Published Online First: June, 2024
https://doi.org/10.21123/bsj.2024.9973
P-ISSN: 2078-8665 - E-ISSN: 2411-7986

Baghdad Science Journal

An Ensemble Model for Predicting Cardiovascular Disease utilizing Nature Inspired Optimization

Annwesha Banerjee Majumder[1], **Somsubhra Gupta**[2], **Sourav Majumder** [3], **Dharmpal Singh** [4]*

[1]Department of Information Technology, JIS College of Engineering, Kalyani, India.
[2]Department of Computer Science and Engineering, Swami Vivekananda University, Barrackpore, India.
[3]Capgemini India, Kolkata, India.
[4]Department of Computer Science and Engineering, JIS College of Engineering, Kolkata, India.
*Corresponding Author.

Received 17/10/2023, Revised 27/02/2024, Accepted 29/02/2024, Published Online First 20/06/2024

Abstract

This paper represents an efficient model for heart disease prediction model utilizing an ensemble mechanism optimized through BAT algorithm. Worldwide mortality rates are widely acknowledged to be significantly influenced by the prevalence of cardiovascular disease, particularly in economically disadvantaged regions. The need to mitigate the potentially severe repercussions associated with this particular health concern highlights the requirement for accurate and timely outcome prediction. Proposed methodology incorporates Mutual Information for feature selection, Inter Quartile Range for outlier removal. The StandardScaler method is used to achieve feature-wise standardisation in order to mitigate any bias resulting from varying scale disparities. Gradient boosting is an ensemble technique used in model construction that is well-known for its capacity to handle missing data and produce precise predictions. The BAT algorithm is implemented, which further improves speed by utilising optimisation inspired by nature. The application of the BAT method in this particular model has yielded a notable improvement in performance, resulting in an accuracy rate of 84.94%. The precision, specificity, and sensitivity scores of the model were 76.47%, 81.88%, and 89.65%, respectively. These metrics collectively suggest a balanced performance.

Keywords: *BAT algorithm, Cardiovascular disease, Ensemble classifier, Gradient Boosting, Inter Quartile Range, Nature Inspired Optimization.*

Introduction

Cardiovascular disease is a leading contributor to mortality worldwide [1]. In order to mitigate the extent of this condition, it is imperative to implement timely and accurate prognostication. Identifying subtle signs that may go unrecognized is a complicated challenge in early disease prediction today. Early-stage diagnosis is a challenge for traditional diagnostic techniques. This is addressed by machine learning (ML) models, which examine a variety of datasets, discover subtle signals suggestive of early disorders, and recognize complex patterns. Because machine learning (ML) can manage multidimensional data, it can facilitate personalized risk assessment and timely interventions by providing a more thorough understanding of individual health concerns. By utilizing sophisticated algorithms and data analytics, machine learning models present a viable way to enhance early disease prediction. More over in many regions worldwide, there is a shortage of healthcare professionals. Machine learning-based models for heart disease prediction offer a scalable solution by automating certain aspects of risk assessment, helping to bridge the gap in areas with limited access to medical practitioners.

This study has successfully constructed a predictive model for heart disease utilizing a machine learning framework. The main aim of this research endeavor is to develop and train a machine learning model capable of proficiently analyzing diverse clinical and demographic characteristics of individuals, with the ultimate goal of properly predicting the existence or likelihood of heart disease.

The model has been trained using heart disease dataset collected from UCI data repository. Mutual Information was used in feature selection during the first stage of model formulation in order to exclude features that were judged unnecessary. Because mutual information is impartial and efficient, it is used in the proposed method. The dataset was cleaned in the next phase, removing any remaining outliers. This has been achieved by using a reliable method called the Inter Quartile Range approach. Differently scaled variables might not have the same effect on a model's fitting and learning process, which could produce biased results. As a result, feature-wise standardization with a mean of zero and a standard deviation of one is commonly employed prior to model fitting in order to mitigate this possible concern. Next, the dataset has been standardized using the StandardScaler method. The StandardScaler approach may be influenced by outliers, a concern that has been previously mitigated in the proposed model by employing the IQR method for outlier removal. The ensemble technique of gradient boosting has been utilized to develop a model using a pre-processed dataset. Gradient boosting is a versatile machine learning technique that effectively manages missing data and generates highly accurate predictions. Subsequently, the BAT method has been implemented on the model in order to attain improved performance. The BAT algorithm, which draws inspiration from nature, possesses a significant advantage in its ability to rapidly converge at a critical point by effectively transitioning from exploration to exploitation. Due to this characteristic, it functions as a highly effective algorithm for tasks akin to categorizations.

This research article aims to fulfill a significant clinical requirement by utilizing machine learning methodologies to support healthcare practitioners in making well-informed decisions and delivering tailored healthcare to those who are either at risk of or have been diagnosed with heart disease. The principal contribution of this research lies in the creation of an innovative machine learning model that is specifically designed for the purpose of predicting cardiac disease.

The novelty of this work distinguishes itself through a meticulous amalgamation of sophisticated techniques, including the integration of Mutual Information for precise feature selection, the robust IQR method for outlier elimination, feature-wise standardization to ensure objective model fitting, and the deliberate application of the BAT algorithm for optimizing overall model performance. The confluence of these methodologies not only advances the state-of-the-art in disease prediction modeling but also elevates the sophistication of techniques employed in this domain.

By conducting a comprehensive assessment utilizing suitable performance indicators and statistical analysis, this study showcases the enhanced predictive accuracy and effectiveness of the suggested model. The validation process offers significant empirical support for the model's practical utility and reliability within real-world healthcare contexts.

Major contributions of the work are as follows:

a. The research makes significant contributions by employing advanced machine learning techniques, including feature selection, outlier removal, and algorithmic enhancements, to create an innovative model tailored for predicting cardiac disease.

b. The implementation of the BAT algorithm further boosts model performance by efficiently transitioning between exploration and exploitation, facilitating rapid convergence at critical points.

c. The study addresses clinical requirements, supporting healthcare practitioners in making informed decisions and delivering personalized healthcare to individuals at risk of or diagnosed with heart disease.

In below Related Work section, few existing works of this field have been analyzed and gaps have been identified. In Method and Material section, details methodological approach has been presented. Details result analysis have put in forth in Result and Discussion section where as the conclusive remarks have given in Conclusion section.

Related Work

To prevent the severity out of any disease, researchers are concentrating on developing intelligent models for early disease prediction. Extensive research has been conducted in this particular domain, a selection of which has been examined in the subsequent section.

A model for heart disease prediction that incorporates Random Forest, Support Vector Machine, Naive Bayes, and Multilayer Perceptron was proposed by Chaimaa Boukhatem et al.[2] Support Vector Machine produced best performance out of all the algorithms. Detail result analysis were not present in the manuscript.

Using K Nearest Neighbor, Logistic Regression, and Random Forest classifiers, Harshit Jindal et al. suggested a model of heart disease prediction. The UCI heart disease dataset was used to develop and test the model. Multifaceted analysis of classifier performance were missing in the manuscript [3].

Another heart disease prediction model using machine learning and deep learning was put forth by Rohit Bharti et al. Isolation Forest and Robust Scaler were used in this model's feature selection and outlier handling [4] . In this study, a comprehensive analysis of heart disease prediction was conducted by applying various machine learning algorithms and deep learning techniques to the UCI Machine Learning Heart Disease dataset. The objective was to compare the performance of different methodologies, utilizing 14 main attributes from the dataset. To enhance predictive accuracy, the Isolation Forest algorithm was employed to address irrelevant features, and the data were normalized. The results were rigorously validated using accuracy metrics and confusion matrices. The effectiveness of diverse algorithms in heart disease prediction was illuminated, emphasizing the significance of preprocessing techniques for improved model performance.

A model using Random Forest with Logistic Regression applied for feature evaluation was proposed by Chang et al. [5] . The paper constructed an AI-based heart disease detection system using machine learning algorithms, with a Python application for reliable healthcare research. It processed data, including handling categorical variables, and implemented key stages: database collection, logistic regression, and attribute evaluation. A random forest classifier achieved approximately 83% accuracy over training data. The discussion emphasized the algorithm's experiments and results, concluding with objectives, acknowledging limitations, and highlighting contributions to advancing heart disease detection through machine learning.

Melillo P et al. suggested a model using an artificial classifier to evaluate risk in congestive heart failure patients. The suggested classifier uses common long-term heart rate variability (HRV) measurements to distinguish between people at reduced risk and those at higher risk [6] . A retrospective analysis on two Holter databases included 12 patients with mild CHF (NYHA I and II, labeled lower risk) and 32 patients with severe CHF (NYHA III and IV, labeled higher risk). Eligibility criteria ensured signal quality, selecting patients with NN/RR intervals greater than 80%. Classification and regression tree (CART) methodology developed classifiers for 30 higher-risk and 11 lower-risk patients. The proposed trees achieved 93.3% sensitivity and 63.6% specificity.

Ramprakash et al. suggested a model for heart disease prediction using Deep Neural Network and χ 2 -statistical analysis. The model's efficacy was measured using DNN and ANN [7] . For model validation authors used accuracy, sensitivity, specificity and Matthews correlation. DNN achieved better results than ANN as per result presented in the manuscript. No other optimization mechanism has been applied.

A model using Chi Square PCA was proposed by Gárate Escamila et al. and tested on the UCI heart disease dataset [8] . Decision Tree, Gradient boosting, Logistic Regression, Multilayer perceptron, Naïve Bayes and Random Forest were applied as classifier, out of which Random Forest achieved highest accuracy.

Divya K.et al. suggested a model using a variety of machine learning algorithms, and it was found to be more accurate than existing methods when using logistic regression with majority voting [9] . This paper focused on the implementation of diverse machine learning algorithms for heart disease prediction, achieving an impressive accuracy of 88.59% through the use of logistic regression with majority voting, surpassing established techniques and emphasizing the historical impact of machine learning on advancing medical prognostication.

Applying Support vector machine (SVM), Gaussian Naive Bayes, logistic regression, LightGBM, XGBoost, and random forest authors proposed another model of heart disease prediction where 80.32%, 78.68%, 80.32%, 77.04%, 73.77%, and 88.5% accuracy have achieved respectively by each method [10] .

A model for COVID 19 prediction using deep transfer learning was proposed by Y. Pathal et al. This model also applies a top-2 smooth loss function with cost-sensitive features [11] . Comparing the proposed deep transfer learning-based COVID-19 classification model to other supervised learning models, experimental results show that it performs more efficiently. Optimization of hyperparameters was not considered in this proposed work.

Kumaret al. represented details analysis of different machine learning algorithms -Random Forest, Decision Tree, Logistic Regression, Support vector machine (SVM), K-nearest neighbors (KNN) in prediction of cardiovascular disease [12] . The analysis, conducted based on their accuracy and AUC ROC scores, revealed that the Random Forest machine learning classifier outperformed others, achieving a higher accuracy rate of 85%, an ROC AUC score of 0.8675.

A model was represented by Kota Pet al. where Support vector machine and Artificial Neural network were used [13] . Utilizing 80% of the data for training and 30% for testing on a dataset of 303 individuals, SVM exhibited an accuracy of 84% with a sensitivity of 78.5% and specificity of 87.8%. In comparison, ANN yielded an accuracy of 87%, along with a sensitivity of 85% and specificity of 88.2%. Overall, both ML and ANN demonstrated favorable accuracy in distinguishing individuals with and without heart disease.

On three separate datasets, Kareem AK, AL-Ani MM, and Nafea AA suggested a 1-D CNN-based system for the detection of autism spectrum disorder. CNN outperforms all other machine learning models in terms of accuracy. The highest documented accuracy rates for adults, children, and adolescents are 99.45%, 98.66%, and 90%, respectively [14] .

A Covid-19 infection diagnostic method using chest X-ray pictures was proposed by Zaki SM, Jaber MM, and Kashmoola MA. Both SVM and neural networks provide an estimated AUC score of 0.999, which is satisfactory for diagnosing Covid-19 [15] .

S. Mitra et al. proposed a model for Lung Cancer prediction. To determine the ideal set of parameters in CNN to reliably detect lung cancer, the suggested model was trained on 1000 CT scan images of malignant and non-cancerous cells. The system that was suggested had the best accuracy, 92.79%. Furthermore, the study discusses 192 observations that were made with the CNN model [16].

Majumder et al [17] proposed a heart disease prediction model applying bagging methods. Logistic Regression, K Nearest Neighbor and Naive Bayes have been used as base learner in this proposed model. The proposed work achieved an average accuracy of 82%. No further optimization mechanism has been applied for performance enhancement.

By analyzing the different existing work in the afore mentioned section it has been observed that proper optimization mechanism has not been utilized in disease prediction for achieving more robust model. In most of the work only accuracy, specificity and sensitivity have been considered as metric for model evaluation which is not enough. Justification of robustness, validity and reliability of the proposed models were missing in most of the manuscript. These gaps have been addressed in our proposed work.

Materials and Methods

In this study, the machine learning method known as Gradient Boosting has been employed for the purpose of predicting heart disease. The utilization of this potent ensemble method has been employed to construct a resilient and precise predictive model. Through the utilization of the collective power of numerous weak learners, Gradient Boosting effectively augments the prediction skills of the model and effectively captures intricate correlations present within the dataset. The dataset utilized in this work was obtained from the UCI data repository, a reputable and generally acknowledged source of diverse and dependable datasets intended for scientific endeavors. The utilization of Mutual Information for feature selection and the Inter Quartile Index for outlier identification has been implemented, resulting in an improvement in the performance of the model. Moreover, the model has undergone enhancements through the utilization of the Bat Algorithm, an optimization methodology that draws inspiration from the natural behavior exhibited by bats. The method presented in this study incorporates intelligent search and exploration algorithms that emulate the echolocation behavior observed in bats.

The urgent need to solve a worldwide health crisis led to the study's decision to concentrate on heart disease prediction. Because heart disease continues to be a major cause of morbidity and death. This novel method improves model performance by using the Inter Quartile Index for outlier identification and Mutual Information for feature selection. Additionally, by including the intelligent search and exploration algorithm of the Bat Algorithm—which draws inspiration from the echolocation behavior of bats—the prediction model is improved. The study's approach to the complicated problem of heart disease prediction not only closes a significant research gap but also has the potential to significantly influence public health policy, interdisciplinary collaboration, and healthcare practices.

The block diagram of the proposed model has shown in Fig. 1.

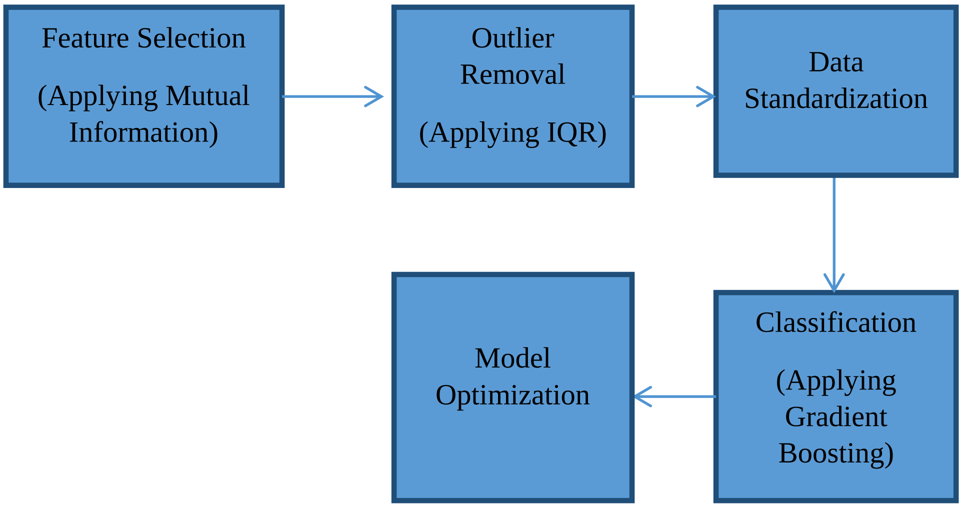

Figure 1. Proposed Model Block Diagram

Detail Description of Methodology Applied

Algorithmic Approach:

Step 1: Data set (Dt) collection from UCI data repository

Step 2: Feature Selection applying Mutual Information (MI): NewDt=MI(Dt)

Step 3: Data standardization applying StandardScaler: StandardScaler(NewDt)

Step 4: Outlier removal applying Interquartile range(IQR):NewDt_IQR=IQR(NewDt)

```
  def outremoval(d):
  sorted(d)
  Q1,Q3=np.percentile(d,[25,75])
  IQR=Q3-Q1
  lr=Q1-(1.5*IQR)
  ur=Q3+(1.5*IQR)
  return lr,ur
```

Step 6: Disease classification applying Gradient Boosting

Step 7 : Model optimization applying Nature Inspired BAT algorithm :

//List of Parameters:

```
param_grid = {
  'n_estimators': range(10, 80, 20),
  'max_depth': [2, 4, 6, 8, 10, 20],
  'min_samples_split': range(2, 8, 2),
  'max_features': ["auto", "sqrt", "log2"]
}
nia_search = NatureInspiredSearchCV(
  g,
  param_grid,
  cv=5,
  verbose=1,
  algorithm='fa',
  population_size=25,
  max_n_gen=100,
  max_stagnating_gen=10,
  runs=5,
  scoring='f1_macro',
  random_state=42,
)
nia_search.fit(X_train, y_train)
clf = GradientBoostingClassifier(**nia_search.best_params_
)
```

Libraries and Environment used:

The proposed model has been developed on Google Colab environment using python machine learning library

Model Performance Metric:

Performance of the proposed model has been measured through Accuracy, Precision, Sensitivity, Specificity and UC score.

Measuring the Robustness of the Model:

The robustness of the proposed model has been justified through variation of error and AUC ROC score over training dataset and validation dataset.

Model Validity and Reliability:

Validity and reliability of the proposed model has been measured through training and test error.

Dataset Description:

The dataset on which the model has developed has been collected from UCI data repository[18]. The data set consisting of 13 independent features age, sex,cp, trestbps, chol, fbs, restecg, thalach, exang ,oldpeak, slope, ca and thal followed by the dependent feature target. Age is a crucial factor, with older adults having a higher risk. Sex is represented by 0 for females and 1 for males. Chest pain (Cp) is a categorical variable indicating different types of angina, with 0 for asymptomatic, 1 for atypical angina, 2 for non-angina pain, and 3 for classic angina. Blood pressure (Tresbps) is continuously

Published Online First: June, 2024
https://doi.org/10.21123/bsj.2024.9973
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

recorded, and extremely high values may indicate the presence of cardiac disease. Cholesterol levels (Chol) in the blood, a waxy substance, contribute to heart disease risk. Resting electrocardiography (Restecg) results, categorized as 0 and 1, provide insights into left ventricular hypertrophy and anomalies in the T wave or ST segment. Thalach represents the maximal heart rate under stress, and Exang is a category field indicating whether a patient experiences angina during a stress test, denoted by 0 and 1. Oldpeak reflects a drop in the ST segment during activity. The Slope field is another categorical variable indicating ST segments during activity, with 0 for descending, 1 for flat, and 2 for rising. Thal, a categorical variable, reflects blood flow observed with radioactive dye, with values 0 for null, 1 for a fixed defect, 2 for normal blood flow, and 3 for a reversible effect. Ca denotes the quantity of blood vessels colored by the radioactive dye. The Target variable in the dataset, represented by 0 and 1, indicates the presence or absence of heart disease.

The dataset histogram has shown in. Fig. 2.

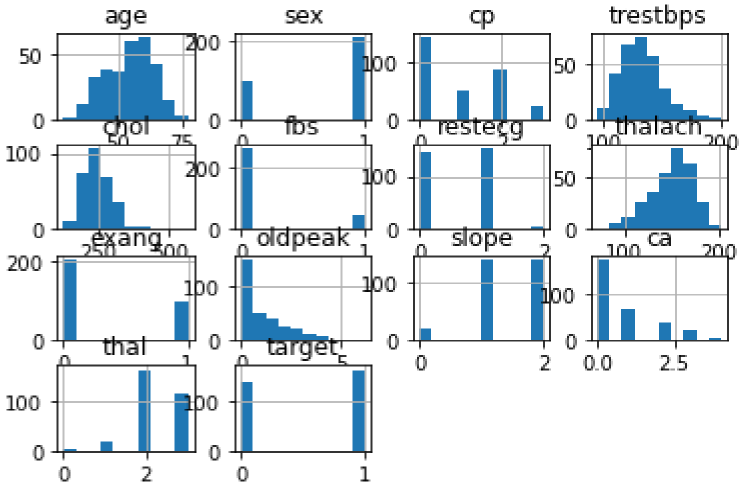

Figure 2. Dataset Histogram

Feature Selection:

In the suggested framework, the elimination of extraneous attributes plays a vital role in tackling the problem of subpar model efficacy. In order to accomplish this objective, the utilization of Mutual Information has been employed as a technique for selecting features. Mutual Information is a filter method that measures the information gained about one random variable by means of another random variable. Mutual Information functions as a metric for evaluating the association or interdependence between two variables, specifically denoted as X and Y.

The equation below is used to express Mutual Information.

$$\int_x \int_y p(x,y)\log\frac{p(x,y)}{p(x),p(y)}dxdy \qquad 1$$

where p(x) and p(y) are the marginal density functions and p(x,y) is the combined probability density function of X and Y. The mutual information determines the degree of similarity between the joint distribution p(x,y) and the children of the factored marginal distributions. If there is no relationship at all between X and Y, then p(x,y) would equal p(x)p(y), and this integral would be negative. In the proposed paradigm, Mutual Information is used as a filter approach for feature selection with the goal of improving the model's performance and predictive accuracy by concentrating on the most pertinent and informative features related to the prediction of heart disease.

Dataset Standardization

The StandardScaler technique has been used in the suggested model to standardise the dataset. A method for standardising or normalising a dataset's features is called StandardScaler. It functions by taking each feature's mean value and scaling it to have a unit variance.

Outlier Removal:

Outliers must be removed from the proposed model in order to guarantee the model's strong performance. Values known as outliers, which deviate significantly from an attribute's typical range, might affect the behaviour and predictions of the model. The interquartile range (IQR) has been used to find and eliminate outliers in the data in order to address this problem.

The interval between a dataset's first quartile (25th percentile) and third quartile (75th percentile) is defined statistically as the IQR. It offers insightful details regarding the distribution and fluctuation of the data.

Eq. 2 and Eq. 3 can be used to compute the upper and lower bounds of the IQR, and they can also be used to display the IQR.

Q1: 25% data lies between minimum and Q1

Q3: 75% of data lies between minimum and Q3.

$$\mathrm{IQR} = \mathrm{Q}_1 - \mathrm{Q}_2 \qquad 2$$

$$\mathrm{LowerBound} = \mathrm{IQR} - 1.5 * \mathrm{Q}_1 \qquad 3$$

$$\mathrm{UpperBound} = \mathrm{IQR} + 1.5 * \mathrm{Q}_3 \qquad 4$$

Any value that is not in range of UpperBound and LowerBound will be treated as outlier.

Classification:

Gradient Boosting an ensemble mechanism has been applied for classification of heart disease. This has implemented by Jerome H. Friedman in 1999 [19] . This process involves initially training a model using training data, followed by a second model that is trained concurrently, which attempts to reduce the mistakes of the first model.

The incorporation of Gradient Boosting in the proposed model is justified by its numerous inherent benefits in comparison to alternative methodologies. First and foremost, Gradient Boosting is widely recognized for its high efficiency and rapid execution, rendering it a highly preferred option for many modeling jobs. The enhanced efficiency facilitates expedited model training and prediction, hence offering notable advantages in scenarios involving extensive datasets or time-critical applications. One additional benefit of Gradient Boosting lies in its capacity to properly manage missing data. The presence of missing data is a prevalent obstacle in real-world datasets, necessitating the implementation of suitable algorithms for managing missing values in Gradient Boosting. The rapidity and effectiveness of Gradient Boosting facilitate expeditious model building and deployment, rendering it highly advantageous in clinical contexts where prompt prognostications are needed.

Model Optimization:

For optimization purpose BAT algorithm has been applied over the classifier. A metaheuristic global optimization technique is called the Bat algorithm. The echolocation behavior of microbats, which can locate prey and distinguish between several bug species even in complete darkness, has inspired Xin She Yang to suggest a bat algorithm [20] . In the BA, a bat searches for prey by flying randomly at position xi with a velocity vi and a set frequency range [fmin, fmax], altering its emission rate r∈ [0, 1]and loudness A0 in response to the distance of its target [21] . New velocity and position can be calculated through Eq6 and Eq.7. e is a uniformly distributed random vector having a [0, 1] range.

$$\mathrm{f}_\mathrm{i} = \mathrm{f}_{\mathrm{min}} + (\mathrm{f}_{\mathrm{max}} - \mathrm{f}_{\mathrm{min}})\mathrm{e} \qquad 5$$

$$\mathrm{v}_\mathrm{i}^{\mathrm{t}+1} = \mathrm{v}_\mathrm{i}^\mathrm{t} + (\mathrm{x}_\mathrm{i}^\mathrm{t} - \mathrm{x}^*)\mathrm{f}_\mathrm{i} \qquad 6$$

$$\mathrm{x}_\mathrm{i}^{\mathrm{t}+1} = \mathrm{x}_\mathrm{i}^\mathrm{t} + \mathrm{v}_\mathrm{i}^\mathrm{t} \qquad 7$$

Results and Discussion

The study that is being suggested is an improved machine learning-based model for heart disease prediction. Mutual Information, IQR, gradient boosting, and Bat Algorithm are all used in the suggested model. The proposed model's performance has been supported by ratings for accuracy, precision, sensitivity, and specificity.

Accuracy is the measure of correct prediction which represents any classifier’s overall performance. Precision is the measure of correct positive classification. Sensitivity represents the percentage of genuine positives that were accurately detected and Specificity demonstrates the model's capacity to forecast a true negative.

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \qquad 8$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \qquad 9$$

$$\text{Sensitivity}=\frac{\text{TP}}{\text{TP+FN}} \qquad 10$$

$$\text{Specificity}=\frac{\text{TN}}{\text{FP+TN}} \qquad 11$$

The importance and impact of each sub components of the model has been justified in below sections

Feature selection through Mutual Information

Total 11 features have been selected after applying Mutual Information. The features are sex,cp, chol, fbs, thalach, exang ,oldpeak, slope, ca ,thal and target. The dataset histogram has been shown in the Fig. 3.

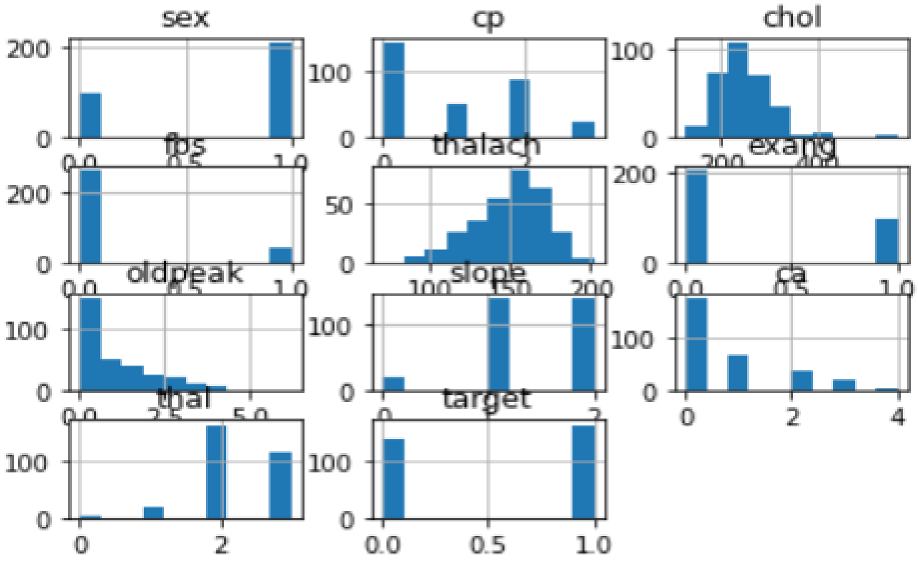

Figure 3. New selected feature set after applying Mutual Information

Outlier Identification and Removal:

The interquartile range has been used to locate and eliminate outliers. The graphs below show that the dataset contains outliers for the variables chol, thalach, and Oldpeak which have been shown in Fig. 4, Fig. 5 and Fig. 6.

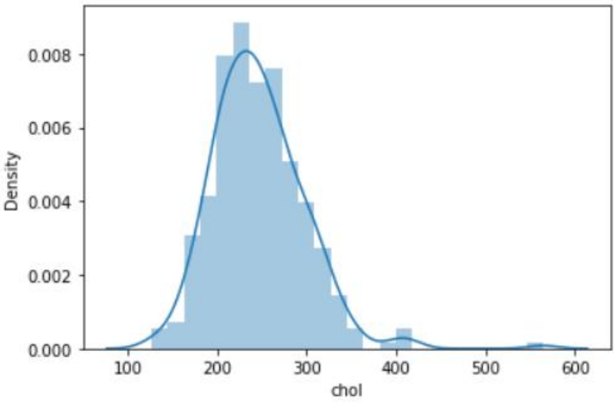

Figure 4. Outlier Observation of chol

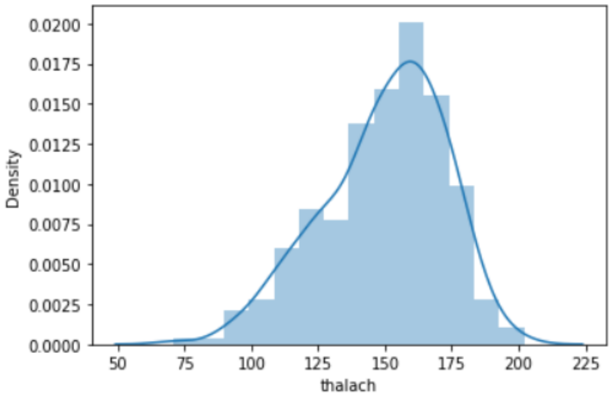

Figure 5. Outlier Observation of thalach

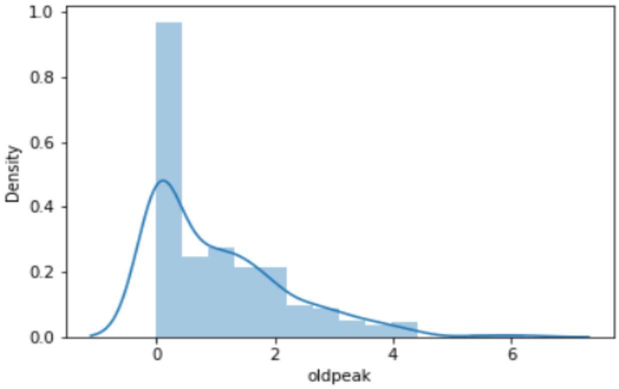

Figure 6. Outlier Observation of oldpeak

After identification of outliers they have been deleted from the dataset.

Prediction of Heart Disease applying Gradient Booting

Gradient Boosting has been used in this suggested model to predict heart disease. The model's success rate is 80.83% through the classifier. Confusion matrix in Fig. 7 illustrates the model performance accuracy.

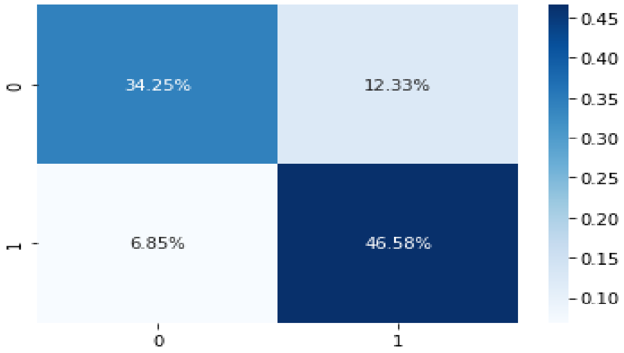

Figure 7. Confusion Matrix for classification of Heart Disease Applying Gradient Boosting.

Achieved AUC value of this classifier in .9072, which has been shown through AUC and ROC curves in Fig. 8.

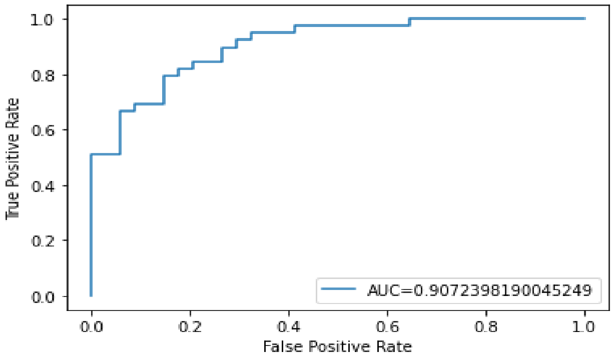

Figure 8. AUC Curve of the proposed model applying Gradient Boosting

Observation: Model Optimization through Nature Inspired BAT Algorithm

The major observation in this phase is the impact of Bat Algorithm in performance enhancement of the proposed classifier. As per the industry standard minimum 70% accuracy should be achieved by a good machine learning based model. It has been observed that the accuracy of the proposed model has been increased to 84.93% and AUC score has increased to 0.9430 after utilizing BAT algorithm. Confusion matrix and AUC ROC curve have shown in the figures below. Measured precision, sensitivity and specificity score for the models are 76.49%, 89.66% and 81.87% respectively. The confusion matrix generated applying BAT optimization has shown in Fig. 9 and the AOC curve has shown in the Fig. 10.

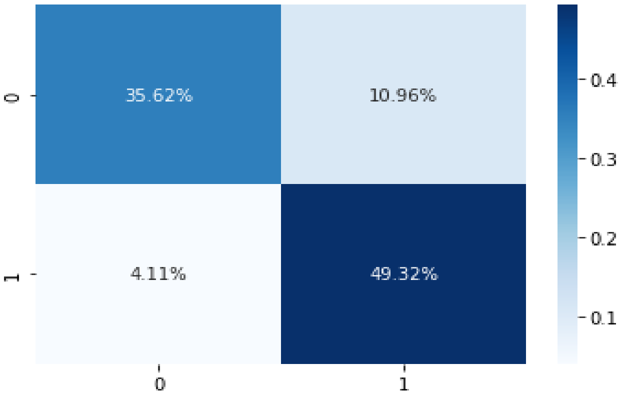

Figure 9. Confusion Matrix for classification of Heart Disease Applying Gradient Boosting and BAT Algorithm

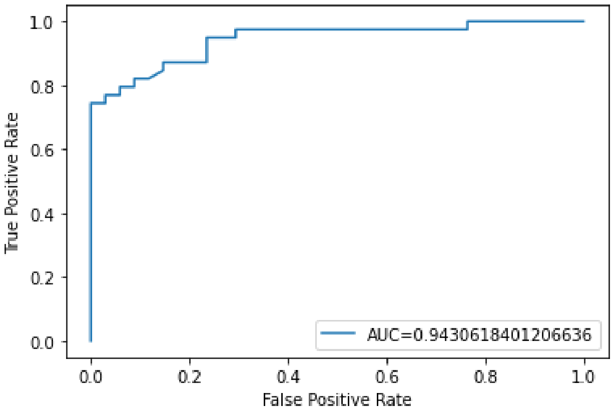

Figure 10. AUC Curve of the proposed model applying Gradient Boosting and BAT algorithm

In the below mentioned table 1 and Fig. 11 the summarized performance has been shown.

Table 1. Summarized Performance of the Proposed Model

Methodology	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	ROC-AUC (%)
Gradient Boosting Classifier	80.83	73.6	83.2	89.1	90.72
Optimized Gradient Boosting applying BAT algorithm	84.93	76.49	89.66	81.87	94.30

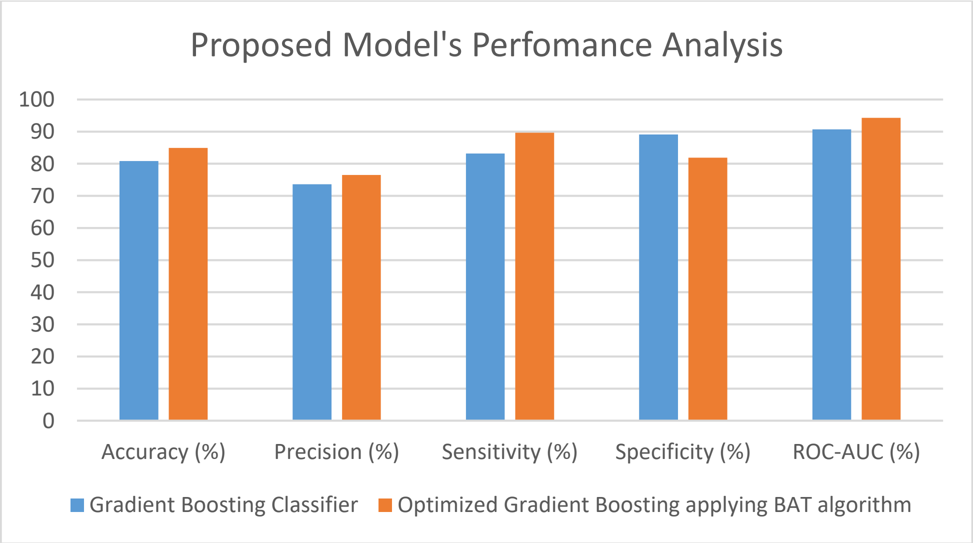

Figure 11. Summarized Performance of proposed model

Measuring Robustness of the Proposed Model

Robustness of any machine learning model can be evaluated by deviation of metrics between training and validation error & deviation between the ROC-AUC score of training and validation set. The training error for this proposed model has been identified as 0.113 and the validation error has been identified as 0.150. The ROC-AUC score of the proposed model over the training dataset has been found as 0.9473 whereas over the validation dataset it has been found as 0.943. Performance of the proposed model over training set and validation set has been shown in the Fig. 12 below.

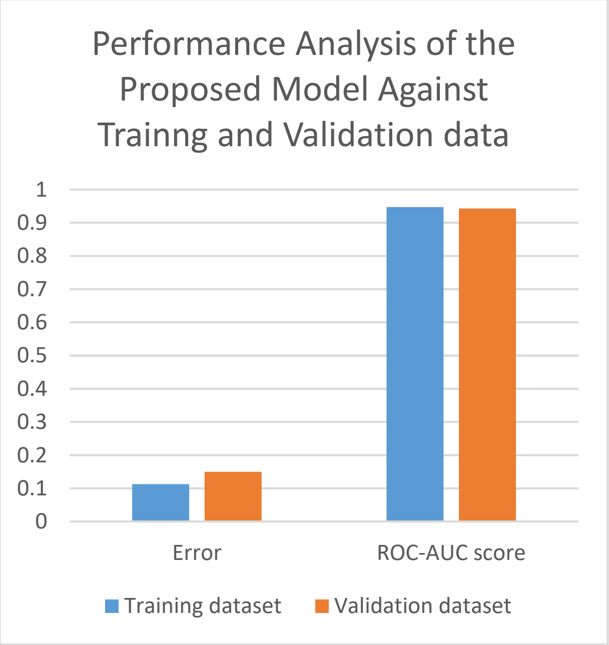

Figure 12. Model Robustness measure.

Based on the aforementioned measures, it has been noticed that the performance of the proposed optimized model exhibits a high degree of consistency across both the training and validation datasets. The observed consistency suggests that the model being offered exhibits robustness and resilience in the face of data variances. Notwithstanding various obstacles and variations in the training and validation sets, the optimized model consistently produces dependable outcomes. The observed stability is a positive indication, implying that the model's performance is not significantly affected by individual data points or random fluctuations, but rather effectively captures the fundamental patterns and exhibits strong generalization capabilities when applied to unseen data. In general, the persistent performance of the suggested optimized model serves to strengthen its robustness in the prediction of heart disease.

Model Validity and Reliability

Validity primarily concerns itself with whether research instruments measure what they are supposed to measure and whether the conclusions and outcomes of data analysis accurately reflect the real world from which the data were gathered. The primary focus of reliability is the persistence of the research findings, data analysis outcomes, and research instrument measurements. In case of machine learning based classifier, validity and reliability are measured through training error and test error respectively. In this proposed model training error has been calculated as 0.113 and test error has been identified as 0.150. These two values are not moderate enough to justify the validity and reliability of a machine learning based model.

In the Table 2 below a comparative analysis of our proposed model with few existing models has been analyzed.

Table 2. Comparative Analysis

Proposed Work	Observation
[2]	Method used: Logistic Regression and KNN No Optimization technique employed Sensitivity, Specificity, ROC-AUC score not measured and analyzed.
[3]	Isolation Forest and RobustScaler used for feature selection Deep learning method used as classifier.
[4]	Random Forest with Logistic Regression were used for feature evaluation. Details description regarding the use of python in machine learning based disease prediction algorithm development was presented. No optimization method incorporated Lack of result analysis with different metrics.

[5]	Classification and Regression Tree were used No optimization method applied High sensitivity score but poor specificity score of 63.3%
[9]	Methodology used: Support vector machine (SVM), Gaussian Naive Bayes, logistic regression, LightGBM, XGBoost, and random forest algorithm No Optimization method used. Booting Model Accuracy-Light GBM-77.04% and XGB 73.77% AUC Score of Boosting Methods: Light LGB 0.76 and XGB 0.72
[11]	Methodology used: Random Forest, Decision Tree, Logistic Regression, Support vector machine (SVM), K-nearest neighbors (KNN) No Optimization method included. Precision: 85% ROC AUC score of 0.8675
[our proposed model]	Gradient Boost applied for classification BAT Algorithm used for model optimization Accuracy: 84.93% Specificity: Sensitivity: ROC-AUC Score:0.94

It has been observed from the afore mentioned comparative analysis that our proposed model has achieved better performance considering different evaluation measures as accuracy, sensitivity and ROC-AUC score. Our model has also been optimized applying nature inspired BAT algorithm as an added feature.

It has been identified through this experiment, that the Mutual Information as feature selection, Interquartile Index as outlier removal, Gradient Boosting as classifier, and BAT algorithm as optimizer can be regarded as a good combination in the prediction of heart disease.

Conclusion

The expeditious and precise identification of cardiovascular disease possesses the capacity to potentially rescue a substantial quantity of individuals. This paper presents a heart disease prediction model that utilizes a gradient boosting classifier, with the objective of facilitating early diagnosis of the ailment. The combination of the Interquartile Index and Mutual Information has been utilized to effectively eliminate outliers and identify relevant attributes. The aforementioned combination has exhibited a degree of precision reaching 80.83%. The implementation of the BAT algorithm has led to a significant improvement in precision, with a noteworthy accuracy rate of 84.93%. This model stands as a significant contribution to machine learning in disease prediction, particularly promising for populations in rural regions with limited access to healthcare facilities. The study's specific aim of facilitating early diagnosis aligns seamlessly with the achieved results, showcasing the practical value of the proposed model in advancing the timely identification of cardiovascular ailments and, consequently, contributing to improved healthcare outcomes.

The practical value of the methodology resides in its direct application to healthcare, rather than just in its technological improvements. The research helps medical professionals make educated judgments by customizing the model to anticipate cardiac disease, which may facilitate early diagnosis and individualized risk assessment.

The heart disease prediction model, while strong, has limitations. The use of the UCI dataset may introduce biases. Realtime data collection and monitoring has not being addressed through this work.

The future scope encompasses addressing real-time implementation assessments, ensemble techniques, and advanced feature engineering. In order to refine the model for practical usage in actual healthcare settings, collaboration with healthcare experts for user feedback is essential.

Acknowledgment

We are grateful to our organization to motivate us to carry out the work.

Authors' Declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been included with the necessary permission for re-publication, which is attached to the manuscript.
- No animal studies are present in the manuscript.
- No human studies are present in the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at JIS College of Engineering.

Authors' Contribution Statement

AB has analyzed data, design the methodology, computed the experiment, tested the model and written the first draft of the paper. SG has checked the model, edited the manuscript and supervised the work. SM and DS have supervised the overall work.

References

1. World Health Organization. World Health Organization home/Health topic/cardiovascular disease. www.who.net. 2021.
2. Boukhatem C, Youssef HY, Nassif AB. Heart disease prediction using machine learning. In: 2022 Advances in Science and Engineering Technology International Conferences (ASET). IEEE; 2022.https://doi.org/10.1109/ASET53988.2022.9734880.
3. Jindal H, Agrawal S, Khera R, Jain R, Nagrath P. Heart disease prediction using machine learning algorithms. IOP Conf Ser: Mater Sci Eng. 2021 Jan 1; 1022(1): 012072.https://doi.org/10.1088/1757-899X/1022/1/012072
4. Bharti R, Khamparia A, Shabaz M, Dhiman G, Pande S, Singh P. Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. Comp Intell Nurosci. 2021 Jul 1; 2021: 1-7. https://doi.org/10.1155/2021/8387680
5. Chang V, Bhavani VR, Xu AQ, Hossain M. An artificial intelligence model for heart disease detection using machine learning algorithms. Healthc Analytics. 2022 Nov; 2: 100016.https://doi.org/10.1016/j.health.2022.100016.
6. Melillo P, De Luca N, Bracale M, Pecchia L. Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability. IEEE J Biomed Health Inform. 2013 May; 17(3): 727-33. Https://doi.org/10.1109/jbhi.2013.2244902.
7. Ramprakash P, Sarumathi R, Mowriya R, Nithyavishnupriya S. Heart disease prediction using deep neural network. In: 2020 International Conference on Inventive Computation Technologies (ICICT). IEEE;

2020.https://doi.org/10.1109/ICICT48043.2020.9112443.

8. Gárate-Escamila AK, Hajjam El Hassani A, Andrès E. Classification models for heart disease prediction using feature selection and PCA. Inform Med Unlock. 2020; 19: 100330.https://doi.org/10.1016/j.imu.2020.100330.
9. Divya K, Akash Sirohi, Sagar Pande, Rahul Malik. An IoMT Assisted Heart Disease Diagnostic System Cognitive Internet of Medical Things for Smart Healthcare. Springer, Cham; 145-161. https://doi.org/10.1007/978-3-030-55833-8_9
10. Karthick K, Aruna SK, Samikannu R, Kuppusamy R, Teekaraman Y, Thelkar AR. Implementation of a Heart Disease Risk Prediction Model Using Machine Learning. Comput. Math Methods Med. 2022 May 2; 2022: 1-14. https://doi.org/10.1155/2022/6517716
11. Pathak Y, Shukla P, Tiwari A, Stalin S, Singh S, Shukla P. Deep Transfer Learning Based Classification Model for COVID-19 Disease. Ing Rech Biomed. 2022 Apr; 43(2): 87-92. https://doi.org/10.1016/j.irbm.2020.05.003.
12. Kumar NK, Sindhu GS, Prashanthi DK, Sulthana AS. Analysis and prediction of cardio vascular disease using machine learning classifiers. In: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE; 2020. https://doi.org/10.1109/ICACCS48705.2020.9074183
13. Nukala BT. Heart disease classification comparison among patients and normal subjects using machine learning and artificial neural network techniques. Int J Biosens Bioelectron. 2021; 7(3):77-79 . https://doi.org/10.15406/ijbsbe.2021.07.00216
14. Kareem AK, AL-Ani MM, Nafea AA. Detection of Autism Spectrum Disorder Using A 1-Dimensional Convolutional Neural Network. Baghdad Sci J. 2023; 20(3(Suppl.): 1182. https://doi.org/10.21123/bsj.2023.
15. Zaki SM, Jaber MM, Kashmoola MA. Diagnosing COVID-19 Infection in Chest X-Ray Images Using Neural Network. Baghdad Sci J. 2022 Dec 1; 19(6): 1356.https://doi.org/10.21123/bsj.2022.5965.
16. Mitra S, Majumder AB, Saha T. An observation and analysis the role of Convolutional Neural Network towards lung cancer prediction. Baghdad Sci J. 2023; 20(6(Suppl.)): 2568.
17. Majumder AB, Gupta S, Singh D. An ensemble heart disease prediction model bagged with Logistic Regression, naïve Bayes and K Nearest Neighbour. J Phys Conf Ser. 2022; 2286(1): 012017.https://doi.org/10.1088/1742-6596/2286/1/012017
18. Janosi A, Steinbrunn W, Pfisterer M, Detrano R. Heart Disease. UCI Machine Learning Repository; 1989.https://doi.org/10.24432/C52P4X.
19. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001; 29(5). https://doi.org/10.1214/aos/1013203451
20. Yang XS. A new metaheuristic bat-inspired algorithm. In Nature inspired cooperative strategies for optimization. NICSO. arXiv:1004.4170. 2010 Nov: 65-74. http://dx.doi.org/10.48550/arXiv.1004.4170
21. Yang X-S, Chien SF, Ting TO. Bio-inspired computation and optimization. In: Bio-Inspired Computation in Telecommunications. Elsevier. 1st Ed 2015; p. 1–21. https://doi.org/10.1016/B978-0-12-801538-4.00001-X.

Published Online First: June, 2024
https://doi.org/10.21123/bsj.2024.9973
P-ISSN: 2078-8665 - E-ISSN: 2411-7986

Baghdad Science Journal

نموذج تجميعي للتنبؤ بأمراض القلب والأوعية الدموية باستخدام التحسين المستوحى من الطبيعة

أنويشا بانيرجي[1]، سومسوبرا غوبتا[2]، سوراف ماجومدير[3]، دارمبال سينغ [4]

[1]قسم تكنولوجيا المعلومات، كلية الهندسة JIS، كالياني، الهند.
[2]قسم علوم و هندسة الحاسوب، جامعة سوامي فيفيكاناندا، باراكبور، الهند.
[3]كابجيميني الهند، كولكاتا، الهند.
[4]قسم علوم و هندسة الحاسوب، كلية الهندسة JIS، كولكاتا، الهند.

الخلاصة

قدم هذا البحث نموذجًا تنبؤيًا مبتكرًا لأمراض القلب، تم تحسينه من خلال تطبيق منهجية تحسين أفضل التقنيات المتاحة. من المعترف به عالميًا أن انتشار أمراض القلب والأوعية الدموية، خاصة في الدول التي تواجه تحديات اقتصادية، هو مساهم كبير في معدلات الوفيات العالمية.يتم التأكيد على ضرورة التنبؤ الدقيق بالنتائج في الوقت المناسب من خلال ضرورة التخفيف من العواقب الوخيمة المحتملة المرتبطة بهذا التحدي الصحي المحدد. تتضمن المنهجية اختيار الميزات من خلال المعلومات المتبادلة، يليها التخلص من العناصر الخارجية باستخدام نهج Inter Quartile Range أثناء تنقية البيانات.لمعالجة التحيز المحتمل من اختلافات المقياس المتغير، يتم تنفيذ توحيد الميزات باستخدام طريقة .Standard Scaler يتم استخدام تقنية تعزيز التدرج لتطوير النماذج، المعروفة بقدرتها على إدارة البيانات المفقودة وإنشاء تنبؤات دقيقة. لمزيد من تعزيز الأداء، تم تقديم خوارزمية BAT، مع الاستفادة من التحسين المستوحى من الطبيعة. وقد أدى تطبيق أسلوب أفضل التقنيات المتاحة في هذا النموذج بالذات إلى تحسن ملحوظ في الأداء، مما أدى إلى معدل دقة قدره 84.94%. وكانت درجات الدقة والنوعية والحساسية للنموذج 76.47%، 81.88%، و89.65%، على التوالي. وتشير هذه المقاييس مجتمعة إلى أداء متوازن.

الكلمات المفتاحية: خوارزمية أفضل التقنيات المتاحة، أمراض القلب والأوعية الدموية، مصنف المجموعة، تعزيز التدرج، المدى الرباعي، التحسين المستوحى من الطبيعة.