RESEARCH ARTICLE OPEN ACCESS

Heart Disease prediction using Random Forest Algorithm

Yash Soni, Akhilesh A Waoo
Department of Computer Science and Engineering
A.K.S university,Satna(M.P),India

1. Abstract

Heart disease is one of the major health issues globally, for which proper prediction techniques are recommended for proper intervention. This paper aims at optimizing the prediction accuracy of heart disease using various machine learning models. Advanced data-mining techniques are essential here because the researchers are combining electronic health records and wearable medical devices to capture continuous data.The methodology involves deep data preprocessing techniques, missing value handling, and feature scaling techniques. The models provided are fine-tuned by experimentation and hyperparameter tuning for superior predictive performance. The research paper is focused on predicting which patients are likely to suffer from heart disease based on various medical attributes. A heart disease prediction system is prepared using some machine learning algorithms, such as logistic regression, to predict and classify patients with heart disease. The proposed model depicts good accuracy against previous classifiers, thereby alleviating the pressure on the probability of correctly identifying heart diseases. It enhances medical care and decreases costs, thereby providing some valuable knowledge in predicting patients having heart disease.

Keywords: Heart disease, machine learning, Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF),Naive Bayes(NB).

2. Introduction

The heart is the most essential organ in the human body. Without a heart, the brain and countless other organs will cease working, and the existence will die in a few twinkles. Due to changes in life, work pressure, and bad dieting, many heart-related issues are on the rise. Heart disease is one of the leading causes of death and a serious global health issue. Grounded on the World Health Organization (WHO), an estimate states that CVDs account for 31% of all deaths worldwide, with 17.9 million deaths annually (1). It encompasses different heart conditions that need to be linked before in life to forestall and treat them effectively, including heart failure and hypertension. Individual threat biographies could not be appropriately captured by the traditional form of threat assessment that is used with variables similar to blood pressure and cholesterol situations. The ability to predict and treat cardiac complaints, especially for those with pre-existing conditions, has been underlined by the COVID-19 preface. In these lines of discussion, it is proved and shown that heart conditions and COVID-19 issues go hand in hand, hence being an intimidating concern that can develop more precise models and strategies for forestalling.

This review has touched on the types of cardiac ails, symptoms, threat factors like age and gender, and adjustable variables like food and smoking. The most integral bodily organ, the heart, circulates blood throughout all passageways of the body; indeed, it is practically the center of every segment of the body. There is a commonly used phrase that refers to heart and blood vessel disease: cardiovascular disease, or CVD.

ML has been shown to be a significant tool for heart disease prediction and management using complex algorithms to analyze complicated data and the choice of high-risk factors. The studies focused on diverse techniques of ML, and all of them present possible applications in the early diagnosis and improvement of the patient's outcome. In fact, the ML models may offer comprehensive information about cardiovascular health through the introduction of a vast number of clinical features, such as age, blood pressure, cholesterol, and lifestyle factors. Future developments in this area promise to open the door to synergy between machine learning techniques and clinical expertise in the transformation of prospects of

managing cardiovascular health. Such an outlook may therefore represent opportunities related to timely intervention and patient-tailored treatment plans that can result in better health outcomes for at-risk patients. Application of a few high-capacity ML algorithms such as LR, RF, SVM, and NB in order to classify heart disease. Several of these algorithms differ in terms of their strengths and are used to analyze a cross-section of the various attributes affecting patients, and it thus identifies subtlety-related correlations between a set of factors that could lead to heart disease.

The research studies discuss the various variables that constituted a patient's demographic, medical history, and lifestyle issues and predict cardiac disease using machine learning algorithms to better enhance outcomes of the patients. This study, hence, aims to assist health professionals in making the right decisions so they can act effectively and promptly.

Types of Heart Disease:

Heart disease occurs in various forms, some of which are summarized below:

Heart Diseases	Description
Stroke	Interruption of blood supply occur damage to the brain.
Peripheral artery disease	The circulatory condition is the narrowed blood vessels, which reduce the flow of blood in the limbs.
High-Blood pressure	It has a situation where the blood's force on the arterial walls is very high.
Coronary artery disease	The illness can develop in the blood or damage the heart's blood arteries.
Congenital heart disease	The heart`s abnormality which develops before birth.
Cardiac arrest	Suddenly, respiration, consciousness, and heart function are all abruptly lost.
Congestive heart failure	The heart does not pump blood as well as it should,it is the condition of chronic.

Risks factor

Important Risk Elements for Heart Disease are:

Threats	Details
Smoking	Increases the threat of heart complaint significantly; damages blood vessels and raises blood pressure.
High-blood Pressure	Puts redundant strain on the heart and highways, leading to increased threat of heart attack and stroke.
High Cholesterol	Elevated LDL cholesterol contributes to shrine make up in highways, adding heart complaint threat.
Diabetes	High blood sugar situations can damage blood vessels and lead to atherosclerosis, adding heart complaint threat.
Rotundity	Redundant body weight is linked to high blood pressure, diabetes, and high cholesterol situations.
Physically Inactivity	Sedentary life contributes to rotundity and other cardiovascular threat factors.

Unhealthy Diet	Diets grandly in impregnated fats, trans fats, and sugars increase the threat of heart complaint.
Habitual Stress	Long- term stress can lead to unhealthy habits(e.g., smoking, poor diet) and directly affect heart health.
Age	Threat increases with age; men are at advanced threat earlier than women.
Family History	inheritable predilection; having a first-degree relative with heart complaint increases individual threat.
Gender	Men generally have an advanced threat at a youngish age; women's threat increases post-menopause.
Race	Some races and ethnic groups have an increased prevalence of heart disease because of both genetic factors and socio-economic conditions, as is the case with African Americans and Native Americans.

Machine learning algorithm for prediction of heart diseases

Machine learning (ML) algorithms have greatly facilitated the vaticination of heart complaints, offering health care professionals important tools for early opinion and intervention. Several colors of algorithms have been used in this sphere. Each has its uniqueness and can be more delicate compared to others. This section will discuss several famous ML algorithms that are nowadays used for heart complaint vaticination, pressing their performance based on the latest studies that appeared.

Naive Bayes (NB)

Naive Bayes is a class of probabilistic bracket algorithms embedded in the operation of Bayes' theorem, therefore making a strong independence supposition between features. It is particularly useful for bracket problems wherein input variables are categorical. Indeed, it can easily handle veritably large datasets containing vast numbers of categorical variables, in fact those that occur most frequently in heart complaint vaticination scripts based on gender, type of casket pain, and other clinical pointers. It is very simple and fast and, therefore, nearly ideal for large-scale operations. Moreover, Naive Bayes requires relatively small amounts of training data to bound the parameters necessary for bracket. Although the major limitation of the independence assumption is that it fails to handle dependencies between correlated features, Naive Bayes surprisingly has shown emotional performance in colorful practical operations. Naive Bayes attained a delicacy around 85 in relating individualities at threat in the realm of heart complaint vaticination.

Naive Bayes has an interpretable benefit since this model can explain very clearly the exact donation of each variable to the liability of a heart complaint by the healthcare professional. This interpretability helps the clinicians to become strong enough to offer their informed views and construct proper treatment strategies. In a nutshell, Naive Bayes is a very informative tool used for fast, rapid-fire predictions and new exploratory analysis for medical datasets. Therefore, it is of extreme value to healthcare.

Support vector machine (SVM)

SVM is one algorithm that has been used intensively for the purposes of retrogression and bracketing. It is supposed to compute the best hyperplane in a high-dimensional space that will adequately classify the points of data into different classes such that the maximum distance is there between the hyperplane and the nearest data points from each class. One of the crucial benefits SVM possesses is that it can fluently reuse data in any number of confines, which can be veritably useful for medical operations where numerous features will presumably be involved. Also, SVM uses kernel functions to map data into high-dimensional spaces, so it can use veritably subtle, non-linear connections that a more introductory model might miss. In the heart complaint soothsaying environment, SVM has shown emotional perceptivity

rates numerous times that surpass 88.52. Its ability to avoid overfitting, especially when dealing with a high-dimensional dataset, justifies its in-fashion ability among researchers. However, performance may be determined by the choice of hyperparameters and kernel functions, thus challenging proper tuning for optimal results. Despite these considerations above in the discussion, SVM is still one of the most highly advocated algorithms in healthcare analytics because of its effectiveness in classifying complex patterns in patient data and because it can, if applied in the right way, give valuable perceptivity into trouble factors related to heart complaints.

Random Forest (RF)

Random Forest is also known as an adaptive ensemble learning method. It uses the combined knowledge of several decision trees to produce very accurate predictions. Many decision trees are grown in training, each trained on a different subset of characteristics and training data. This helps reduce the chance of overfitting but enhances overall prediction performance by promoting variety among the trees. It has the wonderful benefit of not requiring a lot of preprocessing because it can function on a variety of numerical and categorical variables without it. Plus, Random Forest brings very interesting feature-important insights that enable practitioners to know the major risk factors for heart disease. Heart disease prediction: From all the above prediction data for heart disease, it can be seen that Random Forest performs outstandingly and achieves accuracy as high as 91.80%. This is an ensemble technique that does not allow overfitting of a single decision tree to special patterns in the data and leads to better and more reliable predictions at the end. On top of that, Random Forest can be very robust toward missing values, holding accuracy cases as well, in a lot of cases even when a great percentage of the data is missing. Beyond the prediction, there is adaptability in the random forest that is highly suited for diseases such as heart diseases. Very high accuracy, even considering the fact that the model isn't noisy, is offered alongside great interpretability; it is well regarded as one of the best algorithms used for predictive modeling in many clinical settings.

Logistic Regression (LR)

Logistic regression (LR) is one of the most important statistical models for classification and a super-efficient technique in classifying tasks where results can be represented as being present or absent. This function predicts the probability that an input belongs to some specific class by using the logistic function on a linear combination of meaningful features of patient data. One of the major advantages of LR is that it is interpretable; thus, it is always clear how the various risk factors—attributed, for example, to age, cholesterol levels, or blood pressure—function to determine the level of cardiovascular disease risk. Such transparency would support informed clinical decision-making with regard to the patient's care and treatment strategy.It usually comes in the range between 85-89%, depending on the chosen dataset and type of feature extraction technique. Though less precise than some of the more complex algorithms, like random forests or SVMs, at detecting even the slightest relationships, LR is still a viable approach in medical studies due to its simplicity and reliability. Adding L1 or L2 regularization also makes LR very effective against overfitting in high-dimensional datasets. There could be several applications to predict risk in heart disease. Adaptability and strong performance of logistic regression have assured direct interpretability. This tool would suit better healthcare through informed intervention and adjustment of lifestyle. Further updating of research will continue to improve with advanced techniques of machine learning, which may ultimately make prediction models based on heart disease more effective.

Accuracy Table:

Algorithm	Accuracy (%)
Naive Bayes	85
Support vector machine	88.52
Random Forest	91.80
Logistic Regression	85 - 89

Advantages and Disadvantages

Naive Bayes (NB)

Benefits

- ✓ Achieves good delicacy (85-90) in predicting heart complaints predicated on trouble factors.
- ✓ It's applicable for real-time conditioning since it's quick and easy to apply.

Limitations

- ✓ Assumes feature independence, which could not be the case for data on cardiac complaints.
- ✓ Sensitive to imbalanced datasets, potentially turning prognostications towards the maturity class.

Support Vector Machines (SVM)

Benefits

- ✓ It works well in high-dimensional spaces and is applicable for datasets with a large number of characteristics.
- ✓ Suitable to employ kernel functions to manage non-linear connections .

Limitations.

- ✓ computationally ferocious, which results in prolonged training durations, particularly when dealing with huge datasets.
- ✓ The kernel and hyperparameter selection, which might be delicate to acclimate, have a significant impact on performance.

Logistic Regression (LR)

Benefits:

- ✓ Provides interpretable results, allowing clinicians to understand the influence of each predictor on heart complaint trouble.
- ✓ Effective for lower datasets and performs well when the relationship is roughly direct.

Limitations:

- ✓ Assumes a direct correlation between the log chances of the outgrowth and independent factors, which could miss intricate patterns.
- ✓ Limited to double issues unless extended, complicating multi-class prognostications.

Random Forest (RF)

Benefits

- ✓ Quite resistant to overfitting in case enough trees are utilized and proper parameters are fine-tuned.
- ✓ Can handle a lot of features well and can provide some form of insight into point significance in terms of heart complaints prediction.

Limitations

- ✓ It will most likely overfit the training data if not well-tuned, which degrades generalization to new data.
- ✓ Less interpretable than other simpler models such as logistic regression, which complicates clinicians' decisions in deciding perceptivity from predictions.

Discussion

Random Forest is the algorithm that has a fair deal of promise for predictive analytics in heart disease. Mainly, its accuracy is high, and results are powerful from different datasets. It merges multiple decision trees and captures all the patterns that decision trees can, therefore reducing all the chances of overfitting. Results comparable with Random Forest, especially after some proper selection techniques towards the features. The whole generalizing ability makes it a strong tool for predicting heart disease since both

Naive Bayes and logistic regression are not quite accurate in models that predict heart diseases and are sensitive to feature independence and linearity.

Conclusion

This paper discusses the potential capabilities of NB, LR, SVM, and RF algorithms in terms of improving the accuracy of heart disease prediction. Each algorithm has its strengths: it is favored by the probabilistic predictability of NB, the interpretability of LR, the handling capability with high-dimensional data of SVM, and the robustness using ensemble learning of RF. Thus, feature selection and parameter tuning are significant entities that this review emphasizes. Further improvement in reliability can be achieved by using the ensemble approach. Future work should integrate diverse datasets and allow more experiments to enhance generalization and avoid data imbalance problems.

References

1) AN Repaka, SD Ravikanti and RG. Franklin, "Design and implementing heart disease prediction using naives Bayesian", International conference on trends in electronics and informatics (ICOEI), pp. 292-297.
2) VV Ramalingam, A Dandapath and MK. Raja, "Heart disease prediction using machine learning techniques: a survey", International Journal of Engineering & Technology, vol. 7, no. 2.8, pp. 684-7.S
3) Akhilesh A Waoo, Sanjana Chaudhari, Mr Chandra Shekhar Gautam, "Optimizing Heart Disease Prediction Accuracy using Machine Learning Models",International Journal of All Research Education and Scientific Methods (IJARESM).
4) E. I. Elsedimy, S. M. M. AboHashish, and F. Algarni, "New cardiovascular disease prediction approach using support vector machine and quantum-behaved particle swarm optimization," Multimedia Tools and Applications, 2023.
5) Aram S, Sadeghian R, Abdellatif I, et al. Diagnosing Heart Disease Types from Chest X-Rays Using a Deep Learning Approach. 2019 International Conference on Computational Science and Computational Intelligence (CSCI). Las Vegas, NV, USA: IEEE; 2020.
6) Jain AK, Kumar K, Tiwari RG, et al. Machine Learning-Based Detection of Cardiovascular Disease using Classification and Feature Selection. 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT). Bhopal: IEEE; 2023.
7) Bani Hani SH, Ahmad MM. Machine-learning Algorithms for Ischemic Heart Disease Prediction: A Systematic Review. Curr Cardiol Rev 2023;19:e090622205797. [Crossref] [PubMed]
8) Cardiovascular Disease (CVD)-World Heart Federation (accessed Jan 11, 2023). Available online: https://world-heart-federation.org/what-is-cvd/
9) Das CR, Das CM, Hossain MA, et al. Heart Disease Detection Using ML. 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC). Las Vegas, NV, USA: IEEE; 2023.
10) Jahed R, Asser O, Al-Mousa A. Using Personal Key Indicators and Machine Learning-based Classifiers for the Prediction of Heart Disease. 2023 International Conference on Smart Computing and Application (ICSCA). Hail: IEEE; 2023.
11) Chopra S, Karla N, Rani R. Identification of Cardiovascular Disease using Machine Learning and Ensemble Learning. 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA). Uttarakhand: IEEE; 2023.
12) Gola KK, Arya S. Satin Bowerbird Optimization-Based Classification Model for Heart Disease Prediction Using Deep Learning in E-Healthcare. 2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW). Bangalore: IEEE; 2023.
13) R Hasan ,Comparative analysis of machine learning algorithms for heart disease prediction., ITM Web of Conferences, 2021 - itm-conferences.org
14) Nasser AA, Akhloufi MA,Deep Learning Methods for Chest Disease Detection Using Radiography Images. SN Comput Sci 2023;4:388.

15) Shukla A, Khan IR, Sharma V, et al. A Novel Prediction System to Diagnose Heart Disease. 2023 International Conference on Inventive Computation Technologies (ICICT). Lalitpur: IEEE; 2023.
16) M. Yousef and K. Batiha, "Heart Disease Prediction Model Using Naïve Bayes Algorithm and Machine Learning Techniques," International Journal of Engineering & Technology, vol. 10, no. 1, 2021.
17) S. Ambesange, A. Vijayalaxmi, S. Sridevi, and B. Yashoda, "Multiple heart diseases prediction using logistic regression with ensemble and hyper parameter tuning techniques," in 2020 fourth world conference on smart trends in systems, security and sustainability (WorldS4), London, UK, 2020, pp. 827–832: IEEE.
18) A. G, B. Ganesh, A. Ganesh, C. Srinivas, Dhanraj, and K. Mensinkal, "Logistic regression technique for prediction of cardiovascular disease," Global Transitions Proceedings, vol. 3, no. 1, pp. 127–130, 2022.https://doi.org/10.1016/j.gltp.2022.04.008.
19) A. A. Muideen, C. K. M. Lee, J. Chan, B. Pang, and H. Alaka, "Broad Embedded Logistic Regression Classifier for Prediction of Air Pressure Systems Failure," Mathematics, vol. 11, no. 4, 2023.https://doi.org/10.3390/math11041014.
20) V. Sai Krishna Reddy, P. Meghana, N. V. Subba Reddy, and B. Ashwath Rao, "Prediction on Cardiovascular disease using Decision tree and Naïve Bayes classifiers," Journal of Physics: Conference Series, vol. 2161, no. 1, 2022.
21) P. Deepika and S. Sasikala, "Enhanced model for prediction and classification of cardiovascular disease using decision tree with particle swarm optimization," in 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2020, pp. 1068–1072: IEEE.
22) M. Arifuzzaman, M. R. Hasan, T. J. Toma, S. B. Hassan, and A. K. Paul, "An Advanced Decision Tree-Based Deep Neural Network in Nonlinear Data Classification," Technologies, vol. 11, no. 1, 2023.https://doi.org/10.3390/technologies11010024.
23) S. Saeedbakhsh, M. Sattari, M. Mohammadi, J. Najafian, and F. Mohammadi, "Diagnosis of Coronary Artery Disease based on Machine Learning algorithms Support Vector Machine, Artificial Neural Network, and Random Forest," Advanced Biomedical Research, vol. 12, p. 51, 2023.https://doi.org/10.4103/abr.abr_383_21.
24) S. B. Shuvo, S. N. Ali, S. I. Swapnil, M. S. Al-Rakhami, and A. Gumaei, "CardioXNet: A Novel Lightweight Deep Learning Framework for Cardiovascular Disease Classification Using Heart Sound Recordings," IEEE Access, vol. 9, pp. 36955–36967, 2021.https://doi.org/10.1109/ACCESS.2021.3063129.
25) A. Aldelemy and R. A. Abd-Alhameed, "Binary Classification of Customer's Online Purchasing Behavior Using Machine Learning," Journal of Techniques, vol. 5, no. 2, pp. 163–186, 2023.https://doi.org/10.51173/jt.v5i2.1226.
26) X. Zhou, X. Wang, X. Li, Y. Zhang, Y. Liu, J. Wang, S. Chen, Y. Wu, B. Du, X. Wang, X. Sun, and K. Sun, "A novel 1-D densely connected feature selection convolutional neural network for heart sounds classification," Annals of Translational Medicine, vol. 9, no. 24, p. 1752, Dec 2021.https://doi.org/10.21037/atm-21-4962.
27) M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," Healthcare Analytics, vol. 3, 2023.https://doi.org/10.1016/j.health.2022.100130.
28) M. B. Er, "Heart sounds classification using convolutional neural network with 1D-local binary pattern and 1D-local ternary pattern features," Applied Acoustics, vol. 180, 2021.https://doi.org/10.1016/j.apacoust.2021.108152.
29) V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," The Journal of Supercomputing, vol. 77, no. 5, pp. 5198–5219, 2020.https://doi.org/10.1007/s11227-020-03481-x.
30) Ali L, Niamat A, Khan JA, et al. An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure. IEEE Access 2019;7:54007-14.

