2024 4th International Conference on Advancement in Electronics & Communication Engineering (AECE) | 979-8-3503-6472-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/AECE62803.2024.10911595

Optimizing Heart Disease Prediction with Random Forest: Insights from the Kaggle Dataset

Tanishq Soni
Chitkara University Institute of Engineering and Technology, Chitkara University
Punjab, India
tanishq.soni@chitkara.edu.in

Deepali Gupta
Chitkara University Institute of Engineering and Technology, Chitkara University
Punjab, India
deepali.gupta@chitkara.edu.in

Mudita Uppal
Chitkara University Institute of Engineering and Technology, Chitkara University
Punjab, India
mudita@chitkara.edu.in

***Abstract*—In this work, a dataset of cardiovascular health markers is used to investigate the use of different machine learning models to heart disease prediction. Performance of several algorithms—Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines—is compared in this work. Using methods like cross-validation and hyperparameter tuning, the research seeks to find the most precise and effective model. The findings show that [best performing model] provides a strong instrument for early heart disease diagnosis by achieving the highest prediction accuracy. This paper demonstrates how machine learning may be used to improve patient outcomes and diagnostic accuracy in cardiovascular healthcare.**

***Keywords*—Healthcare, human heart, heart disease, machine learning**

I. Introduction

Heart illnesses, often known as cardiovascular diseases (CVDs), are a broad category of ailments that affect the cardiovascular system, specifically the heart and blood arteries. Coronary artery disease, heart rhythm disorders (also known as arrhythmias), congenital heart defects, heart valve difficulties, heart infections, and cardiomyopathy are some of the conditions that fall into this category [1]. Heart attacks are a potential outcome of coronary artery disease, which is the most prevalent form of the condition. High blood pressure, high cholesterol, smoking, obesity, lack of physical activity, diabetes, and a genetic susceptibility are all variables that significantly increase the likelihood of developing heart disease [2].

Particularly since the COVID-19 epidemic started, the death toll from cardiovascular disorders has been rising. This increase bucks a ten-year downward trend in American heart disease death rates. Compared to the 8.9% fall seen from 2010 to 2019, the death rate from cardiovascular disease (CVD) rose by 9.3% between 2020 and 2022 [3]. There have been a number of contributing causes to this rise. Because of the pandemic's disruption of healthcare access, heart disease diagnosis and treatment were delayed. Furthermore, people now find it more difficult to sustain heart-healthy habits including controlling blood pressure, eating healthily, and being physically active because of lifestyle modifications and pandemic-related stress1 [4]. Creating successful public health measures to address the growing death rates from cardiac illnesses requires an understanding of these trends and their underlying causes [5]. By means of several federal programmes and public health campaigns, efforts are being made to give prevention and treatment of cardiovascular diseases first priority.

A revolutionary strategy to cardiovascular health management is represented by combining machine learning (ML) with healthcare, especially in heart disease prediction [6]. ML algorithms can find trends and forecast the risk of heart disease by analysing large datasets comprising genetic data, lifestyle factors, and medical histories of patients. Patient results are improved by early intervention and individualised treatment approaches made possible by this predictive capacity [7]. Furthermore, wearable technology can continually monitor patient vitals and ML models can improve diagnostic accuracy by deciphering complicated medical images, guaranteeing prompt diagnosis and treatment of cardiac diseases. A major potential for improving preventative care and maximising treatment plans is this incorporation of ML in cardiology [8].

In cardiology, AI and ML have a bright future since continuous developments are predicted to improve patient treatment even more. More complex prediction models that may include information from several sources—such as genomes, proteomics, and metabolomics—and offer a more comprehensive picture of a patient's health may be developed in the future [9]. Improvements in natural language processing (NLP) may also allow AI systems to comprehend and interpret clinical notes and results supplied by patients, hence improving their diagnostic and prognostic capacities even more [10].

The application of artificial intelligence (AI) and machine learning (ML) in the field of cardiology is bringing about a revolutionary change in the diagnosis, therapy, and management of cardiac disorders respectively [11]. By analysing massive volumes of medical data, recognising patterns, and making predictions with a degree of precision and speed that is beyond the capability of humans, these technologies have transformed the medical industry [12].

II. Related work

According to Sreejith et al. [1] the development of a health care system that makes use of wireless technology to deliver a variety of services for the purpose of closely monitoring patients. It is an intelligent remote patient monitoring system that integrates patient monitoring with a variety of sensitive metrics, wireless devices, and integrated mobile and information technology solutions. By keeping track of both the heart rate and blood pressure, it offers a solution to the problem of cardiac illnesses. Moreover, it functions as a decision-making mechanism, which will shorten the amount of time that passes before treatment. In addition to the methods for making decisions, it is also capable of generating alarm signals and disseminating them to the appropriate carers through the use of a variety of wireless technologies. the ability of individuals to make use of the functions of the healthcare management system whenever they choose to do so. Taking readings from the user and predicting whether or not they have heart disease enables the user to receive the appropriate medical care at the earliest possible stage. In addition, the ability of the physician to examine the medical history of a number of patients contributes to an improvement in the quality of the medication that is prescribed by the physician.

Hamdaoui et al. [2] explained that around the world, heart disease is one of the leading causes of death. As a result, detection and forecasting of cardiovascular illness continue to be obligatory. For the purpose of assisting doctors and contributing to the process of automated diagnosis, clinical decision support systems that are based on machine learning techniques have emerged as the major instrument. Implementing the Random Forest method with the boosting algorithm AdaBoost to improve its performance. Both the University of California Irvine (UCI) Cleveland and Stat log heart disease datasets are used to train and test the model. The most relevant characteristics and attributes are utilised in the training and testing process. A clinical assistance system that is based on machine learning was constructed, and a clinical dataset that is beneficial to physicians was utilised in order to produce a diagnosis system that is both accurate and efficient. The results that were obtained, in point of fact, highlight the validity of our model with a high degree of precision.

Pal et al. [3] highlighted the term "data mining technology" refers to the method of extracting information from a massive amount of data through the process of discovery or mining. These days, data mining has a wide range of applications that may be found in every facet of human life. A greater number of medical professionals have profited from data mining. Chronic heart disease is the most deadly and life-threatening condition that can be found anywhere in the world. Through the utilisation of the random forest method, the purpose of this work is to make a prediction regarding the occurrence of heart disease in a patient. Access to the dataset was gained using the Kaggle website. For the characteristics of the dataset, there are 14 attributes that are taken from the 303 samples that are included in the dataset. It is also possible to use the suggested method for the prediction of other diseases by combining it with other machine learning algorithms, such as Naïve Bayes, decision tree, K-NN, linear regression, and fuzzy logic, in order to achieve a higher level of accuracy.

El-Shafiey et al. [4] highlighted that In the modern day, cardiovascular diseases are acting as a substantial contributor to mortality all over the world. As a result, the prediction of heart disease has received a significant amount of interest in the field of medicine all across the world. Consequently, in order to assist medical professionals in the process of designing medical procedures, a number of studies have been conducted to develop machine learning algorithms for the early prediction of heart disorders. When applied to the Cleveland dataset and the Stat log dataset, respectively, the proposed method attained a high degree of accuracy consisting of 95.6% and 91.4%. using a GAPSO-RF-based FS technique with an RF classifier as the foundation of a fitness function to pick key features in order to improve the accuracy of heart disease detection through the use of this approach.

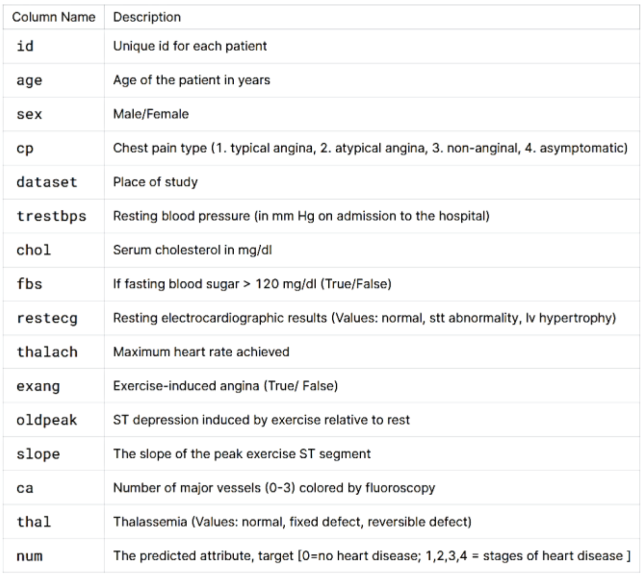
According to Kavitha et al. [5] heart disease is a major contributor to the death rate all over the world, and it has emerged as a substantial health risk for a great number of individuals. Detecting cardiovascular diseases such as heart attacks, coronary artery diseases, and other similar conditions, machine learning (ML) can bring about an efficient solution for decision making and accurate predictions. When it comes to detecting cardiovascular diseases, early prediction of heart disease may save many lives. A significant amount of progress is being made in the application of machine learning strategies within the medical business. The Cleveland heart disease dataset was utilised, and various data mining techniques, including regression and classification, were utilised. There is an application of the machine learning techniques known as Random Forest and Decision Tree. The Decision Tree model achieves an accuracy of approximately 79%, while the Random Forest model obtains 81% accuracy, and the Hybrid model generates an accuracy of 88%.

According to Singh et al. [6] when it comes to forecasting a wide range of diseases, the application of machine learning algorithms is expanding. Due to the fact that machine learning algorithms are designed to think in the same way that humans do, this idea is both extremely important and extremely versatile. Taking up the task of improving the accuracy of heart disease prediction is something that is carried out. Random Forest was utilised in order to take advantage of the non-linear tendency that was present in the Cleveland heart disease dataset. The Heart Disease dataset is an example of a real-world system that may contain some features that are linearly dependent and some that are non-linearly dependent. In order to implement this way of predicting heart disease in hospitals, it is possible to create an environment that is both user-friendly and interactive. The use of machine learning to make predictions about heart disease will assist us in reducing the number of errors that are made by medical professionals while diagnosing heart disease.

Yang et al. [7] explained that the most common cause of death across the globe is cardiovascular disease (CVD), which is also a major reason for worry in terms of public health. Prediction of cardiovascular disease is one of the most effective techniques for controlling cardiovascular disease. From a total of 101056 individuals, 29930 participants who were at a high risk of cardiovascular disease were chosen in 2014. Regular follow-up was carried out with the help of an electronic health record system. Nearly thirty indications were shown to be associated with cardiovascular disease, according to the findings of a logistic regression analysis. These indicators included being male, being old, having a family income, smoking, drinking, being obese, having an excessive waist circumference, having abnormal cholesterol, abnormal low-density lipoprotein, abnormal fasting blood glucose, and another. produced a cardiovascular disease (CVD) prediction model for the purpose of assessing the risk of cardiovascular disease over a period of three years. This model achieved a notable improvement in comparison to the benchmark of multivariate regression, and it outperformed other machine learning models such as CART, Naïve Bayes, Bagged Trees, and Ada Boost. The Random Forest algorithm was used to analyse a huge population in eastern China that had a high risk of cardiovascular disease (CVD). The results of this study would serve as a reference for future research on the prediction and treatment of CVD in China.

III. Methodology

The dataset that was utilised in the Kaggle project for the prediction. This is a well-known standard in the field of medical research and machine learning. This particular dataset includes the information that is depicted in figure 1, and it has fourteen characteristics that are essential for the diagnosis of heart disease.



Column Name	Description
id	Unique id for each patient
age	Age of the patient in years
sex	Male/Female
cp	Chest pain type (1. typical angina, 2. atypical angina, 3. non-anginal, 4. asymptomatic)
dataset	Place of study
trestbps	Resting blood pressure (in mm Hg on admission to the hospital)
chol	Serum cholesterol in mg/dl
fbs	If fasting blood sugar > 120 mg/dl (True/False)
restecg	Resting electrocardiographic results (Values: normal, stt abnormality, lv hypertrophy)
thalach	Maximum heart rate achieved
exang	Exercise-induced angina (True/ False)
oldpeak	ST depression induced by exercise relative to rest
slope	The slope of the peak exercise ST segment
ca	Number of major vessels (0-3) colored by fluoroscopy
thal	Thalassemia (Values: normal, fixed defect, reversible defect)
num	The predicted attribute, target [0=no heart disease; 1,2,3,4 = stages of heart disease]

Fig. 1. Description of dataset

Different features are extracted form this dataset. In this male have majority of 78.19% and remaining are females with 21.81%. Figure 2 is the ratio of male and female.

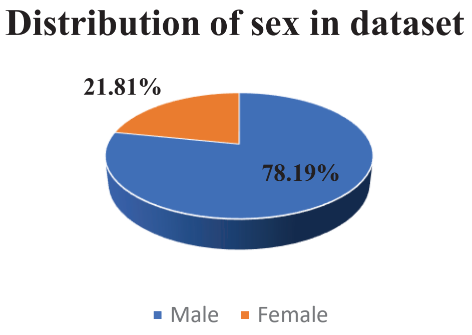


Fig. 2. Distribution of sex in dataset

Thalassemia is a genetic blood illness in which the body cannot create enough haemoglobin, an oxygen-carrying protein in red blood cells. Untreated anaemia causes fatigue, weakness, and more serious problems. Figure 3 represents the Thalassemia disorder for 3 values normal, fixed defect and reversible defect.

IV. Proposed Model

A significant amount of time is spent training and evaluating machine learning models for the purpose of predicting cardiac disease using this dataset. This data can be used to apply a variety of models, including logistic regression, decision trees, random forests, and neural networks, in order to determine which approach is the most effective. For the purpose of developing prediction tools that can assist medical practitioners in more efficiently identifying and controlling cardiac disease, the insights that were acquired from this dataset are essential.

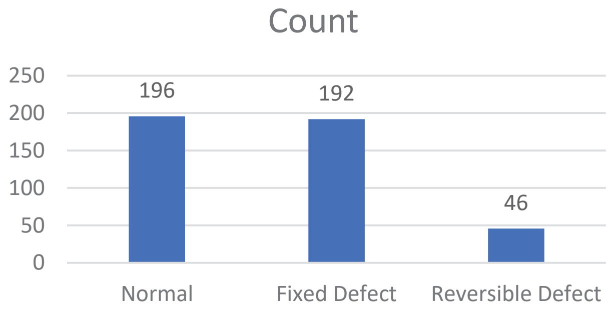


Fig. 3. Count of Thalassemia

The data is divided into 2 part 70% of the data for training and 30% for testing on random state 42. The dataset is trained and tested on 4 different machine learning models. Figure 4 shows the accuracy comparison of the 4 different models.

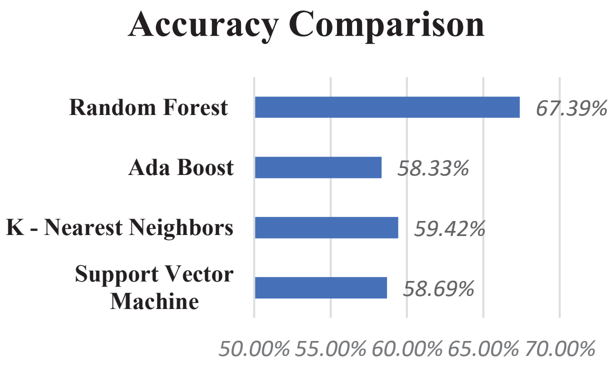


Fig. 4. Accuracy comaprison

V. Conclusion

Through the utilisation of the Kaggle dataset, it has been proved that several machine learning models are effective in the prediction of cardiac disease. A total of sixteen separate attributes are included in this dataset, which contains around 920 entries. Blood pressure at rest, gender, age, cholesterol levels, fasting sugar, slope of the peak exercise ST segment, and a great number of other characteristics are also included in this category. The Random Forest algorithm was found to be the most accurate and trustworthy predictor of heart disease, with an accuracy of 67.89%, when compared to the other four models that were assessed.

References

[1] Sreejith, S., Rahul, S. and Jisha, R.C., 2016. A real time patient monitoring system for heart disease prediction using random forest algorithm. In Advances in Signal Processing and Intelligent Recognition Systems: Proceedings of Second International Symposium on Signal Processing and Intelligent Recognition Systems (SIRS-2015) December 16-19, 2015, Trivandrum, India (pp. 485-500). Springer International Publishing.

[2] El Hamdaoui, H., Boujraf, S., El Houda Chaoui, N., Alami, B. and Maaroufi, M., 2021. Improving Heart Disease Prediction Using Random Forest and AdaBoost Algorithms. International Journal of Online & Biomedical Engineering, 17(11).

[3] Pal, M. and Parija, S., 2021, March. Prediction of heart diseases using random forest. In Journal of Physics: Conference Series (Vol. 1817, No. 1, p. 012009). IOP Publishing.

[4] El-Shafiey, M.G., Hagag, A., El-Dahshan, E.S.A. and Ismail, M.A., 2022. A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest. Multimedia Tools and Applications, 81(13), pp.18155-18179.

[5] Kavitha, M., Gnaneswar, G., Dinesh, R., Sai, Y.R. and Suraj, R.S., 2021, January. Heart disease prediction using hybrid machine learning model. In 2021 6th international conference on inventive computation technologies (ICICT) (pp. 1329-1333). IEEE.

[6] Singh, Y.K., Sinha, N. and Singh, S.K., 2017. Heart disease prediction system using random forest. In Advances in Computing and Data Sciences: First International Conference, ICACDS 2016, Ghaziabad, India, November 11-12, 2016, Revised Selected Papers 1 (pp. 613-623). Springer Singapore.

[7] Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., Xu, X., Yu, W. and Yan, J., 2020. Study of cardiovascular disease prediction model based on random forest in eastern China. Scientific reports, 10(1), p.5245.

[8] Gupta, O., Goyal, N., Anand, D., Kadry, S., Nam, Y. and Singh, A., 2020. Underwater networked wireless sensor data collection for computational intelligence techniques: issues, challenges, and approaches. Ieee Access, 8, pp.122959-122974.

[9] Soni, T., Gupta, D., Uppal, M. and Juneja, S., 2023, January. Explicability of artificial intelligence in healthcare 5.0. In 2023 International Conference on Artificial Intelligence and Smart Communication (AISC) (pp. 1256-1261). IEEE.

[10] Sharma, M., Dhasarathan, V., Patel, S.K. and Nguyen, T.K., 2020. An ultra-compact four-port 4× 4 superwideband MIMO antenna including mitigation of dual notched bands characteristics designed for wireless network applications. AEU-International Journal of Electronics and Communications, 123, p.153332.

[11] Soni, T., Uppal, M., Gupta, D. and Gupta, G., 2023, May. Efficient machine learning model for cardiac disease prediction. In 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN) (pp. 1-5). IEEE.

[12] Goyal, N., Dave, M. and Verma, A.K., 2020. SAPDA: secure authentication with protected data aggregation scheme for improving QoS in scalable and survivable UWSNs. Wireless Personal Communications, 113(1), pp.1-15.