AI-Powered Cardiac Risk Prediction System Using Random Forest and Decision Tree

[1.]S.Pandikumar , [2.]V.M.Vinay
[1] Associate Professor, Dept. of MCA, Acharya Institute of Technology, Bengaluru-560107
[2] Dept. of MCA, Acharya Institute of Technology, Bengaluru-560107

Abstract:

Despite advances in clinical medicine, cardiovascular disease among the primary causes of mortality is still illness. Worldwide, killing one person every minute due to complications stemming from heart disease. The high velocity of life today makes it difficult to catch heart disease early. Machine Learning (ML) holds great potential in diagnosing diseases accurately and early in health care.This revision presents A condition of the heart prediction system that utilizes patient medical records and applies particular Random Forest, Decision Tree, and machine learning algorithms for analysis, containing various clinical parameters, is processed in Python, enabling prediction of cardiac conditions based on historical patient data. Among the models used, the Random Forest algorithm demonstrates superior reliability and early detection cardiac stages. The proposed system reads patient data from CSV files and effectively determines the likelihood of a heart attack. It offers advantages such as high performance, flexibility, and a high success rate in prediction.

Keywords: Machine Learning, Heart Disease Prediction, Decision Tree, Random Forest, Medical Diagnosis, Early Detection, Healthcare Analytics, Patient Records.

I. INTRODUCTION

Cardiovascular disease (which, in particular, primarily affects the heart) is both the important global cause of death and one of the inexorable pressures facing the public health system. WHO, or the World Health Organization identifies that Heart disease causes around 17.9 million deaths annually. annually. , which is almost 32% of full deaths worldwide.[1] The sedentary lifestyle and poor nutritional choices contributing to obesity and elevated stress levels in our modern sedentary lifestyle have been implicated in the initializing of heart-related illness. Early identification of heart disease is important and may save a person's life as often symptoms or signs are ignored or misdiagnosed until its too late. Current diagnostic practices mainly rely on physicians to recognize symptoms, to perform physical assessment, rely on costly tests such as ECG's, angiograms, and blood tests. In many cases, while the diagnoses is accurate, they are all subject to potential sources of error, they can take time to make a diagnosis, and they all heavily depend on hospitals with the capacity to perform comprehensive testing. It is also possible that diagnostic testing is not available at all in poorer regions of the world. [2] Therefore, An urgent need exists for a clearer, more capable, also cost-effective result to assist in the early detection and treatment of heart conditions. Learning Machines (ML) is becoming a game-changing technology in healthcare, that has been developed over the past few years. ML methods can identify relationships and hidden patterns in medical datasets that can allow for faster and more accurate predictions. [3] When utilized in heart disease detection, [4] This research proposes the using two robust classifiers:. The system processes a structured dataset containing medical attributes and outputs how likely it is that a patient has heart disease. [5] The proposed model uses Python for algorithm implementation and Flask as the web interface framework to make it accessible to users such as clinicians, researchers, and patients. [6] The primary goal of this system is to achieve high prediction accuracy while ensuring ease of use and real-time feedback. It seeks to address the drawbacks of existing systemssuch as low accuracy and time-consuming processesby offering a model that is flexible, scalable, and capable of delivering results with minimal delay.[7] By deploying this model, healthcare practitioners can be better equipped to make informed decisions, thereby reducing the burden of late diagnosis and ultimately saving lives.

II. LITERATURE REVIEW

[8] Kaan Uyar and Ahmet İlhan suggested a neural network with recurrent fuzzy learning model in order to diagnose cardiac illness that is based on genetic algorithms and has a high level of precision.

. Inspired by their methodology, this work introduces A system that predicts heart disease using machine learning

. [9] A major cause of death worldwide is still heart disease. and early identification is essential. The suggested model applies Python-implemented Random Forest and Decision Tree classifiers to clinical data. The dataset contains important medical parameters and is formatted as a CSV file. [10]When predicting illness risk, Random Forest demonstrated greater accuracy and dependability. With its great accuracy, adaptability, and efficacy, the system shows promise in identifying cardiac disorders early on. Accuracy, precision, sensitivity, and F-score are evaluation measures that validate the effectiveness of the system.[11]

Y. K. Sharma, Sahil Ahir, Lakhan Ahuja, and Ashish Chhabbi explored employing data mining techniques to predict heart disease, highlighting the importance of feature selection and classifier performance.[12] Inspired by their study, this work presents Predicting heart disease using machine learning model that uses clinical datasets for early detection. Heart disease is still the primary worldwide cause of death, and predictive tools can greatly aid in diagnosis. [13] This system processes medical parameters using Decision Tree and Random Forest algorithms in Python. Patient data stored in CSV format is analyzed to identify disease risk at early stages. Random forest provides high precision and robustness in prediction. [14] Several evaluation F1-score, accuracy, sensitivity, and precision are examples of metrics that validate the effectiveness of the system.The proposed approach ensures flexible and reliable detection to support medical diagnosis.

[15] Berry et al. evaluated the effectiveness between the ATP III risk estimator and the Framingham Risk Score (FRS). in predicting deaths due to coronary heart disease (CHD) in

young men aged Using data from 10,551 participants without prior CHD or diabetes, the study compared predicted risks with observed death rate from CHD over years. Results showed that underestimated long-term risk, especially in individuals under 30, even when they had significant risk factors. [16] The study concluded that existing prediction models are inadequate for younger populations and recommended developing alternative strategies for early CHD risk communication and assessment.

Using data mining techniques, R. J. Thomas and Theresa Princy both developed a system for predicting cardiac disease. With rising heart disease cases due to lifestyle and hereditary factors, their study focuses on predicting individual risk levels based on parameters such as blood pressure, gender, age, cholesterol, and pulse rate. The paper surveys various classification algorithms to analyze medical data and assess heart disease risk. [17] It emphasizes the importance of early detection using data-driven approaches, considering medically accepted normal rangessuch as 120/90 mmHg for blood pressure and 72 bpm for pulse rateas reference points.

Kaur H Beant and William Jeet Singh reviewed various Data mining methods applied to heart conditions prediction. They emphasized that traditional methods are inadequate for processing the large and complex datasets involved in medical diagnosis. [18] Data mining enables faster and more accurate prediction by uncovering patterns in clinical data. The study highlights that certain algorithms, such as neural networks, achieved nearly 100% accuracy in some cases. Overall, the paper concludes that data mining methods significantly improve the efficiency and reliability of heart disease prediction systems.[19]

Kirmani and Ansarullah proposed An algorithm for predicting heart disease that uses decision trees algorithm, a popular data mining technique. Their study focused on classifying patient data based on medical attributes to determine heart disease risk.[20] By using Decision Tree models, the system was able to provide interpretable and accurate results, helping in early diagnosis and treatment planning. The paper highlights the effectiveness of decision trees in extracting meaningful patterns from complex healthcare datasets.[21]

Salam Ismaeel and Ali Miri presented a heart disease diagnosis system utilizing the ELM (extreme learning machine) algorithm. The model analyses key patient factors such as age, sex, cholesterol, and blood sugar to predict the likelihood of heart disease. Using real-world data from the The Cleveland Clinic Foundation involving approximately, the system achieved an accuracy of about 80%. [22] The proposed ELM-based model offers a cost-effective alternative to medical checkups by acting as an early warning tool for potential heart disease.

Tahira Mahboob, Rida Irfan, and Bazelah Ghaffar (2017) proposed an Machine learning with an ensemble approach model for predicting coronary heart disease. The study reviews several techniques, including Markov Hidden Models and Support Vector Machines , genetic algorithms, and feature selection methods, to identify the most suitable approach.[23] Based on this analysis, an Ensemble Model was developed and evaluated using Receiver Operating Characteristics (ROC). The model achieved 94.21% accuracy, 0.981 ROC score, 0.2568 RMSE, and 0.953 precision—outperforming traditional models like These include SVM, K-Nearest Neighbor, and artificial neural networks. The outcomes demonstrate how successfully the ensemble model can classify the risk of heart disease. Ammar Asjad Raja et al. (2016) proposed an intelligent framework for predicting Syncope disease using data mining techniques.[24] The model integrates an ensemble approach combining (SVM) classification systems. The Armed Forces Institute of Cardiology (AFIC & NIHD), Pakistan, provided the patient data. with 31 medical attributes used for analysis. Each classifier applied distinct rules for prediction, and final results were determined through majority voting. The framework demonstrated reliable accuracy, proving its effectiveness and potential for future medical diagnostic applications.[25]

B.L. Deekshatulu, M.A. Jabbar, and Priti Chandra (2016) developed an intelligent heart disease prediction system using a combination of Random Forest and evolutionary feature selection techniques. The model integrates the Chi-square test and genetic algorithm to identify relevant attributes, enhancing prediction accuracy. Random forest is chosen as the primary classifier, illustrating its well-known robustness and consistency of performance in medical usages. Experimental results indicate that the proposed model gives the better accuracy compared to the traditional classifications method. The system provides healthcare practitioners with more reliable diagnoses of heart disease.

III.METHODOLOGY

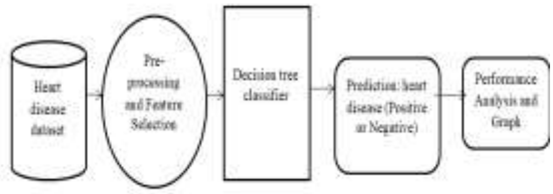
The research process explained here will be based on the application of machine learning methodologies for accurate, early detection of heart disease with structured clinical datasets. The process is designed to be well-defined and efficient, consisting of a number of stages involving data collection through to model evaluation. The main stages are given below:

Data Acquisition We utilized an openly available clinical heart disease dataset (CSV) in this research. This dataset contains critical medical parameters For example, blood pressure, cholesterol, blood sugar, age, gender, and ECG results heart rate, and other items which are used to train and test the machine learning algorithms. Data Preprocessing Prior to training models, we completed data preprocessing to establish data quality and increase performance of the models. The preprocessing consists of: Dealing with Missing Data: all records that were incomplete or missing were either dropped or imputed using valid statistical approaches. Encoding of Categorical

Variables: all categorical outputs (e.g., gender/comorbidity) were encoded to represent categories in a numerical range with either label encoding of one-hot encoding. Data Regularization: all continuous variables were normalized and all unique features values were placed into a uniform range using min-max scaling thus improving convergence for classifiers.Feature Selection We completed a feature selection step to improve model efficiency and mitigate overfitting. A group of decisions called the Random Forest Classifier trees built on random subsets of data and features. This model enhances prediction accuracy and reduces overfitting. Both models were trained using the processed dataset, with Random Forest expected to outperform due to its ensemble nature and robustness. System Implementation The entire model pipeline was integrated into a web-based application using Flask, a Python-based micro web framework. The user interface consists of: Dataset upload functionality. Data preview module. Prediction interface. Performance analytics dashboard. This setup allows end-users such as clinicians to interact with the system in real time. Model Evaluation Standard classification metrics were used to assess the model's performance: Accuracy: sensitivity, recall, and positives: Precise positive forecasts. Recall and precision are harmonically mean to get the F1-Score. An illustration of actual and expected values is called a confusion matrix.

The Random Forest model achieved an accuracy of approximately 96%, indicating high reliability in prediction tasks. Visualization and Reporting A series of graphical modules were created to assist in interpretation: Confusion Matrix

SYSTEM ARCHITECTURE:



IV.MODULE DESCRIPTION

predicted heart disease outcome model that uses the uploaded CSV dataset medical records as input and output's composite measure of prediction success, which will lead to a percentage of probability of heart disease used to display the prediction's results section of the page (i.e. the user's requested prediction results body content section). The results will display the prediction's probability outcome of heart disease based on their dataset medical records. Following the predicted results section is the outcome analysis and display section. This will allow the users to see how the model has predicted probabilities compared to the medical records of the patient. The Predictive Model Page is the last module, which provides an overview of the prediction model that was used to predict the results of heart disease, and how it works (i.e. model overview and process). An overview of the modules demonstrates how interconnected these components are and how the user's interaction takes place, and how data flows and is used throughout the components.

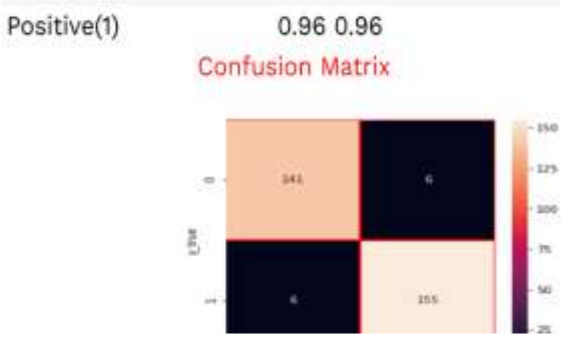


Fig 2: Confusion Matrix

Lastly, the Chart Module enhances interpretability by presenting visual summaries of the results. It uses pie charts, bar graphs, and line charts to display prediction outcomes and performance metrics, making the data more understandable and user-friendly.

V. RESULT

The proposed heart disease prediction system underwent development and testing, using a clinical dataset that includes health parameters of patients. After preprocessing and feature selection were performed, the supplied that was trained to predict whether or not heart disease was present for the given health parameters. The a confusion matrix and other related metrics. The classification performance from a confusion matrix clearly demonstrates strong performance with 141 true negatives and 155 true positives, suggesting the model correctly identified the majority of patients both without and with heart disease, respectively. The false negatives and false positives depicted were minimal given there were only 6 true/false in each category, which illustrate the robustness of the model and ability to avoid false predictions. The overall accuracy achieved was nearly 96%, which illustrates how reliable the model would be to act as part of a clinical diagnosis tool. The dashboard, also provided some graphical insights with a pie chart specifically showing the gender distribution which provided insight into the participants, which included a 60% male and 40% female patient distribution. Each representation further assists users in understanding the data distributions or prediction results.

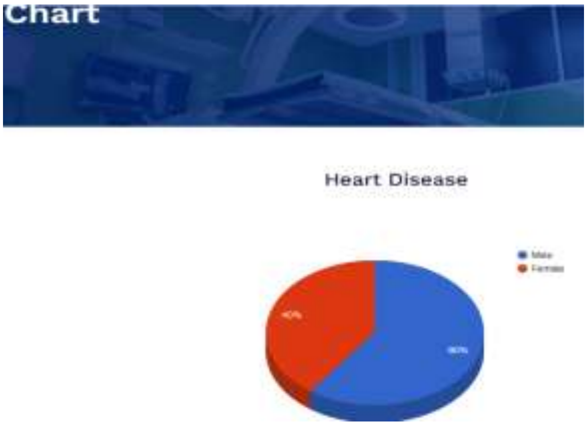


Fig 3: Comparation Between the gender

In conclusion, the system performed efficiently with high accuracy, minimal error rate, and built .It provides quick and reliable heart disease predictions, healthcare clinical decision support.

VI. CONCLUSION

In this project a heart disease prediction was done using ML techniques, particularly Arbitrary Forest and Decision Tree Classifier. The main goal for this particular project was to create a model that was both reliable and accurate in order to assist in the early deterministic process of heart disease based on clinical data. Our approach consisted of taking existing patient records in a CSV format, and using the classification algorithms to analyse them. The result of this process was that the model produced a prediction outcome at the rate of 96% accuracy.

This project served as an example of how machine learning, as well as explained AI, can contribute to the health sector of the world by offering latent information and intervention, while not getting rid of the clinical skills of physicians. The applications justify changes to the diagnostic process which can reduce possibilities of human errors, assist in defining clarity of information with medical professionals determining what is known and informing them of what is suspected. The continuous improvement of assisting health practitioners improves adherence to explanatory metrics, and using performance metrics and understanding visual charts which develop improved understanding to encourage engagement of both technical and non-technical users.

In conclusion, the model has shown that has great potential in being accurate, flexible and scalable, which can be integrated into healthcare systems to be used as a support in the preventive care setting and can assist in early identification of individuals at risk. Furthermore, the next steps in this model may entail implementing more complex datasets at operational levels and bringing to fruition the desire and potentially formalizing the system and advancing it into a complete medical diagnostic tool.

VII.REFERENCES

[1] Kaan Uyar and Ahmet İlhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks " in B.V ICTASC, Elsevier, pp.

[2] Ashish Chhabbi, Lakhan Ahuja, Sahil Ahir and Y. K. Sharma, "Heart Disease Prediction Using Data Mining Techniques ", © IJRAT Special Issue National Conference "NCPC-2016", pp. 104–106, 19 March 2016.

[3] Berry JD, Lloyd-Jones DM, Garside DB, Framingham risk score and prediction of coronary heart disease death in young men. Am Heart J. 2007 ; 154 (1): 80–6.

[4] Theresa Princy and R. J. Thomas, "Human Heart Disease Prediction System using Data Mining Techniques ", © IEEE ICCPCT, 2016.

[5] Kaur H Beant and Williamjeet Singh, "Review on Heart Disease Prediction System using Data Mining Techniques ", © IJRITCC, vol. 2, no. 10, pp. 3003–08, 2014.

[6] Kirmani, M.M., Ansarullah, S.I. : Prediction of heart disease using decision tree a data mining technique. IJCSN Int. J. Comput. Sci. Netw. 5 (6), 885–892 (2016).

[7] Salam Ismaeel, Ali Miri, "Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis ", IEEE Canada International Humanitarian Technology.

[8] Tahira Mahboob, Rida Irfan and Bazelah Ghaffar "Evaluating ensemble prediction of coronary heart disease using receiver operating characteristics " © 2017 IEEE.

[9] Ammar Asjad Raja, Irfan ul-Haq, Madiha Guftar Tamim Ahmed Khan "Intelligence syncope Disease Prediction Framework using DM-techniques " FTC 2016 –Future Technologies Conference 2016.

[10] M.A. Jabbar, B.L. Deekshatulu, and Priti Chandra, "Intelligent heart disease prediction system using random forest and evolutionary approach ", Journal of Network and Innovative Computing, Vol. 4, pp. 174–184, 2016.

[11] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," Machine Learning, (1997).

[12] Ayon Dey, Jyoti Singh, N. Singh "Analysis of supervised machine learning algorithms for heart disease prediction " International journal of computer applications, published in 2006.

[13] Q. K. Al-Shayea, "Artificial neural networks in medical diagnosis," International Journal of Computer Science Issues, published in 2011.

[14] 1.World Health Organization . World Health Statistics 2021. World Health Organization; Geneva, Switzerland: 2021.

[15] Iswisi A.F.A., Karan O., Rahebi J. Diagnosis of Multiple Sclerosis Disease in Brain Magnetic Resonance Imaging Based on the Harris Hawks Optimization Algorithm. BioMed Res. Int. 2021;2021:3248834.

[16] 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; pp. 1–5.

[17] Ternacle J., Côté N., Krapf L., Nguyen A., Clavel M.-A., Pibarot P. Chronic kidney disease and the pathophysiology of valvular heart disease. Can. J. Cardiol. 2019;35:1195–1207.

[18] House A.A., Wanner C., Sarnak M.J., Piña I.L., McIntyre C.W., Komenda P., Kasiske B.L., Deswal A., DeFilippi C.R., Cleland J.G.F. Heart failure in chronic kidney disease: Conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Kidney Int. 2019;95:1304–1317.

[19] Nguyen T., Wang Z.A. Cardiovascular screening and early detection of heart disease in adults with chronic kidney disease. J. Nurse Pract. 2019;15:34–40. doi: 10.1016/j.nurpra.2018.08.004.

[20] Liu R., Ren C., Fu M., Chu Z., Guo J. Platelet Detection Based on Improved YOLO_v3. Cyborg Bionic Syst. 2022;2022:9780569.

[21] 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.

[22] Rahebi J. Fishier mantis optimiser: A swarm intelligence algorithm for clustering images of COVID-19 pandemic. Int. J. Nanotechnol. 2023;20:25–49. doi: 10.1504/IJNT.2023.131111.

[23] Al Shalchi N.F.A., Rahebi J. Human retinal optic disc detection with grasshopper optimization algorithm. Multimed. Tools Appl. 2022;81:24937–24955.

[24] Sameni R., Clifford G.D. A review of fetal ECG signal processing; issues and promising directions. Open Pacing Electrophysiol. Ther. J.

2010;3:4–20. doi: 10.2174/1876536X01003010004.

[25] Al-Safi H., Munilla J., Rahebi J. Patient privacy in smart cities by blockchain technology and feature selection with Harris Hawks Optimization (HHO) algorithm and machine learning. Multimed. Tools Appl. 2022;81:8719–8743. doi: 10.1007/s11042-022-12164-z.