STM JOURNALS

Journal of
Artificial Intelligence
Research & Advances

ISSN: 2395-6720
Volume 12, Issue 1, 2025
January–April
DOI Journal: 10.37591/JoAIRA

https://journals.stmjournals.com/joaira

Review JoAIRA

A Hybrid Machine Learning Approach for Cardiovascular Disease Prediction

K. Purushotam Naidu[1,*], V. Lakshmana Rao[2], Esha Thaniya Malla[3], Indu Kola[3], Renuka Sai Reddi[3], Bharathi Kolluru[3], Raj Tanuja Pentapati[3]

Abstract

Heart disease ranks among the top causes of death globally. Accurately predicting cardiovascular conditions has become a key challenge in the realm of clinical data analysis. It has been shown that machine learning is an effective means of assisting with predicting and decision-making based on the large volume of data produced by the medical industry. In this study, we describe a unique approach that increases the prediction accuracy of heart-related conditions by using machine learning approaches to identify important attributes. The prediction model is displayed using a variety of feature combinations and widely used categorization techniques. With an accuracy level of 85.7%, we attain an enhanced performance level with the Hybrid Random Forest with Logistic Model (HRFLM) prediction model for heart disease. This adds to the continuing conversation in healthcare analytics and lays a solid basis for clinical decision support's use of data-driven predictive models. In this case, machine learning is an effective way to solve the challenges associated with cardiovascular disease prediction. The acquired results highlight the potential of using cutting-edge analytics and novel predictive modelling tools to enhance patient outcomes and well-being.

Keywords: Clinical data analysis, hybrid model, risk assessment, heart disease prediction, cardiovascular disease, healthcare industry

INTRODUCTION

Identifying heart disease can be difficult because it is influenced by various risk factors, such as elevated blood pressure, high cholesterol levels, and diabetes. To address this issue, data mining approaches like neural networks have been employed. Various models such as random forest, k-nearest neighbor, logistic regression (LR), decision trees (DT), gradient boosting (GB), and hybrid models (RF and linear model) are utilized for classifying the severity of the condition. Data mining, along with medical science, allows for the identification of different metabolic syndromes associated with cardiac disease. Decision trees have proven effective in predicting events linked to heart disease. This study explores the relevance of machine learning in the diagnosis and early detection of heart-related ailments.

***Author for Correspondence**
K. Purushotam Naidu
E-mail: purushotam.k30@gmail.com

[1]Assistant Professor, Department of Computer Science and Engineering (AI&ML), Gayatri Vidya Parishad College of Engineering for Women, Visakhapatnam, Andhra Pradesh, India
[2]Assistant Professor, Department of Computer Science and Engineering, Gayatri Vidya Parishad College of Engineering for Women, Visakhapatnam, Andhra Pradesh, India
[3]Student, Department of Computer Science and Engineering, (AI&ML), Gayatri Vidya Parishad College of Engineering for Women, Visakhapatnam, Andhra Pradesh, India

Received Date: October 05, 2024
Accepted Date: December 05, 2024
Published Date: December 30, 2024

Citation: K. Purushotam Naidu, V. Lakshmana Rao, Esha Thaniya Malla, Indu Kola, Renuka Sai Reddi, Bharathi Kolluru, Raj Tanuja Pentapati. A Hybrid Machine Learning Approach for Cardiovascular Disease Prediction. Journal of Artificial Intelligence Research & Advances. 2025; 12(1): 69–75p.

The Importance of Early Detection

The heart, being the most vital organ in the human body, is also prone to various diseases, including heart disease, which is a leading cause of

death worldwide. Identifying cardiovascular diseases in their early stages remains a major challenge in the healthcare field. Therefore, having a method or application that can assist individuals in identifying the occurrence of heart-related ailments early on is crucial. Avoiding the inconvenience and expense of multiple procedures to confirm one's health is desirable. To tackle the complexities surrounding heart disease diagnosis, machine learning algorithms such as the support vector machine algorithm, the decision tree algorithm, and the logistic regression algorithm have shown promising results. Machine learning, a branch of artificial intelligence (AI), empowers computers to learn automatically from historical data and experiences, enabling the identification of patterns and anticipation of outcomes with minimal human involvement.

Machine learning involves training algorithms with a specific dataset to develop a model. This model is then utilized by the trained machine learning algorithm to make predictions when presented with new input data. Machine Learning includes various algorithms like Supervised Learning where machines are trained on labeled datasets, allowing them to anticipate outputs based on the provided training data which are further categorized as classification and regression algorithms. Unsupervised Learning where machines are trained on unlabeled datasets, enabling them to anticipate outputs without human oversight which are categorized as clustering and association. For heart disease prediction we have used algorithms like logistic regression, and random forest.

LITERATURE SURVEY

Heart Disease Prediction using Hybrid Machine Learning Techniques by Sadar *et al*. [1]

The researchers employ various machine learning algorithms on the Cleveland heart disease dataset, with the Extended version of Extreme Machine Learning (EML) achieving the highest accuracy at 93%, emphasizing the importance of advanced techniques in improving heart disease prediction [1].

Hybrid Machine Learning Techniques for Heart Disease Prediction by Sharanyaa *et al*. [2]

This paper explores the use of hybrid machine learning techniques for predicting heart disease, combining methods like Decision Tree, Support Vector Machine, K-Nearest Neighbor, and Random Forest. The integrated approach achieved a notable accuracy of 94%, underscoring the effectiveness of machine learning in the early detection of heart conditions [2].

Heart Disease Prediction Using Hybrid Machine Learning Model by Kavitha *et al*. [3]

Advanced and smart heart disease prediction using hybrid machine learning techniques, focusing on early and precise diagnosis of heart disease, this paper utilizes artificial neural networks (ANN) and hybrid machine learning techniques, demonstrating that the SVMANN hybrid model achieves good accuracy of 91.3% for cardiac disease prediction [3].

Heart Disease Prediction Using Hybrid Random Forest and Linear Model by Khan *et al.* [4]

A hybrid system for the prediction of heart disease using machine learning: Introducing a hybrid random forest with a linear model (HRFLM), this study reports an improved performance with 88.7% accuracy in predicting heart disease using Logistic Regression, Random Forests, Neural network, and other classification algorithms [4].

Predicting Heart Disease Using Hybrid Machine Learning Model by Renugadevi *et al*. [5]

Heart disease prediction using hybrid machine learning emphasizing a hybrid random forest with a linear model and feature score approach, this research focuses on data preparation methods and reveals that the suggested technique accurately predicts heart disease with significantly higher performance [5].

Advances in Mechanical Engineering by Manik *et al.* [6]

Heart disease prediction using hybrid machine learning model, Random Forest achieves the highest accuracy of 90.16% among various machine learning algorithms for predicting cardiac disease, including support vector classification and naive bayes classifier, using a dataset from the UCI laboratory [6].

A Hybrid Machine Learning Algorithm for Heart and Liver Disease Prediction Using Modified Particle Swarm Optimization with Support Vector Machine by Behera *et al*. [7]

A combined machine learning approach for predicting heart and liver diseases using enhanced particle swarm optimization and support vector machine: Introducing a hybrid model combining support vector machine (SVM) and modified particle swarm optimization, this paper evaluates results based on classification accuracy, error, correctness, recall, and F1 score, comparing favorably with SVM and other hybrid algorithms [7].

Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques by Verma *et al*. [8]

Proposing a Decision Tree and Random Forest combination, this study employs the Cleveland Heart Disease database to forecast heart disease, emphasizing effectiveness of proposed scheme [8].

Machine Learning Approach for Heart Disease Prediction: a Survey by Gupta and Sharma [9]

Predicting heart disease using hybrid machine learning model using Random Forest and Decision Tree techniques in a hybrid methodology, this paper achieves an 88.7% accuracy in predicting heart ailment, reinforcing the effectiveness of the proposed scheme using the Cleveland Heart Disease database [9].

Cardiovascular Disease Prediction Using Hybrid-Random-Forest-Linear-Model (HRFLM) by Sathwik *et al*. [10]

This paper presents an innovative approach for predicting heart disease by introducing the Hybrid Random Forest with Linear Model (HRFLM) [10]. By combining the strengths of both the Random Forest algorithm and a Linear Model, the proposed method achieves improved accuracy in predictions, reaching 88.7%. The results highlight the effectiveness of this hybrid technique in enhancing the accuracy of heart disease prediction.

METHODOLOGY

The methodology consists of several steps, including data loading and exploration, data preprocessing, model training and assessment, model comparison, hybrid model building, hyperparameter tweaking, GUI application development, and model persistence. In this work we have collected the data for heart disease prediction from Kaggle. This data collection dates from 1988 comprises of four databases. There are 76 features in all, including the target value, however all published studies use just 14 of them. The "target" field indicates the existence of cardiac disease in the patient. A value of 0 represents the absence of disease, while a value of 1 signifies the presence of disease. Once the data is gathered from various sources, it needs to undergo preprocessing before being used to train the model. The dataset can be preprocessed in several ways, where we have worked on eliminating the missing values and duplicate values from the dataset. Categorical and continuous features are recognized using one-hot encoding for categorical data. The continuous features are then standardized using the Standard Scaler for improved precision. The dataset is subsequently divided into training and testing sets. Different classification models have been built and evaluated using the accuracy scores as metrics. In our research, we have employed various models, including Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees, and Random Forests.

Logistic Regression (LR)

Logistic Regression works by fitting a logistic function to the input features, transforming their linear combination into probabilities. In predicting heart disease, logistic regression calculates the likelihood that a patient has the condition by analyzing the given input features.

K-Nearest Neighbors (KNN)

KNN determines the classification of a data point by looking at the class labels of the closest neighboring points. In heart disease prediction, KNN assesses the similarity of a patient's features to those of its k closest neighbors, attributing the majority class to the patient. It operates effectively when local patterns are essential for accurate predictions.

Decision Trees (DT)

Decision Trees operate by continuously dividing the dataset according to the features that provide the most valuable information. In heart disease prediction, DT identifies critical decision points, such as cholesterol levels or age thresholds, that guide the classification process. It provides interpretability by illustrating the decision-making logic within the tree structure.

Random Forest (RF)

Random Forest works by creating numerous Decision Trees and aggregating their predictions through a voting mechanism. In heart disease prediction, RF creates an ensemble of trees, each capturing different aspects of the dataset's complexity. By averaging the predictions, RF enhances accuracy and mitigates overfitting.

Hybrid Model

Basing on the accuracy scores obtained for each model we have combined two models and form a hybrid model utilizing a voting classifier. Given that the accuracies of Random Forest and Logistic Regression surpass those of the other models we tested, we have decided to combine these two models to enhance overall performance. After that we have used Grid Search to undertake a systematic search for ideal hyperparameters in both the Random Forest and Logistic Regression models. This process ensures that the model parameters are optimized to achieve the highest level of performance, as shown in Figure 1.

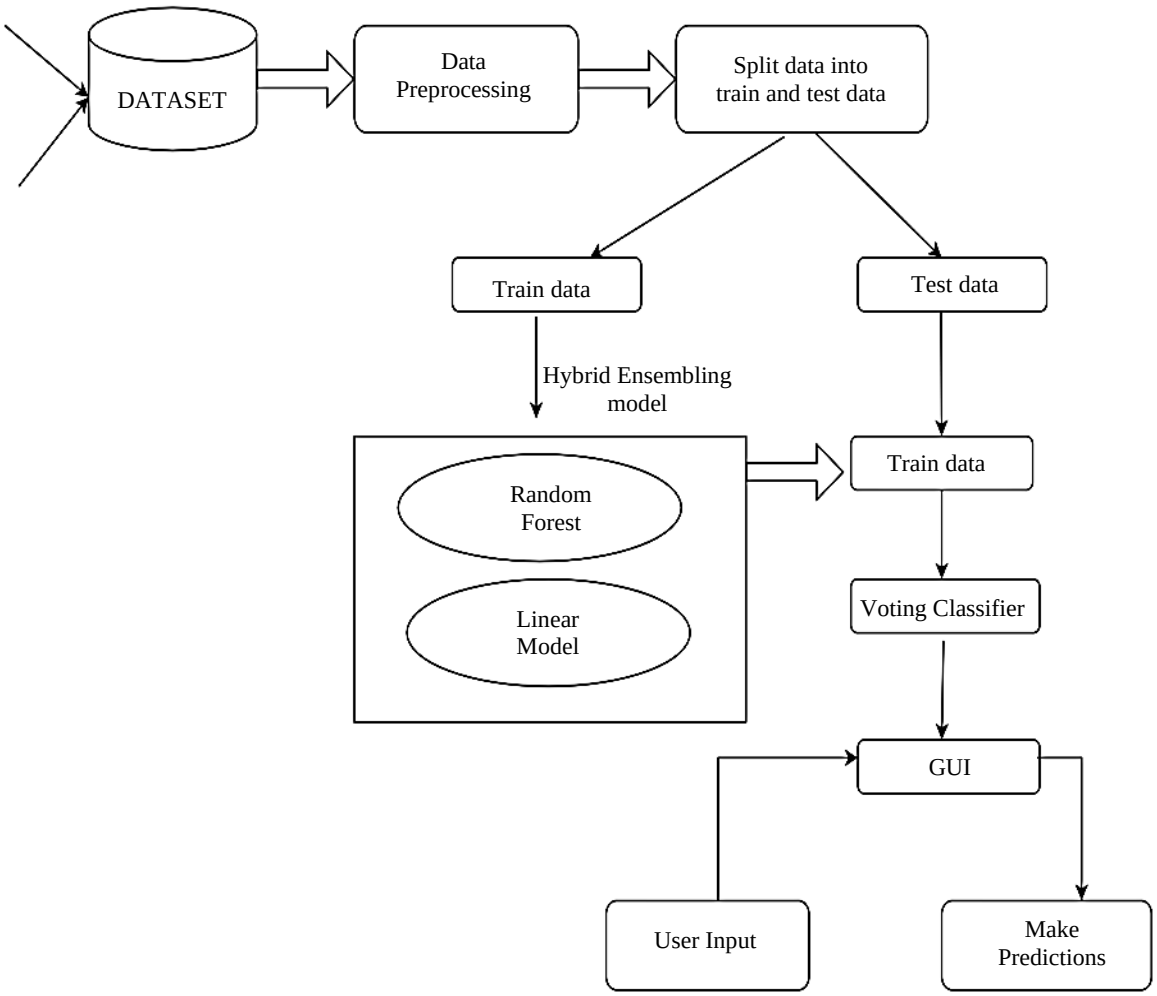

Figure 1. System architecture.

RESULTS AND DISCUSSION

In this study we have employed algorithms like Logistic Regression, KNN, Decision Tree and Random Forest and have obtained the accuracies of 78.6, 73.7, 73.7 and 80.6% respectively. We then merged the models that demonstrated high accuracy, specifically Random Forest and Logistic Regression. The initial training of this hybrid model on the dataset obtained an accuracy of 83.6%. Following hyperparameter tuning, the hybrid model's accuracy improved to 85.24%.

Figure 2 shows the accuracies of all the models used in this project. By this figure we can understand that Logistic regression obtained an accuracy of 78.68%, K-Nearest Neighbor has an accuracy of 73.77%, Decision Tree has obtained an accuracy of 73.77% and Random Forest has obtained an accuracy of 83.60%.

Figure 3 presents a graphical representation of all the accuracies achieved to enhance visualization.

Figure 4 shows the metrics of the hybrid model (Random Forest and Logistic Regression). From the figure we can understand that the hybrid model has produced an accuracy of 85.24%, higher than the accuracies of all the individual models. It also prints confusion matrix which shows the number of actual and predicted outcomes for each class. The diagonal components (24 and 227) denote the number of accurate guesses, whereas the others (8 and 2) indicate the number of wrong predictions. This generated 8 false positives (predicted 1, but actually 0) and 2 false negatives (predicted 0, but actually 1).

Out[4]:

	Models	ACC
0	LR	78.688525
1	KNN	73.770492
2	DT	73.770492
3	RF	83.606557

Figure 2. Accuracies of individual models.

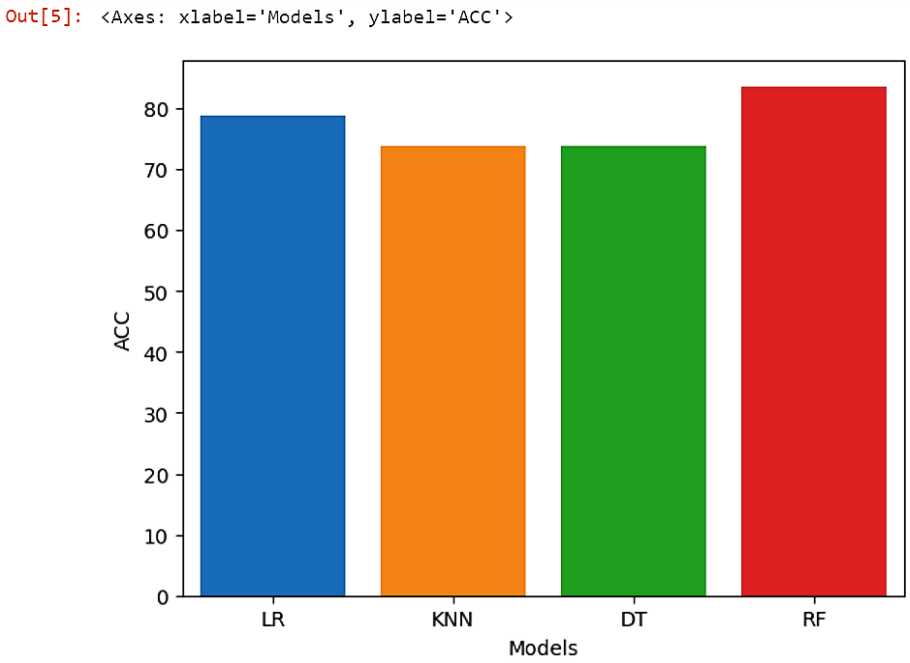

Figure 3. Accuracies visualization.


```
Accuracy: 85.24590163934425
Confusion Matrix:
[[24  8]
 [ 2 27]]
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.75      0.83        32
           1       0.77      0.93      0.84        29

    accuracy                           0.84        61
   macro avg       0.85      0.84      0.84        61
weighted avg       0.85      0.84      0.84        61
```

Figure 4. Metrics of hybrid model.

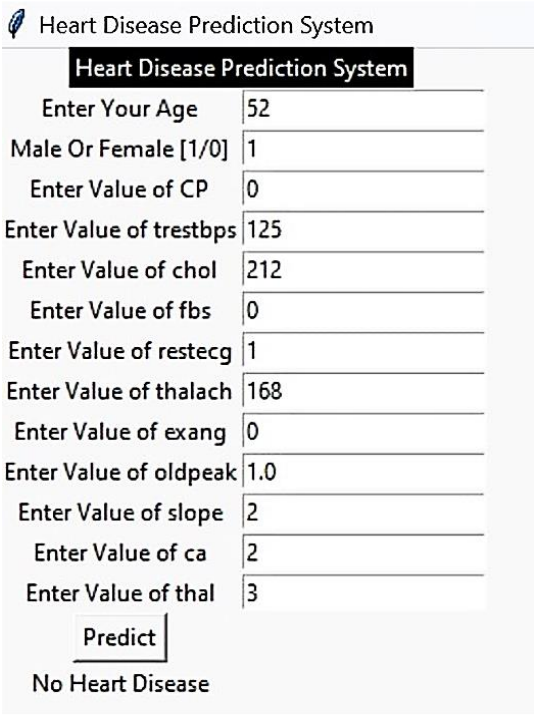

Figure 5. Heart disease prediction system.

Figure 5 shows the heart disease prediction system wherein the user has to enter the details about the input and when clicked on predict button, if the person is prone to heart disease or not.

CONCLUSION

This project harnessed the potential of various machine learning algorithms, including Logistic Regression, K-Nearest Neighbor, Decision Tree, and Random Forest, each contributing distinct strengths with accuracies of 78.6, 73.7, 73.7 and 80.6%, respectively. The hybrid model, a pivotal focus of this endeavor, emerged as a robust amalgamation of Random Forest and Linear Model, achieving an exceptional accuracy of 85.25%. The effectiveness of the hybrid model can be linked to its thoughtful integration of various educational methods. The hybrid model's ability to mitigate individual model weaknesses while leveraging their collective strengths is a key contributor to its success. By integrating the non-linear predictive prowess of Random Forest with the interpretability of the Linear Model, the hybrid approach struck a balance that surpassed the individual models. Importantly, the hybrid model demonstrated resilience to overfitting, a critical consideration in real-world scenarios where data distribution can vary. Furthermore, the hybrid model's accuracy of 85.25% not only surpassed the best-

performing individual model (Random Forest at 83.6%) but also highlighted the synergistic effects of integrating diverse algorithms. This achievement underscores the significance of combining contrasting learning approaches to create a versatile and powerful predictive tool. As the project concludes, it emphasizes the potential of such hybrid models in addressing the complexities of real-world data and enhancing predictive performance across various domains.

REFERENCES

1. Sadar U, Agarwal P, Parveen S, Jain S, Obaid AJ. Heart disease prediction using machine learning techniques. In International Conference on Data Science, Machine Learning and Applications. Singapore: Springer Nature Singapore; 2023 Dec 15; 551–560.
2. Sharanyaa S, Lavanya S, Chandhini MR, Bharathi R, Madhulekha K. Hybrid machine learning techniques for heart disease prediction. Int J Adv Eng Res Sci. 2020; 7(3): 44–8.
3. Kavitha M, Gnaneswar G, Dinesh R, Sai YR, Suraj RS. Heart disease prediction using hybrid machine learning model. In 2021 IEEE 6th international conference on inventive computation technologies (ICICT). 2021 Jan 20; 1329–1333.
4. Khan Z, Anwar S, Sikandar G. Heart Disease Prediction Using Hybrid Random Forest and Linear Model. International Journal of Emerging Engineering and Technology. 2023 Jul 6; 2(1): 6–12.
5. Renugadevi G, Priya GA, Sankari BD, Gowthamani R. Predicting heart disease using hybrid machine learning model. J Phys: Conf Ser. 2021 May 1; 1916(1): 012208. IOP Publishing.
6. Manik GA, Kalia S, Sahoo SK, Sharma TK, Verma OP. Advances in Mechanical Engineering. Singapore: Springer; 2021.
7. Behera MP, Sarangi A, Mishra D, Sarangi SK. A hybrid machine learning algorithm for heart and liver disease prediction using modified particle swarm optimization with support vector machine. Procedia Comput Sci. 2023 Jan 1; 218: 818–27.
8. Shivam Verma, Yash Manshani, Ramesh Kumar Gupta. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. Int Res J Eng Technol. 2021; 08(09): 1754–1759.
9. Gupta S, Sharma P. Machine learning approach for heart disease prediction: a survey. In AIP Conf Proc. 2022 Oct 21; 2555(1): 020009. AIP Publishing.
10. Sathwik AS, Naseeba B, Challa NP. Cardiovascular Disease Prediction Using Hybrid-Random-Forest-Linear-Model (HRFLM). In 2023 IEEE World Conference on Applied Intelligence and Computing (AIC). 2023 Jul 29; 192–197.

