Article

A Hybrid Approach for Cardiovascular Disease Diagnosis: J48 Classifier

Suma T , Preethi S , Allwin Pon Suthers A , Saranraj I , Shubham Malhotra * , Meenu Gupta

Posted Date: 8 May 2025

doi: 10.20944/preprints202505.0530.v1

Keywords: CVD; Machine Learning; J48 algorithm; Prediction

Article

A Hybrid Approach for Cardiovascular Disease Diagnosis: J48 Classifier

Suma T [1], Preethi S [2], Allwin Pon Suthers A [3], Saranraj I [4], Shubham Malhotra [5,*] and Meenu Gupta [6]

[1] Senior Assistant Professor, Department of Mathematics, New Horizon College of Engineering, Karnataka, India

[2] Professor, Department of ISE, Cambridge Institute of Technology, Bengaluru, India

[3] Assistant Professor, Department of Mechanical Engineering, Gojan School of Business and Technology, Chennai, India

[4] Associate Professor, Department of Mechanical Engineering, Mahendra College of Engineering, Salem, India

[5] Research Scholar, Department of Software Engineering, Rochester Institute of Technology, Rochester, NY, 14623

[6] Professor, Department of Computer Science and Engineering, Chandigarh University, Punjab, India

* Correspondence: shubham.malhotra28@gmail.com

Abstract: In today's world, we frequently hear about the heart disease issues such as the cardiovascular disease (CVD), especially in a younger generations due new culture, new virus pandemic, and food quality/habits, which has a high possibility of the leading to death. CVD is also the cause of the high global mortality rate. One of the biggest challenges is detecting cardiovascular illnesses with routine clinical data analysis, as their early detection can save countless lives. Machine learning (ML) algorithms allow for precise forecasting, intelligent decision-making, and exact predictions for data analysis. This research uses a variety of the factors that individual person has presented to a assess if a cardiac arrest has occurred. The employment of numerous ML algorithms to determine to predict CVD is common nowadays, but increasing prediction accuracy is challenge. This work tackles these issues by presenting fresh, ethically obtained using CVD dataset that includes thorough information on risk variables, examination methods, and symptoms. Through utilizing sophisticated ML methods such as SVM, Naive Bayes, LR, KNN, and J48, we were able to attain impressive testing accuracy of approximately 98.54% with J48. The suggested method is the J48 classifier, which uses a ML techniques to integrate these hybrid models and datasets. It gives an adequate diagnosis promptly and fashion designers nutritional advice each person. This study proposes a new way to find the scalable and reliable signs of heart disease by using structured datasets in tandem with sophisticated machine-learning techniques. This might result in greater outcomes for patients and less mortality.

Keywords: CVD; machine learning; J48 algorithm; prediction

Introduction

There has been an increasing appetite for more effective diagnosis and then forecasting techniques such as CVD is an epidemic of mortality globally. By leveraging ML and accomplishing rigorous evaluation of data, anticipated therapies are becoming increasingly influential in the healthcare market. They enable an amelioration in the burden of CVD and aversion of superfluous constraints by delivering correct forecasts for implementation [1]. CVD is a huge acute wellness issue that impacts a substantial percentage of individuals, among other disorders. It is vital to acquire a diagnosis soon after since a faulty diagnosis might lead to catastrophic issues. A pioneering software

doi:10.20944/preprints202505.0530.v1

application employing computer technology and ML methodologies is being created to aid medical practitioners in the early diagnosis of cardiac issues [2].

CVD acute wellness issue that impacts death of worldwide, taking a significant number of lives each year. The incapacity competently flow an enough amount of blood to different organs is the fundamental pathophysiology of cardiac disorders. This illness is one of most deadly and serious chronic conditions in the the environment, and it conveys a major danger to life. CVD interferes with the regular flow of blood by damaging the heart or blood vessels, which hinders the healthy operation of vital bodily organs [3].

The early detection heart disease attempts to decrease death statistics by examining someone's existing cardiac status and correlated risk indicators. Conventional techniques usually rely on pre-processed information with restrained traits such as making the diagnosis and categorization of heart illness a tough issue. A full treatment important because of condition's complexity, encompassing the wellness in general assessments, systemic symptom evaluations, risk evaluations of factors, and detailed testing for diagnosis. On top of that, the restricted usefulness of common datasets in actual-life situations further hampers effective detection and classification. Furthermore, the constrained use of standard datasets in actual scenarios worsens the difficulty of exact identification and categorization [4–6].

This paper addresses these challenges by giving a well-gathered and arranged dataset for illness categorization. The collection, updated by a constant data refinement, comprises more than a 45,000 accurate patient records acquired from government hospitals, diagnostic centers, and then internet libraries. This rich and diverse dataset intends to a increase model training, permitting more exact predictions and boosting diagnosis and categorization of heart illness [7–9].

Through a integration of the cardiac disease diagnosis, classification, and prognosis, the paper's innovative paradigm allows for personalized disease classification and prediction. Complex model utilizing classifiers from the J48, SVM, KNN, Naive Bayes, and LR, together with the developing specialized datasets like CVD, demonstrates a possibility of accurate predictions.

To summarize our efforts in this paper:

- We gathered CVD datasets from Kaggle website and categorized them into various classifications in this study.
- We compared a effectiveness of our advised strategies.
- To compare proposed models to those that currently exist.

Literature Review

Researchers have proposed a several kinds of machine learning-based algorithms diagnosis techniques in the literature to diagnose CVD. To illustrate the significance of this study, a few current ML-based diagnostic methods are presented. Arunima Jaiswal et al. [10] aimed to precisely and early predict cardiac disease. They use deep learning (DL) to apply the different features through a digital patient record assessment. The Convolutional Neural Networks (CNN) method produces a excellent results, that including accuracy, and has lot of potential as a diagnostic tool that can improve diagnostic efficiency and accuracy in healthcare settings with limited resources.

Accurately predicting cardiac disease is the aim of Jian Ping Li's research [11]. Use (ML), employing a range of supervised choices and highlights of significance, to evaluate digital health records. The FCMIM algorithm has produced positive outcomes, including accuracy, and they have developed a successful ML-based diagnostic system for the diagnosis of cardiac disease. According to United Kingdom's Sustainable Development Goal 3 (SDG 3), everyone would be healthy and content. This study claims that CVD can prevented by the monitoring patients' symptoms and performing physical examinations. Some hazards for leads to coronary arrest include tobacco use, elderly status, and heart disease in previous generations, high-fat stages, ranging a lack of consistent physical activity, high blood pressure, obesity, diabetes, and stress [12]. The effectiveness of automated diagnostic system was examined utilizing precision in classification, specificity, as well as sensitivity on an Indian dataset used to diagnose cardiac diseases. The results showed that for medical illness diagnostic applications, using the Sequential Minimization Optimization (SMO) technique for learning in SVM performed more effectively than other strategies [13].

An a synopsis of the essential risk markers for CVD was presented. The research identified ten significant hazards: thrombosis/smoking, kidney dysfunction, genetics/familial high cholesterol levels, demographic data (including aging, race/ethnicity, and gender differences), obesity, high blood pressure, high glucose levels, dyslipidemia, lack of regular exercise, and biological aspects. Patients with CVD may be have several risk factors, necessitating a multimodal preventative strategy. To the lower risk of CVD, this paper also emphasizes the significance of intermittent fasting, nutritional restriction, and physical activity. When every factor is examined, the research underlines how crucial it is to evaluate and treat diverse risk factors connected to CVD to enhance prevention tactics [14]. Abdallah Abdellatif et al. [15] focused on a initiatives to improve the RF classifier's training time and accuracy. They are based on the following: separate segregating of training datasets; generating RF base decision trees using split measures or multiple feature for evaluation and using proportional voting instead of a large proportion voting; generating the most widely distributed classifiers using a wide variety of bootstrap datasets; and using dynamic programming method to identify the greatest portion of Random Forest (RF).

The significance of ML and DL in the diagnosis of CVD is emphasized by the literature study. CNN and SVM are two of the many ML-based techniques These were offered for the early forecasting of CVD. Feature selection techniques and assessments of digitized medical information have improved the effectiveness of ML approaches. ML algorithms can achieve a diagnostic accuracy of up to 89% in CVD. We tried hybrid algorithm techniques to achieve higher accuracy in our research.

Proposed Methodology

A carry out illustration of the proposed researching framework's concept can be provided in "Figure 1". The illustration below represents a complete understanding of the structure as well as the elements of the proposed framework. The corresponding diagram offers a complete explanation of various components as well as the structure of the recommended framework.

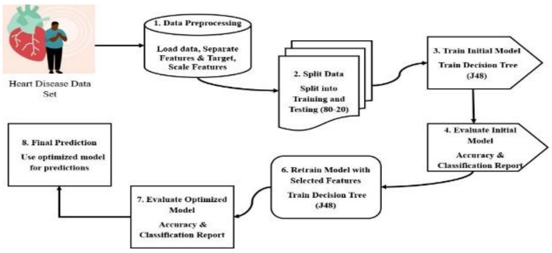
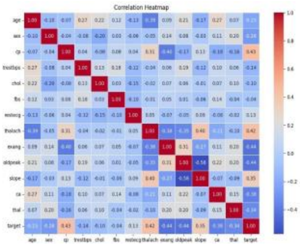

Heart Disease Data Set
1. Data Preprocessing
Load data, Separate Features & Target, Scale Features
2. Split Data
Split into Training and Testing (80-20)
3. Train Initial Model
Train Decision Tree (J48)
4. Evaluate Initial Model
Accuracy & Classification Report
6. Retrain Model with Selected Features
Train Decision Tree (J48)
7. Evaluate Optimized Model
Accuracy & Classification Report
8. Final Prediction
Use optimized model for predictions

Figure 1. CVD Prediction Workflow Diagram.

CVD Dataset Collection

The dataset was gathered from hospitals, healthcare facilities, and reputable online sources. The binary data (1 for Yes, 0 for No) in "Figure 2" indicates symptoms such as dyspnea, raised cholesterol, increased blood sugar, chest discomfort, and health issues. The affected and non-affected individuals represented in "Figure 3". The Ethical Review Board of each institution provided legal authorization for the data retrieval.



doi:10.20944/preprints202505.0530.v1

Figure 2. Heatmap for CVD Dataset.

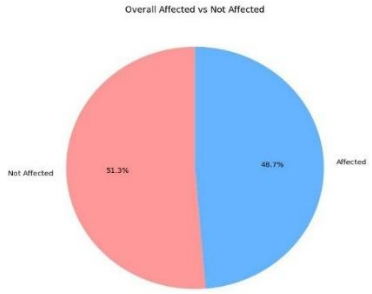


Figure 3. Classification between Overall CVD Affected VS. Non-Affected.

Feature Extraction

According to their diagnostic importance, 15 symptoms and risk factors were chosen as features represented in "Figure 4" and "Figure 5". In this hybrid model technique, CVD is successfully predicted utilizing J48 algorithm approach, attaining a greater accuracy than previous algorithms. This approach ensures exceptional accuracy by optimizing model performance and effectively managing complex connections in the data.

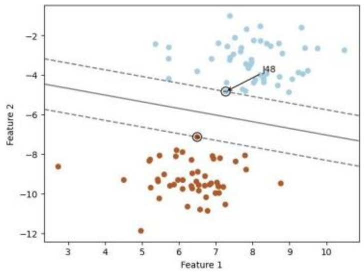


Figure 4. Scatter Plotting for J48 Classifier.

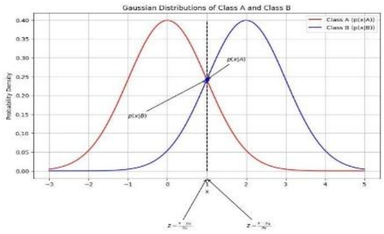


Figure 5. Distribution of Class A and Class B.

Model Selection

A variety of ML algorithms were used in our experimental study, including LR [2], SVM, Naive Bayes, KNN Classifier [7], and J48 Classifier.

<H3> ***Support Vector Machine (SVM)*:** To determine the classes in the provided information, it is also a classification algorithm. The approach is bound by data that has already been classified into two different categories. This constructed model is used to identify the class of fresh data points. It is used to achieve the widest possible boundaries between objects on a plot in addition to classifying them. The support vectors are used for the "equation (1)", where the weight vector, with the same dimensions as the input x, determines the hyperplane's position in the feature space, as well as the bias term shifts the hyperplane from the origin, providing flexibility in its position.

$W_0^T x + b_0 = 1 \; or \; W_0^T x + b_0 = \; -1$ ……….(1)

doi:10.20944/preprints202505.0530.v1

<H3> ***Logistic Regression (LR)*:** LR is a widely utilized supervised machine-learning method applicable to tasks involving regression as well as classification. The logistic regression technique predicts the labelling of classified data using probability. Binary classification probability measures are the basis of LR's learning and prediction technique. Binary classification is required for the class variables in logistic regression technique. With two distinct binary numbers, it is comparable to the target column in our study dataset. Patients in the dataset who are projected to have heart failure are represented by one, while those who have no probability of developing heart failure are represented by a zero. The sigmoid function "equation (2)" and "equation (3)" where q(y = 1∣x) is the likelihood that, given the predictor variable(s) x, the binary result variable y will take the value 1. Any real variable z can be mapped to the interval [0,1] using the logistic function e−z.

$q(y = 1|x) = \frac{1}{1+e^{-z}}$ ………………………...…..(2)

$z = \beta_0 + \beta_1 x_1 + \cdots + \beta_q \beta_q$…………………..….(3)

<H3> ***Naïve Bayes*:** The Naïve Bayes classifier presupposes the existence of an individual feature inside a class is independent of the presence of any other feature. Only the class feature is needed for this classifier, as the other class traits are independent. The specified "equation (4)" is utilized for the computation of this classifier, where it represents a likelihood of a class given the feature X.

$P(C_k|x) = \frac{P(C_k)*P(x|C_k)}{P(X)}$……………………….….(4)

<H3> ***K-Nearest Neighbor (KNN)*:** The K-Nearest Neighbor classification method identifies how a data point is categorized based on how far away it is from the nearest set of points. The Euclidean distance "equation (5)" is used to determine the distance between data points, where x and y stand for the points in n-space. The k-nearest destinations are picked depending on the distance. Data points that are categorized according to Euclidean distance, which is the smallest value between two or more of the closest groups points.

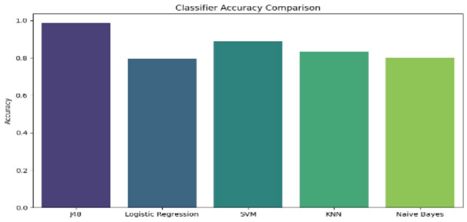
$d(x, y) = \sum_{k=1}^{n} (y_k - x_k)^2$…..…...…..……..….(5)

<H3> ***J48 Algorithm*:** The Naïve Bayes classifier presupposes the existence of an individual feature inside a class is independent of the presence of any other feature. Only the class feature is needed for this classifier, as the other class traits are independent. The specified "equation (4)" is utilized for the computation of this classifier, where it represents a likelihood of a class given the feature X.

$H(S) = -\sum_{i=1}^{k} p_i \cdot \log_2(p_i)$…...…………….(6)

Result and Discussion

This work's outcomes the experiment were obtained using a CVD dataset. The research targets examining the efficacy of several ML algorithms in categorizing and predicting outcomes within this sector. A hybrid model including several ML approaches was developed, and the success rates of these methods are displayed in "Figure 6". The ML algorithms observed for the study were SVM, Naïve Bayes, KNN, LR, and J48. The above methods were adjusted by applying fine-tuning, a contemporary metaheuristic optimization approach, to increase its efficiency. As the dataset size got higher, considerable disparities in accuracy arose between the algorithms. J48, a decision tree- based classifier. "Table 2" presents a complete study of accuracy rates reached by each approach, clearly indicating higher performance of J48. This work stresses potential of hybrid optimization techniques in boosting the reliability and accuracy of machine learning models, which are essential sectors such as CVD prediction.



doi:10.20944/preprints202505.0530.v1

Figure 6. Algorithm Comparison for CVD Prediction.

Table 2. Accuracy Differentiation Report.

Techniques	Accuracy (%)
Logistic Regression	0.79%
Naïve Bayes	0.80%
K-Nearest Neighbor	0.83%
Support Vector Machine	0.88%
J48 Classifier	**0.98%**

Performance Evaluation using Confusion Matrix

The discrepancy between the actual as well as predicted values of data is illustrated by the confusion matrix. As shown in "Figure 7", it is employed to evaluate how well our classification model performs in the ML. You may use the following "equation (7)" to evaluate accuracy. The confusion matrix's constituent parts are:

- True positive: Both the initial data values and the expected values were positive.
- False positive: It is fallacious to anticipate improvements when they were previously negative.
- False negative: When readings in the actual data remained positive, they were mistakenly projected to be negative.
- True negative: Both the starting data quantities and the predicted values were negative.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \ldots\ldots\ldots\ldots\ldots\ldots(7)$$

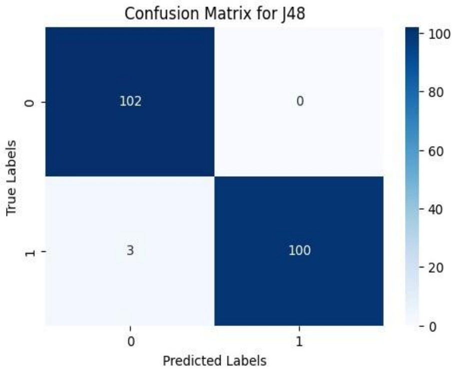


Figure 7. Confusion Matrix for J48 Algorithm.

doi:10.20944/preprints202505.0530.v1

Conclusion

The severity of CVD can recognized by utilizing a hybrid model of ML approaches. As proven by our experimental outcomes, the decision tree technique J48 has obtained greater accuracy (98%) in contrast to ML classifiers such as LR (79%), Naïve Bayes (80%), KNN (83%), SVM (88%), and others. As the dataset develops, the algorithm has done an adequate task of assessing the learned material and properly categorizing it. In this work, five algorithms are tailored, trained, and evaluated using five different models, including LR, KNN, SVM, Naive Bayes, and J48, utilizing and excluding the hyperparameter tweaking strategies. Their precision are evaluated with existing techniques. By changing the requirements of the J48 testing accuracy reached at 98%. J48 Classifier gives The superior precision in testing all of these five approaches at 98%. The J48 Classifier is optimal hyperparameter values for precision. In the future, the research will focus on employing different machine-learning algorithms as well as upgraded selecting features strategies to boost the precision and efficacy of CVD prediction. Optimization approaches will be used to boost the performance of prediction systems, while work will also extend to post- diagnostic features, including managing and treating cardiovascular disorders.

References

1. N. Yanes, L. Jamel, B. Alabdullah, M. Ezz, A. M. Mostafa, and H. Shabana, "Using Machine Learning for Detection and Prediction of Chronic Diseases," *IEEE Access*, p. 1, Jan. 2024, doi: 10.1109/access.2024.3494839.
2. M. Haque *et al.*, "Multi-class heart disease Detection, Classification, and Prediction using Machine Learning Models," *arXiv.org*, Dec. 06, 2024. https://arxiv.org/abs/2412.04792
3. T. Ullah *et al.*, "Machine Learning-Based Cardiovascular Disease Detection Using Optimal Feature Selection," *IEEE Access*, vol. 12, pp. 16431–16446, Jan. 2024, doi: 10.1109/access.2024.3359910.
4. M. Obayya, J. M. Alsamri, M. A. Al-Hagery, A. Mohammed, and M. A. Hamza, "Automated Cardiovascular Disease Diagnosis Using Honey Badger Optimization With Modified Deep Learning Model," *IEEE Access*, vol. 11, pp. 64272–64281, Jan. 2023, doi: 10.1109/access.2023.3286661.
5. M. S. A. Reshan, S. Amin, M. A. Zeb, A. Sulaiman, H. Alshahrani, and A. Shaikh, "A Robust Heart Disease Prediction System Using Hybrid Deep Neural Networks," *IEEE Access*, vol. 11, pp. 121574–121591, Jan. 2023, doi: 10.1109/access.2023.3328909.
6. J. J. Gabriel and L. J. Anbarasi, "Accurate Cardiovascular Disease Prediction: Leveraging Opt_hpLGBM with Dual-Tier Feature Selection," *IEEE Access*, p. 1, Jan. 2024, doi: 10.1109/access.2024.3470537.
7. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," *IEEE Access*, vol. 9, pp. 106575–106588, Jan. 2021, doi: 10.1109/access.2021.3098688.
8. Kumar, K. U. Singh, and M. Kumar, "A Clinical Data Analysis Based Diagnostic Systems for Heart Disease Prediction Using Ensemble Method," *Big Data Mining and Analytics*, vol. 6, no. 4, pp. 513–525, Aug. 2023, doi: 10.26599/bdma.2022.9020052.
9. M. Qadri, A. Raza, K. Munir, and M. S. Almutairi, "Effective Feature Engineering Technique for Heart Disease Prediction with Machine Learning," *IEEE Access*, vol. 11, pp. 56214–56224, Jan. 2023, doi: 10.1109/access.2023.3281484.
10. Jaiswal, M. Singh and N. Sachdeva, "Empirical Analysis of Heart Disease Prediction Using Deep Learning," 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2023, pp. 1-9, doi: 10.1109/ACCAI58221.2023.10201235.
11. J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," *IEEE Access*, vol. 8, pp. 107562–107582, Jan. 2020, doi: 10.1109/access.2020.3001149.
12. H. F. El-Sofany, "Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques," *IEEE Access*, vol. 12, pp. 106146–106160, Jan. 2024, doi: 10.1109/access.2024.3437181.

13. G. N. Ahmad, H. Fatima, S. Ullah, A. S. Saidi, and N. Imdadullah, "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV," *IEEE Access*, vol. 10, pp. 80151–80173, Jan. 2022, doi: 10.1109/access.2022.3165792.
14. S. Ghorashi *et al.*, "Leveraging Regression Analysis to Predict Overlapping Symptoms of Cardiovascular Diseases," *IEEE Access*, vol. 11, pp. 60254–60266, Jan. 2023, doi: 10.1109/access.2023.3286311.
15. Abdellatif, H. Abdellatef, J. Kanesan, C.-O. Chow, J. H. Chuah, and H. M. Gheni, "Improving the Heart Disease Detection and Patients' Survival Using Supervised Infinite Feature Selection and Improved Weighted Random Forest," *IEEE Access*, vol. 10, pp. 67363–67372, Jan. 2022, doi: 10.1109/access.2022.3185129.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.