Proceedings of the Sixth International Conference on Inventive Computation Technologies [ICICT 2021]
IEEE Xplore Part Number: CFP21F70-ART; ISBN: 978-1-7281-8501-9

Heart Disease Prediction using Hybrid machine Learning Model

Dr. M. Kavitha[1*], G. Gnaneswar[1], R. Dinesh[1], Y. Rohith Sai[1], R. Sai Suraj[1]

[1*] Assistant Professor, [1] Student

Department of Computer Science and Engineering,
Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
mkavita@kluniversity.in, gnane.gullapalli@gmail.com, rampallidinesh@gmail.com, rohithsai@gmail.com, rangasaisuraj@gmail.com

2021 6th International Conference on Inventive Computation Technologies (ICICT) | 978-1-7281-8501-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICT50816.2021.9358597

Abstract –
Heart disease causes a significant mortality rate around the world, and it has become a health threat for many people. Early prediction of heart disease may save many lives; detecting cardiovascular diseases like heart attacks, coronary artery diseases etc., is a critical challenge by the regular clinical data analysis. Machine learning (ML) can bring an effective solution for decision making and accurate predictions. The medical industry is showing enormous development in using machine learning techniques. In the proposed work, a novel machine learning approach is proposed to predict heart disease. The proposed study used the Cleveland heart disease dataset, and data mining techniques such as regression and classification are used. Machine learning techniques Random Forest and Decision Tree are applied. The novel technique of the machine learning model is designed. In implementation, 3 machine learning algorithms are used, they are 1. Random Forest, 2. Decision Tree and 3. Hybrid model (Hybrid of random forest and decision tree). Experimental results show an accuracy level of 88.7% through the heart disease prediction model with the hybrid model. The interface is designed to get the user's input parameter to predict the heart disease, for which we used a hybrid model of Decision Tree and Random Forest.

***Key Words*: Cleveland Heart Disease Database, Decision Trees, Random forest, Hybrid algorithm, Machine learning**

I. INTRODUCTION

Data mining is useful for studying and understanding a large amount of data. It is used for the extraction of data and to make the decision for further applications. The most common techniques covered under data mining are clustering, association rule mining, and classifications. There are plenty of algorithms available for implementing these data mining techniques. Though there are tools like weka are available for simulations, Python programming is emerging with these algorithms built with scikit learn packages. Thus, the real-time implementation of data mining concepts is more reliable than ever.

Machine learning usage is growing vastly in the medical diagnosis industry, where the manual error can be reduced with computer analysis, and accuracy is improved. The diagnosis of a disease is highly reliable with machine learning techniques. Disease such as heart disease, liver disease, diabetes, tumor predictions is done through machine learning concepts [18]. classification algorithms such as decision tree, naïve bayes and SVM (Support Vector Machine) are available; similarly, regression algorithms, namely Random forest, lasso, and logistic regressions, were used in the medical industry. In most of the tumor predictions, deep learning algorithms are largely used in the medical diagnosis field.

As per survey reports, each year, nearly 17 million deaths occurred due to cardiovascular diseases (CVD). The early detection of disease may save many lives, and mortality can be reduced if the patients take their treatments on time [19]. Cardiovascular diseases include many threats such as heart disease, and all etc. With the lack of physical activity due to the lifestyle changes, these diseases are becoming very common even in the lesser age groups. Smoking, lack of physical exercise, high cholesterol food, junk food, living habits

are the leading causes of heart disease.

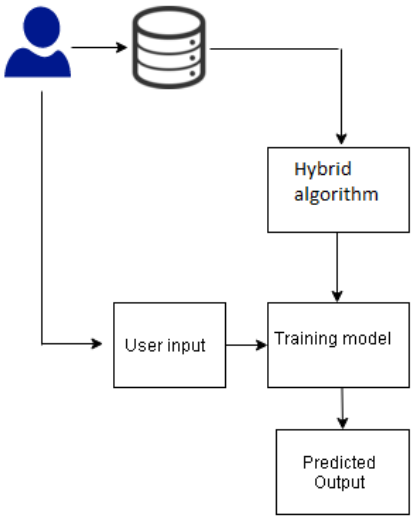

Fig 1: Block diagram of Heart Disease prediction

This study aims to predict heart disease based on machine learning via an automated medical diagnosis method. We use the hybrid model, as it is the finest classification method for predicting heart disease. A hybrid model is a novel technique, which uses the probabilities arrived from one machine learning model is given as input to the other machine learning model.

This hybrid model gives us the better-optimized results based on both machine learning procedure, which is considered for the implementations.

The proposed system is the prediction of heart disease by an automated machine learning diagnosis model with a high novelty-based hybrid model. This hybrid model is used to predict heart disease. Cleveland dataset is utilized here for processing. This dataset is considered commonly by machine learning researchers. This dataset has a entire of 303 instances and around 14 characteristics.

The study aims to classify it as a binary classification type 0(absence of heart disease) to 1 (present of heart disease). Patients can go for treatment based on the result generated through our proposed model. The proposed application helps in taking advance measures for patients.

In the following chapters, literature survey and related effort is studied. In chapter III, the projected system is discussed, and the implementation algorithm and methodology are discussed. In chapter 4, results and discussions are done. In chapter V, this work is concluded, and enhancements are discussed.

II. RELATED WORK

There are many current works studied by the researchers about heart disease prediction and analysis. Some of such works are addressed below.

The author studies heart disease using the random forest in [1] with the Cleveland dataset. The author used the Chi Square feature selection model and genetic algorithm (GA) based feature selection model for the study. They proved in the experimental results that their proposed model with Genetic algorithm feature selection has given high accuracy than the existing models. However, the results are evaluated with existing machine learning models.

In [2], the author has generated specific rules based on this PSO algorithm and evaluated different rules to get a more accurate rule for heart disease identification. After evaluating the rules, C 5.0 is used for the classification of disease based on binary classification. The author used UCI repository data for implementation and evaluated high accuracy using PSO and the Decision tree algorithm.

Backpropagation neural network for heart disease prediction was discussed in [3]. Deep learning model, which is a highly effective learning model for disease prediction. The author used a neural network for learning and prediction. The author used the Cleveland dataset for the study and implemented simulation in Matlab. However, the work can be done with deep learning models and highly accurate, and this can be extended to real world applications.

The author in [8] discussed prediction of heart disease using data mining practices. They studied and evaluated with some techniques such as the KNN algorithm, decision tree algorithm, neural network classifications, and Bayesian classification algorithms. The author also studied the genetic algorithm's use in feature selection for heart disease essential features. and experimented with the study and evaluated high accuracy with the decision tree model.

Heart disease calculation using different machine learning procedures is studied in [9]. Classification and regression models are used for prediction, namely the Decision tree, KNN algorithm, SVM, and linear regression procedure is used for the study. Experiment results proved that the KNN algorithm with the highest accuracy. However, this model can be implemented in a real-time environment or applications.

A cognitive approach is carried out in [10] for heart disease prediction. In this work, five machine learning algorithms are considered for prediction, and all are evaluated with accuracy. Logistic model tree is implemented to get better results in prediction, which used an ADA boost and bagging model to forecast heart disease. Their investigational results have exposed that random forest achieved high accuracy on predictions.

It is inferred from the existing works that there is a need for novelty in the study, and a robust, optimized model is needed for heart disease prediction. The existing works are discussed with the available machine learning algorithms, either implemented with tools such as Weka or MATLAB. Some of the works are also done with the deep learning model. However, the optimized model is not studied. In our proposed model, the novelty of work is done. It is implemented with a hybrid model to give more optimized results.

III. PROPOSED WORK

A hybrid model is a novel technique, which uses the probabilities arrived from one machine learning model is given as input to the other machine learning model. This hybrid model gives us the better-optimized results based on both machine learning algorithm, which is considered for the implementations.

The proposed work is implemented with sklearn libraries, pandas, matplotlib, and other compulsory libraries. We have the dataset downloaded from the uci repository. There are binary groups of heart disease in the downloaded info. The machine learning algorithm is implemented along with the hybrid model, such as decision tree and random forest.

IV. DATASET DETAILS

Dataset collected with attributes sex indicates the gender of the patient, age indicates the age of the patient, trestbps indicates the resting blood pressure, cp indicates the chest pain, fbs indicates the fast blooding sugar, chol indicates cholesterol, thalach indicates the maximum heart rate achieved, restecg indicates the resting electroc. result (1 anomality), oldpeak indicates the ST depression induc. ex, exang indicates the exercise induced angina, ca indicates the number of major vessels, slope indicates the slope of peak exercise ST, pred_attribute, thal indicates the thalassemia. The sample of collected data is shown in the below figure.

The dataset is visualized to get number of Heart disease cases and number of normal cases from the dataset. It is shown as histogram plot as given below.

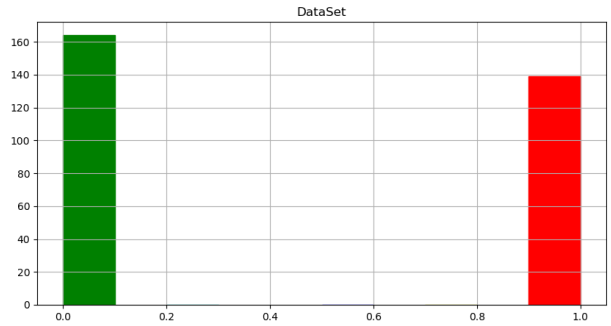

Fig 2: Data Visualization of heart Disease in Cleveland Dataset

The dataset is visualized to get number of heart disease cases and number of normal cases from the dataset. It is shown as histogram plot as given in figure 2.

The proposed workflow has the following advantages

- Implemented two machine learning algorithm and a Hybrid model
- Accuracy of all proposed algorithm is arrived to show the best model
- Implement a hybrid model to make the proposed work as an optimized model.

The execution is carried out with the below given methodologies

a. Dataset is collected from uci.edu
b. Data Visualization is done
c. Splitting dataset into test and train data
d. Apply DT and RF models for training and analysis
e. Train the model
f. Test the trained model and predict values
g. Get single input from user and predict heart disease through hybrid model

Cleveland dataset is considered. It is split into two parts as training and testing sets. We have assumed 70% of the dataset as training input to the machine learning algorithms and fit the model. the remaining 30% as testing data for heart disease prediction.

We exploited the Decision Tree, Random Forest, and Hybrid of the Decision tree. Random forest is used to predict heart disease for 30% test input, and the values predicted to be plotted and compared for accuracy.

.

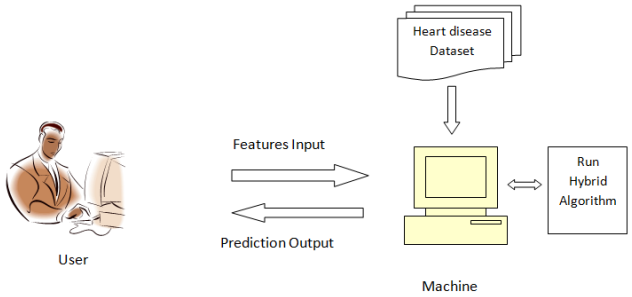

Fig 3: Heart disease prediction system architecture

Figure 3 shows the architecture of proposed system for heart disease prediction through machine learning algorithm models, which is briefly clarified below.

a. Decision Tree

Decision tree is one of the learning models that is used in the problem of classification. We divide the dataset into two or more sets using this technique. In decision tree, internal nodes represent a test on the characteristics, the branch portrays the outcome, and leaves are the decisions generated after subsequent processing.

Decision Tree algorithm as follows

i. Set the dataset's best feature as the root of the tree.

ii. Dataset is split into test and train sets. Subsets should be made in such a way that each subset contains information with the feature attribute like that.

iii. On each subset, the steps above are repeated until we get leaves in the tree.

The prediction for a record of a class label in the decision tree will start from the root. The values are compared with the following record attributes with the root attributes. The corresponding value of the next node to go arrives in this comparison.

b. Random Forest Regression

Random Forest regression aggregates multiple decisions to make a single decision. For training characteristics and then random sub characteristics for sampling nodes, random sampling is done.

Split the dataset into the test set and the train. Subsets should be made in such a way that each subset contains a feature attribute like that.

On each subset, the steps above are repeated until we get leaves in the tree.

The tree building samples are performed by bootstrapping, meaning it can multiple times consider the same feature. The maximum number of node splitting features could be limited by numbers. This algorithm reduces the problem of the fitting.

c. Hybrid Model

We develop a hybrid model using a decision tree and random forest algorithm. The combined model works based on probabilities of random forest. The probabilities from the random forest are added to train data and fed to the decision tree algorithm. Similarly, decision tree probabilities are identified and fed to test data. Finally, values are predicted.

Implementing machine learning on a preprocessed dataset is done; the anticipated cardiovascular disease for the given test dataset is plotted. Figure 4 shows the application we designed for heart disease prediction. The user/ patient can give their own input and detect the threat of heart disease. Disease prediction is classified as a binary prediction type, which means 0 - normal and 1- Heart disease. The application is designed using TkInter in python.

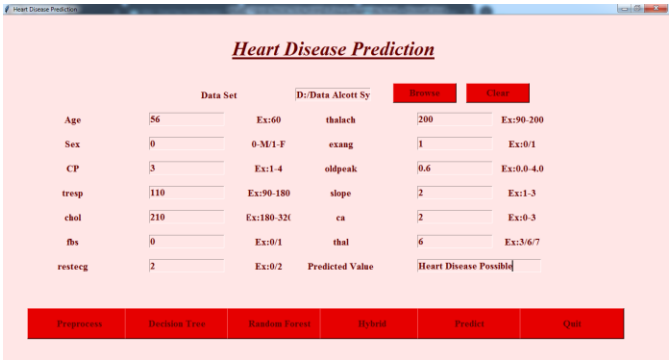

Fig 4: Basic GUI for Heart Disease prediction

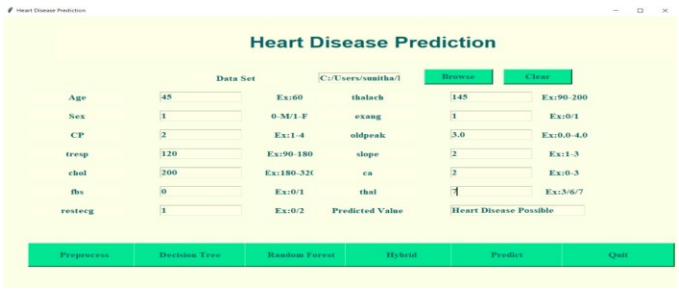

Fig 5: Positive Case of Heart Disease prediction

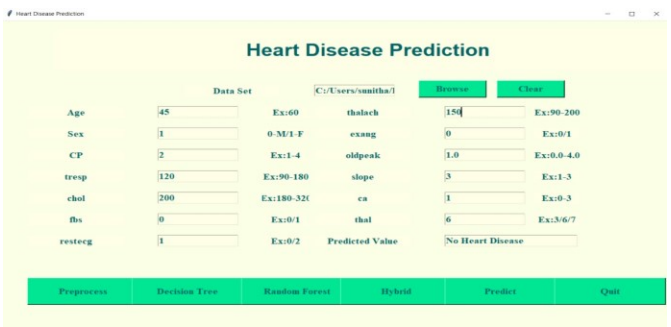

Fig 6: Negative Case of Heart Disease prediction

V. RESULTS AND DISCUSSIONS

In Python 3.7.6, the proposed work is implemented with sklearn libraries, pandas, matplotlib, and other required libraries. The dataset of heart disease downloaded from uci.edu will be considered for the study. Machine Learning algorithms such as Decision Tree and Random Forest were used. These machine learning algorithms were used to predict the heart disease. To improve the work and novelty of the work, we implemented a hybrid model of Decision Tree and Random Forest. The result shows that Heart disease detection is effective using the Random Forest algorithm and a hybrid model. Decision Tree achieves around 79% accuracy, and Random forest achieves 81% accuracy, Hybrid model achieves 88% accuracy.

Table 1: Experimental Results

Algorithm	Accuracy (%)
Decision Tree	79
Random Forest	81
Hybrid (Decision Tree+ Random Forest)	88

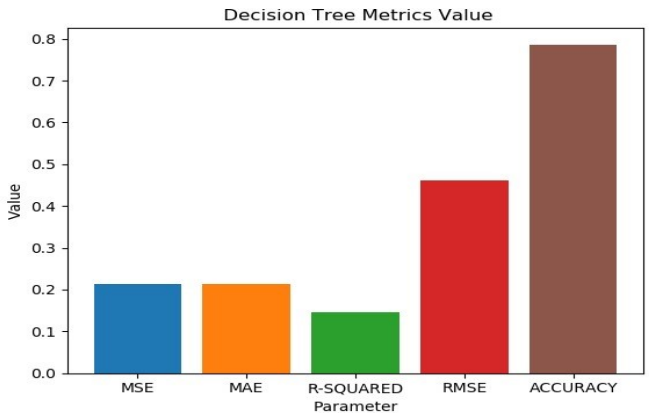

Fig 7: Heart Disease prediction through Decision Tree

Figure 7 shows the mean square error (MSE), mean absolute error (MAE), R-Squared parameter, root mean square error (RMSE) and accuracy for Decision Tree model.

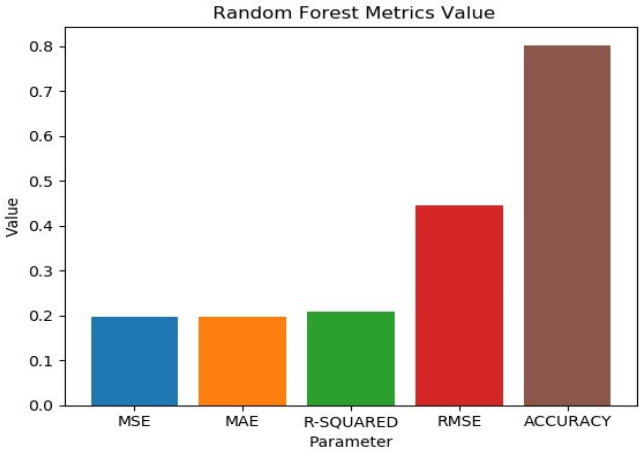

Fig 8: Heart Disease prediction through Random Forest

Figure 8 shows the mean square error (MSE), mean absolute error (MAE), R-Squared parameter, root mean square error (RMSE) and accuracy for Random Forest model.

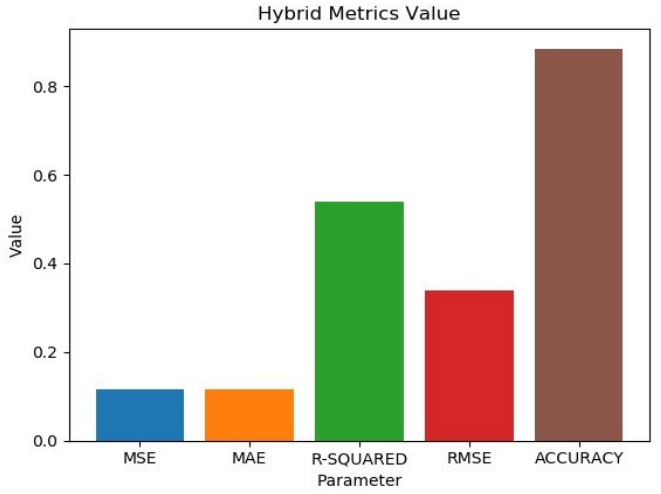

Fig 9: Heart Disease prediction through Hybrid model

Figure 9 shows the mean square error (MSE), mean absolute error (MAE), R-Squared parameter, root mean square error (RMSE) and accuracy for Hybrid model.

VI. CONCLUSION

Heart disease is one of the life-threatening diseases seen around the world. The changing lifestyle and lack of physical activities give more threat to condition. There are many diagnosis processes available in the medical industry. However, in terms of accuracy, machine learning is considered the best choice. The proposed work uses a TkInter Python designed application for the heart disease prediction. The proposed system using combinations of Decision Tree and Random forest for heart disease prediction as a hybrid model. Cleveland database is used for this study.

VII. FUTURE WORK

Deep learning algorithms playing a vital role in health care applications. So, applying deep learning procedures for heart disease prediction may give better outcome. Also, we are interested in classifying it as a multi-class problem to identify the disease's level.

REFERENCES

[1] Jabbar, M. A., B. L. Deekshatulu, and Priti Chandra. "Intelligent heart disease prediction system using random forest and evolutionary approach." *Journal of Network and Innovative Computing* 4.2016 (2016): 175-184.

[2] Alkeshuosh, Azhar Hussein, et al. "Using PSO algorithm for producing best rules in diagnosis of heart disease." *2017 international conference on computer and applications (ICCA)*. IEEE, 2017.

[3] Al-Milli, Nabeel. "Backpropagation neural network for prediction of heart disease." *Journal of theoretical and applied information Technology* 56.1 (2013): 131-135.

[4] Mythili, T., et al. "A heart disease prediction model using SVM-Decision Trees-Logistic Regression (SDL)." *International Journal of Computer Applications* 68.16 (2013).

[5] Detrano R. *VA Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, MD* (Doctoral dissertation, Ph. D., Donor: David W. Aha, 1998.

[6] Xing, Yanwei, Jie Wang, and Zhihong Zhao. "Combination data mining methods with new medical data to predicting outcome of coronary heart disease." *2007 International Conference on Convergence Information Technology (ICCIT 2007)*. IEEE, 2007.

[7] Chen, Jianxin, et al. "Predicting syndrome by NEI specifications: a comparison of five data mining algorithms in coronary heart disease." *International Conference on Life System Modeling and Simulation*. Springer, Berlin, Heidelberg, 2007.

[8] Soni, Jyoti, et al. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." *International Journal of Computer Applications* 17.8 (2011): 43-48.

[9] Singh, A., et.al (2020, February). Heart Disease Prediction Using Machine Learning Algorithms. In *2020 International Conference on Electrical and Electronics Engineering (ICE3)* (pp. 452-457). IEEE.

[10] Hashi, E.K. and Zaman, M.S.U., 2020. Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction. *Journal of Applied Science & Process Engineering*, *7*(2), pp.631-647.

[11] Shouman, Mai, Tim Turner, and Rob Stocker. "Using data mining techniques in heart disease diagnosis and treatment." *2012 Japan-Egypt Conference on Electronics, Communications and Computers*. IEEE, 2012.

[12] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." *IEEE Access* 7 (2019): 81542-81554.

[13] Ramalingam, V. V., Ayantan Dandapath, and M. Karthik Raja. "Heart disease prediction using machine learning techniques: a survey." *International Journal of Engineering & Technology* 7.2.8 (2018): 684-687.

[14] Polat, Kemal, Seral Şahan, and Salih Güneş. "Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing." *Expert Systems with Applications* 32.2 (2007): 625-631.

[15] Palaniappan, Sellappan, and Rafiah Awang. "Intelligent heart disease prediction system using data mining techniques." *2008 IEEE/ACS international conference on computer systems and applications*. IEEE, 2008.

[16] Das, Resul, Ibrahim Turkoglu, and Abdulkadir Sengur. "Effective diagnosis of heart disease through neural networks ensembles." *Expert systems with applications* 36.4 (2009): 7675-7680.

[17] Jonnavithula, et.al (2020, October). Role of machine learning algorithms over heart biseases prediction. In AIP Conference Proceedings (Vol. 2292, No. 1, p.040013). AIP Publishing LLC.