PREDICTION OF CARDIOVASCULAR DISEASE USING GRADIENT BOOSTING

Gurappagari Likhith Sai[1], Kathi Venu[2], Dr.R.Venkata Ramana Chary[3]

[1] *Student, Information Technology, B V Raju Institute of Technology, Telangana, India*
[2] *student, Information Technology, B V Raju Institute of Technology, Telangana, India*
[3] *Associate Professor, Information Technology, B V Raju Institute of Technology , Telangana, India*

ABSTRACT

Heart monitoring is one of the most important aspect of healthcare monitoring. For detecting risk indicators, machine learning algorithms are a potential way. To acquire accurate cardiac disease prediction, this work introduces a gradient boosting (GB) approach. The aim of this project is to use the Machine Learning (ML) technology on medical data to predict the chance of cardiovascular disease(CVD). The medical data available these days is often a valuable resource of quality information. This project aims to use supervised machine learning algorithms like Gradient Boosting, Logistic Regression, Decision Tree to obtain a CVD classifier. However, the challenge usually lies in the quantum of the data since there aren't many datasets available regarding CVD, the correlations and the accuracy of the model when it comes to prediction using conventional techniques. This project proposes the use multiple datasets available in UCI repository namely Cleveland, Hungary, Statlog, Long Beach, Switzerland in deriving classifiers using above mentioned algorithms. The classifiers obtained using the datasets are then used to build a web application for users for them to know whether they are vulnerable to CVD or not based upon the details provided by them as input i.e., features that are used in classification of CVD. Depending on the prediction of the model the output is provided to the user, according to which the user should undergo lifestyle changes. The final goal is identification of the possibility of heart disease that can reduce the risk of mortality from cardiovascular disease (CVD).

Keyword: *Cardio vascular disease, Gradient boosting, Patients health records.*

1. INTRODUCTION

Arrhythmia refers to an abnormal heart rhythm, which can occur due to various factors such as heart disease, electrolyte imbalances, drug toxicity, or genetic factors. It is characterized by an irregular heartbeat that can be too slow (bradycardia), too fast (tachycardia), or irregular (fibrillation).To find Arrythmia we use an algorithm known as Gradient Boosting.

Gradient Boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differential loss function.

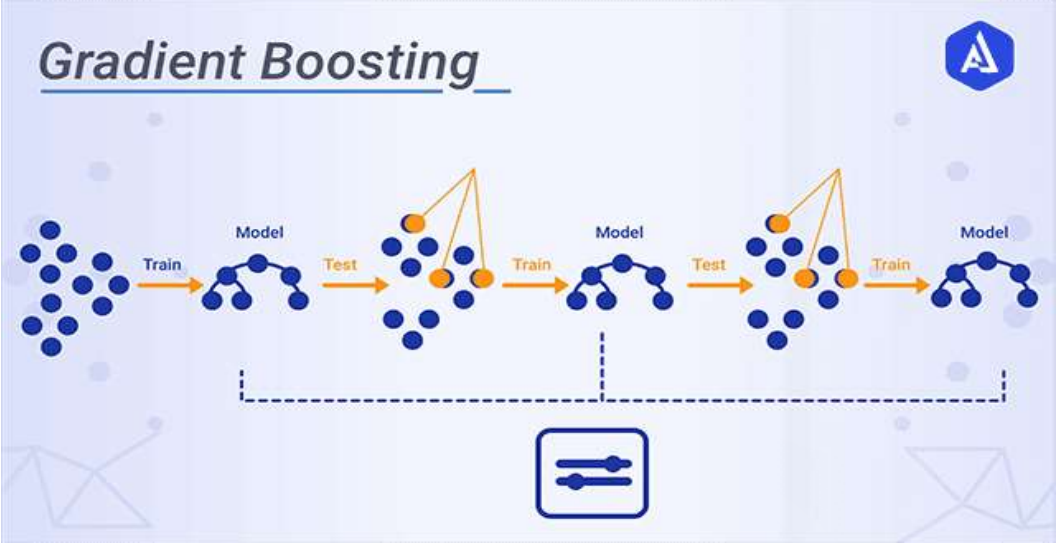

Fig-1*:* Gradient Boosting Algorithm

Fig. 1 displays the flow in gradient boosting algorithm which are used to prepare a model that could identify the chances of getting effected with heart disease known as arrhythmia.
There are other algorithms that are used for determining the arrhythmia. But in our project we are going to compare the with those algorithms and we are going to prove that the gradient boosting is the most accurate algorithm for finding arrhythmia. This can be done by checking the accuracy of each algorithm. The algorithms that are used for comparison are:

1.1 Decision Tree Algorithm

The Decision Tree algorithm is used for solving both classification and regression problems but is mostly involved in classification. The Decision Tree algorithm creates a training model that is used to predict the target value by learning some simple decision rules.

1.2 Logistic Regression Algorithm

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorial dependent variable. Therefore the outcome must be a categorial or discrete value. It can be either Yes or No, 0 or 1, True or False, etc. But instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression can be used to classify the observation using different types of data and can easily determine the most effective variables used for the classification.

2. METHODOLOGY

The raw dataset is the data that is acquired from UCI Machine Learning Repository from credible sources. Then the data pre-processing phase starts where the dataset is cleaned of any noise and duplicate values. In this case, the dataset also contains some missing values. These missing values were replaced with mean value of the respective column data since it provided better accuracy compared to that of the remaining methods of handling missing data. Finally, 918 instances were used in this project after data pre-processing. This phase is also responsible for feature scaling to make the dataset appropriate for distance-based algorithms, Gradient Boosting in this case. The feature extraction or attribute selection also takes place in this phase itself. The heart disease dataset consists of 13 attributes in total. After the attribute selection using filter-based method named missing value ratio two attributes were

removed. After the data pre-processing the data splitting was done, where the data was split into training data and testing data in 3:1 ratio which means 688 instances were used for training and remaining 230 were used for testing the models for evaluation purpose. Now, multiple supervised algorithms like Gradient Boosting, Decision Tree and Logistic Regression were used to build the classifiers using training data. These trained models were then tested using testing data and different evaluation measures like specificity, sensitivity, F1 score etc. Lastly, the classifier with highest accuracy is taken to build a web application for users to input values know whether they are prone to have heart disease or not. This process is be carried to the server to be used with the web application carried out using a concept the model built and tested will known as serialization where the model built and tested will known as serialization where the model built and tested will be carried to the server to be used with the web application.

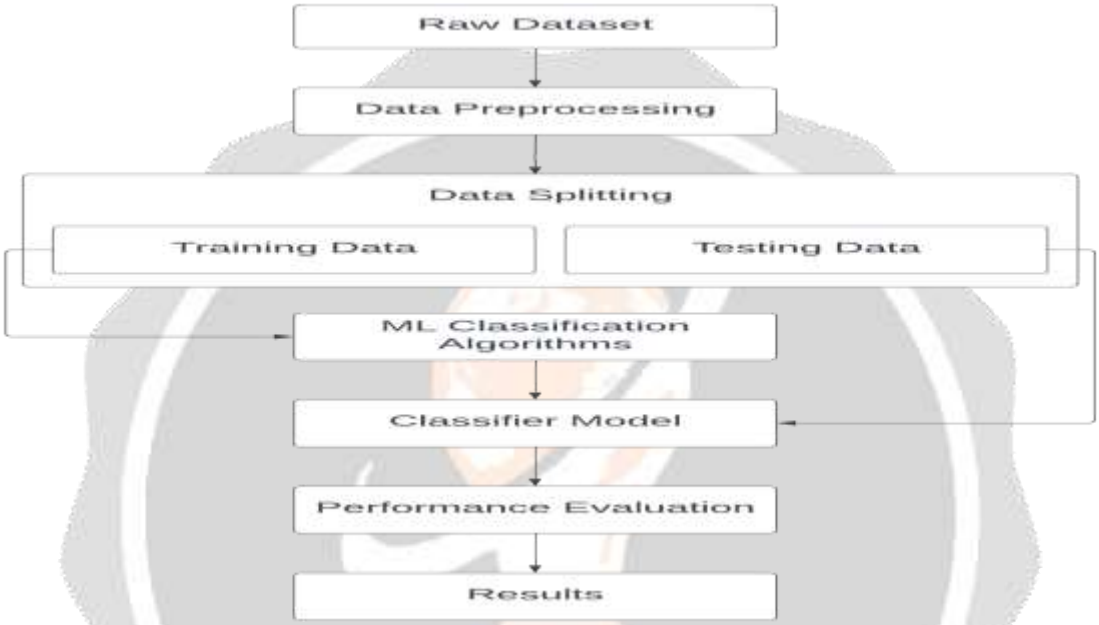

Fig-2: Architecture of Proposed System

Fig. 2 shows the proposed System's architecture.

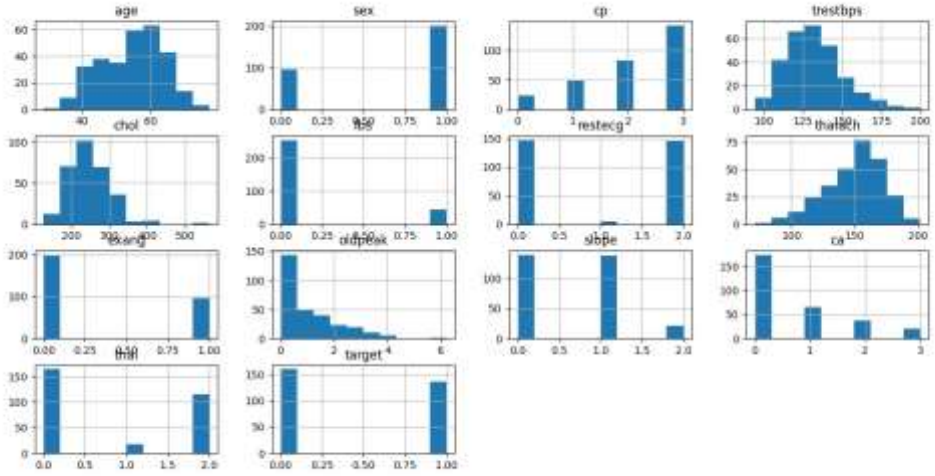

Fig 3: Graphical representation of attributes

2.1 DESCRIPTION OF MODULES

Modules present in predicting the possibility of heart disease project are:

1. **User module:** It is the only part visible to the user. Here the user enters the required inputs in the attributes provided by us. Takes the input from the users the features/attributes considered for predicting the possibility of heart disease. When user clicks the submit button provided at the lower end of the web page then the input are taken and it will move to next screen which shows the results that there is any possibility of getting affected with heart disease or not.

2. **Building Module:** In building module our machine learning model will be build by using Gradient Boosting Algorithm. We split the dataset for training and testing the model.

3. **Prediction module:** Here predictions are made by the model. This prediction module is responsible for building the classification model that is used for predicting the possibility of heart disease. From here the results are sent and shown to the user about the possibility of heart disease.

3. RESULTS

After entering the inputs when we click submit button then the results will be visible to the user, whether he/she is going to be affected with heart disease or not. The proposed system's results are shown in the figures below.

Fig 4: Results of Heart Disease Prediction

The proposed system results are shown below in a graphical representation. Here the graph shows accuracy of every algorithm and how best is our Gradient Boosting algorithm compared to other algorithm such as Decision Tree and Logistic Regression.

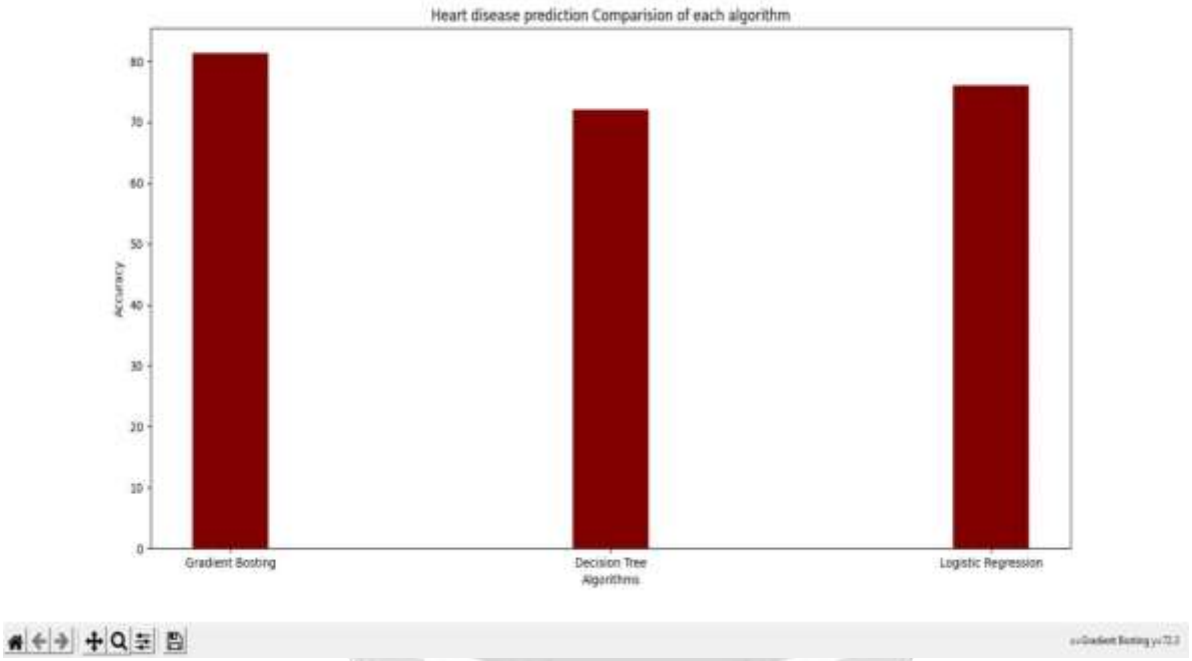

Fig 5: Graphical Representation of Results

4. CONCLUSIONS

For any machine learning model, the first thing that comes to our mind is how we can build an accurate & 'good fit' model and what the challenges are that will come during the entire procedure. Identifying and processing raw data of heart disease will help in the long term in early detection of abnormalities in heart conditions. Although there are several limitations to overcome to use Machine Learning in healthcare industry, overall, these supervised machine learning algorithms showed good results. Early identification of heart disease is challenging and very important in the medical field. It also shows how best our algorithm Gradient Boosting is when compared to other algorithms accuracy.

CLASSIFIER	ACCURACY
Gradient boosting	81.3
Logistic Regression	76.0
Decision Tree	73.3

Fig 6: Accuracy of Each Algorithm

5. REFERENCES

[1]. King, M. A. (2018). Dementia could be detected via routinely collected data, new research shows. Retrieved from University of Plymouth
Website:https://www.plymouth.ac.uk/news/dementiacould-bedetected-via-routinelycollected-data-new-research-shows.

[2]. World Health Organization. (2021, June 11). "Cardiovascular Diseases CVDs". Retrieved from: www.who.int/en/news-room/factsheets/detail/cardiovasculardiseases-(cvds).

[3]. Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2020). Secure and robust machine learning for healthcare: A survey. IEEE Reviews in Biomedical Engineering, 14, 156-180.

[4]. Hassan, S. A., & Khan, T. (2017). A Machine Learning Model to Predict the Onset of Alzheimer Disease using Potential Cerebrospinal Fluid (CSF) Biomarkers. International Journal of Advanced Computer Science and Applications, 8(12), 124-131.

[5]. Sun W, Zhang P, Wang Z. Prediction of cardiovascular diseases based on machine learning. ASP Transactions on Internet of Things. 2021 May 29;1(1):30-5.

[6]. Hossen, M.D., Tazin, T., Khan, S., Alam, E., Sojib, H.A., Monirujjaman Khan, M., & Alsufyani, A. (2021). Supervised Machine Learning-Based Cardiovascular Disease Analysis and Prediction. *Mathematical Problems in Engineering*.

[7]. K. Bourzac, "The computer will see you now", *Nature*, vol. 502, no. 3, pp. S92S94, 2013.

[8]. Louridi N, Amar M, El Ouahidi B. Identification of cardiovascular diseases using machine learning. In2019 7th mediterranean congress of telecommunications (CMT) 2019 Oct 24 (pp. 1-6). IEEE.

[9]. Gonsalves AH, Thabtah F, Mohammad RM, Singh G. Prediction of coronary heart disease using machine learning: an experimental analysis. InProceedings of the 2019 3rd International Conference on Deep Learning Technologies 2019 Jul 5 (pp. 51-56).