by machine learning community from novices to experts to understand data empirically. Various academic papers and researches have been conducted using this repository.

There are actually 76 attributes in the dataset but only 14 attributes are used for this study, these 14 attributes are in Table 1.

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

The 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).

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

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

Proposed algorithm is an ensemble learning technique, used for classification and regression problems. It can produce an effective model consisting of weak learners, usually decision trees. The basic idea of the proposed method is to build and generalize the ensemble model in a stage wise fashion by optimizing an objective arbitrary loss function. The proposed technique constructs its model from the previous loss function of negative gradient in an iteration manner. In the ML, minimizing the loss function is an important issue and needs to be optimized. In other words, the loss function represents the difference between the predicted output and the target. A low value of loss function means a high prediction or classification result. When the loss function decreases sequentially and iteratively, the model goes consecutively along a specific direction, which is the Gradient of loss function.

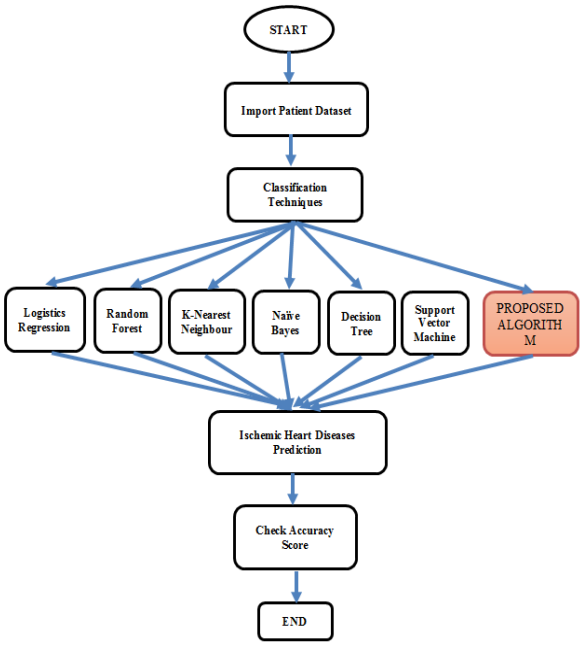

Fig. 5: Flow chart of Proposed Algorithm

4. SIMULATION RESULT

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.

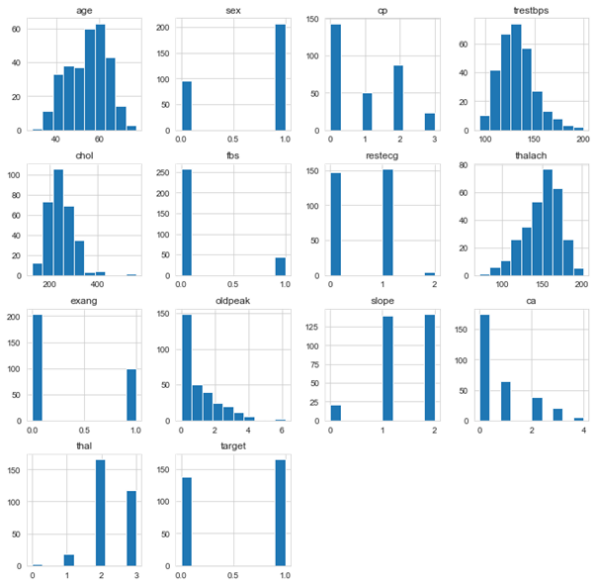

Fig. 6: Histogram of Dataset

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 ANN to achieve an accuracy of 92.615%. The Majority vote-based model as demonstrated by Saba Bashir et al. which comprises of Naïve Bayes, 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.

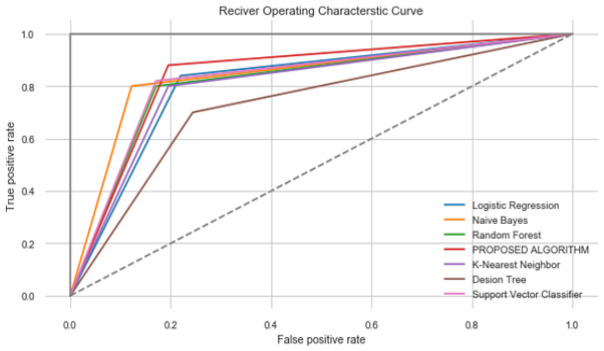

Fig. 7: Roc Curve for Accuracy

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.

5. CONCLUSION

The ongoing appropriation of electronic advances across medical care enterprises expands quick aggregation of clinical data. The electronic wellbeing record may contain patient's important data, for example, finding, test results, clinical history of patient.

The way toward investigating and mining the wellbeing information gives deliberately critical advantages to the wellbeing area. It doesn't just aides the medical services experts additionally give various advantages to every single patient.

The cycle of wellbeing information mining is profoundly inescapable across unsafe infections, for example, cardiovascular sicknesses. The cycle of quicker finding, forecast, and arrangement of ideal treatment, assurance of danger factors, treatment length and its result in an earlier way can spare numerous individuals live from unsafe infections, for example, cardiovascular sicknesses.

6. REFERENCES

[1] 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.

[2] 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.

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

[14] 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.