2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE) | 978-1-6654-9260-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/IITCEE57236.2023.10091028

Comparative Analysis of different Heart Disease Prediction Models

Kumar Rethik
School of Computer Science and Engineering, *Lovely Professional University*
Phagwara, Punjab, India
rethikguptakumar@gmail.com

Arun Singh
Student, IEEE member, School of Computer Science and Engineering, Lovely Professional University
Phagwara, Punjab,India
arunmandiarun2001@gmail.com

Dalwinder Singh
Assitant Professor, School of Computer Science and Engineering, Lovely Professional University, Phagwara, India
dalwinder.singh@lpu.co.in

Manik Rakhra
Assitant Professor, School of Computer Science and Engineering, *Lovely Professional University ,* Phagwara, India
rakhramanik786@gmail.com

***Abstract*— In today generation, heart disease is most common disease among most of the people in the world. There are many types of heart disease, but CAD is most common and it can block or reduced the flow of blood to heart muscle that can cause a heart attack or stroke. To detect cardiac disease, doctors are advised you to take an EKG and exercise stress test which are very costly. As a result, people do not take a test. The main cause of cardiovascular disease is Diabetes, high blood pressure, smoking, high cholesterol etc. In this study, machine learning techniques are used to detect whether a patient has heart disease or not. The algorithms which are used in this study are Random Forest, Logistic Regression, Naïve Bayes, Decision Tree and K-Nearest Neighbors. After the experimentation, it was concluded that RF gave out the best accuracy, F1 score and precision score.**

Keywords—Heart Disease, Machine Learning, Cardiovascular Disease, Prediction.

I. INTRODUCTION

During the outbreak of the Covid-19, the amount of research in the field of healthcare was boosting. According to world health organization, cardiovascular disease is one of the worse diseases in the world. Researchers said that people who have covid recently has increased risk of 20 cardiovascular disease including heart attack and stroke. We also get to know that more than (>) 24% of people in India are died due to several form of Heart Disease[1]. Therefore, it is essential to create a system for early identification of cardiac disease.

The heart disease is also known as cardiac disease, mostly caused on the walls of arteries due to the deposit of atheroma (fatty deposits) around the heart. The flow of blood to heart disease is reduced or blocked due to the accumulation of fatty deposits caused the arteries to constrict. Angiography is the most common method used for detecting the CAD but it is the costliest method and caused a reaction to the patient. Consequently, it is necessary to create an automated (self-operating) system that can identify cardiovascular disease based on a variety of human medical parameters.

In this paper we use a machine learning algorithm like Random Forest, Decision tree, Naïve Bayes, K Nearest Neighbor (KNN) and Logistic Regression. This model choice is based on earlier research that has already undergone algorithmic testing and produced predictions with a comparatively high level of accuracy.

In this paper we use a machine learning algorithm like Random Forest, Decision tree, Naïve Bayes, K Nearest Neighbor (KNN) and Logistic Regression. This model choice is based on earlier research that has already undergone algorithmic testing and produced predictions with a comparatively high level of accuracy.

The rest of the paper is discussed as follows: The heart-related works, existing methods, and strategies covered in Section II are discussed. Section 111 discuss about the dataset description. Section 1V consists of Machine learning algorithms used in this research.

II. LITERATURE REVIEW

D.P.Yadav, Prabhav Saini, Pragya Mittal [2] applied machine Learning techniques like K-Nearest Neighbour, Support Vector Machine (SVM), Naïve Bayes and Random Forest on the dataset to predict Heart Disease. Among these model Naïve bayes using 3-fold cross validation get the highest accuracy of 87.9%. A feature optimization technique Genetic algorithm is implemented for increasing the model performance. After applying optimization technique Naïve Bayes achieved accuracy 96%.

Surai Shinde, Juan Carlos Martinez-Ovando [3] combines the features of recurrent neural network and convolutional neural network to create the deep learning based hybrid model, which helped to attain better accuracy. The dataset used in this paper is a combination of PASCAL Classifying Heart Sounds Challenge dataset [3] and a Kaggle competition dataset has total of 832 heartbeat audios divided in to two classes: healthy and sick. The MFCC (Mel frequencies cepstral coefficients) and the DFT (discrete Fourier transform) techniques are used for pre-processing the audio data. This hybrid architecture based on deep learning that combines the ability to extract features from a CNN with the capacity to identify patterns over time from a LSTM (Long Short-Term Memory) Recurrent Neural Network (RNN), both networks receiving separate inputs, and the results of the two networks are combined to produce a specific prediction.

As per the result, the hybrid neural network model performed better than the most recent model in terms of performance.

Narendra Mohan, Vinod Jain, Gauranshi Agrawal [4] apply the machine learning models like Random Forest, KNN (K Nearest Neighbour), Logistic Regression and Naïve Bayes for predicting the heart disease. Among these models Logistic Regression achieves highest accuracy of 90.2%.

Indrajani Sutedja [5] predicted the heart disease using Deep Learning and Machine Learning Algorithms. The dataset used is taken from Kaggle and has 14 attributes. From Machine Learning Model, Support Vector Machine attains the maximum accuracy level by 88%. Meanwhile, for Deep learning model Recurrent neural network has the greatest accuracy level of 90%.

S. Nithyavishnupriya, R. Mowriya, R. Sarumathi, P. Ramprakash constructed a model using Deep neural Network and X^2 statistical model[1] to predict the Heart Disease. The onlineUCI archive dataset is used in this study that is fed

in to the proposed model which contains 2 hidden layers and in the output layer sigmoid activation function is used. The main objective of this framework is to remove the overfitting and underfitting problem. The X^2-DNN achieve high accuracy than the conventional DNN.

Geetha S., Santhana Krishnan J.[6] predicted the Heart diseases using data mining techniques. The algorithms used in this study are Naïve Bayes and Decision Tree. An accuracy level of 91% was achieved by the decision tree model in predicting heart disease patients, while an accuracy level of 87% was achieved by the Naive Bayes classifier.

Deepak Kumar Chohan, Dinesh C Dobhal [7] proposed a model to compare the supervised ML algorithms for the prediction of Heart Disease. They used a dataset containing 13 attributes and 1025 samples. Decision Tree provided the best accuracy which was 98.7% among all the Machine Learning Algorithms used in this experiment.

Gautam Srivastava , Chandrasegar Thirumalai, Senthilkumar Mohan [8] proposed a novel HRFLM (Hybrid Random Forest with Linear Model) [8] for improving the accuracy in the prediction of cardiovascular disease. In the study a dataset consisting of 13 attributes and 303 patient record from which 6 of them contains a missing value. After pre-processing the dataset, the final dataset consists of 297 patients record with 13 attributes. The study conducted a comparative analysis experiment on various Machine Learning models and a Hybrid Model and as a result it was concluded that Hybrid Model with an accuracy of 87.8% was the best performing model.

Ke Yuan, Longwei Yang, Yabing Huang, Zheng Li [7] proposed a new hybrid gradient boosting decision tree with logistic regression (HGBDTLR) [9] model for improving the accuracy in the prediction of heart disease. The Cleveland heart disease dataset contain 14 attributes. By using pre- processing techniques missing values is filled in the dataset. The HGBDTLR provided the accuracy of 91.8% which is better from all the other machine learning model.

M-Tahar Kechadi, Abdelkamel Tari, and Dhai Eddine Salhi [10] predicted a heart disease using three data analytics technique like K-Nearest Neighbour (KNN), Support Vector Machine (SVM) and Neural Network. In this study, dataset was collected manually and it contains 20 attributes. For pre- processing the dataset, Pearson Correlation Matrix was applied for feature selection and finally only 13 attributes remained that are highly correlated with each other. The three algorithms were tested on the same dataset while changing the testing size. The best performing model with an accuracy of 93% found to be Neural Network.

Sakshi Bhoyar, Nikki Wagholikar, Kshitij Bakshi, Sheetal Chaudhari [11] proposed a model named as Multilayer Perceptron for the prediction of heart disease. In this study, two types of datasets are used namely UCI heart disease (13 attributes) and cardiovascular disease (12 attributes) dataset. These two datasets are chosen because number of missing values present in the dataset is less. After that dataset is fed in to the Multilayer perceptron (MLP) which consists of 2-hidden layers and 8-unit neuron each are present in the layers. In this study, MLP scored accuracy of 85.71% on UCI dataset and 87.3% on CVD dataset higher than the existing model accuracy by a great margin[11].

III. Research Methodology

A. Dataset

The dataset used was precured from Kaggle. The Heart disease UCI dataset has 14 attributes and 1025 samples. Figure 1 is a correlation heatmap plotting the dependencies between different attributes. Blocks belonging to the same sector are correlated.

Figure 2 depicts the dependencies between different attributes using histogram (bar graph). Horizontal x-axis depicts the classes of the dataset and the vertical y-axis depicts the range of occurrences a class observes.

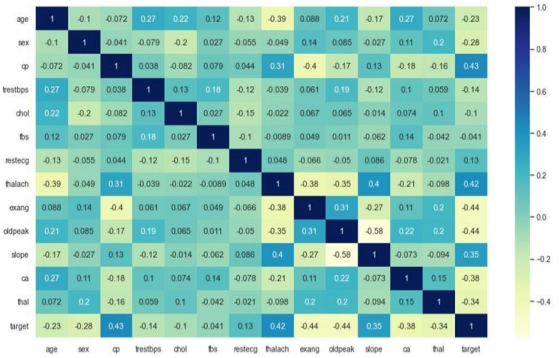

Figure 1 Heatmap Of 14 attributes

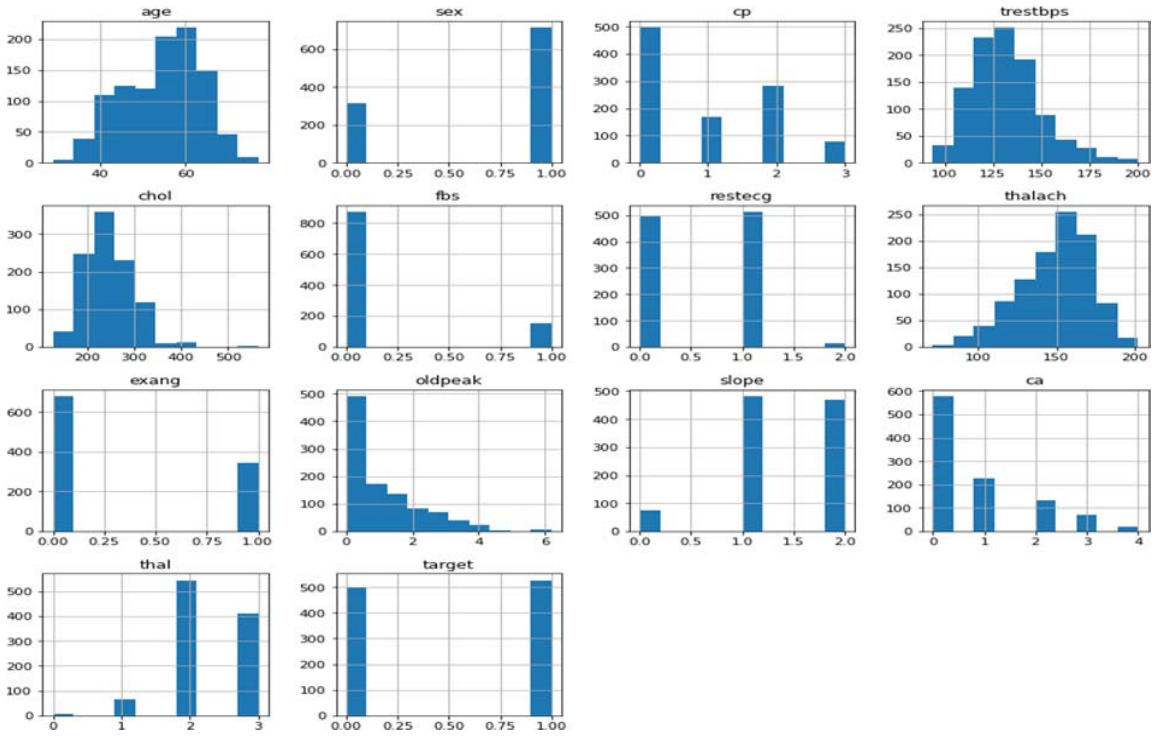

Figure 2 Histogram On each attribute

B. Decision Tree

Decision tree is also known as non-parametric supervised learning technique for it is used for regression and classification problems while also creating a model capable to predict values of a target variable. The randomness of data is an essential part of decision tree which is also known as Entropy. If a selection of an attribute occurs the change in entropy is also known as Information Gain.

$$\text{Entropy} = -\sum_{i=1}^{m} p_i \log_2 p_i$$

$$\text{Information Gain} = Entropy - \sum_{i=0}^{n} p(x) \times H(x)$$

C. K Nearest Neighbour

KNN is an algorithm which is useful for both regression and classification problems [12] for K-NN is a supervised ML algorithm. K-NN algorithm assumes similarities between the new case besides already available cases moreover places the new data into the category that is most parallel to the already existing class. Moreover, it's another name is lazy learner algorithm since instead of immediately learning from the training set, it saves the dataset, performing an action on the dataset only when it is time to classify. The main goal of the KNN is to find that new data point belongs to which category. Euclidean Distance has been calculated to determine which data points are closest to a given query point.

The Distance Equation:

$$\sqrt{((x_1 - x_2)^2 + (y_1 - y_2)^2)}$$

D. Logistic Regression

LR is also a ML algorithm that is used for binary classification problems. It is an algorithm which is used to predict categorial or discrete values, it gives a probabilistic value while using logistic function and the output lies between 0 and 1 given an input variable.

$$\text{Logistic Function} = \frac{1}{(1+e^{-x})}$$

E. Naïve Bayes

NB algorithm is a probabilistic supervised ML technique used for Bayes theorem-based classification problems. Bayes Theorem is used for calculating conditional probability. The possibility of an event happening given that another action has already happened is called conditional probability.

Mathematical Representation of Bayes theorem:

$$P\left(\frac{A}{B}\right) = \left(P\left(\frac{B}{A}\right) * P(A)\right)/P(B)$$

P(D/E) = Possibility (P) of D happening while indicating E has already happened.

P(B/A) = possibility of E happening while indicating D has already happened.

P(D) = D's possibility of happening. P(E) = E's possibility of happening

F. Random Forest

Random Forest is a ML algorithm used for regression and classification algorithm mainly used for classification purpose. It's an algorithm which makes use of many decisions tree applied on different subsets of the same input dataset, the average of all these values increases the dataset predicted accuracy. The accuracy of the models in the Forest depends on the number of trees used, greater the no. of trees higher the accuracy and it also helps in overcoming overfitting issues.

IV. Result and Discussion

This project's goal is to determine if the patient suffering from cardiovascular disease or not. The entries in the datasets will be split into training subset and test subset. X- axis shows the test size taken for the training while Y-axis show accuracy of various models. Total of 5 different algorithms are compared and the result is plotted in the **Figure 3,4,5 and 6**.

In the experiment conducted the dataset was splitted in different ratios and then the accuracy of all the models is calculated to better compare different models.

A we can see in the **Figure 3** overall the best performing model was RF (Random Forest) with an average accuracy of 93.7%. While RF was the best model the worst performing model was NB with an average accuracy of 83.36% the second worst model was DT with an average accuracy of 83.98%, while KNN and LG performed average with the accuracy of 86.6%, 85.92% respectively.

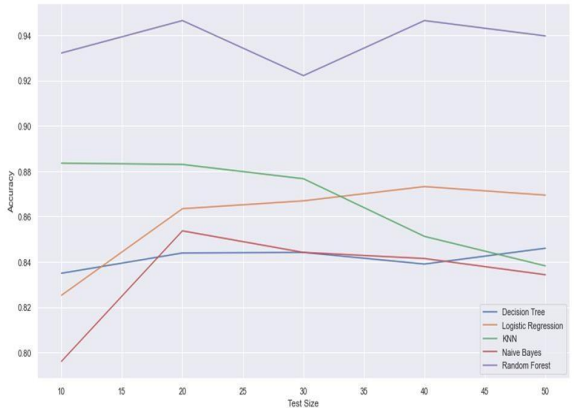

Figure 3 Comparing the Accuracy of all Models

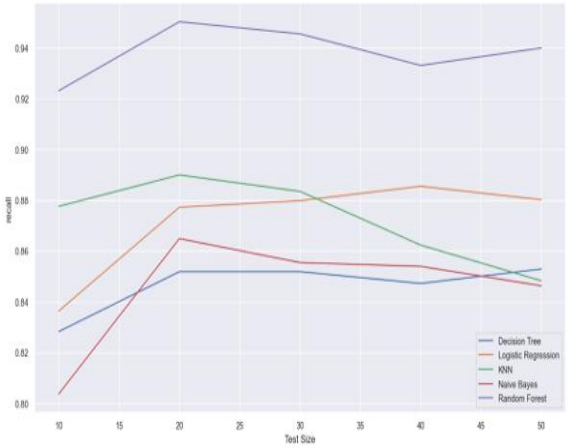

Figure 4 Comparing the F1 score of all Models

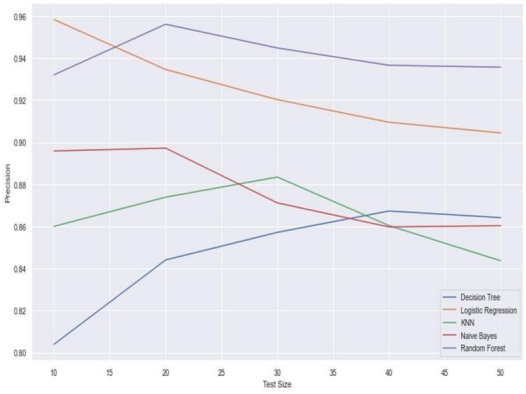

Figure 5 Comparing the Precision of all models

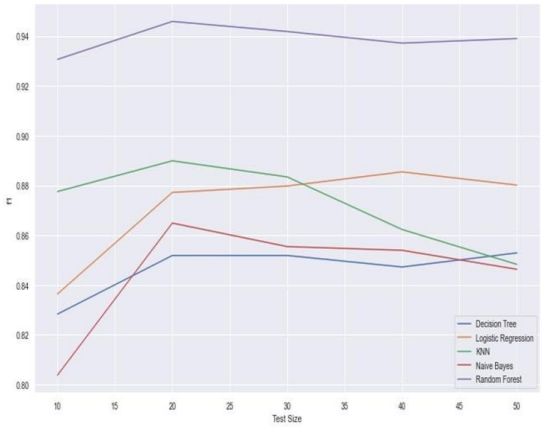

Figure 6 Comparing the F1-score of all Models

Table 1 Comparative Analysis of 5 different models with 0.2 test size

Models	*Accuracy*	*Recall*	*F1-score*	*Precision*
Decision Tree	*84.3%*	*85.18%*	*85.18%*	*85.7%*
Logistic Regression	*86.3%*	*87.7%*	*87.71%*	*93.45%*
KNN	*88.29%*	*88.99%*	*88.99%*	*87.38%*
Naïve Bayes	*85.36%*	*86.48%*	*86.48%*	*89.71%*
Random Forest	*95.12%*	*95.45%*	*95.45%*	*95.12%*

V. CONCLUSION

With the help of heart disease dataset containing 1025 samples, 5 different models were implemented and for each model tunning of test size hyperparameter was done. The conclusion was made that RF is the best classifier for this kind of Dataset. With an overall higher average accuracy than others model as well as RF has the best F1 and precision score. While the worst model was Naïve Bayes model. For all model's best test size percentage was calculated as 20%. Future research area would be to predict what kind of disease the patient is suffering because this models only predicts if the patient has a heart related disease or not.

VI. REFERENCES

[1] R. Mowriya, P. Ramprakash, S. Nithyavishnupriya and R. Sarumathi, "Heart Disease Prediction Using Deep Neural Network," Proc. 5th Int. Conf. Inven. Comput. Technol. ICICT 2020, pp. 666-670, 2020, doi: 10.1109/ICICT48043.2020.9112443

[2] D. P. Yadav, P. Mittal and P. Saini, "Feature Optimization Based Heart Disease Prediction using Machine Learning," 2021 5th Int. Conf. Inf. Syst. Comput. Networks, ISCON 2021, pp. 1-5, 2021, doi: 10.1109/ISCON52037.2021.9702410.

[3] J. C. Martinez-Ovando and S. Shinde, "Heart Disease Detection with Deep Learning Using a Combination of Multiple Input Sources," ETCM 2021 - 5th Ecuador Tech. Chapters Meet., pp. 2021-2023, 2021, doi: 10.1109/ETCM53643.2021.9590672.

[4] S. Priyanka, G. Thilagavathi, J. S. Shri and V. Roopa, "Heart Disease Prediction using Machine Learning Algorithms," Proc. - Int. Conf. Appl. Artif. Intell. Comput. ICAAIC 2022, pp. 494-501, 2022, doi: 10.1109/ICAAIC53929.2022.9793107.

[5] I. Sutedja, "Descriptive and Predictive Analysis on Heart Disease with Machine Learning and Deep Learning," 3rd Int. Conf. Cybern. Intell. Syst. ICORIS 2021, 2021, doi: 10.1109/ICORIS52787.2021.9649585.

[6] Geetha.S and S. Krishnan.J, "Prediction of heart disease using machine learning algorithms," Proc. Int. Conf. Smart Technol. Comput. Electr. Electron. ICSTCEE 2020, pp. 197-202, 2020, doi: 10.1109/ICIICT1.2019.8741465.

[7] D. C. Dobhal and D. K. Chohan, "A Comparison Based Study of Supervised Machine Learning Algorithms for Prediction of Heart Disease," pp. 372-375, 2022, doi: 10.1109/cises54857.2022.9844328.

[8] G. Srivastava, C. Thirumalai and S. Mohan , "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[9] Z. Li, L. Yang, K. Yuan and Y. Huang, "Heart Disease Prediction Algorithm Based on Ensemble Learning," Proc. - 2020 7th Int. Conf. Dependable Syst. Their Appl. DSA 2020, pp. 293-298, 2020, doi: 10.1109/DSA51864.2020.00052.

[10] D. E. S. M-Tahar Kechadi, Abdelkamel Tari, "Heart Disease Prediction using Machine Learning," in 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2022, 2021, no. Febraury, pp. 2633-2635. doi: 10.1007/978-3-030-69418-0_7.

[11] K. Bakshi, N. Wagholikar, S. Bhoyar and S. Chaudhari, "Real-time heart disease prediction system using multilayer perceptron," 2021 2nd Int. Conf. Emerg. Technol. INCET 2021, pp. 5-8, 2021, doi: 10.1109/INCET51464.2021.9456389.

[12] F. Mendonca, R. Manihar, A. Pal, and S. U. Prabhu, "INTELLIGENT CARDIOVASCULAR DISEASE ESTIMATION".