INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

An Advanced Model to Predict Heart Disease Applying Random Forest Classifier and Whale Optimization Algorithm

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

1 Department of Information Technology, JIS College of Engineering, Kalyani, West Bengal, India
2 Department of Science, Swami Vivekananda University, Kolkata, West Bengal, India
3 Department of Computer Science and Engineering, JIS College of Engineering, Kalyani, West Bengal, India
4 Capgemini India, Kolkata, West Bengal, India

OPEN ACCESS

Received: 13-07-2023
Accepted: 11-09-2023
Published: 11-11-2023

Citation: Majumder AB, Gupta S, Singh D, Majumder S (2023) An Advanced Model to Predict Heart Disease Applying Random Forest Classifier and Whale Optimization Algorithm. Indian Journal of Science and Technology 16(43): 3679-3690. https://doi.org/10.17485/IJST/v16i41.1756

* **Corresponding author**.

annwesha.banerjee@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background: In this article a model for heart disease prediction has been proposed using machine learning to address the significant concern of lives lost due to this disease, especially in remote and underprivileged areas where access to proper medical support is lacking. Identifying the disease at an early stage is crucial for preventing unnecessary fatalities. **Methods:** The dataset has been collected from UCI data repository. The raw data set consists of 14 features, which are age, sex, cp, tresbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, thal, ca, target. The collected dataset has been standardized applying StandardScaler and then feature selection has been carried out applying SelectKBest to select 9 most competent features. The selected features are age, sex, cp, thalach, exang, oldpeak, slope, ca, thal and target. 80% of the dataset has been used for training and remaining for testing. For classification Random Forest has been applied and the model performance has been increased applying Whale Optimization Algorithm. **Findings:** The WOA, inspired by the social behavior of humpback whales, has been utilized to optimize the parameters and configurations of the model. The proposed model has achieved 86.53% accuracy. **Novelty:** The proposed model combines the Random Forest classifier, with the SelectKBest to select optimal number of features through experiments which ensures that the model is trained only with effective features. This dimensionality reduction makes the model computationally efficient. More over application of Whale Optimization has increased the performance of model with 0.93 AUC score. This model's novelty lies in its targeted focus on undeserved populations, aiming to improve early identification of heart disease and reduce fatalities in resource constrained areas.

Keywords: Heart Disease Prediction; Random Forest Classifier; Standardscaler; Whale Optimization; SelectKBest

1 Introduction

Machine learning can indeed be utilized to address the problem of identifying and predicting heart disease. It can be used to develop models that classify patients as healthy or diseased, predict the risk of developing heart disease, and assist healthcare professionals in making treatment decisions. Additionally, machine learning can aid in remote monitoring of patients and provide decision support in areas with limited medical support. A plenty of work has already been done in the field of disease prediction using machine learning methods. In this section some of the works have analyze to understand and identify the working principals and impact of the works in this field. In [(1)] authors have proposed a model using Random Forest, SVM, Naive Bayes and Decision Tree for prediction of heart disease. The proposed work was carried out on the dataset collected from UCI data repository. Authors also experimented the correlation among features. It was identified that Random Forest achieved highest accuracy in lesser time. A novel heart disease prediction model has been proposed in [(2)]. Authors used Weighted Associative Rule Mining for calculating feature score. The model was trained by data collected from UCI data repository. In this work the detailed analysis of different features were represented graphically. Rohit Bharti et al. proposed another efficient model using deep learning approaches. The authors presented detailed data analysis in this paper. Random Forest, Logistic Regression, K Nearest Neighbors, Support Vector Machine, Decision Tree and XGBoost were used for classifications. By applying feature selection and outlier removal the proposed work achieved an average AUC score of 80% [(3)]. Another Intelligent heart disease prediction model was proposed by V Chaig et al. in [(4)]. Different machine learning methods applied and analyzed through their experiments were Logistic Regression, Support Vector Machine, Decision Tree, Random Forest and K Nearest Neighboring [(5)]. Another efficient model was proposed where authors applied lazy, meta, tree, bayes and rules based different machine learning models. Hyper parameter tuning was done for the K value of nearest neighbor algorithm. The best accuracy achieved through the model was 86.46%. In [(6)] another intelligent model was proposed by authors using KNN, Random Forest, Naïve Bayes and Decision Tree. In this work the missing value issue was addressed. Six rows were deleted from the dataset as they had missing values. Rina S. Patil & Mohit Gangwar proposed a model for prediction of heart disease using Recurrent Neural Network and deep learning techniques Authors applied many feature selection methods also [(7)]. In [(8)] an efficient algorithms was proposed by authors for heart disease prediction. Particle Swarm Optimization was being used for identifying cardiac sick people. Along with that K Means clump and K Nearest Neighbour were also being utilized. In [(9)] another efficient model was proposed by authors applying Support Vector Machine, Logistic Regression, Random Forest and Naïve Bayes. The data set was collected from the Kaggle. Authors did not apply any dataset preprocessing. Accuracy, Sensitivity and Specificity were considered as performance measuring matrices. The achieved accuracy was in between 58.71% to 77.06%. Senthilkumar M, Chandra Segar T, Srivastava G proposed a model of heart disease prediction where they applied Hybrid Random Forest with linear Model and achieved 88.7% accuracy. Different feature selection methods were also applied in this work to achieve a better accuracy [(10)]. Fahed Shah A. et al. proposed another heart disease prediction model applying Decision Tree, Random Forest, Logistic Regression, Naïve Bayes and Support vector Machine. The model was trained with Claveland dataset. Highest accuracy was achieved 86.88% through Naïve Bayes and least accuracy of 78.69% was achieved through Decision Tree [(11)] . C.M Bhatt et al. proposed a model for heart disease prediction by applying k-modes

clustering with Huang. Random Forest (RF), decision tree classifier (DT), multilayer perception (MP), and XGBoost (XGB) were used. GridSearchCV machine learning model was also utilized in this proposed model. It was claimed in the work that applying cross validation the accuracy was increased for each classifiers [12]. Niloy Biswas et al. proposed a model for heart disease prediction applying different feature selection mechanisms. Chi-square, ANOVA, and Mutual Information were being applied for feature selection where as Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Random Forest, Naive Bayes and Decision Tree were applied as classifier. It was identified through the proposed work that Mutual Information worked best among other as feature selector. [13]. Taher M. Ghazal et al. proposed a heart disease prediction model applying Support Vector Machine and K Nearest Neighbour. It was claimed by the authors that their proposed model worked better than NeiveBayes used in their previous work [14]. A machine learning based model was proposed by Subramani S et al. applying stacking methods. Random Forest, Logistic Regression, Multilayer Perceptron, Extra Tree, and CatBoost classifiers were used as base learner. The performances of the model were being justified through Precision, Recall, F1 score, and Area Under the Curve (AUC). Average AUC score achieved through the model was 0.80 [15]. By combining Logistic Regression, Naive Bayes, K Nearest Neighbor, Support Vector Machine, Kernel SVM, Random Forest, and Artificial Neural Networks, Majumder et al. proposed an explainable hybrid technique for heart disease prediction. The best classifier was determined by accuracy, sensitivity, and specificity. SHAP and LIME were used to get detailed insight of the proposed model [16]. Annwesha Banerjee et al. proposed an estimating model of bagging techniques-based cardiac disease prediction. As its basic learner, this proposed model used Naive Bayes, K Nearest Neighbor, and Logistic Regression. The data gathered from the UCI data repository were used to train the suggested model. The goal of the work was to apply numerous bagged classifiers in order to improve the prediction outcome. Gaussian Naive Bayes, K Nearest Neighbor, and Bagged Logistic Regression all showed accuracy of 82.8%, 82.5%, and 83.2%, respectively [17]. Ay, Ş et al. proposed a heart disease prediction model applying different machine learning methods. The novelty of the work was combination of different meta heuristic optimization methods applied for feature selection. Authors used Cuckoo Search, FlowerPollination Algorithm, Whale Optimization algorith and Harris Hawks optimization. For classification of disease authors applied K Nearest Neighbour, Gaussian Naïve Bayes, Random Forest, Logistic Regression and Support Vector machine [18]. Madhumita Pal et al. proposed a heart disease prediction model using Random Forest. The dataset was collected from Kaggle to train the model. The dataset correlation was put in the paper graphically. No feature selection or data standardization methods were applied. This proposed model achieved 86.9% accuracy [19]. L.M. Luthimath et al. implemented a model for heart disease prediction applying Random Forest. The paper likely included details about the methodology used to collect and preprocess the data, the implementation of the Random Forest algorithm for heart disease prediction, and the evaluation of the model's performance [20]. P. Dhaka et al. proposed advanced model for heart disease prediction utilizing deep learning mechanism. Deep Bidirectional Long Short Term Memory was used as classifier where the tuning of hyper parameter was done by Whale-Marine Optimization. This proposed model also had applied Elliptic Curve Cryptography dependent Diffi-Huffman algorithm for maintaining data security [21]. Xi Wei proposed a model for heart disease prediction applying Categorical Boosting. Over that Sparrow search algorithm based on salp swarm algorithm, OBL and Lateral mutation strategy were used for parameter optimization. The proposed model achieved an average accuracy of 85% [22]. Asadi S et al. proposed a model for heart disease prediction applying Random Forest and Multi-Objective Particle Swarm Optimization. Applying this proposed model on different dataset on an average 86% accuracy was achieved [23]. Manyala Naga Sailaja proposed an optimized model for heart disease prediction. Different machine learning algorithms like support vector machines. (SVM), K Nearest Neighbors (KNN), Naive Bayes (NB), Artificial Neural Networks (ANN), and Random Forest (RF) were applied over which Genetic Algorithm with Particle Swarm Optimization was applied [24].

Upon careful examination of the research methodologies outlined above, it becomes evident that certain approaches primarily prioritized feature selection, while others employed a combination of multiple classifiers followed by the selection of the most suitable one. Moreover, noteworthy achievements in terms of accuracy were attained by some through the strategic implementation of deep learning techniques. However, it is noteworthy that a subset of the proposed models fell short in achieving commendable levels of accuracy and AUC scores. Nevertheless, a crucial aspect that remains conspicuously absent across these methodologies is the incorporation of a secondary layer of optimization aimed at further enhancing performance. This additional layer of optimization could potentially yield significant improvements in the predictive capabilities of the models, leading to heightened accuracy and more robust AUC scores. Addressing this issue an Optimized machine learning based model utilizing Whale Optimizer, Random Forest and SelectKBest has been proposed in this paper. The proposed study has implemented the practice of data standardization, reduced the dimensionality of the data, and improved the process. This work makes significant contributions to the field of heart disease prediction. An advanced machine learning-based model has been introduced specifically designed for this purpose. One of the key contributions is the use of feature selection technique, which help identify and select the most relevant features from the dataset. This not only improves accuracy but also reduces the dimensionality of the data. The model also utilizes the Random Forest algorithm, an ensemble method that combines

multiple decision trees for classification. This approach enhances the robustness and accuracy of the predictions. Additionally, the application of the Whale Optimization Algorithm further enhances the model's performance by optimizing its parameters. The proposed model achieves accuracy rate of 86.53% in predicting heart disease, demonstrating its effectiveness and potential impact in early detection and prevention efforts. Overall, this work provides a novel and comprehensive approach to heart disease prediction, contributing to the advancement of machine learning applications in the healthcare domain.

In section 2 the proposed model has been described in details. Result analysis and model performance has been discussed and depicted in section 3 and in section 4 conclusion and future scope has been discussed in.

2 Methodology

The proposed model focuses on the early detection of heart disease by utilizing several key techniques. Firstly, the Random Forest algorithm applied for classification, which leverages an ensemble of decision trees to improve accuracy and robustness. Secondly, the SelectKBest method has been employed for feature selection, ensuring that only the most relevant features are considered, thereby enhancing the model's performance. Lastly, the Whale Optimization Algorithm has been utilized to achieve enhanced accuracy by optimizing the model's parameters. By combining these approaches, the proposed model aims to provide an effective and accurate solution for early detection of heart disease.

The detail block diagram of the model is shown in the Figure 1 below.

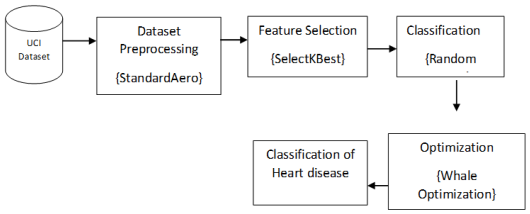

Fig 1. Model Block Diagram

2.1 Data Collection and Description

The dataset on which the model has built and tested has been collected from UCI[25]. The dataset consists of 14 features as described in below mentioned Table 1.

Table 1. Dataset Description

Features	Descriptions
Age	This is continuous field that holds the age of patient
Sex	Categorical field identifies male or female
Cp	This stands for Angina and consisting values for asymptomatic, atypical angina, non angina pain and typical angina
Tresbps	Continuous field of blood pressure.
Chol	This is a continuous field represents the cholesterol values
Fbs	This represents the blood sugar level.
Restecg	This field describes the echocardiogram when patient is at rest.
Thalach	This represents patient's heart rate at stress situation.
Exang	This categorical field represents if a patient has angina during stress or not.
Oldpeak	This represents decrease of ST segment during exercise time.
Slope	This field shows ST segments during exercise.
Thal	This represents the blood flow through radioactive dye.
Ca	vessels colored by the radioactive dye is represented by the field.
Target	It is the identification of whether a patient has heart disease or not.

2.2 Data Preprocessing

In the presented phase, the data underwent standardization using the StandardScaler method. StandardScaler is a data preprocessing technique that transforms the data by subtracting the mean and scaling it to have a variance of 1.

For each feature (column) in the dataset, the StandardScaler subtracts the mean and scales the values through dividing by the standard deviation.

The formula for StandardScaler can be represented mathematically as equation no 1:

$$z = \frac{(x-u)}{s} \tag{1}$$

Where:

z is the standardized value of the feature

x is the original value of the feature

u is the mean (average) of the feature

s is the standard deviation of the feature

By applying this equation to each feature in the dataset, the StandardScaler ensures that the transformed data has a mean of 0 and a standard deviation of 1, resulting in a standardized distribution.

2.3 Feature Selection

In the proposed model, out of the initial 13 independent features, SelectKBest has been utilized to choose the eight most significant independent features. By selecting these relevant features, the model's accuracy and performance are enhanced as they capture the most informative and impactful aspects related to the target variable.

Overall, feature selection is a crucial step in building accurate models as it helps to focus on the most relevant features, improves interpretability, and reduces the dimensionality of the dataset. In this model, SelectKBest is specifically applied to select the eight most significant independent features from the original set of 13 features.

The features are selected based on K highest score. The feature scores have shown in the Figure 2.

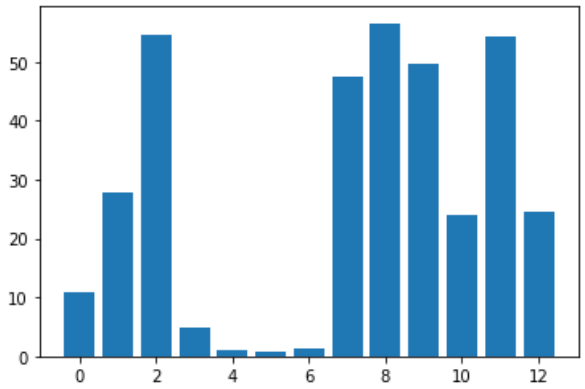

Fig 2. Feature Importance based on SelectKBest

The list of feature selected are age, sex, cp, thalach ,exang, oldpeak, slope, ca, thal. The selected feature histogram has shown in Figure 3.

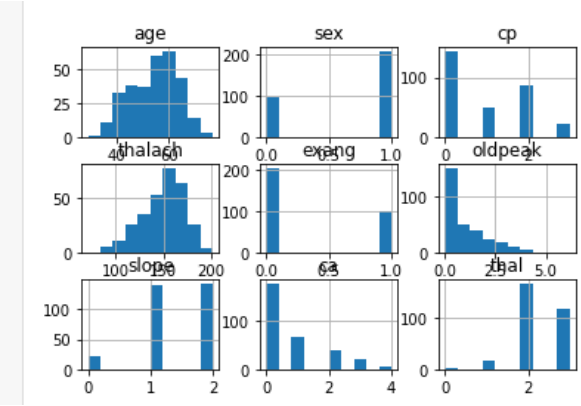

Fig 3. Selected Feature set Histogram

2.4 Applying Random Forest for Classification

In this step of the proposed model, the Random Forest method has been applied to the selected feature set. Random Forest is a supervised ensemble machine learning technique that belongs to the bagging category of ensemble mechanisms.

Ensemble learning combines the predictions of multiple individual models to make a final decision. Bagging and Boosting are the two main categories of ensemble mechanisms. In the case of Random Forest, it falls under the bagging category.

Random Forest builds multiple decision trees on randomly sampled subsets of the dataset. Each decision tree is trained independently on a different subset of the data. This process is performed in parallel, making efficient use of computing power. The true strength of the Random Forest algorithm lies in the combination of multiple decision trees, resulting in improved accuracy, robustness, and generalization capability compared to a single decision tree.

2.5 Applying Whale Optimization Algorithm

In this phase, the Whale Optimization Algorithm (WOA) has been applied to enhance the performance of the model. The Whale Optimization Algorithm is a metaheuristic optimization algorithm inspired by the social behavior of humpback whales.

The goal of applying the Whale Optimization Algorithm is to optimize the parameters or configurations of the model, ultimately improving its accuracy and efficiency. The algorithm mimics the hunting behavior of whales, where they collaborate and adapt their movements to find the best feeding spots.

The Whale Optimization Algorithm iteratively updates the solutions based on three main operators:

1. Encircling prey: In this operator, a whale adjusts its position towards a better solution by encircling the potential prey (i.e. the optimum solution). This helps explore the search space efficiently.

2. Bubble-net feeding: This operator simulates the cooperative behaviour of whales. Whales form a group and create a bubble-net to enclose and capture prey. In the context of optimization, this operator promotes exploitation of the best solutions found so far.

3. Searching for prey: This operator mimics the exploration behaviour of whales. Whales move randomly to search for better solutions and avoid local optima.

By applying the Whale Optimization Algorithm, the model iteratively optimizes its parameters or configurations to find the best possible solution within the search space. This optimization step aims to enhance the model's performance, accuracy, and ability to find an optimal solution for heart disease prediction.

In this proposed experiment N no of whale position has been chosen where N is 5. Each position of whale represents each feature subset. A sample feature subset were as follows

Subset 1:{age,sex,cp,thalach, exang}

Subset 2:{age,sex,cp,thalach, oldpeak}

…

Subset N:{exang, oldpeak, slope, thal ,age}

The accuracy function has been chosen as fitness function for the Whale Optimization experiments which has been represented in equation no 2.

$$Accuracy = \frac{TN + TP}{TN + FN + TP + FP} \tag{2}$$

Where:

TP (True Positives) is the number of correctly predicted positive instances.
TN (True Negatives) is the number of correctly predicted negative instances.
FP (False Positives) is the number of incorrectly predicted positive instances.
FN (False Negatives) is the number of incorrectly predicted negative instances

3 Results and Discussion

Overall, the proposed model follows a systematic approach by first standardizing the data, then selecting the most important independent features using SelectKBest, and finally employing Random Forest for the classification task. This combination of techniques aims to improve the accuracy and reliability of the model in identifying the presence of heart disease in patients.

Accuracy, recall, precision, and F1 score are essential metrics for evaluating the performance of classification models. Accuracy measures the overall correctness of predictions, representing the ratio of correct predictions to the total number of predictions made. It provides a general assessment of the model's performance across all classes. Recall, also known as sensitivity or true positive rate, focuses on the model's ability to correctly identify positive instances. It measures the ratio of true positives to the sum of true positives and false negatives, indicating how well the model avoids false negatives. Precision quantifies the accuracy of positive predictions, evaluating the ratio of true positives to the sum of true positives and false positives. It reflects the model's ability to avoid false positives. The F1 score combines precision and recall into a single metric by calculating the harmonic mean of the two measures. It provides a balanced assessment of a model's performance, particularly useful in imbalanced datasets. Together, these metrics provide a comprehensive understanding of a classification model's effectiveness, considering accuracy, correctness, sensitivity, and precision in its predictions.

3.1 Observation 1: Model Performance applying Random Forest Classifier (Without optimization)

For feature selection in this proposed model SelectKBest has been applied. Numbers of experiments have been carried out to find the most optimal K values (no of features).

Initially the model was tested with different number of K values i.e no of features staring from 4 to 13. It has been observed that best score of 81.58% has achieved with K=9 and the features are age, sex, cp, thalach,exang, oldpeak,slope, ca, thal.

In the below mentioned Figure 4 the sample performance of the model with different number of features have been shown.

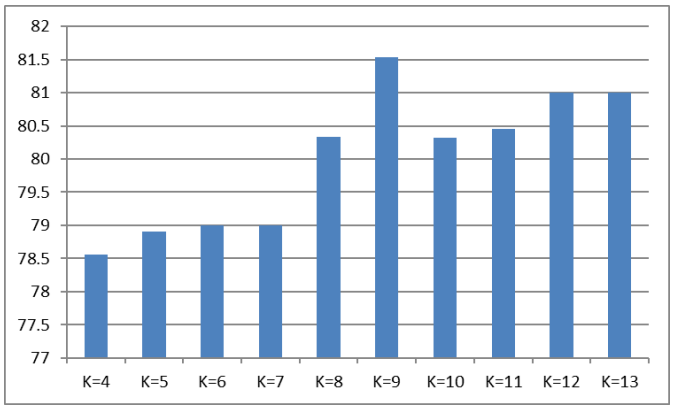

Fig 4. Selection of best K value

Applying Random Forest for heart disease prediction the model has achieved 81.58% accuracy. The confusion matrix has shown in the Figure 5.

The Accuracy, Recall, Precision and F1 score have been shown in the Figure 6 below through sample program execution.

Table 2 summarized performance of the model has been shown based on which a graph has been put in Figure 7.

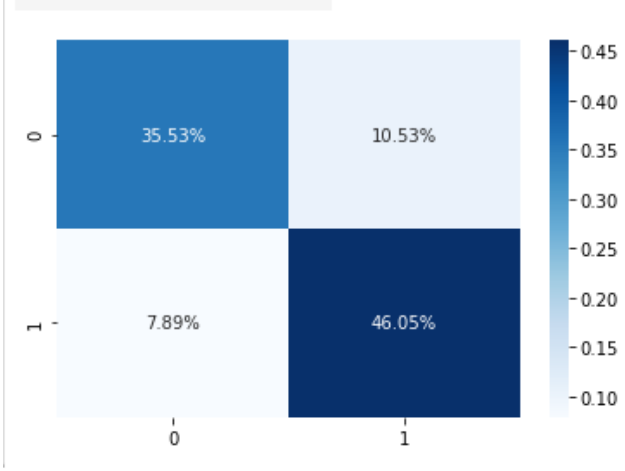

Fig 5. Confusion Matrix of Proposed Model

```
accuracy = accuracy_score(y_test, y_pred)
recall = recall_score(y_test, y_pred, average='weighted')
precision = precision_score(y_test, y_pred, average='weighted')
f1 = f1_score(y_test, y_pred, average='weighted')
y_pred_proba = RForest.predict_proba(X_test)[::,1]
auc = roc_auc_score(y_test, y_pred_proba)
print(f"Accuracy: {accuracy:.4f}")
print(f"Recall: {recall:.4f}")
print(f"Precision: {precision:.4f}")
print(f"F1 Score: {f1:.4f}")
print(f"AUC Score: {auc:.4f}")
```

```
Accuracy: 0.8158
Recall: 0.8182
Precision: 0.7714
F1 Score: 0.7942
AUC Score: 0.8894
```

Fig 6. Performance Measure–Sample Execution

Table 2. Summarized Performance of Model

Accuracy	Recall	Precision	F1 Score	AUC Score
0.8158	0.8182	0.7714	0.7942	0.8894

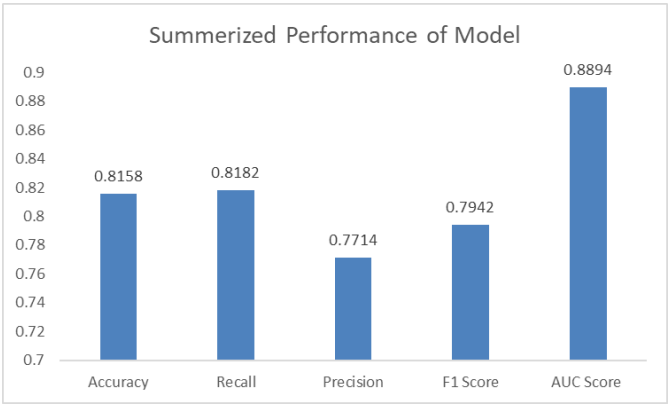

Fig 7. Graphical presentation of Random Forest Model's Performance (without optimization)

3.2 Observation 2: Applying Random Forest Classifier and Whale Optimization Algorithm

In this phase Whale Optimization Algorithm has been applied over Random Forest for performance enhancement.

It has been observed that accuracy score which has been used as fitness score for that optimization has been increased to 86.53%. The achieved AUC score of the optimized model is 0.93.

In the Figure 8 below sample execution on different iteration the whale positions (different subset of features) has been shown.

```
def easom(variables_values = [0, 0]):
    RForest.fit(X_train, y_train)
    y_pred = RForest.predict(X_test)
    accuracy = np.mean(y_pred == y_test)
    return accuracy
woa = whale_optimization_algorithm(target_function = easom, **parameters)
```

```
Iteration =  0  f(x) =  0.8320123452453522
Iteration =  1  f(x) =  0.8393185632193421
Iteration =  2  f(x) =  0.8431995739953211
Iteration =  3  f(x) =  0.8505532667321996
Iteration =  4  f(x) =  0.8578753219656390
Iteration =  5  f(x) =  0.8601882356174527
Iteration =  6  f(x) =  0.8653212188235611
Iteration =  7  f(x) =  0.8653212188235611
```

Fig 8. Sample Execution of Optimization

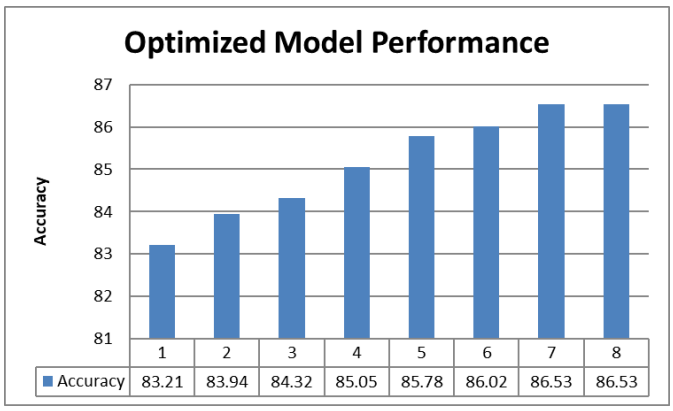

Fig 9. Sample execution of Whale Optimization –AccuracyScore

The ROC curve and AUC are commonly used to evaluate and compare the performance of different classification models. They provide a clear visual representation of the model's ability to correctly classify positive instances while minimizing false positives across various thresholds. The ROC curve of our proposed model is shown in Figure 10.

The proposed model utilized the effectiveness of Random Forest, SelectKBest and Whale Optimization for achieving a good level of performance in predicting heart disease.

SelectKBest has been applied for feature selection where K stands for no of features. SelectKBest is a key feature selection technique in machine learning, enhancing model performance by identifying and retaining essential features while discarding irrelevant ones. This reduces complexity, boosts efficiency, and prevents overfitting, resulting in improved accuracy and faster training. Careful selection of the right number of features (K) is vital for optimal results. In this work the model's performance has been analyzed on each and every K values from 4 to 13 and best accuracy achieved on K=9. Then Random Forest has been applied for classification. Random Forest is a potent machine learning technique that leverages multiple decision trees to achieve high accuracy. It adeptly handles diverse data, reduces overfitting and provides valuable feature insights. Its outlier resilience, capacity to grasp complex relationships, and user-friendly implementation add to its allure. Then Whale Optimization Algorithm has been used to achieve better performance over the Random Forest classifier. The Whale Optimization Algorithm (WOA), which mimics humpback whales' hunting behaviour, provides a particular edge in machine learning. This method effectively explores the solution space while adjusting to different optimization issues. In both exploitation and exploration, WOA succeeds, establishing a balance between enhancing local optima and looking for global optima. It is a potential option for optimization problems in machine learning due to its simplicity, lack of hyperparameters, and efficacy in handling complex,

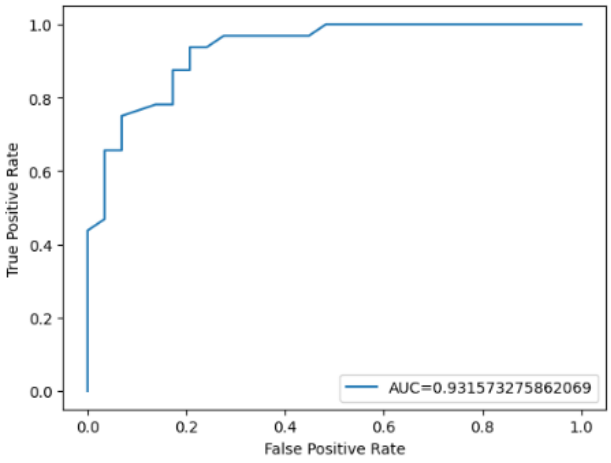

Fig 10. ROC Curve of proposed model

multimodal functions.

The proposed optimized model has achieved 86.53% accuracy and proven its significance to heart disease prediction over other different models available in this domain.

Table 3. Comparative Analysis with current work in this field

Existing work	Discussion
"Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning" [3]	Applied Isolation Forest for feature selection Achived Accuracy Random Forest is 80.3%, Logistic Regression is 83.31%, K Nearest Neighbors is 84.86%, Support Vector Machine is 83.29%, Decision Tree is 82.33%, and XGBoost is 71.4%.
"Prediction of Cardiovascular Disease using Machine Learning Algorithms" [9]	Applied Support Vector Machine , Logistic Regression, Random Forest and Naïve Bayes. No data preprocessing carried out. Accuracy achieved : 58.71% to 77.06%.
"Effective Heart Disease Prediction Using Machine Learning Techniques" [12]	Applied Rrandom Forest, Decision Tree Classifier, Multilayer Perceptron ,XGBoost (XGB) and GridSearchCV. Accuracy: ecision Tree: 86.37% (with cross-validation) and 86.53% (without cross-validation), XGBoost: 86.87% (with cross-validation) and 87.02% (without cross-validation), Random Forest: 87.05% (with cross-validation) and 86.92% (without cross-validation), Multilayer Perceptron: 87.28% (with cross-validation) and 86.94% (without cross-validation).
"Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques" [13]	Applied Chi-Square, ANOVA, and Mutual Information for feature selection and Logistic Regression, Support Vector Machine , K-Nearest Neighbor, Random Forest, Naive Bayes, and Decision Tree for classification Mutual Information was identified as most effective one.
"Heart Disease Prediction Using Machine Learning" [14]	Applied K Nearest Neighbour and Support Vector Machine. It was proven through the work that both the methods achieved better performance than Naïve Bayes
"Cardiovascular diseases prediction by machine learning incorporation with deep learning" [15]	Applied Random Forest, Logistic Regression, Multilayer Perceptron, Extra Tree, and CatBoost classifiers. Average AUC score : 0.80
"An Explainable Hybrid Intelligent System for Prediction of Cardiovascular Disease" [16]	Applied Bagging methodology with base learners Naïve Bayes, K Nearest Neighbor, and Logistic Regression. Average Accuracy : 82.5%
Our Proposed Model	Applied Chi Square Test for feature selection which helps to select accurate number of features , reduce dimensionality of dataset in turn enhance computatoion and predictive efficiency. Applied Random Forest for classification over which an layer of optimization applied using Whale Optumization. Accuracy : 86.53% AUC Score : 0.93

The preceding discourse firmly establishes the efficacy of the proposed model as a robust approach for early-stage heart disease prediction, capable of effectively reducing associated risks. The model surpasses contemporaneous alternatives in terms of both accuracy and AUC score, thereby demonstrating its superior capacity to predict the disease with heightened efficiency.

This work presents a notable contribution in proposing a machine learning-based model for predicting heart disease, specifically tailored to address the critical issue of lives lost due to the ailment, particularly in underserved regions with limited medical access. The paper underscores the crucial role of early disease identification in preventing unnecessary fatalities. The dataset collected from UCI repository, standardized and then applied SelectKBest for selecting most relevant 9 features. Training on 80% of this refined dataset and employing Random Forest for classification, the model's effectiveness is further boosted by the innovative Whale Optimization Algorithm, inspired by humpback whale behaviour. This comprehensive approach yields a notable 86.53% accuracy rate. The work's uniqueness lies in merging Random Forest and SelectKBest for efficient feature selection, leading to computational efficiency. Additionally, the introduction of the Whale Optimization significantly enhances model performance, resulting in an AUC score of 0.93. Importantly, this research pioneers a focus on underserved populations, aiming to revolutionize early heart disease detection and minimize fatalities in resource-constrained areas. In sum, this study offers a compelling blend of advanced techniques with a socially impactful mission.

4 Conclusion

The model presented in this study makes a valuable contribution to the field of machine learning in the context of disease prediction, with a specific emphasis on heart disease. The model attained an accuracy of 81.58% by employing the Random Forest classifier and preprocessing the dataset using StandardScaler. The utilization of SelectKBest for feature selection has resulted in an improvement in the performance of the model by prioritizing pertinent features. Additionally, the deployment of the Whale Optimization Algorithm has further boosted its efficacy, leading to an accuracy rate of 86.53%. The model has attained an AUC score of 0.93. All the evidence suggests that this model has the capability to accurately forecast cardiac disease.

The utilization of this model exhibits promising capabilities in facilitating the early-stage prognosis of cardiac disease, hence facilitating the implementation of timely therapies. In the future, our objective is to prioritize the acquisition of local datasets and employ deep learning techniques to enhance accuracy and propel advancements in the domain of heart disease prediction.

Acknowledgement

The authors are thankful to their institution for providing support to conduct the research work

References

1) Sharma V, Yadav S, Gupta M. Heart Disease Prediction using Machine Learning Techniques. *2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. 2020;p. 177–181. Available from: https://doi.org/10.1109/ICACCCN51052.2020.9362842.
2) Yazdani A, Varathan KD, Chiam YK, Malik AW, Ahmad WAW. A novel approach for heart disease prediction using strength scores with significant predictors. *BMC Medical Informatics and Decision Making*. 2021;21(1):194. Available from: https://doi.org/10.1186/s12911-021-01527-5.
3) Bharti R, Khampari A, Shabaz M, Dhiman G, Pande S, Singh P. Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. *Computational Intelligence and Neuroscience*. 2021;p. 11. Available from: https://doi.org/10.1155/2021/8387680.
4) Chang V, Vallabhanentrupabhavani A, Qianwenxu, Hossain M. An artificial intelligence model for heart disease detection using machine learning algorithms. *Healthcare Analytics*. 2022;2. Available from: https://doi.org/10.1016/j.health.2022.100016.
5) Reddy KVV, Elamvazuthi I, Aziz AA, Paramasivam S, Chua HN, Pranavanand S. Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators. *Applied Sciences*. 2021;11(18):8352. Available from: https://doi.org/10.3390/app11188352.
6) Karpagam S, Kaleeswari M, Kavitha K, Priyadarsini S. Heart Disease Prediction Using Machine learning Algorithm. *International Journal of Scientific Development and Research*. 2020;5(8). Available from: https://www.ijsdr.org/papers/IJSDR2008043.pdf.
7) Patil RS, Gangwar M. Heart Disease Prediction Using Machine Learning and Data Analytics Approach. In: Proceedings of International Conference on Communication and Artificial Intelligence;vol. 435. Springer Nature Singapore. 2022;p. 351–361. Available from: https://doi.org/10.1007/978-981-19-0976-4_29.
8) Nanthini K, Preethi S, Venkateshwaran S. Heart Disease Prediction Using Machine Learning Algorithms. *International Journal of Advanced Science and Technology*. 2020;29(3):9965–9965. Available from: http://sersc.org/journals/index.php/IJAST/article/view/26971.
9) Muktevisrivenkatesh. Prediction of Cardiovascular Disease using Machine Learning Algorithms. *International Journal of Engineering and Advanced Technology (IJEAT)*. 2020;(9):2249–8958. Available from: https://www.ijeat.org/wp-content/uploads/papers/v9i3/B3986129219.pdf.
10) Senthilkumar M, Segar TC, Srivastava G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *Institute of Electrical and Electronics Engineers* ;(7):81542–81554. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8740989&tag=1.
11) Saleh F, Alotaibi. Implementation of Machine Learning Model to Predict Heart Failure Disease. *International Journal of Advanced Computer Science and Applications(IJACSA)*. 2019;10(6). Available from: https://thesai.org/Publications/ViewPaper?Volume=10&Issue=6&Code=IJACSA&SerialNo=37.
12) Bhatt CM, Patel P, Ghetia T, Mazzeo PL. Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms*. 2023;16(2):88. Available from: https://doi.org/10.3390/a16020088.
13) Biswas N, Ali MM, Rahaman MA, Islam M, Mia MR, Azam S, et al. Precision Medicine and Big Data Research Progress in Inflammatory Diseases. 2023. Available from: https://doi.org/10.1155/2023/6864343.
14) Ghazal TM, Ibrahim A, Akram AS, Qaisar ZH, Munir S, Islam S. Heart Disease Prediction Using Machine Learning. In: 2023 International Conference on Business Analytics for Technology and Security (ICBATS). IEEE. 2023;p. 1–6. Available from: https://doi.org/10.1109/ICBATS57792.2023.10111368.
15) Subramani S, Varshney N, Anand MV, Soudagar MEM, Al-Keridis LA, Upadhyay TK, et al. Cardiovascular diseases prediction by machine learning incorporation with deep learning. *Frontiers in Medicine*. 2023;10:1150933. Available from: https://doi.org/10.3389/fmed.2023.1150933.

16) Majumder AB, Gupta S, Singh D, Majumder S. An Explainable Hybrid Intelligent System for Prediction of Cardiovascular Disease. 2023. Available from: https://www.informaticsjournals.com/index.php/jmmf/article/view/34171/22526.
17) Majumder AB, Gupta S, Singh D. An Ensemble Heart Disease Prediction Model Bagged with Logistic Regression, Naïve Bayes and K Nearest Neighbour. 2017. Available from: https://doi.org/10.1088/1742-6596/2286/1/012017.
18) Ay Ş, Ekinci E, Garip Z. A comparative analysis of meta-heuristic optimization algorithms for feature selection on ML-based classification of heart-related diseases. *The Journal of Supercomputing*. 2023;79(11):11797–11826. Available from: https://doi.org/10.1007/s11227-023-05132-3.
19) Pal M, Parija S. Prediction of Heart Diseases using Random Forest.1 and Smita Parija.v2021. *Journal of Physics: Conference Series*. 2009. Available from: https://doi.org/10.1088/1742-6596/1817/1/012009.
20) Lutimath NM, Sharma N, Byregowda BK. Prediction of Heart Disease using Random Forest. In: 2021 Emerging Trends in Industry 4.0 . IEEE. 2021;p. 1–4. Available from: https://doi.org/10.1109/ETI4.051663.2021.9619208.
21) Dhaka P, Nagpal B. WoM-based deep BiLSTM: smart disease prediction model using WoM-based deep BiLSTM classifier. *Multimedia Tools and Applications*. 2023;82(16):25061–25082. Available from: https://doi.org/10.1007/s11042-023-14336-x.
22) Wei X, Rao C, Xiao X, Chen L, Goh M. Risk assessment of cardiovascular disease based on SOLSSA-CatBoost model. *Expert Systems with Applications*. 2023;219:119648–119648. Available from: https://doi.org/10.1016/j.eswa.2023.119648.
23) Asadi S, Roshan S, Kattan MW. Random forest swarm optimization-based for heart diseases diagnosis. *Journal of Biomedical Informatics*. 2021;115:103690–103690. Available from: https://doi.org/10.1016/j.jbi.2021.103690.
24) Naga SM. Detection of Cardiovascular Disease using Machine Learning, Genetic Algorithms and Particle Swarm Optimization. *International Journal Of Engineering Research & Technology* . 2023;12(03). Available from: https://www.ijert.org/research/detection-of-cardiovascular-disease-using-machine-learning-genetic-algorithms-and-particle-swarm-optimization-IJERTV12IS030072.pdf.
25) Heart Disease Data Set. . Available from: https://archive.ics.uci.edu/ml/datasets/heart+disease.