Heart Disease Prediction Using XAI

1st Shimul Mandal Tamo
Department of Computer Science and Engineering
BRAC University
66 Mohakhali, Dhaka 1212, Bangladesh
shimul.mondol.tamo@g.bracu.ac.bd

2nd Md. Fazle Rabbi Linkon
Department of Computer Science and Engineering
BRAC University
66 Mohakhali, Dhaka 1212, Bangladesh
fazle.rabbi.linkon@g.bracu.ac.bd

3rd Shahria Hoque
Department of Computer Science and Engineering
BRAC University
66 Mohakhali, Dhaka 1212, Bangladesh
shahria.hoque@g.bracu.ac.bd

4th MD. Mustakin Alam
Department of Computer Science and Engineering
BRAC University
66 Mohakhali, Dhaka 1212, Bangladesh
md.mustakin.alam@g.bracu.ac.bd

5th Md Sabbir Hossain
Department of Computer Science and Engineering
BRAC University
66 Mohakhali, Dhaka 1212, Bangladesh
md.sabbir.hossain1@g.bracu.ac.bd

6th Annajiat Alim Rasel
Senior Lecturer
Department of Computer Science and Engineering
BRAC University
66 Mohakhali, Dhaka 1212, Bangladesh
annajiat@bracu.ac.bd

***Abstract*—Machine learning is a way of manipulating and extracting implicit, previously unknown/known and potential useful information about the data. Machine learning Includes various classifiers Supervised, unsupervised, and ensemble learning that are used to predict and search for accuracy given data file. We can use this knowledge in our heard diseases prediction project. Cardiovascular diseases are very common nowadays, they describe a number of conditions that could affect them our heart The World Health Organization estimates that 17.9 million global deaths from CVD(cardiovascular disease). It is the leading cause of death in adults. Our project can help predict who is likely to diagnose with heart disease using their medical history. It recognizes who all has some symptoms heart disease, such as chest pain or high blood pressure, and can help diagnose disease with less medical means tests and effective treatments so that they can be cured accordingly.**

***Index Terms*—machine learning, heart diseases, KNN, Random forest, Logical Rigression**

I. Introduction

Our work mainly focuses on three data mining techniques, namely: (1) Logistic Regression, (2) KNN and (3) Random Forest Classifier. The accuracy of our project is 87.5%, which is better than the previous one system where only one data mining technique is used. So using more data mining techniques has increased HDPS accuracy and efficiency. Logistic regression falls under the category of supervised learning. Only discrete values are used in logistic regression. The aim of this project is to verify whether a patient is likely to be diagnosed with one cardiovascular heart disease based on their medical attributes such as gender, age, chest pain, fasting sugar level etc. A data-set with patient history and attributes is selected from the UCI repository. According to we use this data set to predict whether a patient may have heart disease or not. To predict this, we use 12 medical characteristics of the patient and classify it if the patient is likely to have heart disease. These medical attributes are trained according to three algorithms: **Logistic Regression**, **KNN** and **Random Forest Classifier**. Most the efficient one among these algorithms is KNN, which gives us an accuracy of 90.63%. And finally we classify patients who are at risk of heart disease or not and also this method is completely cost effective.

II. Literature Review

Lots of work has been carried out to predict heart disease using UCI Machine Learning data-set. Different levels of accuracy have been attained using various data mining techniques which are explained as follows. Avinash Golande studies various different ML algorithms that can be used for classification of heart disease. Research was carried out to study Decision Tree, KNN and K-Means algorithms that can be used for classification and their accuracy were compared[1]. This research concludes that accuracy obtained by Decision Tree was highest further it was inferred that it can be made efficient by combination of different techniques and parameter tuning. T.Nagamani have proposed a system [2] which deployed data mining techniques along with the MapReduce algorithm. The accuracy obtained according to this paper for the 45 instances of testing set, was greater than the accuracy obtained using conventional fuzzy artificial neural network. Here, the accuracy of algorithm used was improved due to use of dynamic schema and linear scaling. Fahd Saleh Alotaibi has designed a ML model comparing five different algorithms [3]. Rapid Miner tool was used which resulted in higher accuracy compared to Matlab and Weka tool. In this research the accuracy of Decision Tree, Logistic Regression, Random

forest, Naive Bayes and SVM classification algorithms were compared. Decision tree algorithm had the highest accuracy. Anjan Nikhil Repaka, proposed a system in [4] that uses NB (Naïve Bayesian) techniques for classification of dataset and AES (Advanced Encryption Standard) algorithm for secure data transfer for prediction of disease. Theresa Princy. R, executed a survey including different classification algorithm used for predicting heart disease. The classification techniques used were Naive Bayes, KNN (K-Nearest Neighbour), Decision tree, Neural network and accuracy of the classifiers was analyzed for different number of attributes [5]. Nagaraj M Lutimath, has performed the heart disease prediction using Naive bayes classification and SVM (Support Vector Machine). The performance measures used in analysis are Mean Absolute Error, Sum of Squared Error and Root Mean Squared Error, it is established that SVM was emerged as superior algorithm in terms of accuracy over Naive Bayes [6]. The main idea behind the proposed system after reviewing the above papers was to create a heart disease prediction system based on the inputs as shown in Table 1. We analysed the classification algorithms namely Decision Tree, Random Forest, Logistic Regression and Naive Bayes based on their Accuracy, Precision, Recall and f-measure scores and identified the best classification algorithm which can be used in the heart disease prediction.

III. Methodology

A. Dataset Description

The dataset that we have used for our project is a public dataset named heart-diseases-dataset that has been taken from the platform Kaggle. This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The ”target” field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

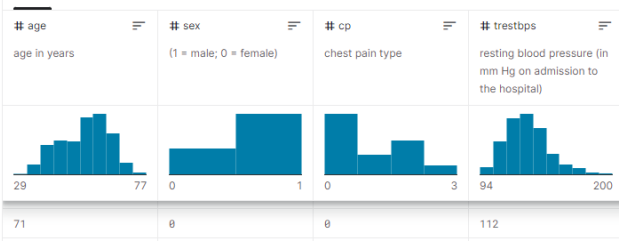

Fig. 1. Heart Disease Dataset

B. Data pre-processing

itemize Heart disease data is pre-processed after collection of various records. The dataset contains a total of 1189 patient records, where no records found with any missing values. So these 1189 patient records are used in pre-processing. The multi- class variable and binary classification are introduced for the attributes of the given dataset. The multi-class variable is used to check the presence or absence of heart disease. In the instance of the patient having heart disease, the value is set to 1, else the value is set to 0 indicating the absence of heart disease in the patient. The pre-processing of data is carried out by converting medical records into diagnosis values. The results of data pre-processing for 1189 patient records indicate that 628 records show the value of 1 establishing the presence of heart disease while the remaining 561 reflected the value of 0 indicating the absence of heart disease.

Sl	Attribute	Distinct Values of Attribute
1	*Age*	Values between 29 and 71
2	*Sex*	0,1
3	*CP*	0,1,2,3
4	*Rest BP*	80 to 200
5	*cholesterol*	65 to 603
6	*FBS*	0,1
7	*RestECG*	0,1,2
8	*Heartbeat*	60 to 202
9	*Exang*	0,1
10	*Oldpeak*	0,1
11	*Slope*	1,2,3
12	*Target*	0,1

C. Classification

The attributes mentioned in Table 1 are provided as input to the different ML algorithms such as Logical Regression, KNN and Random Forest classification techniques. The input data-set is split into 80% of the training data-set and the remaining 20% into the test data-set. Training data-set is the data-set which is used to train a model. Testing data-set is used to check the performance of the trained model. For each of the algorithms the performance is computed and analysed based on different metrics used such as accuracy, precision, recall and F-measure scores as described further. The different algorithms explored in this paper are listed as below.

- Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.
  A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted or not to a particular college. These binary outcomes allow straightforward decisions between two alternatives.
- The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.The KNN algorithm hinges on this assumption being true enough for the algorithm to be useful. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics we might have learned in our childhood— calculating the distance between points on a graph.
- Random Forest algorithms are used for classification as well as regression. It creates a tree for the data and makes prediction based on that. Random Forest algorithm can

be used on large data sets and can produce the same result even when large sets record values are missing. The generated samples from the decision tree can be saved so that it can be used on other data. In random forest there are two stages, firstly create a random forest then make a prediction using a random forest classifier created in the first stage.

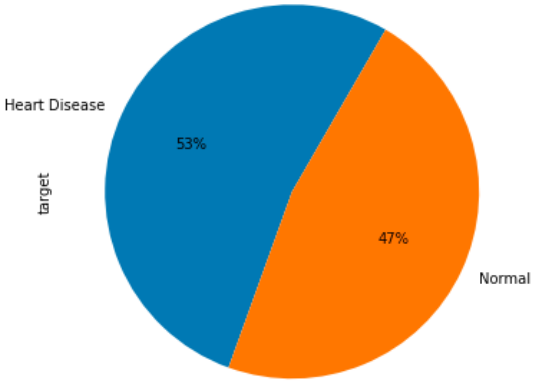

Fig. 2. Percentage of Heart disease patient from data set

IV. Methodology

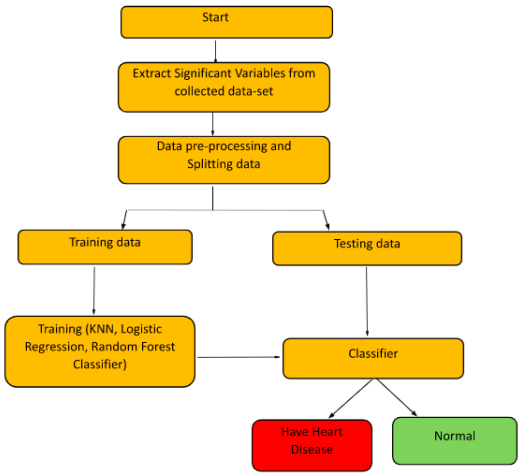

Fig. 3. Proposed model

This paper shows the analysis of various machine learning algorithms, the algorithms that are used in this paper are K nearest neighbors (KNN), Logistic Regression and Random Forest Classifiers which can be helpful for practitioners or medical analysts for accurately diagnose Heart Disease. This paperwork includes examining the journals, published paper and the data of cardiovascular disease of the recent times. Methodology gives a framework for the proposed model [1]. The methodology is a process which includes steps that transform given data into recognized data patterns for the knowledge of the users. The proposed methodology (Figure 1.) includes steps, where first step is referred as the collection of the data than in second stage it extracts significant values than the 3rd is the preprocessing stage where we explore the data. Data preprocessing deals with the missing values, cleaning of data and normalization depending on algorithms used [15]. After pre-processing of data, classifier is used to classify the pre-processed data the classifier used in the proposed model are KNN, Logistic Regression, Random Forest Classifier. Finally, the proposed model is undertaken, where we evaluated our model on the basis of accuracy and performance using various performance metrics. Here in this model, an effective Heart Disease Prediction System (EHDPS) has been developed using different classifiers. This model uses 13 medical parameters such as chest pain, fasting sugar, blood pressure, cholesterol, age, sex etc. for prediction [17].

```
models
```

```
[('LR_L2', LogisticRegression()),
 ('KNN7', KNeighborsClassifier(n_neighbors=7)),
 ('KNN5', KNeighborsClassifier()),
 ('RF_Gini100', RandomForestClassifier())]
```

Fig. 4. Models

V. Result Analysis

From these results we can see that although most of the researchers are using different algorithms such as Decision tree for the detection of patients diagnosed with Heart disease, KNN, Random Forest Classifier and Logistic regression yield a better result to out rule them [23]. The algorithms that we used are more accurate, saves a lot of money i.e. it is cost efficient and faster than the algorithms that the previous researchers used. Moreover, the maximum accuracy obtained by KNN and Logistic Regression are equal to 88.5% which is greater or almost equal to accuracy's obtained from previous researches. So, we summarize that our accuracy is improved due to the increased medical attributes that we used from the dataset we took. Our project also tells us that Logistic Regression and KNN outperforms Random Forest Classifier in the prediction of the patient diagnosed with a heart Disease. This proves that KNN and Logistic Regression are better in diagnosis of a heart disease.

We use 80% of the total data set for training and the rest 20% for test. After train the model using the training data-set we have found that True Negative and True Positive . **Sensitivity** or **true positive rate** is a measure of the proportion of people have hear disease who got predicted correctly as hear disease patient. In other words, the person who really a heart disease patient (positive) actually got predicted as heart disease patient. Mathematically,sensitivity or true positive rate can be

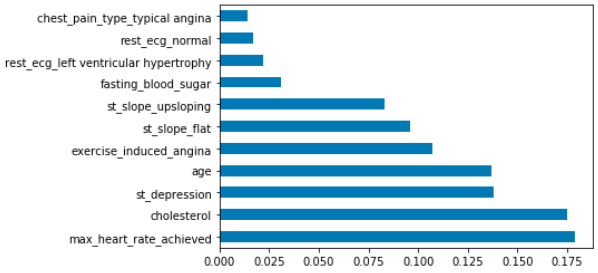

Fig. 5. Feature importance

calculated as the following: *Sensitivity = (True Positive)/(True Positive + False Negative)* A high sensitivity means that the model is correctly identifying most of the positive results, while a low sensitivity means that the model is missing a lot of positive results. The following are the details in relation to True Positive and False Negative used in the above equation which represents the fig5.

- **True Positive:**The person who is diagnose as heart disease patient (positive) actually got predicted as heart disease patient.
- **False Negative:**The false-negative represents the number of persons who are not heart disease patient and got predicted as heart disease patient. Ideally, we would seek the model to have low false negatives as it might prove to be business threatening.

The higher value of sensitivity would mean a higher value of the true positive and a lower value of false negative. The lower value of sensitivity would mean a lower value of the true positive and a higher value of false negative. For the business and financial domain, models with high sensitivity will be desired.

As same **Specificity = (True Negative)/(True Negative + False Positive)** Then we calculate the **accuracy** which is 90.47% for this data-set.

$$Accuracy = \frac{\sum TP + \sum TN}{\sum TP + \sum TN + \sum FP + \sum FN} \quad (1)$$

$$Accuracy = \frac{1041}{1159} = 89.21\%$$

	precision	recall	f1-score	support
0	0.91	0.85	0.88	112
1	0.87	0.93	0.90	123
accuracy			0.89	235
macro avg	0.89	0.89	0.89	235
weighted avg	0.89	0.89	0.89	235

Fig. 6. Classification Report for Random Forest classifier

VI. Conclusion

A cardiovascular disease detection model has been developed using three ML classification modelling techniques.

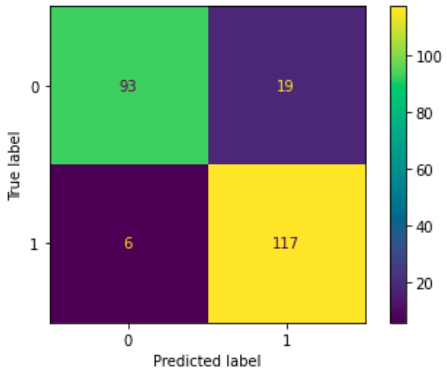

Fig. 7. Confusion Matrix for Random Forest classifier

This project predicts people with cardiovascular disease by extracting the patient medical history that leads to a fatal heart disease from a dataset that includes patients' medical history such as chest pain, sugar level, blood pressure, etc. This Heart Disease detection system assists a patient based on his/her clinical information of them been diagnosed with a previous heart disease. The algorithms used in building the given model are Logistic regression, Random Forest Classifier and KNN [22]. The accuracy of our model is 90.21.5%. Use of more training data ensures the higher chances of the model to accurately predict whether the given person has a heart disease or not . By using these, computer aided techniques we can predict the patient fast and better and the cost can be reduced very much.

//

References

[1] Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.

[2] T.Nagamani, S.Logeswari, B.Gomathy," Heart Disease Prediction using Data Mining with Mapreduce Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.

[3] Fahd Saleh Alotaibi," Implementation of Machine Learning Model to Predict Heart Failure Disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, "Design And Implementation Heart Disease Prediction Using Naives Bayesian", International Conference on Trends in Electronics and Information(ICOEI 2019).

[5] Theresa Princy R,J. Thomas,'Human heart Disease Prediction System using Data Mining Techniques', International Conference on Circuit Power and Computing Technologies,Bangalore,2016.

[6] Nagaraj M Lutimath,Chethan C,Basavaraj S Pol.,'Prediction Of Heart Disease using Machine Learning', International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477, 2019.