Heart Disease Diagnosis Using Machine Learning

Anusha G C, Apoorva M S, Deepthi N, Dhanushree V
Student, Department of Computer Science & Engineering, GSSSIETW, Mysuru

Rummana Firdaus
Assistant Professor, Department of Computer Science & Engineering, GSSSIETW, Mysuru

***Abstract*---- The scope of Machine Learning algorithms are increasing in predicting various diseases. The nature of machine learning algorithms to think like a human beings are making this concept important and versatile. Here the challenge of increasing the accuracy of Heart disease prediction is taken upon. In the proposed work, decision support system is made by two supervised machine learning models namely Random Forest and Logistic Regression. The method of predicting heart diseases using Random Forest with well-set attributes predicts the stage of heart failure of the patient. Logistic Regression Analysis is used in the model to infer how confident can the predicted value be actual value when given as test data of a patient details in predicting the stages of heart failure. By the proposed algorithm for heart disease prediction, many lives could be saved in the near future.**

Keywords **- Heart Disease, Machine learning, Random Forest, Logistic Regression.**

I. INTRODUCTION

In the near future, artificial intelligence (AI) techniques, such as machine learning, deep learning, and cognitive computing, may play a critical role in the evolution of cardiovascular (CV) medicine to facilitate precision CV medicine. CV clinical care currently faces practical challenges in predicting the stages of heart failure. Physicians have long needed to identify, quantify, and interpret relationships among variables to improve patient care. Hence identifying or predicting the disease at the earliest is very important to avoid any unwanted casualties.
Machine learning techniques comprise a variety of methods that allow computers to algorithmically learn efficient representations of data. Machine learning techniques can be broadly classified into either unsupervised or supervised learning. These have different goals. Unsupervised learning focuses on discovering underlying structure or relationships among variables in a dataset, whereas supervised learning often involves classification of an observation into 1 or more categories or outcomes. Supervised learning thus requires a dataset with predictor variables and labeled outcomes.
The proposed system is used to predict the possible risk of heart disease in patient by analyzing the stages of heart failure with the help of Random Forest algorithm. Random forest is a supervised learning algorithm that operates by constructing multiple decision trees during training phase. The decision of the majority of the trees is chosen by the random forest as final decision Random forest develops lots of decision tree based on random selection of data and random selection of variables with that it provides class of dependent variable based on many trees.

II. RELATED WORKS.

Over the years, a range of works have been done related to heart disease prediction system using different data mining algorithms by different authors. They tried to attain efficient methods and accuracy in finding out diseases related to heart by their work including datasets and different algorithms along with the experimental results and future work that can be done on the system to achieve more efficient results.

Shadab Adam Pattekari et al. [1] explains that the main objective of the research is to develop an Intelligent System using data mining modeling technique, namely, Naive Bayes. It is implemented as web based application in this user answers the predefined questions. It retrieves hidden data from stored database and compares the user values with trained data set. It can answer complex queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. By providing effective treatments, it also helps to reduce treatment costs.

Ganesha B et al. [2] describes that it is important to monitor heart rate during cycling. By monitoring heart rate during cycling, Cyclists can control the cycling session such as cycling cadence to determine the intensity of exercise. By controlling the intensity of cycling, cyclists can avoid risks

of over training and heart attack. Exercise intensity can be measured by heart rate of cyclist. The heart rate can be measured by wearable sensor. But there are data that are not recorded by the sensor at a regular time for example, one second, two second, etc. So we need a prediction model of heart rate to complete the missing data. The purpose of this study is to create a predictive model for heart rate based on cycling cadence using Feed forward Neural Network. The inputs are heart rate and cadence on the second. The output is the predictive value if heart rate on the next second. Feed forward Neural Network is used as a mathematical model of the relationship between heart rate and cycling Cadence.

Kuspraspta Mutijarsa et al. [3] reports that the rapid growth is seen in health care services over the past few years. Heart disease causes millions of death worldwide. Many wireless communication technologies have been developed for heart disease prediction. Data mining algorithms are very useful in the detection and diagnosis of heart disease. In this paper, survey is carried out on several single and hybrid data mining algorithm in order to identify the algorithm that best suits the heart disease prediction with high level of accuracy.

AK Srivas et al. [4] relates that the identification of diseases are very challenging task in field of medical science. Heart disease is very critical issues facing by the people. In our proposed work we have used data mining based classification techniques for analysis and classification of different level of heart disease namely Cleveland, Switzerland, Hungarian and Long Beach. They have used WEKA (Waikato Environment for Knowledge Analysis) and Rapid miner data mining tools for analysis of heart disease data set and compared the performance of different classification techniques with four heart disease data set using WEKA and Rapid Miner data mining tool. The proposed SVM gives better accuracy as 66.67% with Hungarian data set in case of WEKA data mining tool while Decision Stump gives better accuracy as 63.94% with same Hungarian data set in case of Rapid miner data mining tool. The Hungarian data set gives better performance with our proposed data mining tools and classification techniques which can help the people to predict effective factors about Coronary Heart Disease.

Thomas Ernest Perry et al. [5] details that the electronic health records possess critical predictive information for machine-learning- based diagnostic aids. However, many traditional machine learning methods fail to simultaneously integrate textual data into prediction process because of its high dimensionality. In this paper, they present supervised method using Laplacian Eigenmaps to enable existing dimensional representations of textual data and accurate predictors based on these low- dimensional representations at the same time. This preserves the local similarities among textual data in high- dimensional space. The proposed implementation performs alternating optimization using gradient descent. For the evaluation, they applied Laplacian Eigenmaps method to over 2000 patient records from a large single- center pediatric cardiology practice to predict if patients were diagnosed with cardiac disease.

Yeshvendra K. Singh et al. [6] States that the scope of Machine Learning algorithms are increasing in predicting various diseases. The nature of machine learning algorithm to think like a human being is making this concept so important and versatile. Here the challenge of increasing the accuracy of Heart disease prediction is taken upon. The non-linear tendency of the Cleveland heart disease dataset was exploited for applying Random Forest to get an accuracy of 85.81%. The method of predicting heart diseases using Random Forest with well-set attributes fetches us more accuracy. Random Forest was built by training 303 instances of data and authentication of accuracy was done using 10- fold cross validation. By the proposed algorithm for heart disease prediction, many lives could be saved in the future.

Prof. Priya R. Patil et al. [7] conveys that the accurate diagnosis of a heart disease, is one of the most important biomedical problems whose administration is imperative. In their proposed work, decision support system was made by three data mining techniques namely Classical Random Forest, Modified Random Forest and Weighted Random Forest. The classical random forests constructs a collection of trees. In Modified Random Forest, the tree is constructed dynamically with online fitting procedure. A random forest is a substantial modification of bagging. Forest construction is based on three step process. 1. Forest construction,2. The polynomial fitting procedure and 3. The termination criterion In Weighted Random Forest, The Attribute Weighting Method is used for improving Accuracy of Modified Random Forest. There are Two Techniques are used in Attribute Weighting: 1. Averaged One- Dependence Estimators (AODE) 2. Decision Tree-based Attribute Weighted Averaged One-dependence Estimator (DTWAODE).

Evanthia E Tripoliti et al. [8] states that, the accurate diagnosis of diseases with high prevalence rate, such as Alzheimer, Parkinson, diabetes, breast cancer, and heart

diseases, is one of the most important biomedical problems whose administration is imperative. In this paper, they present a new method for the automated diagnosis of diseases based on the improvement of random forests classification algorithm. More specifically, the dynamic determination of the optimum number of base classifiers composing the random forests was addressed. The proposed method was different from most of the methods reported in the literature, which followed an overproduce-and-choose strategy, where the members of the ensemble were selected from a pool of classifiers, which was known a priori. In their case, the number of classifiers were determined during the growing procedure of the forest. Additionally, the proposed method produced an ensemble not only accurate, but also diverse, ensuring the two important properties that should characterize an ensemble classifier. The method was based on an online fitting procedure and it was evaluated using eight biomedical datasets and five versions of the random forests algorithm (40 cases). The method correctly decided the number of trees in 90% of the test cases.

Liaw et al, [9] have noted from their experiments that the number of trees necessary for good performance grows with the number of predictors. The best way to determine how many trees are necessary is to compare predictions made by a forest to predictions made by a subset of a forest. When the subsets work as well as the full forest, it means that there are enough trees. A lot of trees are necessary to get stable estimates of variable importance and proximity. However, their experience has been that even though the variable importance measures may vary from run to run, the ranking of the importance is quite stable. For classification problems where the class frequencies are extremely unbalanced (e.g., 99% class 1 and 1% class 2), it was necessary to change the prediction rule to other than majority votes. For example, in a two-class problem with 99% class 1 and 1% class 2, they suggested that one may predict the 1% of the observations with largest class 2 probabilities as class 2, and use the smallest of those probabilities as threshold for prediction of test data (i.e., use the type='prob' argument in the predict method and threshold the second column of the output). They have routinely done this to get ROC curves. By default, the entire forest was contained in the forest component of the random forest object. It took up quite a bit of memory for a large data set or large number of trees. If prediction of test data was not needed, they set the argument keep. forest=FALSE when running Random Forest. This way, only the tree was kept in memory at any time, and thus lots of memory (and potentially execution time) were saved. Since the algorithm falls into the "embarrassingly parallel" category, one can run several random forests on different machines and then aggregate the votes component to get the final result.

Vrushali Y Kulkarni et al. [10] surveys that Ensemble methods aim at improving classification accuracy by aggregating predictions from multiple classifiers. More diverse the base classifiers and less are they correlated; the more is accuracy of the ensemble. Random Forest algorithm uses- (i) Sub-sampling the examples/cases as in bagging, (ii) Sub-sampling the features known as feature selection. Both these strategies were used in Random Forest to introduce randomization and achieve diversity. Also, there was no pruning in the base decision trees to ensure diversity among them. They have stated that the use of Fuzzy decision trees and Semi supervised learning with Random Forest is the recent development and there is future scope for semi supervised learning with Random Forest due to capability of handling both labeled and unlabeled data; especially for scenarios where getting labeled data is a problem. They have presented Taxonomy of Random Forest algorithm and have performed analysis of various algorithms and techniques based on Random Forest algorithm. Their analysis is presented as Comparison chart to serve as a guideline for pursuing future research related to Random forest classifier.

III. PROBLEM FORMULATION

Heart plays a crucial role in circulatory system. When heart does not function properly then it will lead to serious health conditions including death. According to the World Health Organization, in the year 2016, 17.9 million people died from heart disease representing 31% of all global deaths.

The aim of this paper is to present an intelligent predicting system that predicts the stage of heart failure in patients which could help in the earlier detection and treatment effectively. The main objective of this paper is to develop an Intelligent Heart disease prediction system using machine learning techniques.

IV. METHODOLOGY

Medical experts convert the data into Electronic Health Record (EHR). Data labeling is the collection of data which consists of features and target value.

Training dataset of heart disease is taken from UCI repository. The dataset consists of 303 instances taken from the observations made at the Cleveland Clinic Foundation (cleveland.data). The target values 1,2,3,4 corresponds to stages of heart failure as defined by American Heart Association (AHA) and 0 specifies no heart disease.

The dataset consists of 76 attributes. According to the researchers, 13 attributes are considered as best fit for the prediction system.

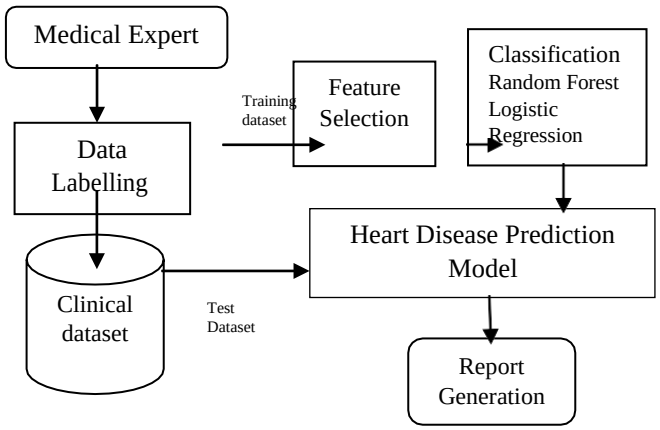

Figure 1: System Architecture

In the above Figure 1, the medical expert is the user of the system who is responsible for loading the data of patients into the system in the desired format as clinical dataset against which the prediction is made. The datasets are taken through feature selection process of selecting the most appropriate features into the classification and prediction algorithm. Feature selection is a process of identifying and removing redundant, irrelevant features and increasing accuracy. The motivation for applying the feature selection has been increased for model building.

The algorithm trains the model against test dataset and a detailed report of the heart disease analysis is generated. This report is used as an advice sheet to the doctors or medical experts to help faster and more accurate diagnosis of patient's health and save their lives.

The algorithms proposed in this paper are Random Forest and Logistic Regression. Random Forest is a supervised learning algorithm that builds multiple decision trees and merges them together to get a more accurate and stable prediction.

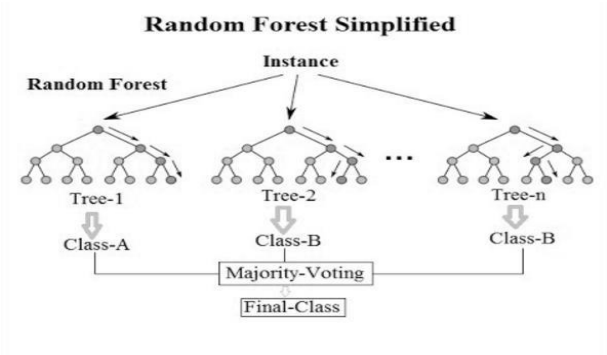

Figure 2: Random Forest Algorithm

The Random Forest algorithm operates by constructing multiple decision tress during training phase. The decision of the majority of the trees is chosen by the random forest as final decision. The "forest" it builds, is an ensemble of Decision Trees, most of the time trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.

Thus, Random Forest is an ensemble classifier which combines bagging and random selection of features. Here, the algorithm predicts the stage of heart failure of the patient details given as test data through random selection and bagging method of 14 features that are considered in the training data set.

Logistic Regression is a machine learning algorithm based on supervised learning. This is used to infer how confident can predicted value be actual value when given as test data of a patient details in predicting the stages of heart failure.

Thus these above mentioned two algorithms under the class of supervised learning are by far the most appropriate and scalable algorithms for our problem specified.

V. RESULTS

The results obtained after the implementation of the model specified are mentioned below.

For every test data that contains patient details that has the attributes specified in exact reflection of the training data, we obtain the stage of heart failure in the patient as predicted by the model that uses random forest and logistic regression.

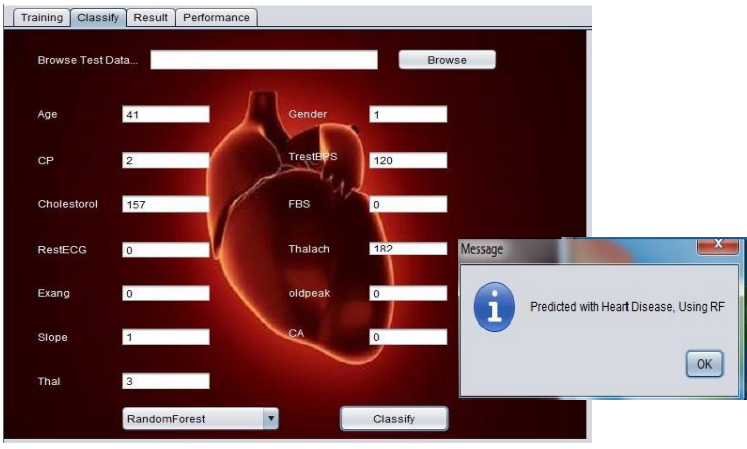

Figure 3: Model detects the presence of heart disease

This phase of the model obtains patient details as test data and with the classify option it predicts whether the patient has heart disease or not.

At this stage the model gives only the presence of heart disease in the patient specified.

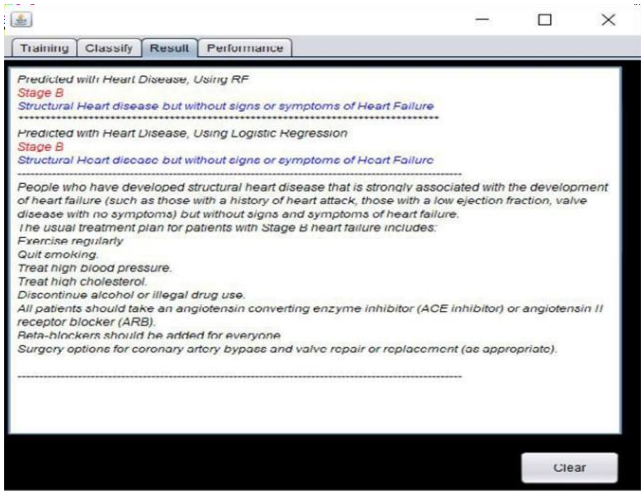

Figure 4: Stages of heart failure in patient along with the treatment suggestions

At this stage the model generates the report about the stage of heart failure in the patient predicted by the two algorithms along with the Generic characteristics of patients of the corresponding stage and its treatment details and suggestions.

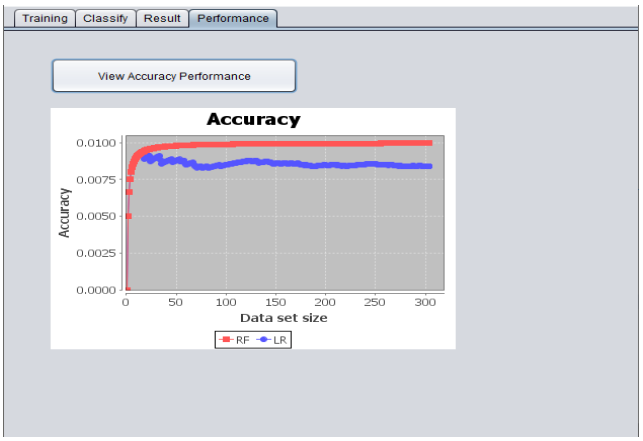

Figure 5: The graph of accuracy of the model

The figure 5 is the graph of accuracy obtained by true positive values of the confusion matrix plotted for every 303 instances in the training dataset in the intervals of 50 instances for the random forest and Logistic regression algorithms.
This shows that the performance of random forest is remarkable over the logistic regression.

VI. DISCUSSION

The problem statement is considered from the real world human health oriented issue in our intention to provide enhanced effectiveness towards the treatment through a technological solution based on the current technology of machine learning.
This paper approaches with the valid dataset taken from the UCI repository of Cleveland Clinic Foundation as the training data to construct the model. The stages of heart failure and its generic patient characteristics, treatment details and goals considered are according to the specification made by the American Heart Association (AHA).
The choice of supervised learning algorithms were based on the hardware and software requirements along with the nature of the training dataset and the solutions that were intended to be provided. Thus, the paper details the application of random forest and logistic regression algorithms for designing this prediction model.

VII. CONCLUSION

The Intelligent Heart disease Prediction system developed can lead to proper selection of treatment methods for a patient diagnosed with stage of heart failure. This system helps physicians in early diagnosis that makes cost reductions in prevention and treatment of the heart condition in patients, and avoids mortality rates.

VIII. FUTURE ENHANCEMENTS

The model produces the report with stage of heart failure and the cumulative characteristics of patients with corresponding stage, its treatment details and goals. According to Fakhraei et al. [11] data obtained during the four phases of data analysis during the drug development process can be constructed as training data for each stage of heart failure which shall lead to the prediction of necessary pharmaceutical discoveries for individual stages that can be predicted along with the stage of heart failure . Thus an addition to the report would be the pharmaceutical discoveries made for each stage of heart failure along with the current details.

REFERENCES

[1] Pattekari, Shadab Adam, and Asma Parveen. "Prediction system for heart disease using Naïve Bayes." *International Journal of Advanced Computer and Mathematical Sciences* 3, no. 3 (2012): 290-294.
[2] Gnaneswar, B., and MR Ebenezar Jebarani. "A review on prediction and diagnosis of heart failure." In *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, pp. 1-3. IEEE, 2017.
[3] Mutijarsa, Kusprasapta, Muhammad Ichwan, and Dina Budhi Utami. "Heart rate prediction based on cycling cadence using feedforward neural network." In *2016 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, pp. 72-76. IEEE, 2016. PR
[4] Shrivas, A. K., and Rajat Kumar Yadu. "An Effective Prediction Factors for Coronary Heart Disease using Data Mining based Classification Technique." *International Journal on Recent and Innovation Trends in Computing and Communication* 5, no. 5 (2017): 813-816.
[5] Perry, Thomas Ernest, Hongyuan Zha, Ke Zhou, Patricio Frias, Dadan Zeng, and Mark Braunstein. "Supervised embedding of textual predictors with applications in clinical diagnostics for pediatric cardiology." *Journal of the American Medical Informatics Association* 21, no. e1 (2013): e136-e142.
[6] Yeshvendra K. Singh, Nikhil Sinha,Sanjay K. Singh “Heart Disease

Prediction System Using Random Forest" International Conference on Advances in Computing and Data Sciences no .el(22 July 2017)

[7] Patil R Priya, Kinariwala A S, "Automated Diagnosis of Heart Disease using Random Forest Algorithm" International Journal of Advance Research, Ideas and Innovations in Technology ijariit.

[8] Tripoliti, Evanthia E., Dimitrios I. Fotiadis, and George Manis. "Automated diagnosis of diseases based on classification: dynamic determination of the number of trees in random forests algorithm." IEEE transactions on information technology in biomedicine 16, no. 4 (2012): 615-622.

[9] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2, no. 3 (2002): 18-22.

[10] Kulkarni, Vrushali Y., and Pradeep K. Sinha. "Pruning of random forest classifiers: A survey and future directions." In 2012 International Conference on Data Science & Engineering (ICDSE), pp. 64-68. IEEE, 2012.

[11] Fakhraei, Shobeir, Eberechukwu Onukwugha, and Lise Getoor. "Data Analytics for Pharmaceutical Discoveries." (2015): 599-623.