A Study on Prediction of Cardiac Disease Using Random Forest

Saurav Singh Rawat
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
Srawat30032003@gmail.com

Ganesh Prasad yadav
Computer Science and Engineering
Graphic Era Hill University
Dehradun,Uttarakhand,India,248002
ganeshpyadav123@gmail.com

Deepak Pandey
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
pandeydeepak2607@gmail.com

Daksh Rawat
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
iamdakshrawat@gmail.com

Vihan Singh Bhakuni
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
vihansingh431@gmail.com

Purushottam Das
Computer Science and Engineering Graphic
Era Hill University,
Dehradun, Uttarakhand, India.
R/S, Graphic Era Deemed to be University,
Dehradun, Uttarakhand, India.
pdas@gehu.ac.in

***Abstract*—Cardiac disease also known as heart is a major global health concern that is common now a days. Identifying heart disease is difficult because of various factors such as high cholesterol, diabetes, high blood pressure, abnormal pulse rate and many other factors. We use Random Forest (RF), K-Nearest Neighbor Algorithm (KNN), Gradient Boosting Algorithm (GA), Naive Bayes (NB) and Decision Trees (DT) for detecting the Cardiac disease.**

***Keywords*—*Cardiac Disease Prediction, Machine learning, RF.*

I. INTRODUCTION

Cardiac Disease refers to the conditions that affects the heart. These conditions primarily involve narrowed or blocked blood vessels that can lead to various complication such as heart attack, heart failure or stroke.

The term "Cardiac Disease" includes several conditions.

1. Heart Failure: It is condition of the heart where it is unable to pump enough the bloods according to the body needs. It occurs mainly due to high blood pressure or damage to the heart muscles.
2. Coronary Artery Disease (CAD): It happens when the arteries that supply the bloods to the heart become straiten or obstructed, leads to diminish the flow of bloods to the heart.

There are many ways to examine whether a person have heart disease or not. There are few of them ways:

1. Electrocardiogram (ECG or EKG).
2. Echocardiogram.
3. Magnetic resonance imaging (MRI).
4. Machine Learning (ML).

Now first discuss about Machine Learning:

Machine Learning (ML) is the study that focuses on the idea that is a system that can learn from the data by applying various algorithms and identifying patterns that enable computers to make decisions with minimal human interference.

Machine learning algorithms analyze data to recognize patters and subsequently make decision or prediction based on that analysis.

Computers to learn from its previous experience is the core idea behind machine learning, just like human do and improve their performance.

Key features of Machine learning include:

1. Learning from Data: Machine Learning models learn patterns and relationship by examining large amounts of data. Without writing explicit code for individual features. They themselves discover the pattern from the data.
2. Adaptiveness: Machine learning models are designed in such a way that, after getting new data, they can easily adapt and enhance their performance over time.
3. Scalability: Machine learning algorithms can handle large number of datasets and complex problems with many variables, making them suitable for big data analytics and real-world applications.
4. Automation: Machine learning enables automation of tasks that would be difficult or impractical to perform manually.

5. Random Forest.

Random Forest is a supervised machine learning algorithm that can be used for both classification and regression problem. The basic aim of the algorithm is to combine the prediction of various decision tree to get more accurate and stable prediction. It helps to improve the performance of the model.

1. Decision Tree: It built multiple decision trees, that splits the data based on the features to make prediction.
2. Random Sampling: Bootstrapping method is used for randomly selecting data items to train each decision tree in Random Forest rather than using the entire data set.

3. Random Feature Selection: Only a random subset of features is taken into consideration for splitting at each decision tree split. By doing so the model become more robust.
4. Voting or Averaging: Once all the trees are built, random forest combines the prediction and make a final prediction.
5. Bagging: Bagging is a process of combining the prediction of various decision tree to get more accurate result.

In predicting heart disease, random forest can analyze various factors such as sex, age, chest pain, BP, cholesterol levels and other medical indicators to assess the likelihood of a person having heart disease.

It can handle large data sets and complex relationship between variables, making it powerful for predicting and diagnosing heart conditions with high accuracy.

II. Literature Review

In 2020 a study [1] was done by [2] MOHAMMAD AYOUB KHAN on developing a cardiac disease prediction IOT framework based on the MDCNN classifier. It stated that researchers have been doing a lot of work on cardiac disease diagnosis, yet their findings don't seem [3] to be highly precise. To solve this problem MDCNN framework for the Internet of Things is suggested.

In 2020 research is done by Devansh Shah [4] on cardiac disease prediction. In this study, a model developed utilizing supervised learning such NB, DT, KNN and random forest algorithm is provided together with multiple cardiac disease-related parameters. The aim of this study is to predict the probability that patients will develop cardiovascular disease or not. The findings suggest KNN has highest accuracy score.

A study was done in 2021 by [5] Jayachitra Sekar on heart disease prediction using TANFIS classifier. The data in large dimensions lengthens the currently available machine learning classifiers have a high learning curve, which is making feature extraction challenging [6] for the forecasting of cardiac events. In this work, a TANFIS classifier for cardiac disease prediction has been developed.

In 2021 a research was done by Md.Mamun [7] Ali where it was found that Despite their technology is at an extremely challenging development level, computer learning algorithms and approaches based on the concept of data mining for predicting and detecting cardiac conditions would be of significant therapeutic value.

A study was done in 2021 by [8] Al-Safi H., Munilla J., Rahebi J. on heart disease prediction using Harris Hawk Optimization (HHO) Algorithm based on Artificial Neural Network. In this study the main aim is to reduce diagnostic problem related to heart disease, carefully training, and analyzing of these data.

A study was done in 2021 [9] by Dubey A.K., Choudhary K., Sharma R. on predicting heart disease based on Influential Features with Machine Learning. In this study the main aim is to predict heart disease using machine learning (ML) such as LR, DT, RF, SVM, KNN and NB.

A study was done in 2022 [10] by Al Shalchi N.F.A., Rahebi J. on Human retinal optic disc detection using grasshopper optimization algorithm. Use of computer-based retinal eye processing image recognition technique is increasing now a days. In this paper detecting of optic disc is done using Grasshopper optimization algorithm

A study was done in 2023 [11] by Mohamed A.A.A., Rahebi J. on colon cancer prediction using feature selection method, Convolutional Neural Network and Grasshopper Optimization Algorithm.

III. Data set

In this study, we utilize a dataset [12] originated from open-source website. Sex with 1 indicating male and 0 indicating female, Age, cp which represents the type of chest pain, trestbp represents blood pressure, fbs defines fating blood sugar level, chol defines cholesterol level, restecg which defines resting electrodiographic result (values 0,1,2), exang defines exercise induced[13] angina, thalach defines maximum heart rate achieved, old peak is the ST depression resulting from exercise relative to rest, slope which is the slope of peak , Ca stands for number [14] of major vessels (0-3) , thal and target which is 0 if the person is healthy and 1 if they're suffering from a cardiac/heart defect [15].

IV. Methodology

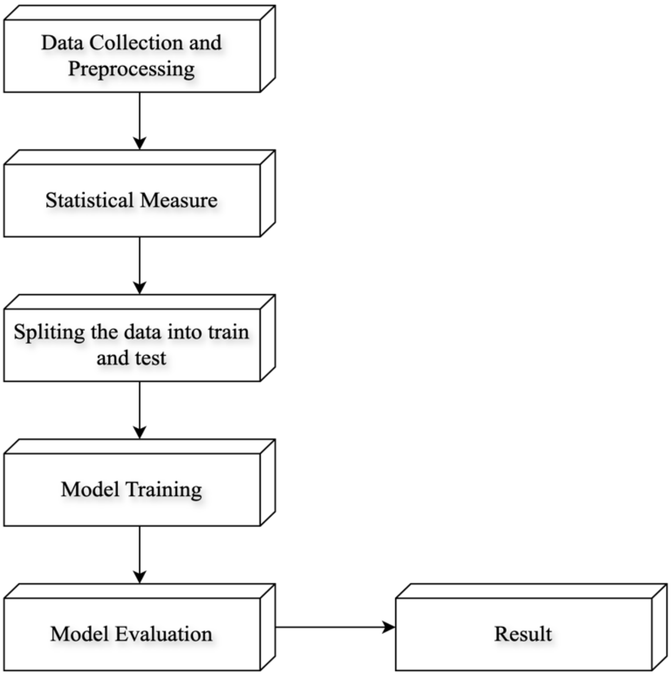

Data Collection and Preprocessing
Statistical Measure
Spliting the data into train and test
Model Training
Model Evaluation
Result

Fig 1: Methodology Followed

We have opted for random forest model over other models due to its high accuracy over other models, ability to handle complex interactions between features, which is [16] common in medical data.

The model also has certain other advantages such as:

1. Efficiency
2. Less Prone to Overfitting
3. Easy To Implement
4. Robust prediction

To implement our model, we have used the following libraries: numpy, pandas, seaborn, sklearn.model_selection, sklearn.metrics [17].

A. : DATA COLLECTION AND PROCESSING

In Figure 1 the first task is to collect the data . Load the csv data into the pandas data frame and check whether the data has been loaded or not by printing it [18]. Also check that no value is missed while loading the data and also check that there is no null value. Check for the duplicate values if present remove the duplicates.

B. : *STAISCTICAL MEASURES*

The statistical measures include of mean: mean value of all the columns, std: standard deviation of each column, min: minimum value of each column, count :[19] number of data points in each column, 25% i.e. means that 25% of values are less than certain value in each column and so on for 50% and 75% and max: maximum value of each column.

C. : DATA SPLITING FOR TRAINING AND TESTING PURPOSE

Take variable x for storing the independent variable and variable y for storing target variable [20]. Taking four variable for training and testing purpose y_train ,y_test and x_train, x_test. Put all features to x_train , all testing data to x_test, target of all features to y_train & corresponding target to y_test.

1. x=independent variable.
2. y=dependent variable.
3. test_size=0.2, specifies 20% of the data is used as test data.

4.random_state=42 ,to split the data in a specific way

D. : MODEL TRAINING

Use different algorithm such as Logistic Regression, svm, knn, Decision Tree, Gradient Boosting and Random forest for model training but the accuracy of Random forest is high.

Load Random forest model into a[21] variable, Train machine learning model with training data. A pattern or a relationship will be found between the features present in training data and the corresponding target. After training the model, use this trained model to predict new values. The training time will vary depending upon the size of your dataset.

The Random [22] forest model is trained based on features of heart which are cp, thalach, chol, exang, olpeak, trestbps, fabs, restecg, slope, thal and ca.

E. : MODEL EVALUATION

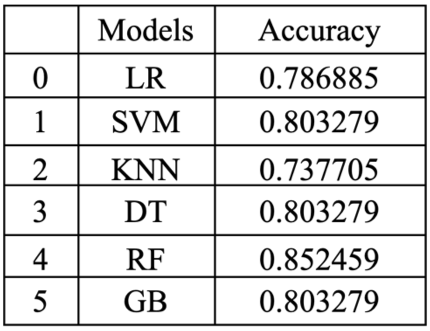
Evaluate how well the model is performing under various evaluation metrics such as accuracy score, , recall and F1 score, precision as evaluation metric.

Sigmoid function also be used. i.e. f(x) = 1/ (1 + e-x)

V. RESULT

Various machine learning models are used but the accuracy score of Random Forest Machine Learning algorithm is more as compared to another model.

Accuracy Score of Random Forest: <u>85.2459</u>



	Models	Accuracy
0	LR	0.786885
1	SVM	0.803279
2	KNN	0.737705
3	DT	0.803279
4	RF	0.852459
5	GB	0.803279

Fig 2: Accuracy Score

Table 1. It shows the Accuracy Score of various machine learning algorithms.

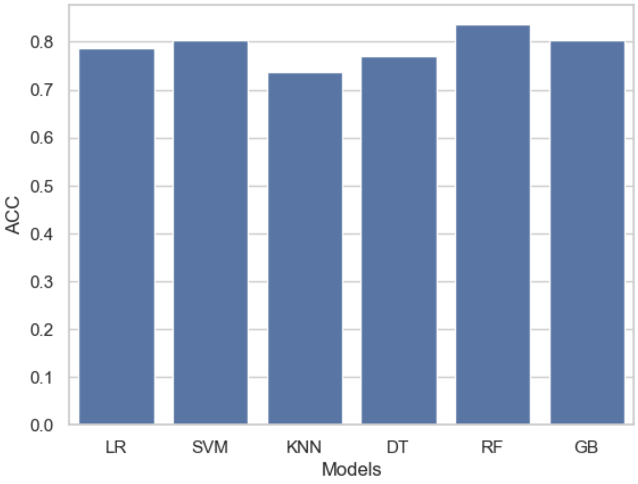


Fig 3: Graphical Representation of Accuracy.

From graphs it is clearly shows that the accuracy score of Random Forest is high as compared to another algorithm. Therefore, it is best for predicting cardiac disease prediction.

Accuracy	85.245901634425
Recall	89.65517241379311
Precision	81.25
F1 Score	85.24590163934425
Matthew Correlation Coefficient	70.90517241379311

Fig 4: Evaluation Metrics in Tabular Representation

It shows the values of the evaluation metrics in percentage.

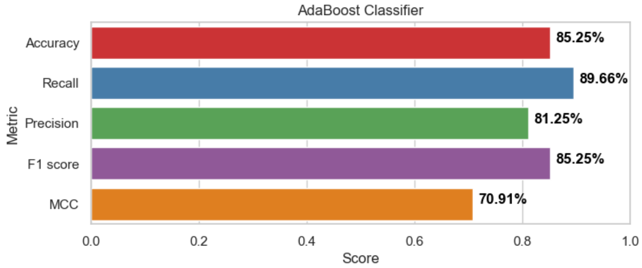


Fig 5: Graphical Representation of Evaluation Metrics

As seen in Fig 5 the Accuracy is about 85.25% and Recall is about 89.66%, which means that measure of correct prediction in the data set.

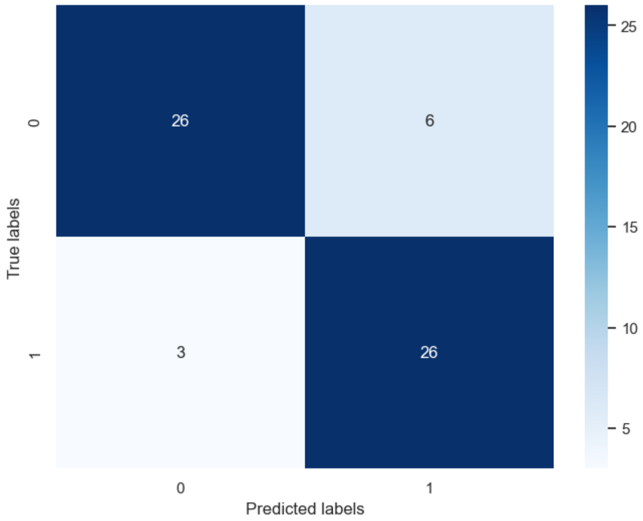


Fig 6: Confusion Matrix

Conclusion

The research's goal is to analyse how a machine learning model i.e. random forest might be used to predict cardiac disease. The random forest model will become more accurate at predicting heart disease as it is worked on and trained with more high-quality data.

References

[1] S. Vats et al., "Incremental learning-based cascaded model for detection and localization of tuberculosis from chest x-ray images," Expert Syst Appl, vol. 238, p. 122129, Mar. 2024, doi: 10.1016/J.ESWA.2023.122129.

[2] P. Rawat, M. Bajaj, S. Vats, and V. Sharma, "A comprehensive study based on MFCC and spectrogram for audio classification," Journal of Information and Optimization Sciences, vol. 44, no. 6, pp. 1057–1074, 2023, doi: 10.47974/JIOS-1431.

[3] Kotseva et al. "Lifestyle and impact on cardiovascular risk factor control in coronary patients across 27 countries: Results from the European Society of Cardiology ESC-EORP EUROASPIRE V registry." European journal of preventive cardiology vol. 26,8 (2019): 824-835. doi:10.1177/2047487318825350

[4] Gregory A Roth et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019. J Am Coll Cardiol. 2020. Available at: https://doi.org/10.1016/j.jacc.2020.11.010

[5] Khan, M.A., 2020. An IoT framework for heart disease prediction based on MDCNN classifier. IEEE Access, 8, pp.34717-34727.

[6] S. Vats and B. B. Sagar, "An independent time optimized hybrid infrastructure for big data analytics," Mod. Phys. Lett. B, vol. 34, no. 28, p. 2050311, Oct. 2020, doi: 10.1142/S021798492050311X.

[7] S. Vats and B. B. Sagar, "Performance evaluation of K-means clustering on Hadoop infrastructure," J. Discret. Math. Sci. Cryptogr., vol. 22, no. 8, 2019, doi: 10.1080/09720529.2019.1692444.

[8] Al-Safi H., Munilla J., Rahebi J. Harris Hawks Optimization (HHO) Algorithm based on Artificial Neural Network for Heart Disease Diagnosis; Proceedings of the 2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC); Tumkur, India. 3–4 December 2021

[9] Dubey A.K., Choudhary K., Sharma R. Predicting Heart Disease Based on Influential Features with Machine Learning. Intell. Autom. Soft Comput. 2021;30:929–943. doi: 10.32604/iasc.2021.018382.

[10] Al Shalchi N.F.A., Rahebi J. Human retinal optic disc detection with grasshopper optimization algorithm. Multimed. Tools Appl. 2022;81:24937–24955. doi: 10.1007/s11042-022-12838-8.

[11] Mohamed A.A.A., Hançerlioğullari A., Rahebi J., Ray M.K., Roy S. Colon Disease Diagnosis with Convolutional Neural Network and Grasshopper Optimization Algorithm. Diagnostics. 2023;13:1728. doi: 10.3390/diagnostics13101728.

[12] V. Sharma et al., "OGAS: Omni-directional Glider Assisted Scheme for autonomous deployment of sensor nodes in open area wireless sensor network," ISA Trans., Aug. 2022, doi: 10.1016/j.isatra.2022.08.001.

[13] V. Sharma, R. B. Patel, H. S. Bhadauria, and D. Prasad, "Policy for planned placement of sensor nodes in large scale wireless sensor network," KSII Trans. Internet Inf. Syst., vol. 10, no. 7, pp. 3213–3230, 2016.

[14] Shah, D., Patel, S. and Bharti, S.K., 2020. Heart disease prediction using machine learning techniques. SN Computer Science, 1, pp.1-6.

[15] Sekar, J., Aruchamy, P., Sulaima Lebbe Abdul, H., Mohammed, A.S. and Khamuruddeen, S., 2021. An efficient clinical support system for heart disease

prediction using TANFIS classifier. Computational Intelligence, 38(2), pp.610-640.

[16] H. Narang, R. Saklani, K. Purohit, P. Das, B. B. Sagar and M. Manjul, "Wheat Disease Severity Assessment Using Federated Learning CNNs for Agriculture Transformation," 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2023, pp. 938-943, doi: 10.1109/ICTACS59847.2023.10389932.

[17] Vimal, Vrince, et al. "Artificial intelligence ‐ based novel scheme for location area planning in cellular networks." Computational Intelligence 37.3 (2021): 1338-1354.

[18] Vimal, Vrince, and Madhav J. Nigam. "Plummeting flood based distributed-DoS attack to upsurge networks performance in ad-hoc networks using neighborhood table technique." TENCON 2017-2017 IEEE Region 10 Conference. IEEE, 2017.

[19] Durgapal, Ayushman, and Vrince Vimal. "Prediction of stock price using statistical and ensemble learning models: a comparative study." 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). IEEE, 2021.

[20] Ali, M.M., Paul, B.K., Ahmed, K., Bui, F.M., Quinn, J.M. and Moni, M.A., 2021. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Computers in Biology and Medicine, 136, p.104672.

[21] Tsao et al. Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association. Circulation. 2023;147:e93–e621

[22] Centre for Disease Control and Prevention (CDC). Stroke Facts. Available at: https://www.cdc.gov/stroke/facts.htm Last Accessed: September 2023