Jawalkar *et al.*
Journal of Engineering and Applied Science (2023) 70:122
https://doi.org/10.1186/s44147-023-00280-y

Journal of Engineering and Applied Science

RESEARCH Open Access

Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting

Anil Pandurang Jawalkar[1*], Pandla Swetcha[1], Nuka Manasvi[1], Pakki Sreekala[1], Samudrala Aishwarya[1], Potru Kanaka Durga Bhavani[1] and Pendem Anjani[1]

*Correspondence: anil.jawalkar022@gmail.com

[1] Department of Information Technology, Malla Reddy Engineering College for Women (UGC-Autonomous), Maisammaguda, Hyderabad, India

Abstract

Heart diseases are consistently ranked among the top causes of mortality on a global scale. Early detection and accurate heart disease prediction can help effectively manage and prevent the disease. However, the traditional methods have failed to improve heart disease classification performance. So, this article proposes a machine learning approach for heart disease prediction (HDP) using a decision tree-based random forest (DTRF) classifier with loss optimization. Initially, preprocessing of the dataset with patient records with known labels is performed for the presence or absence of heart disease records. Then, train a DTRF classifier on the dataset using stochastic gradient boosting (SGB) loss optimization technique and evaluate the classifier's performance using a separate test dataset. The results demonstrate that the proposed HDP-DTRF approach resulted in 86% of precision, 86% of recall, 85% of F1-score, and 96% of accuracy on publicly available real-world datasets, which are higher than traditional methods.

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

Introduction

One person dies due to cardiovascular disease every 36 s in every country. Coronary heart disease is the leading cause of mortality in the USA, accounting for one out of every four fatalities that occur each year. This disease claims the lives of about 0.66 million people annually [1]. The expenditures associated with cardiovascular disease are significant for the healthcare system in the USA. In the years 2021 and 2022, it resulted in annual costs of around $219 billion owing to the increased demand for medical treatment and medication and the loss of productivity caused by deaths. Table 1 provides the statistics of the heart disease dataset with total heart disease cases, deaths, case fatality rate, and total vaccinations. A prompt diagnosis also aids in preventing heart failure,

Table 1 Statistics of heart disease dataset

Country	Total cases	Total deaths	Case fatality rate	Total vaccinations
USA	44,752,659	720,581	1.61%	401,670,644
India	34,157,813	453,996	1.33%	1,031,906,566
Brazil	21,534,894	600,185	2.78%	239,756,958
Russia	8,073,318	222,853	2.76%	99,150,000
Turkey	7,052,488	64,049	0.91%	110,838,084
UK	7,005,365	137,322	1.96%	113,391,940
France	6,939,471	116,512	1.68%	100,355,009
Iran	6,261,269	126,711	2.02%	21,543,821
Argentina	5,301,830	115,662	2.18%	52,038,168
Colombia	4,985,923	126,245	2.53%	43,999,110
Spain	4,988,029	87,928	1.76%	77,561,325

which is another potential cause of mortality in certain cases [2]. Since many traits put a person at risk for acquiring the ailment, it is difficult to diagnose heart disease in its earlier stages while it is still in its infancy. Diabetes, hypertension, elevated cholesterol levels, an irregular pulse rhythm, and a wide variety of other diseases are some risk factors that might contribute to this [3]. These ailments are grouped and discussed under "heart disease," an umbrella word. The symptoms of cardiac disease can differ considerably from one individual to the next and from one condition to another within the same patient [4]. The process of identifying and classifying cardiovascular diseases is a continuous one that has a chance of being fruitful when carried out by a qualified professional with appropriate knowledge and skill in the relevant sector. There are a lot of different aspects, such as age, diabetes, smoking, being overweight, and eating a diet high in junk food. There have been several variables and criteria discovered that have been shown to either cause heart disease or raise the risk of developing heart disease [5].

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 settings 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 complex medical data. This categorization process is a critical step towards enabling early diagnosis of cardiovascular diseases, potentially contributing to improved patient outcomes 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 hyperparameters 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.

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 datasets. However, the potential impact of the SMOTE method on individual patient subgroups 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.

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.

Figure 1 shows the proposed HDP-DTRF block diagram. The initial step in the proposed 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

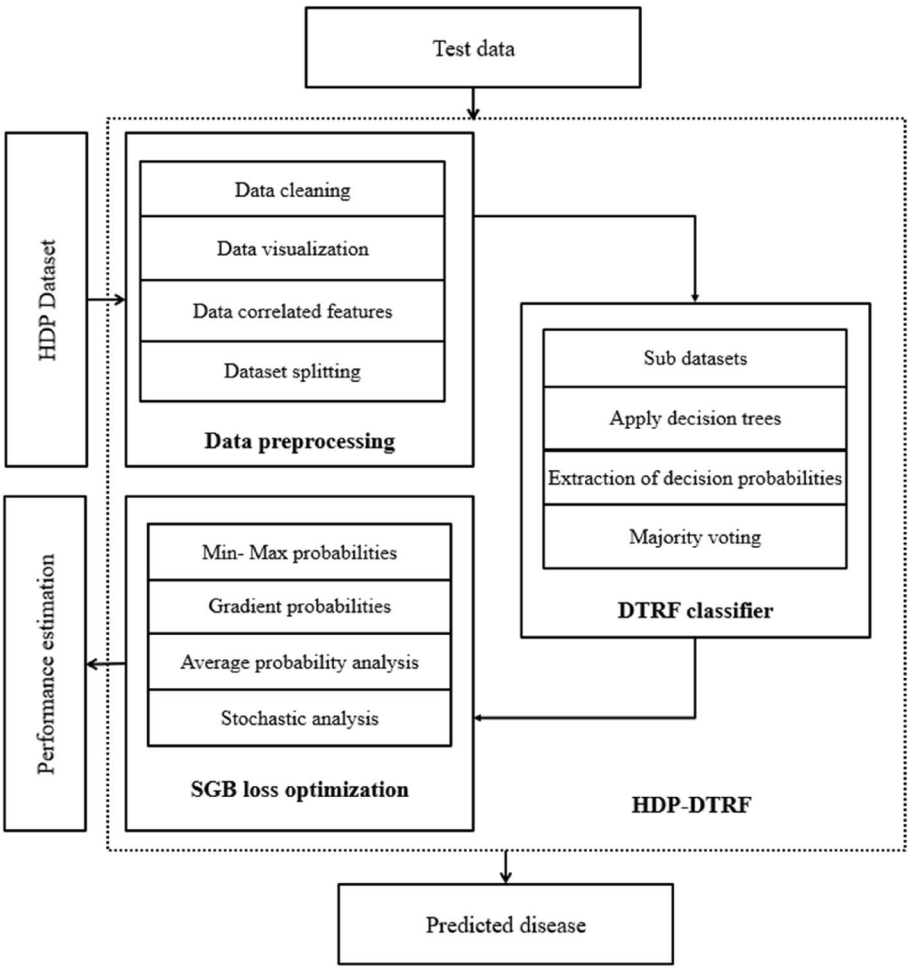

Fig. 1 Block diagram for the proposed HDP-DTRF system

with missing values), and eliminate irrelevant or redundant features. Encode categorical variables (like gender) into numerical values using techniques like one-hot encoding. Scale numerical features to bring them to a common scale, which can prevent 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, determined using metrics like Gini 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 need to be adjusted to minimize the loss. Train a new decision tree using the negative gradient as the tar-

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.

DTRF classifier

The DTRF classifier, an ensemble learning model, centers around the decision tree as its core component. As illustrated in Fig. 2, the DTRF block diagram depicts a framework comprising multiple trained decision trees employing the bagging technique. During the classification process, when a sample requiring classification is input, the ultimate classification outcome is determined through a majority vote from the output of an individual decision tree [26]. In classifying high-dimensional data, the DTRF model outperforms standalone decision trees by effectively addressing overfitting, displaying robust resistance to noise and outliers, and demonstrating exceptional scalability and parallel processing capabilities. Notably, the strength of DTRF stems from its inherent parameter-free nature, embodying a data-driven approach. The model requires no prior knowledge of classification from the user and is adept at training classification rules based on observed instances.

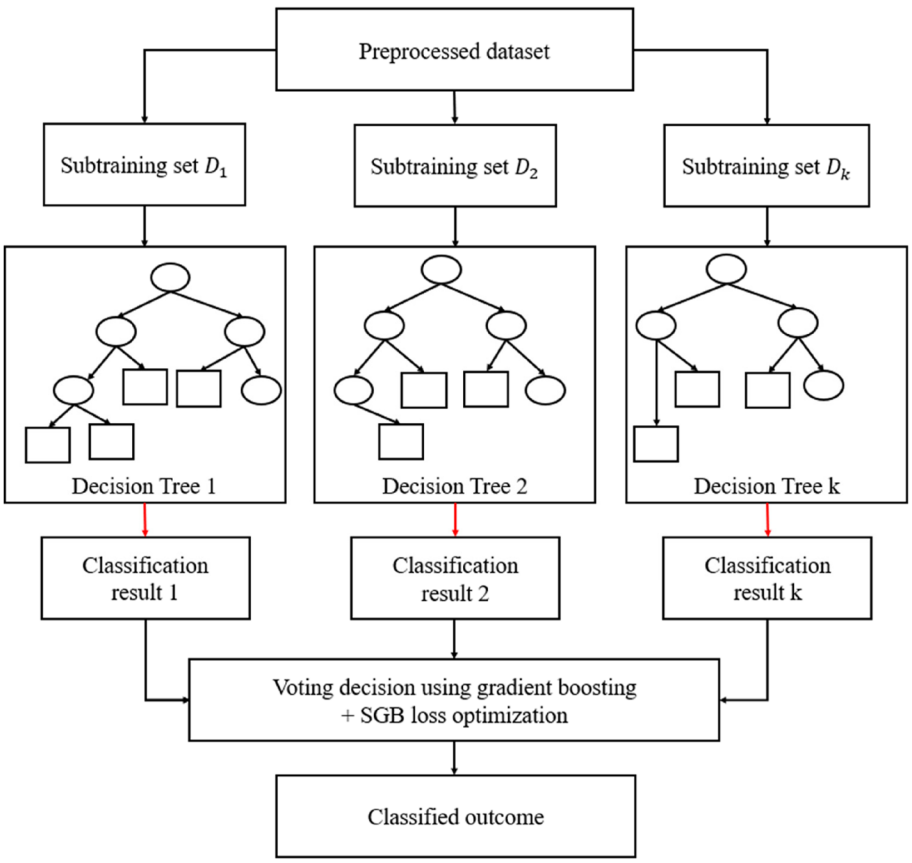

Fig. 2 Block diagram of DTRF

This data-centric attribute enhances the model's adaptability to various data scenarios. The DTRF model's essence lies in utilizing K decision trees. Each of these trees contributes a single "vote" towards the category it deems most fitting, thereby participating in determining the class to which the independent variable X, under consideration, should be allocated. This approach effectively harnesses the collective wisdom of multiple trees, facilitating accurate and robust classification outcomes that capitalize on the diverse insights provided by each decision tree. The mathematical analysis of DTRF is as follows:

$$\{h(X, \theta_K), k = 1, 2, \cdots, K\} \tag{1}$$

Here, K represents the number of decision trees present in the DTRF. In this context, θ_k is a collection of independent random vectors uniformly distributed amongst themselves. Here, K individual decision trees are generated. Each tree provides its prediction for the category that best fits the independent variable X. The predictions made by the K decision trees are combined through a voting mechanism to determine the final category assignment for the independent variable X. It is important to note that the given Eq. (1) indicates the ensemble nature of the DTRF model, where multiple decision trees work collectively to enhance predictive accuracy and robustness. The collection of θ_k represents the varied parameter sets for each decision tree within the ensemble.

The following procedures must be followed to produce a DTRF:

Step 1: The K classification regression trees are generated by randomly selecting K samples from the original training set as a self-service sample set, using the random repeated sampling method. Extracting all K samples requires repeating this procedure.
Step 2: Each node in the trees will include m randomly selected characteristics from the first training set (m n). Only one of the m traits is employed in the node splitting procedure, and it is the one with the greatest classification potential. DTRF calculates how much data is included in each feature to do this.
Step 3: A tree never has to be trimmed since it grows perfectly without help.
Step 4: The generated trees are built using DTRFs, and the freshly received data is categorized using DTRFs. The number of votes from the tree classifiers determines the classification outcomes.

There are a lot of important markers of generalization performance that are inherent to DTRFs. Similarity and correlation between different decision trees, mistakes in generalization, and the system's ability to generalize are all features *t*. A system's decision-making efficacy is determined by how well it can generalize its results to fresh information that follows the same distribution as the training set [27]. The system's performance and generalizability benefit from reducing the severity of generalization mistakes. Here is a case of the overgeneralization fallacy in action:

$$PE^* = P_{X,Y}(mr(X, Y) < 0) \tag{2}$$

Here, PE^* denotes the generalization error, the subscripts X and Y point to the space where the probability is defined, and $Mr(X, Y)$ is the margin function. The following is a definition of the margin function:

$$mr(X,Y) = Yavg_k(I(h(X,\theta_k) = Y) - max(J))) + Yavg_k(I(h(X,\theta_k) = J)) \quad (3)$$

If it stands for the input sample, Y indicates the correct classification, and J indicates the incorrect one. Specifically, $h(g)$ is a representation of a sequence model for classification, $I(g)$ indicates an indicator function, and $avg_k(g)$ means averaging. The margin function determines how many more votes the correct classification for sample X receives than all possible incorrect classifications. As the value of the margin function grows, so does the classifier's confidence in its accuracy. The term "convergence formulation of generalization error" as follows [28]:

$$limk \to \infty PE^* = P_{X,Y}(P_\theta(I(h(X,\theta_k) = Y)) - \max(J) \neq YP_\theta(I(h(X,\theta_k) = J))) \quad (4)$$

As the number of decision trees grows, the generalization error will tend toward a maximum, as predicted by the preceding calculation, and the model will not overfit. The classification power of each tree and the correlation between trees is used to estimate the maximum allowed generalization error. The DTRF model aims to produce a DTRF with a small correlation coefficient and strong classification power. Classification intensity (S) is the sample-space-wide mathematical expectation of the variable $mr(X,Y)$.

$$S = E_{X,Y} * mr(X,Y) \quad (5)$$

Here, θ and $\theta\prime$ are independent and identically distributed vectors of estimated data $E_{X,Y}$, correlation coefficients of $mr(\theta,X,Y)$ and $mr(\theta,\prime X,Y)$:

$$\overline{\rho} = \frac{covX,Y(mr(\theta,X,Y),mr(\theta\prime,X,Y))}{sd(\theta)sd(\theta\prime)} \quad (6)$$

Among them, $sd(\theta)$ can be expressed as follows:

$$sd(\theta) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(mr(x_i,\theta) - \frac{1}{N}\sum_{i=1}^{N}mr(x_i,\theta)\right)^2} \quad (7)$$

Equation (7) is a metric that is used to quantify the degree to which the trees $h(X,\theta)$ and $h(X,\theta\prime)$ on the dataset consisting of X, Y are correlated with one another. The correlation coefficient increases in magnitude in direct proportion $\overline{\rho}$ to the size of the chi-square. The upper limit of generalization error is obtained using the following formula, which is based on the Chebyshev inequality:

$$P_{X,Y}(mr(X,Y) < 0) \leq \frac{\overline{\rho}(1-S^2)}{S^2} \quad (8)$$

The generalization error limit of a DTRF is inversely proportional to the strength of the correlation P between individual decision trees and positively correlated with the classification intensity S of a single tree. That is to say, the stricter the category S, the lower the degree of linkage P. If the DTRF is to improve its classification accuracy, the threshold for generalization error must be lowered.

SGB loss optimization

The SGB optimization approach has recently received increased use in various deep-learning applications. These applications call for a higher degree of expertise in learning than what can be provided by more conventional means. During the whole training process, the learning rate that SGB uses does not, at any time, experience any fluctuations. The SGB uses one learning rate, which is alpha. The SGB algorithm maintains a per-parameter learning rate to increase performance in scenarios with sparse gradients (for example, computer vision challenges). It maintains per-parameter learning rates that are updated based on the average of recent magnitudes of the gradients for the weight, and it does so based on averaging recent gradient magnitudes (for example, how rapidly it is changing). In addition, it does this based on averaging recent gradient magnitudes for the weight. It illustrates that the strategy is effective for online and non-stationary applications (for example, noisy). The chain rule applied calculus to compute the partial derivatives. To calculate the loss gradient about the weights and biases, it will allow us to determine how the loss varies as a function of the weights and biases. Let us assume that we have a training dataset with N samples, denoted as $\{x_i, y_i\}$ for i = 1, 2, ..., N, where x_i is the input, and y_i is the true label or target value. It uses a decision tree with parameters θ to predict the output $\widehat{y}_i$for input x_i. The output can be any function of the parameters and the input, represented as $\widehat{y}_i = f(x_i, \theta)$. The goal is to minimize the difference between the predicted output $\widehat{y}_i$ and the true label y_i. It is typically done by defining a loss function $L(\widehat{y}_i, y_i)$ that quantifies the difference between the predicted and true values. The total loss over the entire dataset is then defined as the sum of the individual losses over all samples:

$$L_{total} = \Sigma_i L\left(f(x_i, \theta), y_i\right) \tag{9}$$

The optimization algorithm focused on estimating the values of the parameters θ that minimize this total loss. It is typically done using gradient descent, which updates the parameters θ in the opposite direction of the gradient of the total loss concerning the parameters:

$$\theta_{new} = \theta_{old} - \alpha \nabla_\theta L_{total} \tag{10}$$

Here,α is the learning rate, which controls the size of the parameter update, and $\nabla_\theta L_{total}$ is the gradient of the total loss concerning the parameters θ. The SGB can sometimes oscillate and take a long time to converge due to the noisy gradients. Momentum is a technique that helps SGB converge faster by adding a fraction of the previous update to the current update:

$$v_t = \beta v_t - 1 + (1 - \beta)\nabla_\theta L_{minibatch} \tag{11}$$

$$\theta_t = \theta_t - 1 - \alpha v_t \tag{12}$$

Here, v_t is the momentum term at iteration t, β is the momentum coefficient, typically set to 0.9 or 0.99, and the other terms are as previously defined.

Results and discussion

This section gives a detailed performance analysis of the proposed HDP-DTRF. The performance of the proposed method is measured using multiple performance metrics. All these metrics are measured for proposed methods as well as existing methods. Then, all the methods use the same publicly available real-world dataset for performance estimations.

Dataset

The Cleveland Heart Disease dataset contains data on 303 patients who were evaluated for heart disease. The dataset is downloaded from open-access websites like the UCI-ML repository. Each patient is represented by 14 attributes, which include demographic and clinical information such as age, sex, chest pain type, resting blood pressure, serum cholesterol level, and exercise test results. The dataset has 303 records, each corresponding to a unique patient. The data in each record includes values for all 14 attributes, and the diagnosis of heart disease (present or absent) is also included in the dataset. Table 2 provides a detailed description of the dataset. Researchers and data scientists can use this dataset to develop predictive models for heart disease diagnosis or explore relationships between the different variables in the dataset. With 303 records, this dataset is relatively small compared to other medical datasets. However, it is still widely used in heart disease research due to its rich attributes and long history of use in research studies.

Table 2 Description of dataset

Column	Description	Min value	Max value
Age	Age of the patient	29	77
Sex	1 = male, 0 = female)	0	1
Chest pain type	1 = typical angina, 2 = atypical angina, 4 = asymptomatic, 3 = non-anginal pain	1	4
Resting blood pressure (mm Hg)	Resting blood pressure	94	200
Serum cholesterol (mg/dl)	Serum cholesterol	126	564
Fasting blood sugar	Fasting blood sugar (> 120 mg/dl or not) of the patient (1 = true, 0 = false)	0	1
Resting electrocardiographic results	Results of resting electrocardiogram (0 = normal, 1 = ST-T wave abnormality, 2 = left ventricular hypertrophy)	0	2
Maximum heart rate achieved	Maximum heart rate achieved (in beats per minute) during exercise	71	202
Exercise-induced angina	Whether exercise-induced angina or not (1 = yes, 0 = no)	0	1
Oldpeak	ST depression induced by exercise relative to rest	0	6.2
Slope	The slope of the peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping)	1	3
Number of major vessels	Number of major vessels (0–3) colored by fluoroscopy	0	3
Thal	Thallium stress test result (3 = normal, 6 = fixed defect, 7 = reversible defect)	3	7
Target	Whether the patient has heart disease or not (0 = no, 1 = yes)	0	1

EDA

EDA is essential in understanding and analyzing any dataset, including the Cleveland Heart Disease dataset. EDA involves examining the dataset's basic properties, identifying missing values, checking data distributions, and exploring relationships between variables. Figure 3 shows the EDA of the dataset. Figure 3 (a) shows the count for each target class. Here, the no heart disease class contains 138 records, and the heart disease presented class contains 165 records. Figure 3 (b) shows the male and female-based record percentages in the dataset. Here, the dataset contains 68.32% male and 31.68% female records. Figure 3 (c) shows the percentage of records for chest pain experienced by the patient in the dataset. Here, the dataset contains 47.19% of records in typical angina, 16.50% in atypical angina, 28.71% in non-anginal pain, and 7.59% in the asymptomatic class. Figure 3 (d) shows the percentage of records for fasting blood sugar in the dataset. Here, the dataset contains 85.15% of records in the fasting blood sugar (>120 mg/dl) class and 14.85% of records in the fasting blood sugar (<120 mg/dl) class. Figure 4 shows the heart disease frequency by age for both no disease and disease classes. The output contains histogram levels that show the frequency of heart disease by age. Here, the counts of patients with and without heart disease are shown in red and green colors. The overlap between the bars shows how the frequency of heart disease varies with age, with a peak in the frequency of heart disease occurring around the age of 29–77 years.

Figure 5 shows the frequencies for different columns of the dataset, which contains the frequencies of chest pain type, fasting blood sugar, rest ECG, exercise-induced angina, st_slope, and number of major vessel columns. Exploring the frequencies

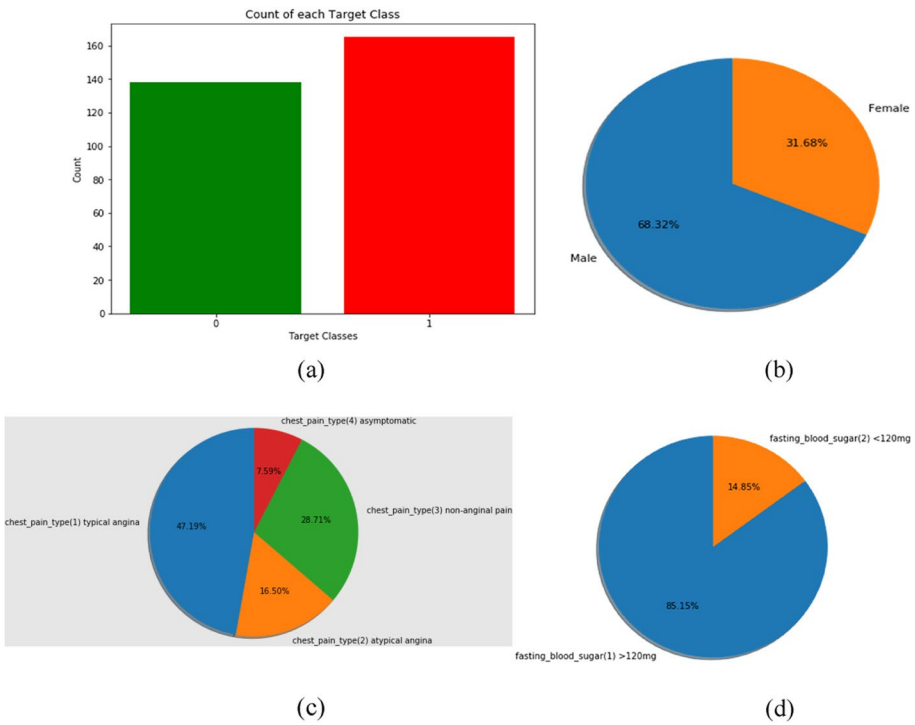

Fig. 3 EDA of the dataset. **a** Count for each target class. **b** Male–female distribution. **c** Chest pain experienced by patient distribution. **d** Fasting blood sugar distribution

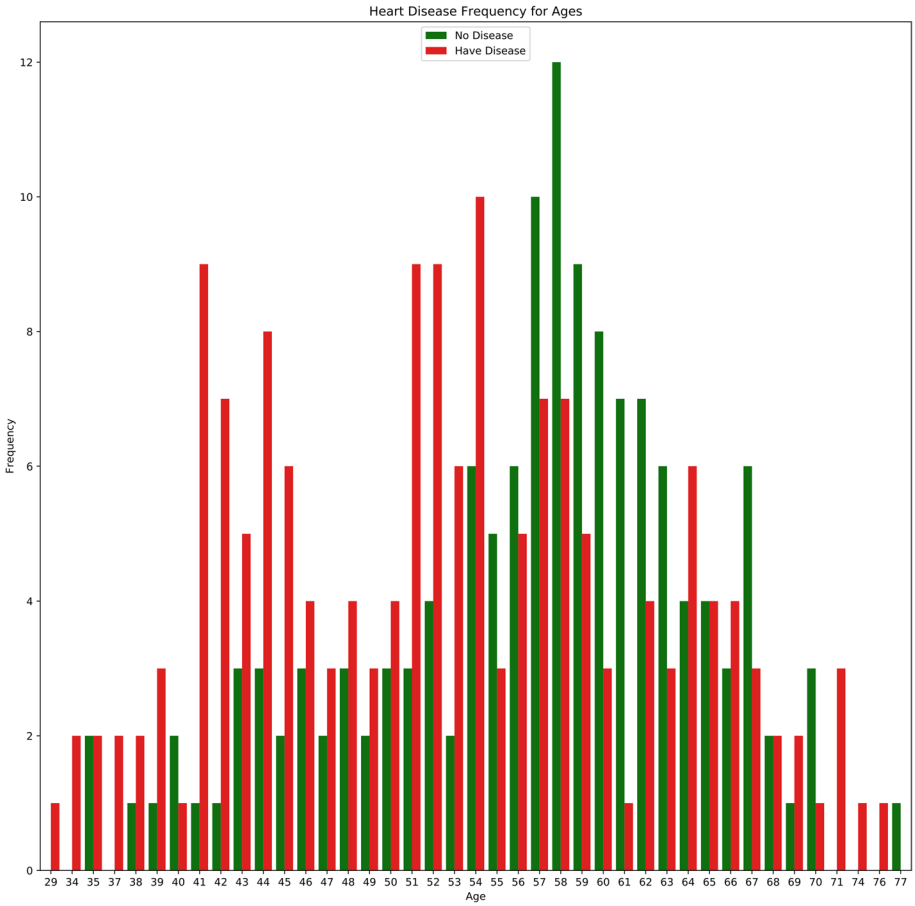

Fig. 4 Heart disease frequency by age

of different variables in a dataset is crucial in understanding the data and gaining insights about the underlying phenomena. By analyzing the frequency of values in each variable, we can better understand the data distribution and identify potential patterns, relationships, and outliers that are important for further analysis. For example, understanding the frequency of different chest pain types in a heart disease dataset reveals whether certain types of chest pain are more strongly associated with the disease than others. Similarly, analyzing the frequency of different fasting blood sugar levels helps to identify potential risk factors for heart disease. Overall, exploring the frequencies of variables is an important step in the EDA process, as it provides a starting point for identifying potential relationships and patterns in the data.

Performance evaluation

Table 3 shows the class-specific performance evaluation of HDP-DTRF. Here, the performance was measured for class-0 (no heart disease) and class-1 (heart disease presented) classes. Further, macro average and weighted average performances were also measured. Macro average treats all classes equally, regardless of their size. It calculates the average performance metrics across all classes, giving each class an equal weight. It means that the performance of smaller classes will have the same impact on the metric

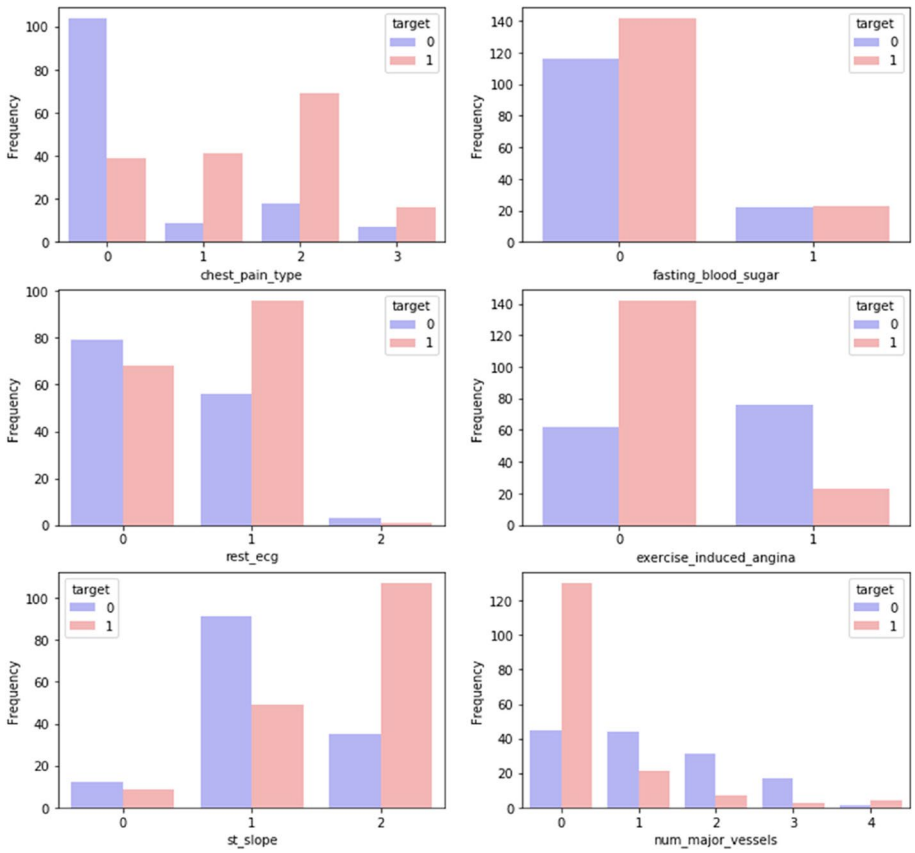

Fig. 5 Frequencies for different columns of the dataset

as larger classes.Then, the weighted average considers the size of each class. It calculates the average performance metric across all classes but gives each class a weight proportional to its size. It means that the performance of larger classes will have a greater impact on the metric than smaller classes.

Table 4 shows the class-0 performance comparison of various methods. Here, the proposed HDP-DTRF improved precision by 5.75%, recall by 1.37%, F1-score by 6%, and accuracy by 2.45% compared to KNN [15]. Then, the proposed HDP-DTRF improved precision by 3.45%, recall by 0.63%, F1-score by 3.61%, and accuracy by 1.45% compared to ILM [16]. Then, 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 LRC [20]. Table 5 shows the class-1 performance comparison of various methods. Here, KNN [15] shows a 2.35% lower precision, a 4.40% lower recall, a 3.53% lower F1-score, and a 1.03% lower accuracy than the proposed HDP-DTRF method. Then, ILM shows a 2.35% lower precision, a 5.49% lower recall, a 1.14% lower F1-score, and a 1.03% lower accuracy than the proposed HDP-DTRF method. Then, LRC [20] shows a 4.71% lower precision, an 11.11% lower recall, a 2.27% lower F1-score, and a 1.03% lower accuracy than the proposed HDP-DTRF method.

Table 3 Class-specific performance evaluation of proposed HDP-DTRF

Method	Precision	Recall	F1-score	Accuracy
Class-0	0.87	0.79	0.83	0.98
Class-1	0.85	0.91	0.88	0.97
Macro average	0.86	0.85	0.85	0.95
Weighted average	0.86	0.86	0.85	0.96

Table 4 Class-0 performance comparison of various methods

Method	Precision	Recall	F1-score	Accuracy
KNN [15]	0.75	0.80	0.77	0.92
ILM [16]	0.80	0.70	0.75	0.91
LRC [20]	0.85	0.75	0.80	0.95
Proposed HDP-DTRF	0.87	0.79	0.83	0.98

Table 5 Class-1 performance comparison of various methods

Method	Precision	Recall	F1-score	Accuracy
KNN [15]	0.83	0.87	0.85	0.96
ILM [16]	0.81	0.85	0.87	0.96
LRC [20]	0.79	0.81	0.85	0.96
Proposed HDP-DTRF	0.85	0.91	0.88	0.97

Table 6 shows the macro average performance comparison of various methods. For KNN [15], the percentage improvements are 7.5% for precision, 13.3% for recall, 10.4% for F1-score, and 6.7% for accuracy. For ILM [16], the percentage improvements are achieved as 2.4% for precision, 6.1% for recall, 6.0% for F1-score, and 3.2% for accuracy. For LRC [20], the percentage improvements are achieved as 3.4% for precision, 10.0% for recall, 6.0% for F1-score, and 4.3% for accuracy archived by the proposed method. Table 7 shows the weighted average performance comparison of various methods. For

Table 6 Macro average performance comparison of various methods

Method	Precision	Recall	F1-score	Accuracy
KNN [15]	0.80	0.75	0.77	0.90
ILM [16]	0.85	0.82	0.83	0.93
LRC [20]	0.81	0.80	0.83	0.92
Proposed HDP-DTRF	0.86	0.85	0.85	0.95

Table 7 Weighted average performance comparison of various methods

Method	Precision	Recall	F1-score	Accuracy
KNN [15]	0.81	0.79	0.79	0.90
ILM [16]	0.83	0.82	0.81	0.93
LRC [20]	0.85	0.81	0.83	0.93
Proposed HDP-DTRF	0.86	0.86	0.85	0.96

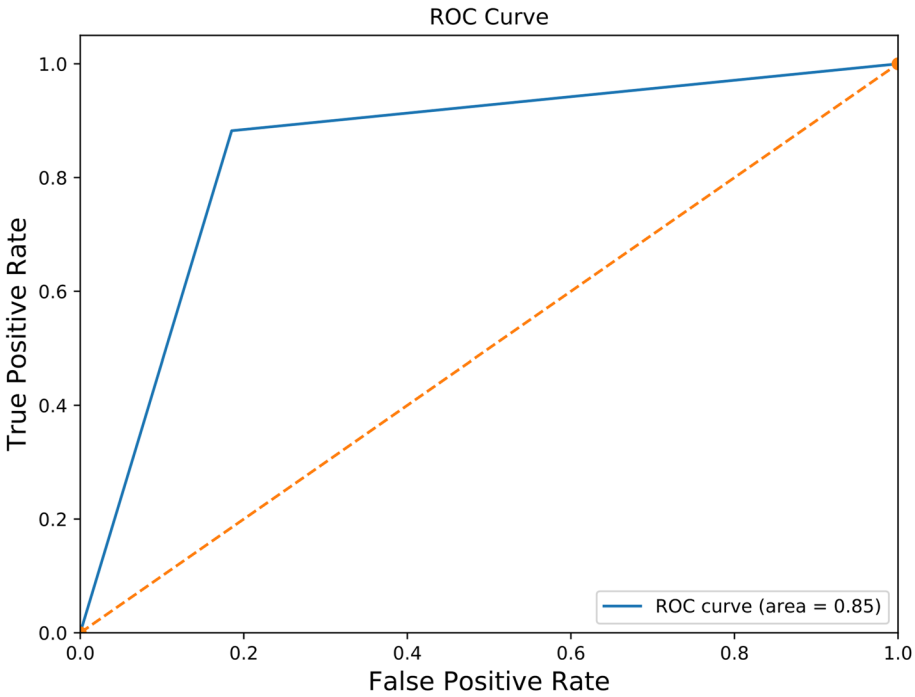

Fig. 6 ROC curve of proposed HDP-DTRF

KNN [15], the percentage improvements are 6.5% for precision, 3.3% for recall, 1.4% for F1-score, and 6.7% for accuracy. For ILM [16], the percentage improvements are achieved as 2.4% for precision, 5.1% for recall, 6.0% for F1-score, and 3.2% for accuracy. For LRC [20], the percentage improvements are achieved as 1.4% for precision, 1.0% for recall, 6.0% for F1-score, and 4.3% for accuracy archived by the proposed method.

The ROC curve of the proposed HDP-DTRF is seen in Fig. 6. The true positive rate (TPR) is shown against the false-positive rate (FPR) on the ROC curve, which considers various threshold values. In the context of the HDP-DTRF technique, the ROC curve illustrates the degree to which the model can differentiate between positive and negative heart disease instances. The model's performance is greater when it has a higher TPR and a lower FPR. The ROC curve that represents the HDP-DTRF approach that has been suggested is used to find the best classification threshold, which strikes a balance between sensitivity and specificity in the diagnostic process. If there is a point on the ROC curve that is closer to the top left corner, this implies that the model is doing better.

Conclusions

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 .

Abbreviations

HDP	Heart disease prediction
DTRF	Decision tree-based random forest
SGB	Stochastic gradient boosting
FP	False positive
FN	False negative
TN	True negative
TP	True positive

Acknowledgements
Not applicable.

Authors' contributions
A.P.J, P.S., and N.M. contributed to the technical content of the paper, and P.S. and S.A. contributed to the conceptual content and architectural design. P.K., D.B., and P.A. contributed to the guidance and counseling on the writing of the paper.

Funding
No funding was received by any government or private concern.

Availability of data and materials
Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 31 May 2023 Accepted: 5 September 2023
Published online: 10 October 2023

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. https://doi.org/10.1007/s44174-022-00060-x
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 Intelligent Systems and Computing, vol 1154. Springer, Singapore. https://doi.org/10.1007/978-981-15-4032-5_41.
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, and

Information Technologies. Lecture Notes in Electrical Engineering, vol 569. Springer, Singapore. https://doi.org/10.1007/978-981-13-8942-9_42
14. Annamalai M, Muthiah P (2022) An Early Prediction of Tumor in Heart by Cardiac Masses Classification in Echocardiogram Images Using Robust Back Propagation Neural Network Classifier. Brazilian Archives of Biology and Technology. 65. https://doi.org/10.1590/1678-4324-2022210316
15. Shah D et al (2020) Heart disease prediction using machine learning techniques. SN Comput Sci 1:345
16. Guo C et al (2020) Recursion enhanced random forest with an improved linear model (RERF-ILM) for heart disease detection on the internet of medical things platform. IEEE Access 8:59247–59256
17. Ahmed H et al (2020) Heart disease identification from patients' social posts, machine learning solution on Spark. Future Gen Comp Syst 111:714–722
18. Katarya R et al (2021) Machine learning techniques for heart disease prediction: a comparative study and analysis. Health Technol 11:87–97
19. Kannan R et al (2019) Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease. Springer, Soft Computing and Medical Bioinformatics
20. Ali MM et al (2021) Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Comput Biol Med 136:104672
21. Mienye ID et al (2020) An improved ensemble learning approach for the prediction of heart disease risk. Inform Med Unlocked 20:100402
22. Dutta A et al (2020) An efficient convolutional neural network for coronary heart disease prediction. Expert Syst Appl 159:113408
23. Latha CBC et al (2019) Improving the accuracy of heart disease risk prediction based on ensemble classification techniques. Inform Med Unlocked 16:100203
24. Ishaq A et al (2021) Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. IEEE Access 9:39707–39716
25. Asadi S et al (2021) Random forest swarm optimization-based for heart diseases diagnosis. J Biomed Inform 115:103690
26. Asif D et al (2023) Enhancing heart disease prediction through ensemble learning techniques with hyperparameter optimization. Algorithms 16(6):308
27. David VAR S, Govinda E, Ganapriya K, Dhanapal R, Manikandan A (2023) "An Automatic Brain Tumors Detection and Classification Using Deep Convolutional Neural Network with VGG-19," 2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 2023, pp. 1-5. https://doi.org/10.1109/ICAECA56562.2023.10200949
28. Radwan M et al (2023) MLHeartDisPrediction: heart disease prediction using machine learning. J Comp Commun 2(1):50-65

Publisher's Note