IJETIR

ISSN (Online) 2583-0554

International Journal of Emerging Technologies and Innovative Research (IJETIR)

Volume 5, Issue 5, May 2025

Impact Factor: 5.731

Effective Heart Disease Prediction Using Machine Learning

Vimalraj K C[1] and Dr. A. Poongodi[2]

MCA Student[1]

Assistant Professor[2]

Vels Institute of Science Technology and Advanced Studies, Chennai

vimalrajchandru1234@gmail.com and apoongodi.scs@vistas.ac.in

Abstract: *Heart disease remains one of the leading causes of mortality worldwide, emphasizing the need for early and accurate prediction to improve patient outcomes. This project presents an effective machine learning-based approach for heart disease prediction using the Random Forest algorithm. By analyzing key clinical parameters such as age, cholesterol level, blood pressure, and lifestyle factors, the model identifies patterns associated with heart disease risk. Random Forest, an ensemble learning method, enhances prediction accuracy by combining the results of multiple decision trees, thereby reducing over fitting and improving robustness. The proposed system undergoes thorough evaluation on benchmark medical datasets, demonstrating high reliability and consistency in distinguishing between healthy and at-risk individuals. This approach supports healthcare professionals in early diagnosis and decision-making, contributing to better patient management and preventive care.*

Keywords: Heart disease, Machine learning, Decision tree, Random forest, Stochastic gradient boosting, Loss optimization

I. INTRODUCTION

Heart disease is one of the most critical health concerns globally, accounting for a significant number of deaths each year. Early detection and preventive measures are essential to reduce the mortality rate and improve the quality of life of individuals at risk. Traditionally, heart disease diagnosis relies heavily on manual evaluation by medical experts, which can be time-consuming and prone to human error. As the volume of healthcare data continues to grow, there is an increasing demand for automated and accurate prediction systems that can support clinicians in making informed decisions.

Machine learning (ML) offers powerful tools to analyze complex medical data and uncover hidden patterns that may not be immediately evident through traditional diagnostic methods. Among various ML algorithms, the Random Forest technique has emerged as a highly effective and reliable approach for classification problems in healthcare. It operates by constructing a multitude of decision trees during training and outputs the class that is the mode of the classes predicted by individual trees. This ensemble approach enhances prediction accuracy and reduces the risk of overfitting, making it well-suited for medical diagnostics.

In this project, we propose a heart disease prediction model based on the Random Forest algorithm. The model utilizes patient data such as age, gender, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, and other clinical attributes. Through careful data preprocessing and feature selection, the system ensures high-quality input for model training and evaluation. The model is trained and tested on well-known datasets like the Cleveland Heart Disease dataset, which contains labeled examples of both healthy and affected individuals.

Most hospitals use management software to monitor the clinical and patient data they

collect. It is well-known these days, and these kinds of devices generate a vast quantity of information on patients. These data are used for decision-making help in clinical set- tings rather seldom. These data are precious, yet a significant portion of their knowledge is left unused [6]. Because of the sheer volume of data involved in the process, the translation of clinical data that has been acquired into information that intelligent systems can use to assist

healthcare practitioners in making decisions is a process fraught with difficulties [7]. Intelligent systems put this knowledge to use to enhance the quality of treatment provided to patients. As a result of this issue, research on the processing of medical photographs was carried out. Because there were not enough specialists and too many instances were misdiagnosed, an automated detection method that was both quick and effective was necessary [8].

The primary objective of the research is centered around the effective utilization of a classifier model, which aims to categorize and identify vital components within com- plex medical data. This categorization process is a critical step towards enabling early

diagnosis of cardiovascular diseases, potentially contributing to improved patient out- comes and healthcare management [9]. However, the pursuit of disease prediction at an early stage is not without its challenges. One significant factor pertains to the inherent

complexity of the predictive methods employed in the classification process [10]. The intricate nature of these methods can lead to difficulties in interpreting the underlying decision-making processes, which might impede the integration of these models into clinical practice. Furthermore, the efficiency of disease prediction models is impacted by the time they take to execute. Swift diagnosis and intervention are crucial in medical conditions, and time-intensive models might not align with the urgency required for timely medical decisions. Researchers [11] have investigated various alternative strategies to forecast cardiovascular diseases. Perfect treatment and diagnosis have the potential to save the lives of an infinite number of individuals. The novel contribution of this work is as follows:

•Preprocessing of HDP dataset with normalization, exploratory data analysis (EDA), data visualization, and extraction of top correlated features.

•Implementation of DTRF classifier for training preprocessed dataset, which can accurately predict the presence or absence of heart disease.

•The SGB loss optimization is used to reduce the losses generated during the training process, which tunes the flyer parameters of DTRF.

The rest of the article is organized as follows: Sect. 2 gives a detailed literature survey analysis. Section 3 gives a detailed analysis of the proposed HDP-DTRF with multiple modules. Section 4 gives a detailed simulation analysis of the proposed HDP-DTRF. Section 5 concludes the article.

II. LITERATURE SURVEY

Rani et al. [12] designed a novel hybrid decision support system to diagnose cardiac ailments early. They effectively addressed the missing data challenge by employing multivariate imputations through chained equations. Additionally, their unique approach to feature selection involved a fusion of genetic algorithms (GA) and recursive feature reduction. Notably, the integration of random forest classifiers played a pivotal role in significantly enhancing the accuracy of their system. However, despite these advancements, their hybrid approach's complexity might have posed challenges in terms of interpretability and practical implementation. Kavitha et al. [13] embraced machine learning techniques to forecast cardiac diseases. They introduced a hybrid model by incorporating random forest as the base classifier. This hybridization aimed to enhance prediction accuracy; however, their decision to capture and store user input parameters for future use was intriguing but yielded suboptimal classification performance. This unique approach could be viewed as an innovative attempt to integrate patient-specific information, yet the exact impact on overall performance warrants further investigation.

Mohan et al. [14] further advanced the field by employing a hybrid model that combined random forest with a linear model to predict cardiovascular diseases. Through this amalgamation of different classification approaches and feature combinations, they achieved commendable performance with an accuracy of 88.7%. However, it is worth noting that while hybrid models show promise, the trade-offs between complexity and interpretability could influence their practical utility in real-world clinical settings. To predict heart diseases, Shah et al. [15] adopted supervised learning

techniques, including Naive Bayes, decision trees, K-nearest neighbor (KNN), and random forest algorithms. Their choice of utilizing the Cleveland database from the UCI repository as their data source added a sense of universality to their findings. However, the lack of customization in data sources might limit the applicability of their model to diverse patient populations with varying characteristics. Guoet et al.

[16] contributed to the field by harnessing an improved learning machine (ILM) model in conjunction with machine learning techniques. Integrating novel feature combinations and categorization methods showcased their dedication to enhancing performance and accuracy. Nonetheless, while their approach exhibits promising results, the precise impact of specific feature combinations on prediction accuracy could have been further explored. Hager Ahmed et al. [17] presented an innovative real-time prediction system for cardiac diseases using Apache Spark and Apache Kafka. This system, characterized by its three-tier architecture—offline model building, online prediction, and stream processing pipeline—highlighted its commitment to harnessing cutting-edge technologies for practical medical applications. However, the scalability and resource requirements of such real-time systems, especially in healthcare settings with limited computational resources, could be an area of concern. Kataria et al. [18] comprehensively analyzed and compared various machine learning algorithms for predicting heart disease. Their focus on analyzing the algorithms' ability to predict heart disease effectively sheds light on their dedication to identifying the most suitable model. However, their study's outcome might have been further enriched by addressing the unique challenges posed by individual attributes, such as high blood pressure and diabetes, in a more customized manner. Kannan et al. [19] meticulously evaluated machine learning algorithms to predict and diagnose cardiac sickness. By selecting 14 criteria from the UCI Cardiac Datasets, they showcased their dedication to designing a comprehensive study. Nevertheless, a deeper analysis of how these algorithms perform with specific criteria and their contributions to
accurate predictions could provide more actionable insights.

Ali et al. [20] conducted a detailed analysis of supervised machine-learning algorithms for predicting cardiac disease. Their thorough evaluation of decision trees, k-nearest neighbors, and logistic regression classifiers (LRC) provided a well-rounded perspective on the strengths and limitations of each method. However, a more fine- grained analysis of how these algorithms perform under various parameter configurations and feature combinations might offer additional insights into their potential use cases. Mienye et al. [21] introduced an enhanced technique for ensemble learning, utilizing decision trees, random forests, and support vector machine classifiers. The voting system they employed to aggregate results showcased their innovative approach to combining various methods. However, the potential trade-offs between ensemble complexity and the robustness of predictions could be considered for future refinement. Dutta et al. [22] revolutionized the field by introducing convolutional neural networks (CNNs) for predicting coronary heart disease. Their approach, leveraging the power of CNNs on a large dataset of ECG signals, showcased the potential for deep learning techniques in healthcare. However, the requirement for extensive computational resources and potential challenges in model interpretability could be areas warranting further attention. Latha et al. [23] demonstrated ensemble classification approaches. Combined with a bagging technique, their utilization of decision trees, naive Bayes, and random forest exemplified their determination to achieve robust results. Nevertheless, the potential interplay between different ensemble techniques and their effectiveness under various scenarios could be explored further.

Ishaq et al. [24] introduced the concept of using the synthetic minority oversampling technique (SMOTE) in conjunction with efficient data mining methods to improve survival prediction for heart failure patients. Their emphasis on addressing class imbalance through SMOTE showcased their awareness of real-world challenges in healthcare data-sets. However, the potential impact of the SMOTE method on individual patient sub- groups and its implications for model fairness could be areas of future exploration. Asadi et al. [25] proposed a unique cardiac disease detection technique based on random forest swarm optimization. Their use of a large dataset for evaluation underscored their dedication to robust testing. However, the potential influence of dataset characteristics and the algorithm's sensitivity to various parameters on prediction performance could be investigated further.

III. PROPOSED METHODOLOGY

Heart disease is a significant health problem worldwide and is responsible for many deaths every year. Traditional methods for diagnosing heart disease are often time-consuming, expensive, and inaccurate. Therefore, there is a need for more accurate and efficient methods for predicting and diagnosing heart disease. The article aims to provide a detailed analysis of the proposed HDP-DTRF approach and its performance in accurately predicting the presence or absence of heart disease. The results demonstrate the effectiveness of the proposed approach, which can lead to improved diagnosis and treatment of heart disease, ultimately leading to better health outcomes for patients.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 5, Issue 5, May 2025
Impact Factor: 5.731

Figure 1 shows the proposed HDP-DTRF block diagram. The initial step in the pro- posed approach is the preprocessing of a dataset consisting of patient records with known labels indicating the presence or absence of heart disease. The dataset is then used to train a DTRF classifier with the SGB loss optimization technique. The performance of the trained classifier is evaluated using a separate publicly available real-world test dataset, and the results show that the proposed HDP-DTRF approach can accurately predict the presence or absence of heart disease. Using decision trees in the random forest classifier enables the algorithm to handle nonlinear data and make accurate predictions even with missing or noisy data. Applying the SGB loss optimization technique further enhances the algorithm's performance by improving the convergence rate and avoiding overfitting. The proposed approach can be useful in clinical decision-making processes, enabling medical professionals to predict the likelihood of heart disease in patients and take appropriate preventive measures.

The detailed operation of the proposed HDP-DTRF system is illustrated as follows:

Step 1: Data preprocessing: Gather a dataset containing patient records, where each record includes features such as age, blood pressure, and cholesterol levels, along with labels indicating whether the patient has heart disease. Remove duplicate records, handle missing values (e.g., imputing missing data or removing instances

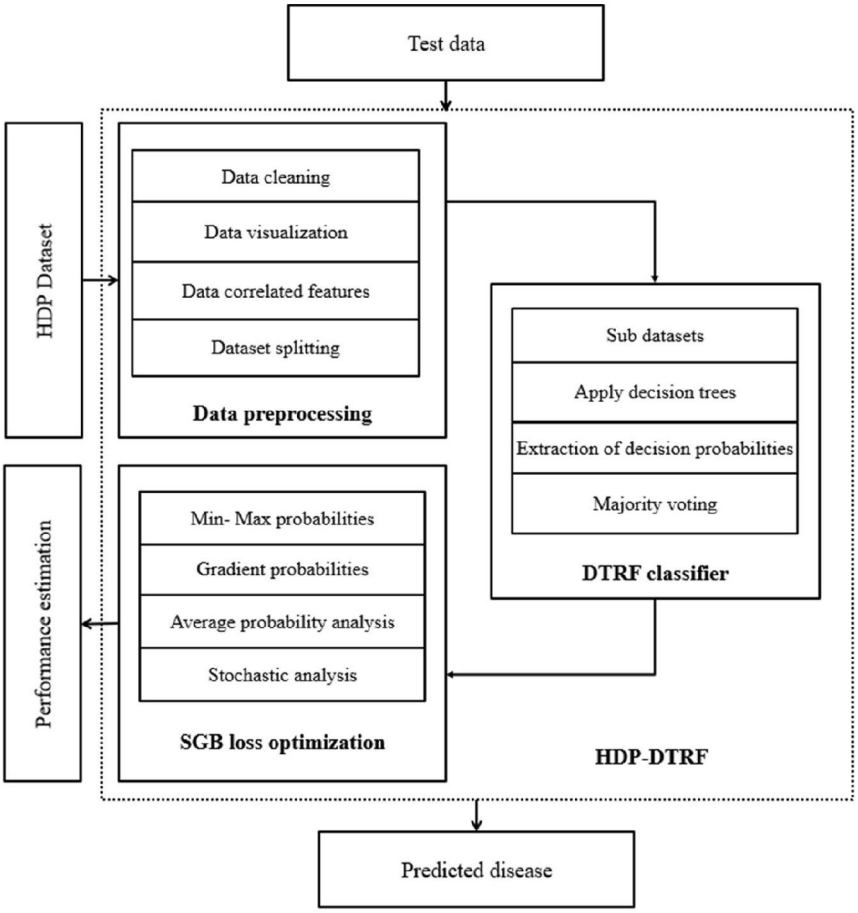

Fig. 1 Block diagram for the proposed HDP-DTRF system

With missing values), and eliminate irrelevant or redundant features. Encode categorical variables (like gender) into numerical values using techniques like one-float encoding. Scale numerical features to bring them to a common scale, which can pre- vent features with larger ranges from dominating the model.

Step 2: Training the DTRF classifier: Initialize an empty random forest ensemble. For each tree in the ensemble, randomly sample the training data with replacement. It creates a bootstrapped dataset for training each tree, ensuring

diversity in the data subsets. Construct a decision tree using the bootstrapped dataset. At each node of the tree, split the data based on the feature that provides the best separation, deter- mined using metrics like impurity or information gain. Add the constructed decision tree to the random forest ensemble. Repeat the process to create the ensemble's desired number of decision trees.

Step 3: SGB optimization: Initialize the model by setting the initial prediction to the mean of the target labels. Calculate the negative gradient of the loss function (such as mean squared error or log loss) concerning the current model's predictions. This gradient represents the direction in which the model's predictions

get. This new tree will help correct the errors made by the previous model iterations. Update the model's predictions by adding the predictions of the new tree, scaled by a learning rate. This step moves the model closer to the correct predictions. Repeat the process for a predefined number of iterations. Each iteration focuses on improving the model's predictions based on the errors made in the previous iterations.

Step 4: Performance evaluation: Use a separate real-world test dataset that was not used during training to evaluate the performance of the trained HDP-DTRF classifier.

IV. CONCLUSION

This article proposes a machine-learning approach for heart disease prediction. The approach uses a DTRF classifier with loss optimization and involves preprocessing a dataset of patient records to determine the presence or absence of heart disease. The DTRF classifier is then trained on the SGB loss optimization dataset and evaluated using a separate test dataset. The proposed HDP-DTRF improved class-specific performances and a macro with weighted average performance measures. Overall, the proposed HDP- DTRF improved precision by 2.30%, recall by 1.27%, F1-score by 3.61%, and accuracy by 1.03% compared to traditional methodologies. Further, this work can be extended with deep learning-based classification with machine learning feature analysis .

REFERENCES

[1]. Bhatt CM et al (2023) Effective heart disease prediction using machine learning techniques. Algorithms 16(2):88

[2]. Dileep P et al (2023) An automatic heart disease prediction using cluster-based bi-directional LSTM (C-BiLSTM) algorithm. Neural Comput Appl 35(10):7253–7266

[3]. Jain A et al (2023) Optimized levy flight model for heart disease prediction using CNN framework in big data application. Exp Syst Appl 223:119859

[4]. Nandy S et al (2023) An intelligent heart disease prediction system based on swarm-artificial neural network. Neural Comput Appl 35(20):14723–14737

[5]. Hassan D et al (2023) Heart disease prediction based on pre-trained deep neural networks combined with principal component analysis. Biomed Signal Proc Contr 79:104019

[6]. Ozcan M et al (2023) A classification and regression tree algorithm for heart disease modeling and prediction. Healthc Anal 3:100130

[7]. Saranya G et al (2023) A novel feature selection approach with integrated feature sensitivity and feature correlation for improved heart disease prediction. J Ambient Intell Humaniz Comput 14(9):12005–12019

[8]. Sudha VK et al (2023) Hybrid CNN and LSTM network for heart disease prediction. SN Comp Sc 4(2):172

[9]. Chaurasia V, et al (2023) Novel method of characterization of heart disease prediction using sequential feature selection-based ensemble technique. Biomed Mat Dev 2023;1–

[10]. Ogundepo EA et al (2023) Performance analysis of supervised classification models on heart disease prediction. Innov Syst Software Eng 19(1):129–144

[11]. de Vries S et al (2023) Development and validation of risk prediction models for coronary heart disease and heart failure after treatment for Hodgkin lymphoma. J Clin Oncol 41(1):86–95

[12]. Vijaya Kishore V, Kalpana V (2020) Effect of Noise on Segmentation Evaluation Parameters. In: Pant, M., Kumar Sharma, T., Arya, R., Sahana, B., Zolfagharinia, H. (eds) Soft Computing: Theories and Applications. Advances in Intel- ligent Systems and Computing, vol 1154. Springer, Singapore.

[13]. Kalpana V, Vijaya Kishore V, Praveena K (2020) A Common Framework for the Extraction of ILD Patterns from CT Image. In: Hitendra Sarma, T., Sankar, V., Shaik, R. (eds) Emerging Trends in Electrical, Communications