Classification Comparison of Different Boosting Algorithms to Predict and Classify Conditions of Heart Disease

Vijay Mane, Pranav Belgaonkar, Harshad Dabhade, Ameya Gandhe

Department of Electronics and Telecommunication Engineering, Vishwakarma Institute of Technology, Pune, India

Corresponding author: Vijay Mane, Email: vijay.mane@vit.edu

In the modern era, heart diseases account for the majority of the casualties. Timely and more efficient identification of every particular heart disease proves to be a vital factor in the healthcare sector. This paper presents a system which can predict and classify conditions of heart disease using several ML techniques called boosting algorithms. Cleveland Dataset is used to train this model. This particular dataset consists of 14 variables measured on 1025 individuals who have different health conditions and some having heart disease. The models used include boosting algorithms such as AdaBoost, XGboost, Gradient Boosting, CatBoost and LightGBM. We have compared all these boosting algorithms and calculated their respective accuracy and confusion matrices.

Keywords: Heart Disease, Machine Learning, Gradient Boosting, XGBoost, ADABOOST, CATBOOST, Light GBM.

2023. In Satyasai Jagannath Nanda & Rajendra Prasad Yadav (eds.), *Data Science and Intelligent Computing Techniques*, 869–876. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.56155/978-81-955020-2-8-74

1 Introduction

Heart Diseases (HD) are often considered as major critical health issues since lately many numerous people have been massively agonizing from this disease all around the world. In the last couple of years, almost one out of three has died due to heart disease. One of the major types of heart disease is referred to as the coronary artery disease, which is caused due to the decreased blood flow. This decreased blood flow affects the blood flow to the heart which might lead to a severe heart attack. It can cause a lot of complications in the human body. There are a couple symptoms for HD which includes physical body pain or weakness, breath shortness, swelling of feet, etc. In general, any individual cannot state or distinguish these symptoms of heart disease at any given hour. Thus, the World Health Organization (WHO) plays a vital responsibility in spreading information about heart diseases. It also advises people all over the world to look out for the symptoms of heart disease. They also advise people to visit the doctors and carry out routine check-ups as a precautionary measure.

The diagnosis and conventional medical care of heart disease is quite vigorous when experts, as well as modern technologies, are not readily available. The diagnosis of Heart Disease is carried out by different parameters of a particular patient and his or her medical history. Although there are many cases where the results pre-vailed from this method are not absolutely precise in pinpointing a patient who has the Heart Disease. Moreover, it is exorbitant as well as quite computationally perplex-ing to scrutinize. Machine learning is one of the optimistic technologies which can be used for identifying people with severe heart disease issues. Nowadays, precise-ness and exactness in heart disease diagnosis is playing an important role in the control and therapy of heart failure. Hence, we have developed a model and highlighted a comparison of boosting algorithms based on machine learning (ML) to provide us with the system/ model with the highest accuracy as well as fastest rate.

We have used Cleveland HD Dataset for training our model. This dataset consists of 14 parameters namely Age, Gender, Chest pain, Cholesterol level, Resting blood pres-sure, Resting ECG, Fasting Blood Sugar, Max Heart Rule, Exercise Angina, Oldpeak, ST Slope, Fluoroscopy, Thalassemia and Heart Disease. Our classification model also uses various types of boosting algorithms such as Ada-boost, XG-boost, gradient boosting, cat-boost as well as LGBM. These boosting algorithms can be called meta-algorithms for reducing bias, and variance in supervised learning. We thought of building this model to classify which boosting algorithm is fastest and has the majority of accuracy. This would lead us to identify and use the boosting algorithm to the requirement.

2 Literature review

Senthilkumar Mohan et al.[1] proposed heart disease prediction system using various kind of ML Techniques. The proposed system used a standard UCI dataset. The hybrid approach is used for combining the characteristics of Random Forest as well as Linear Method. These Machine Learning Techniques proved to be useful for the improvement of the accuracy in the prediction of various cardiovascular diseases. Norma Latif Fitriyani et al. [2] proposed a heart disease prediction model which can be widely used for clinical decision support. The proposed system/ model used two datasets: Statlog and Cleveland. The HDPM model is integrated with DBSCAN outlier detection and SMOTE-ENN. The prototype of HDCDSS is helpful for the diagnosis of patients' heart disease status depending on their current state/condition.

Nikhil Gawande et al. [3] proposed a heart disease classification system while making the use of CNN. The proposed system used 1D Convolution Neural Network to give Electrocardiography (ECG) classification. The final output comprises of four classes names 1(Normal), 2 Left bundle branch block (LBBB), 3 Right bundle branch block (RBBB) and 4 (Premature ventricular contraction). Layer Sampling of CNN (7 layers) for the proposed system was carried out. The ECG signal considered in this model was taken from the MIT-BIH dataset. Jian Ping Li et al. [4] proposed an identification method which uses ML techniques in the E-Healthcare sector. The method used the Cleveland Dataset. The

proposed method and model used feature selection algorithms. This feature selection is carried out using various FCMIM FS algorithms. FCMIM-SVM proved too beneficial and achieves good accuracy.

Pranav Motarwar et al. [5] proposed a machine learning model which can be used to analyze the possibility of heart disease using various ML algorithms/techniques . The substructure is executed using 5 algorithms namely Random Forest, Naïve Bayes(NB), Support Vector Machine(SVM), Hoeffding Decision Tree, and Logistic Model Tree (LMT). Cleveland dataset is used in the mod el. The study undermined and compared all of these algorithms . Result of this comparison was found by comparing on the basis of best prediction time . SVM proved to be the algorithm with the best prediction time due to its nature. Alberto Palacios Pawlovsky [6] proposed an ensemble which is hinged on kNN (k Nearest Neighbor) method. It showed results of its application. The ensemble has been carried out with two configurations. The proposed system used a standard UCI heart disease Cleveland dataset. Use of normalization was carried out to shrink the effect of various features which were included and had different ranges of values. Accuracies were calculated on the basis of : Raw data, Standardized data and Normalized data. Various types of distances such as Euclid, Manhattan and Mahalanobis were also used in this study .

Chittampalli Sai Prakash et al. [7] proposed an effective heart-disease prediction system with addition of visualizations on the medical records of the dataset. The proposed system used numerous techniques such as RF, VM, LR and xgboost, and conducted five classifications of heart disease prediction. In the paper it is demonstrated that Support Vector Machine(SVM) and Logistic Regression(LR) perfomed well in terms of heart disease dataset classification. M. H. Abu Yazid et al. [8] proposed the parameter tuning framework for the artificial neural network. Two types of HD datasets namelyStatlog and Cleveland are used to assess the performance of the proposed model or substructure. The substructure gives high accuracy. Accuracies for both datasets were divided into 3 phases of dataset accuracy. The estimated accuracy for Cleveland dataset was found out to be 90.9% whereas it was 90% for Statlog dataset.
SeyedaminPouriyeh et al. [9] proposed a paper that aims to investigate and compare the accuracy of various classification techniques. Ensemble ML Techniques are employed for the predictions. Cleveland dataset is used in this system. Various ML classifiers were compared in this study. All these classifiers were compared under the experimentation of ensemble learning methods such as bagging, stacking as well as boosting. From the study as well as experimentation, SVM can be concluded as the machine learning classifier with maximum accuracy. M. Kavitha et al. [10] proposed an innovative approach to predict different kinds of heart diseases while making use of ML techniques. The proposed work used different classifiers such as RF, DT and their hybrid combination. Cleveland dataset along was implemented in this substructure/system. Results determine accuracy of approximately around 88.7% with Hybrid system.

T.P.Naidu et al.[11] proposed a hybrid model which manoeuvred several Machine Learning techniques. Cleveland Dataset was used by the authors to implement the model. Feature extraction was carried out using Genetic and PSO algorithms. Moreover, a neural network was used in the prediction model. Prediction model proved to be helpful in estimating the accuracy as the prediction model was applied on the testing data.

These literature papers provide us with various detailed information. Every paper had their different approach to study and deploy models related to heart disease identification, classification as well as support systems.

3 Methodology

3.1 Dataset

To train the model Cleveland HD Dataset is used. This dataset consists of 14 parameters/ variables namely Age, Gender, Chest pain, Cholesterol level, Resting blood pressure, Resting ECG, Fasting Blood Sugar, Max Heart Rule, Exercise Angina, Oldpeak, ST Slope, Fluoroscopy, Thalassemia and Heart Disease. The individuals are further grouped into three levels of heart disease as shown in Table 1.

Table 1: Parameters of Cleveland Dataset.

Age	Patient's age in years
Gender	Patient's gender(Male, Female)
Chest Pain	chest pain type Value 1- Typical angina Value 2- Atypical angina Value 3- Non-anginal pain Value 4- Asymptomatic
Resting Blood Pressure	blood pressure (in mmHg)
Cholesterol	Cholesterol (unit being mg/dl)
Fasting Blood Sugar	Is the quantity or unit of fasting blood sugar > 120 mg/dl
Resting ECG	ECG measurements
Max Heart Rate	Maximum heart rate recorded for a patient
Exercise Angina	Whether the patient shows sign of angina while exercising
Oldpeak	ST depression instigated by exercise correlative to rest
ST Slope	the slope of the peak exercise ST segment -- Value 1: upsloping -- Value 2: flat -- Value 3: downsloping
Fluoroscopy	No. of major vessels (0-3) colored by fluoroscopy
Thalassemia	This disease causes the body to have less hemoglobin than normal
Heart Disease	Target variable - it determines the condition of patient i.e if a particular patient is suffering from heart disease or not

3.2 Workflow of the model

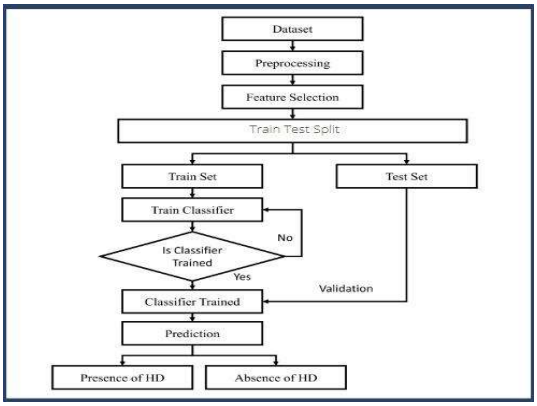

Fig 1: Workflow of the model

The proposed methodology of the system is presented as shown in Fig. 1. First step involved in the process is exploring the data. It can be termed as understanding what the variables mean and gaining some insights to get a deeper understanding of the problem statement. In our model, we have considered various data such as distribution of people suffering from heart disease arranged according to age groups, distribution of patients having heart disease according to gender and maximum and minimum health indicators for patients with and without fasting blood sugar. Below are some insights that were obtained from the data as shown in Fig. 2 to 4.

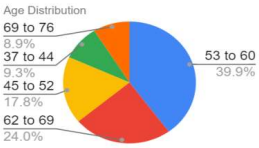

Fig 2: Distribution of patients having heart disease according to age

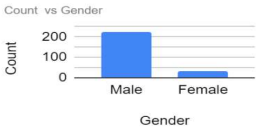

Fig 3: Distribution of patients having heart disease according to gender

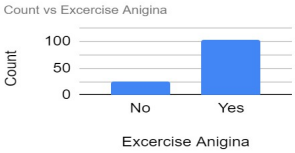

Fig 4: Distribution of patients who have heart disease, and have maximum heart rate not in the range from 130 to 180

Table 2: Maximum and minimum health indicators for patients with and without fasting blood sugar

Fasting Blood Sugar	Minimum Blood Pressure	Maximum Blood Pressure	Minimum Cholesterol	Maximum Cholesterol	Count of Patients
No	94	200	85	564	436
Yes	101	190	123	603	95

After studying and exploring the data, we have followed the traditional training, testing as well as splitting method. The data was split into 2 halves - testing data and training data. The training data was utilized to train the ml model and the testing data was used to check how good the model is using metrics such as accuracy and confusion matrices. 60 % of the data was used for model training and the remaining 40 % data was used for model testing the data. The data was split randomly with a parameter called seed. The seed method is used to initialize random numbers so that the split is exactly the same, each and every time the train test function is executed.

Initially, after train test and split,we have used Boosting algorithms on trained data . Boosting algorithms like catboost, adaboost, gradient boosting, XGboost as well as LGBM were used on the trained data. We have explained all of the boosting algorithms to get a brief idea about all the algorithms and their respective use. All of the boosting algorithms used in our model are as follows:

ADABOOST: The very fundamental concept that Adaboost follows is setting the weights of various classifiers and then training the sample in each of the iterations. This method is carried out in such a way that certifies the more accurate predictions of the unusual observations. Any ML algorithm can be used as a base classifier if it credits the weights on the training set. Coming to the advantages of adaboost, adaboost is very less susceptible to overfitting as the input parameters are not jointly optimized. The accuracy of various weak classifiers can also be improved by using Ada-boost. Nowadays, Ada-boost is generally being used to classify text and images rather than binary classification problems [12].

XGBOOST: XGBoost can be termed or classified as a decision-tree-based aggregated ML algorithm that makes use of a gradient boosting substructure. During the case of prediction problems which involve means of various unstructured data (images, text, etc.), artificial neural networks favouringly excel all of the other types of algorithms and frameworks. It uses more accurate approximations to discover the best tree model. XGBoost is implemented for supervised learning problems, where we maneuver the training data to estimate an intended entity. xgboost is well suited for classification problems, especially those which are emphasized on various business problems like fraud detection .

CATBOOST: CatBoost is an algorithm which is widely used for gradient boosting on different kinds of decision trees. It is easy to use and works efficiently well with heterogeneous data and even with relatively small data. It can be also referred to as an open source which is used for supervised learning / text in Machine Learning.

LightGBM: LGBM is often referred to as a gradient boosting substructure which is deployed on decision trees to intensify or improve the efficiency of the system and reduce memory usage. We have also carried out Hyperparameter Tuning in our model. Selection of optimum hyperparameters is very important to fit models for this particular dataset, to obtain maximum accuracy.

The last step involved in the methodology section is Evaluation. Accuracy and confusion matrices were obtained for these 5 boosting algorithms, comparing and concluding the best algorithm amongst. Accuracy checks the number of correct predictions.

4 Results and Discussion

We have used a confusion matrix in our model as well. Confusion matrix is used to obtain the following parameters as shown in Fig. 5:
1. True positives (TP): True positives can be simply stated as the cases in which we made an assumption or predicted yes (i.e. the patient has the disease), and it eventually turns out to be the case that the patient is infected with disease.
2. True negatives (TN): True negatives can be simply stated as the cases in which we predicted the result no, and the person doesn't have the disease.
3. False positives (FP): False positives can be stated as the cases in which we predicted the result, yes, but the patient is not infected with the disease. ("Type I error.")
4. False negatives (FN): False Negatives can be stated as the cases in which we predicted results as no, but the patients actually are infected with the disease. ("Type II error.")

Confusion matrices for each of the boosting algorithms were as follows:

AdaBoost	
184	5
0	221

XGBoost	
179	10
6	215

Gradient Boosting	
184	5
4	217

LGBM	
187	2
2	219

CatBoost	
186	3
0	221

Fig 5. Confusion matrices for all the trained models

After testing the boosting algorithms, the accuracies obtained were as follows:
Adaboost: 98.78 %
XGboost: 96.09 %
Gradient Boosting: 97.80 %
LGBM: 99.02 %
Catboost: 99.26 %

CatBoost prevents leakage as well as overfitting. It also supports all forms of features such as numeric , categoric or may it be text. It saves time as well as effort which are involved in preprocessing. ThusCatboost proves to be best as it has prediction time faster than all other algorithms. Thus, it can be concluded from the results that the model with the best accuracy for this particular dataset was CatBoost after hyperparameter tuning and the second best model was LightGBM.

5 Future and Scope

The future view of the paper would be to add extra input parameters as well as features and henceforth analyze their results using proposed models. We will also compare different kinds of ml classifiers /algorithms like Naive Bayes, Decision tree, SVM, etc. We are also planning to deploy a model based on the most accurate machine classifier as well as most accurate boosting model so that it could be used for clinical support.

6 Conclusion

Recognition of Heart Disease is carried out by the different parameters of the particular patient and his or her medical history. Although there are many cases where the results prevailed from this approach

are not precise in pinpointing the patient who has the Heart Disease. Moreover, it is exorbitant as well as quite computationally perplexing to scrutinize. Thus we have developed a system/model consisting of various boosting algorithms and presented systematically the comparison of these algorithms.

References

[1] Senthilkumar Mohan, ChandrasegarThirumalai, Gautam Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", Volume7, ISSN: 2169-3536

[2] Norma Latif Fitriyani, Muhammad Syafrudin, GanjarAlfian, Jongtae Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System", Volume 8, ISSN: 2169-3536

[3] Nikhil Gawande, Alka Barhatte, "Heart diseases classification using convolutional neural network", IEEE, 9-20 October 2017, INSPEC Accession Number: 17650687, DOI: 10.1109/CESYS.2017.8321264

[4] Jian Ping Li, Amin UlHaq, Salah Ud Din, Jalaluddin Khan, Asif Khan, Abdus Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare", IEEE Access (Volume: 8), ISSN: 2169-3536

[5] P. Motarwar, A. Duraphe, G. Suganya and M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1-5, doi: 10.1109/ic-ETITE47903.2020.242.

[6] A. P. Pawlovsky, "An ensemble based on distances for a kNN method for heart disease diagnosis," 2018 International Conference on Electronics, Information, and Communication (ICEIC), 2018, pp. 1-4, doi: 10.23919/ELINFOCOM.2018.8330570.

[7] C. S. Prakash, M. Madhu Bala and A. Rudra, "Data Science Framework - Heart Disease Predictions, Variant Models and Visualizations," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), 2020, pp. 1-4, doi: 10.1109/ICCSEA49143.2020.9132920.

[8] M. H. Abu Yazid, M. Haikal Satria, S. Talib and N. Azman, "Artificial Neural Network Parameter Tuning Framework For Heart Disease Classification," 2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2018, pp. 674-679, doi: 10.1109/EECSI.2018.8752821.

[9] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," 2017 IEEE Symposium on Computers and Communications (ISCC), 2017, pp. 204-207, doi: 10.1109/ISCC.2017.8024530.

[10] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.

[11] Naidu, K Amar Gopal, Sk Rameez Ahmed, R Revathi, SkHasaneAhammad, V Rajesh, Syed Inthiyaz, K Saikumar "A Hybridized Model for the Prediction of Heart Disease using ML Algorithms," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2021, pp. 256-261, doi: 10.1109/ICAC3N53548

[12] Ying, Cao, et al. "Advance and prospects of AdaBoost algorithm." Acta AutomaticaSinica 39.6 (2013): 745-758.