CINEFORUM
ISSN: 0009-7039

Vol. 65. No. 3, 2025

Heart Disease Prediction Using Optimized Feature Selection and Classification Techniques

Uzama Sadar[1], Parul Agarwal[*,2], Suraiya Parveen[3]
[1,2,3] Department of Computer Science and Engineering, School of Engineering Sciences and Technology, Jamia Hamdard, New Delhi, India
Email: [1]life.uzma@gmail.com; [*2]pagarwal@jamiahamdard.ac.in;
[3]suraiya@jamiahamdard.ac.in
*Corresponding Author

Abstract:
Machine learning (ML) in healthcare is gaining popularity, particularly for enhancing the accuracy and timeliness of diagnoses. Machine learning can forecast diseases by analysing enormous volumes of medical data, providing patients and medical professionals with the knowledge they need to make informed decisions regarding disease prevention. Heart disease prediction is crucial as early detection of the condition may save lives and expenses. Therefore, this study aims to extract the most significant variables from a high-dimensional dataset that aid in precisely and accurately classifying heart related disease. The authors of this work propose a hybrid heart disease prediction model by integrating AdaSyn, Particle Swarm Optimization, and Machine Learning techniques. This work uses the Cleveland dataset to assess and validate the system's performance. Firstly, preprocessing of the dataset is accomplish by min-max normalization to standardize the dataset. Thereafter, an Adaptive Synthetic Sampling Approach (ADASYN) was used to balance the dataset. Furthermore, to select optimal features, Particle Swarm Optimization (PSO) was used. Lastly, seven different ML models are trained on both full and optimized feature subsets as experimental analysis inputs. The proposed model outperformed conventional models with 91.8% accuracy, 90% precision, 92.8% sensitivity, 91.16% F1 measure, 90.9% Specificity, and 92% Area under the Receiver Operating Characteristic Curve. The classification results demonstrated the strong influence of pertinent characteristics on classification accuracy. When comparing models trained on the whole feature set to those trained on a smaller number of features, the classification models' performance increased noticeably with less training time.
Keywords: Machine Learning, Heart disease Prediction, PSO, Optimisation Techniques, Feature Selection

1. Introduction

The heart ranks among the significant organs in the entirety of humanity. It circulates blood that is rich in oxygen to all other bodily components via a system of veins and arteries. Heart disease accounts for approx31% of all fatalities and is a topmost cause of medical conditions and imposes a global economic burden. In the United States, heart disease accounts for one in four fatalities(*Everything You Need to Know about Heart Disease. Medical News Toda*, n.d.). The diagnosis of heart disease is frequently made after a physical examination and observation

of the patient's symptoms. Cardiovascular disease risk factors include smoking, advanced age, hypertension, obesity, diabetes, stress, elevated cholesterol, and inactivity. The disease's invasive diagnostic techniques are costly and uncomfortable. Thus, a less expensive, non-invasive method of diagnosing cardiac disease is required.

Clinical data can be utilized to create a decision support system that quickly and affordably detects cardiac disease using machine learning approaches. Many features in a dataset can lead to some being useless and producing undesirable outcomes. A patient's medical history consists of a wide range of details. All of these characteristics might not be equally important, and some might even be redundant. Furthermore, diagnosing with all of the features at once degrades performance. The majority of heart disease prediction research concentrated on two aspects: picking the most relevant variables and eliminating the rest, as well as selecting a suitable classifier. As a result, choosing the best characteristics and classifier is the goal of the prediction techniques. (Adler et al., 2020). Lately, advances in machine learning have enhanced our quality of life, particularly in the field of medicine(Sharma et al., 2024).

Receiving and self-learning policies through algorithm manipulation are the main foci of Machine Learning (ML). Regarding handling the wide range of data, labelled or unlabelled, ML is likewise quite understanding. The ML algorithm requires optimization techniques to minimize computation and extract significant features from the dataset. Optimization is a fundamental facet of ML. Use of machine learning models and its popularity are heavily impacted by the efficacy as well as efficiency of numerical optimization techniques in era of massive data. Optimization models are categorized into four combinations based on the problem's traits and the techniques used to solve it: unconstrained optimization, mathematical programming, heuristic algorithms, and simulation-based techniques. Among heuristic algorithms, bio-inspired algorithms are gaining popularity for optimization problems in ML. Metaheuristics are optimization algorithms employed to determine the optimal solution to a particular issue. The innate actions of animals or birds are replicated or inspired by these algorithms.(Darwish, 2018). Some bio-inspired algorithms are Genetic Algorithm (GA), PSO, Ant Colony Optimization, Grey Wolf Optimization, Whale Optimization, Greylag Goose Optimization (El-Kenawy et al., 2024), Puma optimizer used for feature selection(Abdollahzadeh et al., 2024) and many more in the series. Intelligent optimization algorithms are created by imitating or exposing specific natural occurrences, and due to their adaptability, they are utilised extensively in numerous study domains(Gazzaz et al., 2015). Intelligent algorithms for optimisation are employed extensively in numerous study domains due to their adaptability and are created by mimicking or exposing certain natural occurrences(Kamkar et al., 2010). PSO is famous among them; it is a population-based stochastic optimisation technique with a small number of parameters, making it easy to use(Eberhart & Kennedy, 1995). PSO is an important part of the ML model as it is used to adjust the parameters of the algorithm(Tharwat & Hassanien, 2019) and also used to select optimal features(Isik, 2024).

1.1 The paper's primary accomplishment is outlined as follows**:**

- This study suggested a hybrid model for predicting heart illness that combines ML methods, Adasyn, and PSO.

- The Adasyn technique balances the dataset, and PSO is used for feature selection.
- Cleveland Heart Disease dataset used for research.
- The efficacy of the suggested model is verified via performance metrics - accuracy, F-measure, specificity, precision, sensitivity, and AUC-ROC.

.

The remaining portion of the paper is organized as follows: Section 2 highlights the literature review and background research on heart disease prediction. Section 3 includes material and methods with a comprehensive description of the suggested system, dataset description, and classification techniques. In Section 4, Experimental findings, analysis, and comparison of the previously developed framework with the proposed system are discussed. Section 5 addresses the conclusion, limitations, and future scope.

2. LITERATURE SURVEY

Cardiovascular disease, also referred to as heart disease, is one of the deadliest illnesses that greatly increases fatality worldwide. The majority of current coronary heart disease prediction methods have been developed and tested using datasets from the UCI Repository, which are made up of risk factors (such as variables) besides angiograms(da Costa et al., 2016). Based on clinical data in hospitals, these methods may automatically detect
and execute a preliminary examination of patients and are easier, less costly, reproducible, and objective in their diagnosis. The focus of recent studies has been on Feature Selection techniques, optimization techniques, and increasing the precision of heart disease forecasts. This section provides an overview of recent relevant research publications.

In the paper(Jabbar et al., 2016) the authors proposed a technique to predict cardiac disease using a random forest. Relevant attributes were chosen by feature selection using the chi-square approach. The authors' method outperformed the decision tree in terms of accuracy. The C4.5 algorithm was used by the authors in the paper (Liu et al., 2017)to create a system for predicting cardiac disease. Boosting was used to improve the system's functionality. Similarly, in the paper(Haq et al., 2018) In order to diagnose cardiac problems, the authors developed a hybrid machine learning approach. Seven well-known ML classifiers are used with 3 feature selections. Logistic Regression with Relief feature selection shows better performance with 89% accuracy. PSO for feature selection was used in the paper(Vijayashree & Sultana, 2018) along with SVM, and achieves good accuracy. Subsequently, in the paper(Amin et al., 2019) the authors use various feature combinations with seven classification techniques and create prediction models. Based on the experiment results, the model was developed to forecast heart disease using the best data mining technique, and the pertinent features discovered achieved 87.4%accuracy. To create a system for forecasting cardiac illness, the authors of this study(Khourdifi & Baha, 2019) used various ML classifiers improved via ant colony and PSO. The best accuracy was obtained using random forests and K-nearest neighbors. Likewise, in the paper (Eskandari & Hassani, 2019), the authors increase the accuracy by using optimization approaches such as the hybrid Whale Optimization and Dragonfly algorithm, which improve feature selection. Using nine chosen characteristics, Support Vector Machine

in this study achieves 88.89% accuracy. The Enhanced Whale optimization technique for feature selection was used in the paper (Lakshmi & Devi, 2023) by the authors on the Framingham dataset and then trained on Machine learning and hybrid classifiers. In the paper (Yaqoob et al., 2022), the authors developed a prediction model and a Binary artificial Bee colony for feature selection. They achieved an accuracy of 92.4% with the KNN classification technique. Firefly Features selection techniques were used in the paper(Natarajan et al., 2024) to optimize features, and trained on ensemble classification techniques by voting and stacking, and obtained 86.79% accuracy. To forecast cardiac disease, the study in Ref.(Asif et al., 2023) offered a machine learning model that incorporates some preprocessing processes, ensemble learning methods, and hyperparameter optimisation strategies. They produced a complete dataset for study by combining three Kaggle datasets with comparable features. These findings demonstrate the model's competence to predict disease with 98.15%accuracy, which might greatly improve efforts at early prevention, diagnosis, and treatment and lower the mortality and morbidity associated with heart disease.
Table 1 discusses further studies on the prediction of disease using feature selection and other classification methods.

Table 1 Research work on disease prediction using Feature selection.

Reference	**Year**	**Dataset**	**Feature Selection**	**Classification Techniques**	**Findings**
(Ahmadi et al., 2018)	2017	Cleveland	Yes	Neural Network and aC5.0 Decision Tree	This study proposed a neural network and C5.0 decision tree paradigm for the prediction of coronary artery disease.
(Saqlain et al., 2019)	2018	Cleveland, Hungary, Switzerland, SPECTF	Yes	Mathews Correlation Coefficient &Fisher score are used as a Feature selection algorithm & SVM for binary classification.	The proposed approach performed better in terms of prediction results than the other strategies that were studied.

(Verma & Mathur, 2019)	2019	Cleveland	Yes, Hybrid Correlation & cuckoo search	Multi-Layer Perceptron is used as a classification technique	An intelligent CAD diagnosis decision support model that uses many classification methods and the Cuckoo approach for feature set analysis
(Latha & Jeeva, 2019)	2019	Cleveland	Yes	Ensemble technique by combining Naïve Bayes, Random Forest, Bayesian Network, and MLP using voting	Ensemble technique finds an accuracy of 85.48%
(Abdar et al., 2019)	2019	Z-Alizadeh Saini	Hybrid Genetic Algorithm (GA) and PSO	SVM and its variants	An accuracy of 93.08% was achieved, which outperformed
(Tama et al., 2020)	2020	Cleveland, Hungarian, Statlog, Z-Alizadeh Saini	Particle Swarm Optimisation for feature selection	This ensemble technique uses two tiers: Random Forest, Gradient, and Extreme Gradient Boosting classifiers.	Using an ensemble approach, an accuracy of 85.715 has been achieved .
(Abdellatif et al., 2022)	2022	Statlog, Heart disease clinical record	Supervised infinitefeature selection	Improved weighted Random Forest & Bayesian Optimisation	Achieves an accuracy of 98.3 on statlog and 97.2 on Heart Disease clinical data

				to tune hyperparameters	
(Biswas et al., 2023)	2023	Cleveland	chi-square, ANOVA, and mutual information	logistic regression, support vector machine, K-nearest neighbour, random forest, Naive Bayes , and decision tree	Random forest uses mutual information feature selection strategies to attain a maximum accuracy of 94.51%.
(Alghamdi et al., 2024)	2024	Cleveland	Arithmetic Optimization algorithm	Multilayer perceptron neural network classification technique	Achieves an accuracy of 88.89 %
(Wang et al., 2024)	2024	Statlog, Cleveland,Hungary	Pearson Correlation Coefficient	Twelve ML classifiers — LGBM, Adaboost, XGB, RF, DT, KNN, LR, GNB, ET, SVC, GB, and Bagging	Achieves an accuracy of 97.48%
(Al-Mahdi et al., 2025)	2025	Cleveland	GA	Ensemble Deep Learning approach, which is optimised by Tunicate Swarm Algorithm (TSA)	97.5% accuracy

3. Materials and Methods

3.1 Methodology

The authors of this work have suggested a hybrid approach for predicting heart disease. The three stages: data gathering, pre-processing, and model construction. The study uses the UCI repository's Cleveland Heart Disease dataset. In the pre-processing phase, data is cleaned and checked for missing values, normalized using a Min-max scalar. The coefficient of each feature is increased to an equal value by applying a standard scalar in order to ensure that each feature possesses a mean of 0 and a standard deviation of 1. Within the dataset, there are 139 cases of class 1 and 164 instances of class 0. An Adaptive Synthetic Sampling Approach (ADASYN) is performed as the dataset is imbalanced. Feature selection is performed using Particle Swarm Optimisation. Implementing the Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Random Forest (RF), Decision Tree (DT), AdaBoost, and Gradient Boosting classifiers, classification is carried out on specifically chosen features. Additionally, hyperparameter optimization using RandomizedSearchCV is performed on all the classifiers. Fig. 1 demonstrates the proposed methodology.

3.2 Data Preprocessing

Data pre-processing involves cleaning, organizing, and preparing raw data so that machine learning models can be built and trained. Data gathered from many sources is typically raw data that is unsuitable for immediate analysis. In the preprocessing step, firstly, data is cleaned and missing values are removed. After that, min-max normalisation is used for numerical features to avoid feature dominance because of distance. To normalize the data using min-max scaling via equation (1) is used

$$x^{nor} = (x - x_{min}) / (x_{max} - x_{min}) \quad (1)$$

Here x^{nor} - normalize data, and the dataset's minimum- x_{min} and maximum - $x_{max.}$

3.2.1. The ADASYN Technique is an adaptive synthetic sampling approach. It is an entirely novel technique for addressing unequal class distribution(Sadar et al., 2024). The dataset comprises 138 cases of class 0 and 165 examples of class 1. Figure 2 illustrates that the dataset is unbalanced, so we employ the ADASYN technique to balance the dataset.

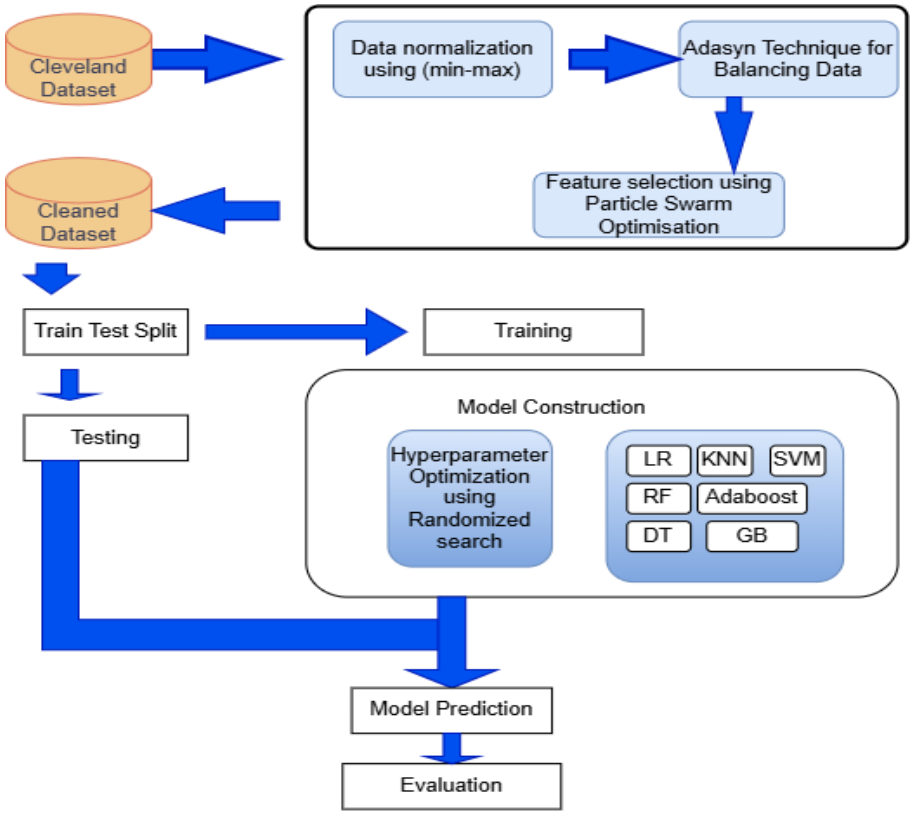

Figure 1. Proposed Methodology

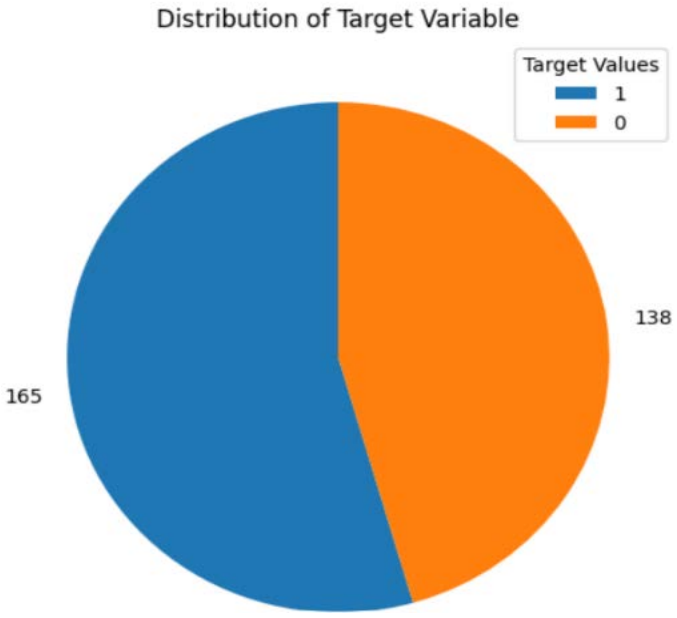

Figure 2. Dataset distribution

3.2.2. Particle Swarm Optimization for Feature Selection (FS) Feature selection is a technique of selecting the best subset of features from the original feature set. Removing extraneous features reduces memory and processing expenses. Using a certain optimization criterion, a subset of characteristics is chosen from the entire features. PSO was used for feature selection and optimization problems(Kazerani, 2024). PSO or global optimization technique is adaptive and has been adaptively used in a variety of fields to solve search and optimization problems. PSO, on the other hand, requires fewer parameter adjustments and is simpler to implement than a Genetic Algorithm. PSO draws inspiration from the coordinated actions of fish schools and bird flocks. PSO, another population-based search technique, begins with an initial particle population made up of randomly produced solutions. Every particle of PSO is associated with a position and velocity. During the search process, PSO keeps track of both the best position discovered by all particles and the best positions discovered by individual particles. Figure 3. Depicts the flowchart of PSO.

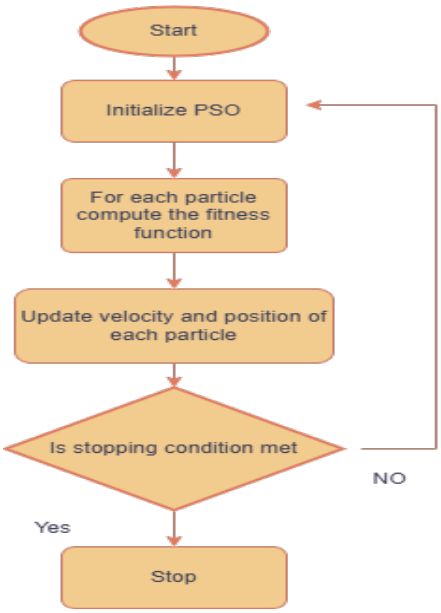

Figure 3. PSO Flowchart

3.3 Dataset Collection

The Cleveland dataset applied in our research was sourced from the online platform UCI repository (*Cleveland Heart Disease Dataset*, n.d.). It has 303 records, and six of these records have missing data. There are 76 variables for each participant in the dataset, although prior studies have revealed that for detecting disease, 13 features are helpful. There are numerical and categorical features in the dataset. The target variable is the num variable with values 1 and 0, showing the disease or no disease. The detailed description of the dataset is shown in Table 2.

Index No	Variable Name	Variable Code	Type	Description
1	Patients age	Age	Numerical	Age in terms of years
2	Patients gender	Sex	Categorical	Male-1, Female- 0
3	Chest Pain	Cp	Categorical	Typical Angina-1, Atypical angina-2, Non Angina 3, Asymptomatic-4
4	Blood Pressure of Resting	tretbps	Numerical	94-200 measured in mm Hg
5	Cholesterol	Chol	Numerical	126-564 calculated in mg/dl
6	Blood Sugar Fasting	FBS	Categorical	Fbs> 120 mg/dl
7	Resting Electrocardiograph	Restecg	Categorical	Normal:0, Abnormal:1, Hypertrophy:2
8	Maximum rate on heart achieved	thalach	Numerical	71 to 202
9	Exercise Induced Angina	exang	Categorical	Yes-1, No-0
10	ST depression induced by exercise	olpeak	Numerical	Upslope= 1, Flat= 2, downslope=3
11	Slope exercise ST segment	slope	Categorical	0 to 6.2
12	Major vessels colored by fluoroscopy	ca	Numerical	0 to 3
13	Thallium	thal	Categorical	Normal= 3, Fixed defect= 6, Defect Reversible=7
14	Target	num	Numerical	Disease =1, No Disease=0

3.4 Machine Learning Classifier

3.4.1. Logistic Regression (LR) Logistic regression predicts a dependent data variable by analysing the correlation between one or more independent variables that already exist(Ambrish et al., 2022). A binary result, like yes or no, can be predicted using the statistical analysis method of logistic regression based on prior observations of a dataset. The cost function, sometimes known as the sigmoid function, is used by LR.

3.4.2. Support Vector Machine (SVM) is an effective classification algorithm that can handle linear and nonlinear data(Elsedimy et al., 2024). SVM was first developed for binary and linear issues, but it may also be used for nonlinear and multiclass problems. SVM searches an n dimensional space for a hyperplane that uniquely qualifies a data point, where n is quantity of features. The hyperplane is only a line when there are only two input features, and it is a two-dimensional plane when there are three. The points that the SVM algorithm determines are closest to the lines from both classes are known as support vectors.

3.4.3. Decision Tree (DT) is a form of Supervised ML non-parametric technique used for tasks involving both regression and classification(Ozcan & Peker, 2023). It looks like a tree structure with a root node serving as a decision node and an internal node, a subtree that indicates dataset features, a leaf node indicating class label or result, and branches showing decision rules. The Attribute Selection Measure approach is used to pick attributes in the decision tree.

The following two metrics are used

(a) Information Gain: The rate at which the entropy of an attribute changes.
(b) Entropy is used to determine attributes impurity.

3.4.4. Random Forest (RF) integrates numerous decision trees using the bagging principle to give one single output tree and improve prediction abilities(Sumwiza et al., 2023). In the RF, several regression and classification trees are generated using randomly chosen training datasets and feature subsets to develop models. The output of several weak decision trees is combined to produce precise results. As compared to decision trees, an RF frequently offers higher accuracy.

3.4.5. K Nearest Neighbour (KNN) is an ML algorithm used to resolve problems in regression and classification. It calculates the distance between the training and test data points(Barry et al., 2024). Three distance metrics are used in KNN: Euclidean and Manhattan distances are used for regression problems, while Hamming distances are used for classification problems.

3.4.6. Adaboost The adaptive boosting algorithm Adaboost employs boosting, an ensemble technique, that strengthens the weak learners' performance(Akinola et al., 2024). The AdaBoost method generates a group of poor learners and adaptively modifies them following each weak learning cycle. Current weak learners' incorrectly categorised training samples will have their weights raised, while correctly classified training samples will have their weights lowered.

3.4.7. Gradient Boosting(GB) T technique seeks to minimise the disparity between expected and real values by forecasting the residual errors of earlier estimators(Ganie et al., 2023). After training the weak learners one after the other, all estimators are introduced progressively by adjusting the weights.

3.5 Hyperparameter Optimization

In machine learning, hyperparameter tweaking is an essential step for maximising model performance.

3.5.1. RandomisedSearchCV Hyperparameter tuning maximises model accuracy and minimises errors by choosing a set of ML algorithm hyperparameters that yield the highest

performance(Valarmathi & Sheela, 2021). It optimises hyperparameters by sampling from predefined distributions rather than attempting every possible combination, and is more effective than GridSearchCV in situations where processing power or time is restricted(Rimal et al., 2024).

4 Experiment Results and Analysis

To forecast cardiac conditions from a dataset, we utilize Jupiter Notebook 7. The experiment is done in a Python environment using different libraries like Pandas, Scikit Learn, Matplotlib, NumPy, and PySwarms for PSO Optimization. Cleaning the Cleveland Heart Disease dataset with NumPy library is the first stage of this study. The dataset is then pre-processed using the StandardScaler function from Python's Scikit-learn package. Thereafter Adasyn technique is applied to balance the dataset using the imblearn library. In the third step, we perform feature selection using PSO, and dataset is then split into 70:30 segments, with 70% going towards training and 30% towards testing.. Lastly trained all ML classifiers to forecast cardiac disease, and we chose the approach that performed the best.

4.1 Evaluation Parameters

The proposed method uses a variety of performance parameters to forecast cardiac disease. The accuracy, specificity, precision, F1 score, ROC curve, and recall scales were used to assess the classifiers' performance.

The following explains the performance measures.

- True Positive (Tp): The quantity of positive cases when the model accurately calculates the likelihood of heart disease.
- True Negative (Tn): The number of negative cases when the model accurately predicts that heart disease doesn't exist.
- False Positive (Fp): The number of negative cases when the model forecasts the existence of heart disease inaccurately.
- False Negative (Fn): The quantity of positive cases where the model forecasts the absence of heart disease wrongly.

4.1.1. Accuracy is a measure of how well a prediction model divides people into two groups: those who have heart disease (positive class) and those who do not (negative class). It can be defined as follows in equation (2).

$$Accuracy = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \tag{2}$$

4.1.2. Precision is the proportion of accurately predicted positive instances to all expected positive instances. It is determined by the following equation (3).

$$Precision = \frac{Tp}{Tp + Fp} \tag{3}$$

4.1.3. Recall or sensitivity assesses its capacity to make precise positive predictions. It is determined by the following formula in equation (4).

$$Recall = \frac{Tp}{Tp + Fn} \tag{4}$$

4.1.4. Specificity This performance metric assesses the system's capacity to provide accurate negative forecasts. It is defined by the following formula equation (5).

$$Specificity = \frac{Tn}{Tn + Fp} \tag{5}$$

4.1.5. F-measure A statistical metric known as the F-measure is used to assess a classification model's effectiveness. It accomplishes this by calculating the accuracy and recall measures' harmonic mean and assigning equal weight to each of these metrics. It is measured by the following formula in equation (6).

$$F - measure = \frac{2 * (Recall * Precision)}{Recall + Precision} \tag{6}$$

4.1.6. AUC-ROC curve AUC is a metric used to assess the way a model can distinguish across classes, while ROC Receiver Operating Characteristics is a probability curve.It is listed in equation (7).

$$Fpr = \frac{FP}{FP + TN} \tag{7}$$

4.2 Results and Discussion

This work aimed to predict heart disease by using ML techniques and optimized Feature selection. A range of classification methods were employed to predict heart disease, including Adaboost, Random Forest, Decision Tree, K Nearest Neighbor, Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boosting. The research study made use of the Cleveland heart disease dataset from the repository maintained by UCI., which consists of 303 records and 13 mostly used attributes in the dataset. Several medical attributes included in the dataset were used to identify heart disease. Classification was performed using these parameters, where class 0 denotes that a person is disease-free and class 1 denotes that a person has an illness. The performance of the model was evaluated using Accuracy, Precision, Recall, Specificity, F-measure, and AUC-ROC curve scales.

4.2.1 Classifiers' Performance using all features

In the first step, after applying the standard scaling technique, the performance of all classifiers is evaluated using all 13 feature sets. The Adaboost classifier performed the best on the whole

feature set, whereas the DT classifier performed the worst. Table 3 displays the evaluation metrics' outcomes.

Table 3 Performance evaluation using all features

Classifiers	Accuracy	Precision	Sensitivity	Specificity	F- measure	AUC-ROC
LR	78	72	96	57	82	91
SVM	73	69	92	52	79	88
KNN	78	75	88	66	81	84
DT	69	67	84	52	75	69
RF	73	69	88	52	77	89
Adaboost	80	75	96	88	84	92
GB	78	74	92	61	82	90

4.2.2 Classifiers' performance after applying the Adasyn technique

By balancing the classes in the second stage, the classifier's performance further increased. To balance the classes, the Adasyn approach was used. In all, it generated 165 occurences of both classes 0 and 1. Again, the Adaboost classifier outperformed the others, although LR's performance declined in comparison to the former. Little improvement is shown by SVM, DT, RF, and GB, indicating that class balancing has a positive impact on the classifiers' performance. The effect of class balancing on classifier performance is displayed in Table 4.

Table 4. Performance evaluation after Class Balancing using the Adasyn technique

Classifiers	Accuracy	Precision	Sensitivity	Specificity	F- measure	AUC-ROC
LR	73	69	92	57	79	89
SVM	76	73	92	61	80	88
KNN	76	75	88	57	79	76
DT	78	72	92	66	82	82
RF	71	69	88	52	77	88
Adaboost	84	75	90	88	87	94
GB	79	74	88	66	79	89

4.2.3 Classifiers: Performance enhancement after Feature Selection

To further improve classifier accuracy, PSO for feature selection was applied. Seven features out of thirteen characteristics were chosen using this technique, namely, sex, cp, fbs, restecg, oldpeak, slope, and thal. After feature selection in the training part, Hyperparameter optimisation using RandomisedSearchCV was performed on all seven classifiers. As shown in Table 5, Feature selection has improved the performance of all classifiers. The SVM achieved Maximum accuracy of 91.8%, which outperformed other classifiers, followed by KNN and RF with 91% accuracy.

Table 5 Performance evaluation after feature selection

Classifiers	Accuracy	Precision	Sensitivity	Specificity	F- measure	AUC-ROC
LR	82	86	96	59	83	91
SVM	91.8	90.9	93.7	89.65	92.3	92
KNN	91	91	84	70	90	85
DT	88	87	84	52	89	74
RF	91	90	92	72	91	89
Adaboost	86	82	90	71	86	88
GB	84	85	92	67	85	89

Feature selection enhances the accuracy, precision, specificity, sensitivity, F-measure, and AUC-ROC of classifiers, as illustrated in Figures 4, 5, 6, 7,8, and 9. Therefore, we conclude that SVM with PSO and Adasyn techniques gives the highest performance. Figures 10 and 11 display the suggested model's confusion matrix and ROC curve, respectively. With an AUC of 0.92, this curve demonstrates the model's effectiveness across all classification thresholds. AUC values greater than 0.9 suggest that the model is quite accurate and might be utilised with assurance in decision-making procedures.

Comparative analysis of the proposed model with previously developed models is discussed in Table 6.

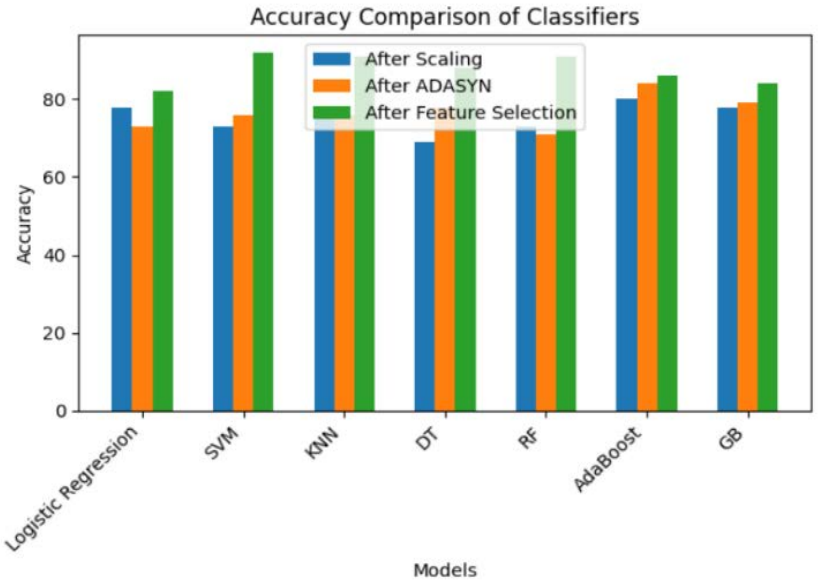

Fig 4. Improvement in classifier accuracy by feature selection using PSO

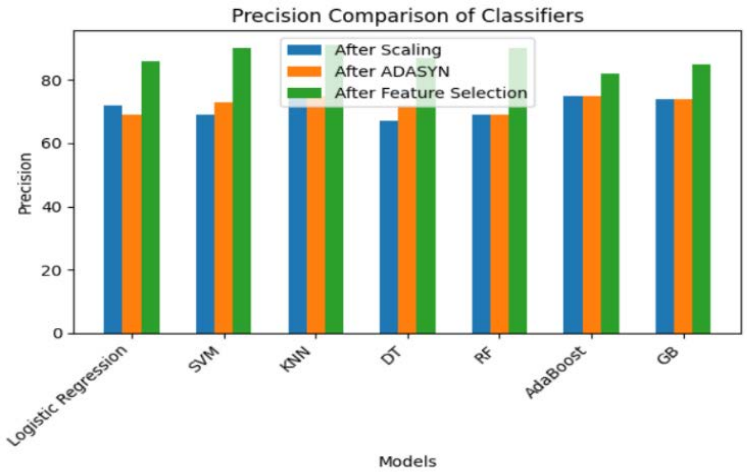

Fig5. Improvement in classifier precision by feature selection using PSO

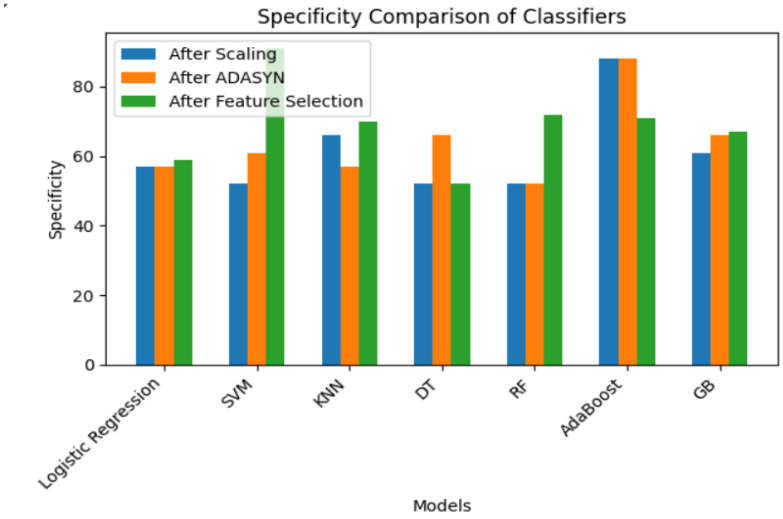

Fig 6. Improvement in classifier Specificity by feature selection using PSO

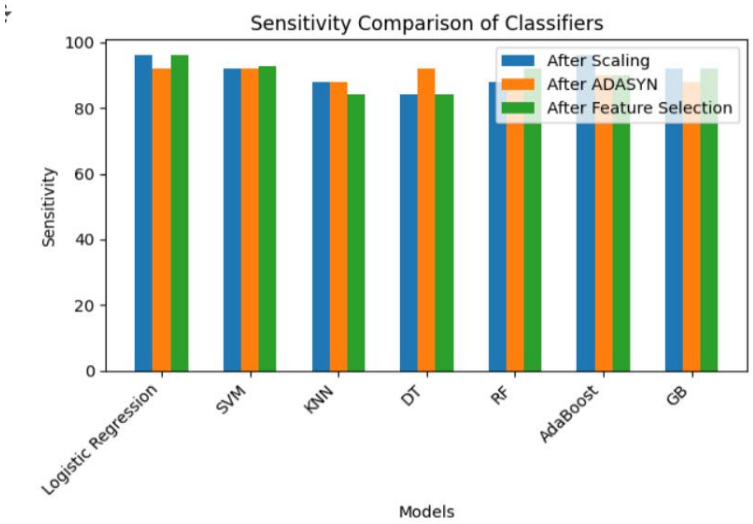

Fig 7. Improvement in classifier Sensitivity by Feature Selection using PSO

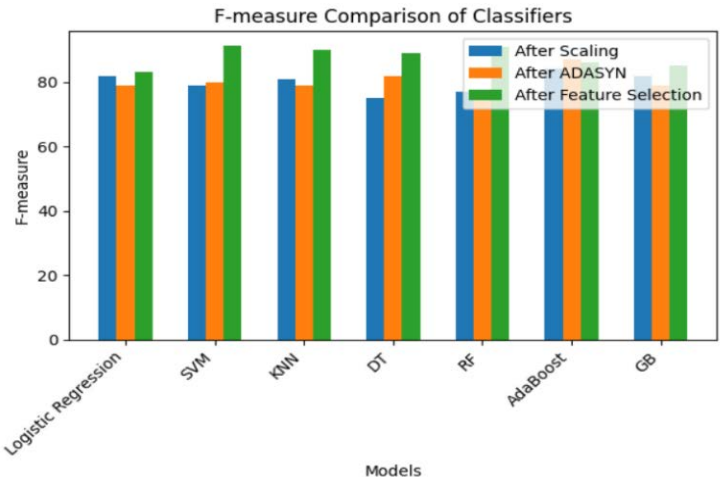

Fig 8. Improvement in classifier F-Measure by Feature Selection using PSO

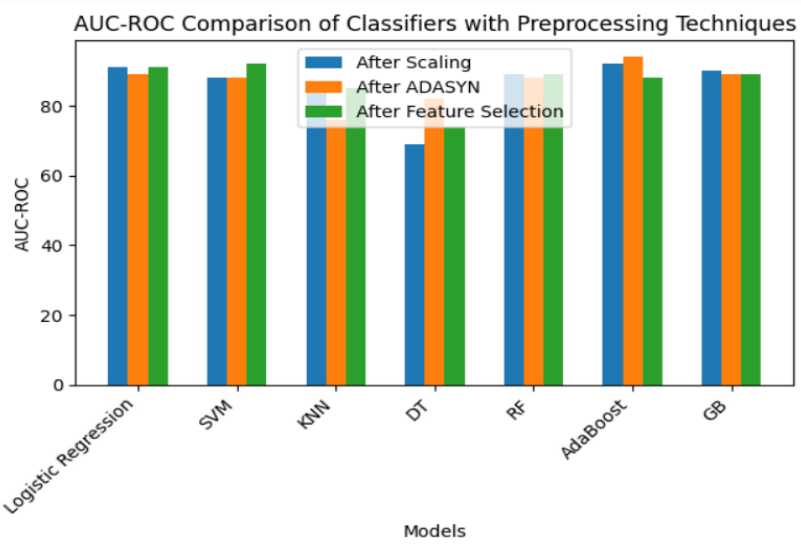

Fig 9. Improvement in classifier AUC-ROC by Feature Selection using PSO

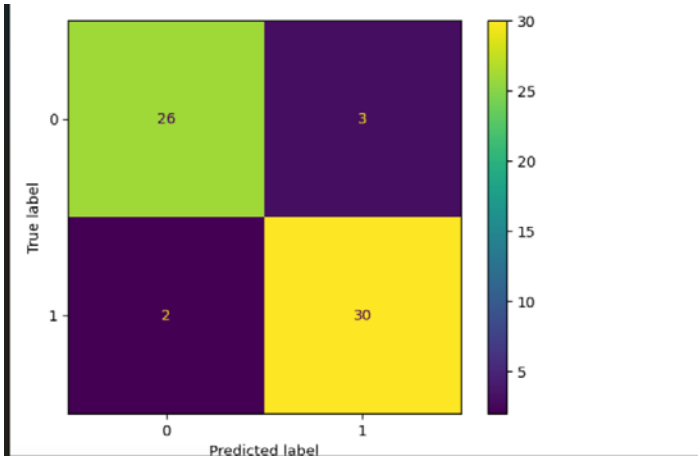

Fig 10. Confusion Matrix of the model

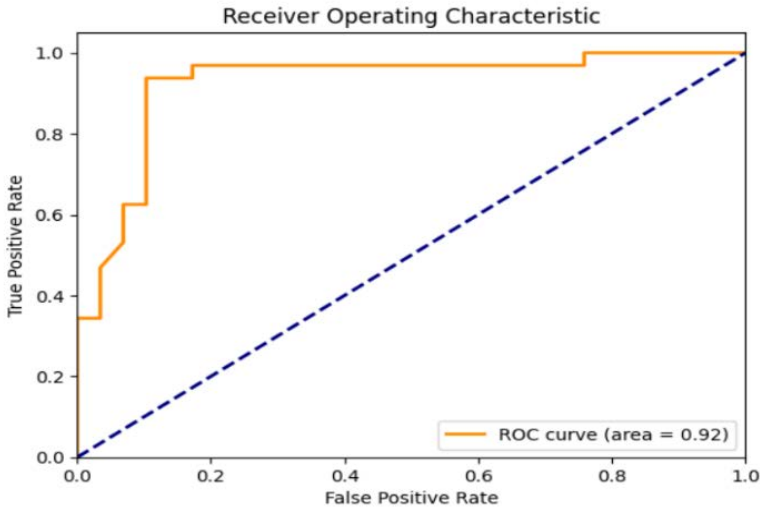

Fig 11. AUC and ROC curve of the model

Table 6. Compares the proposed method with previously developed techniques based on classification accuracy.

Study	Preprocessing	Dataset	Feature Selection	Classification	Accuracy
(Paul et al., 2016)	Missing Values	Cleveland	GA & Correlation Coefficient	Fuzzy Rules	80
(El-Bialy et al., 2015)	-	Cleveland	Manual	C4.5 & fast DT	78.54
(Saqlain et al., 2019)	Data Standardization	Cleveland	Fisher score Mathews Correlation	SVM	81.19
(Shah et al., 2017)	Normalization	Cleveland	PCA & parallel probabilistic	SVM	82.18
(Suresh & Ananda Raj, 2018)	Handling Missing Values	Cleveland	GA	NB	83.12
(Rani et al., 2021)	MICE imputes missing value	Cleveland	Hybrid GA& RFE	RF	86.6
(Tama et al., 2020)	Min Max Normalisation	Cleveland	Hybrid GA & PSO	RF	87.8

(Alghamdi et al., 2024)	Normalization	Cleveland	Arithmetic Optimization Algorithm	Multilayer Perceptron Neural Network	88.89
Proposed model Adasyn+ PSO+ SVM	Min Max Normalisation	Cleveland	PSO	SVM	91.8

5. Conclusion and Future Scope

5.1 Conclusion

Heart disease has grown into an increasing cause of death worldwide and remains an urgent health issue. AI systems with machine learning algorithms can accurately identify illnesses and forecast patients' long-term health. ML models can be pretty helpful in this critical situation by forecasting illnesses in their early stages. The primary contribution of this research is the proposal of an optimized disease prediction system that can more accurately diagnose cardiac disease than current methods.

This study used the hybrid model to predict cardiovascular disease. Initially, data preparation was carried out by applying efficient scaling algorithms and transforming nominal data into numerical data. Thereafter, the Adasyn approach balances the data and selects optimal features; PSO feature selection techniques are applied. Seven features, namely sex, cp, fbs, restecg, oldpeak, slope, and thal, are selected after performing feature selection using PSO. Seven distinct cutting-edge machine learning algorithms, namely Logistic Regression, SVM, KNN, RF, DT, Adaboost, and GB, were then deployed on optimised features. The Cleveland dataset was used for the analysis of the study experiments. RandomisedSearchCV has been used for hyperparameter optimisation of algorithms. The findings show that the suggested model with SVM surpassed all other algorithms, achieving 91.8% accuracy, 90% precision, 92.8% sensitivity, 91.16% F-measure, 90.9% specificity, and 92% AUC-ROC. This model might help practitioners and researchers with tasks on heart disease.

5.2 Limitations and Future scope

Notwithstanding the suggested ML-based methodology's encouraging outcomes and potential benefits for heart disease prediction, certain limitations require attention.

Dataset availability: The caliber and accessibility of datasets affect the ML models' performance and dependability. In our study, we used the Cleveland Heart Disease. The quality, representativeness, and availability of the data may be restricted. This restriction could make it challenging to use the suggested method on a larger sample, like some real-world datasets that include a range of other sources.

Algorithm selection: The researchers employed an array of machine learning approaches to identify the best algorithm for HD prediction. However, the selection of algorithms is arbitrary and might influence the result. Other algorithms that weren't taken into account for this study might produce better accuracy or other trade-offs.

These issues emphasise the necessity of ongoing efforts to maintain the model's robustness, interpretability, and applicability in various healthcare situations by striking a balance between interpretability and model performance.

In the future, the proposed work will be further assessed and examined on a range of medical-based datasets to show its potential for healthcare and medical diagnostics. Future research by the authors aims to further enhance system performance by experimenting with other feature selection techniques, including metaheuristic algorithms. Furthermore, other hyperparameter optimization techniques need to be explored. Additional recommendations for further study include gathering data and contrasting how well different swarm algorithms work using different heart disease datasets and a real-time clinical dataset. In addition, the authors plan to develop a system for diagnosing heart problems using ensemble methods and deep learning.

Resolving these issues and implementing these recommendations in further research might enhance prediction models and clinical practice applications.

Acknowledgment

The authors would like to acknowledge DST-FIST (Department of Computer Science & Engineering, Jamia Hamdard) No SR/FST/ET-11/2019/313(C) for providing the facilities to conduct the research.

References

Abdar, M., Książek, W., Acharya, U. R., Tan, R.-S., Makarenkov, V., & Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. *Computer Methods and Programs in Biomedicine*, *179*, 104992.

Abdellatif, A., Abdellatef, H., Kanesan, J., Chow, C.-O., Chuah, J. H., & Gheni, H. M. (2022). Improving the heart disease detection and patients' survival using supervised infinite feature selection and improved weighted random forest. *IEEE Access*, *10*, 67363–67372.

Abdollahzadeh, B., Khodadadi, N., Barshandeh, S., Trojovský, P., Gharehchopogh, F. S., El-kenawy, E.-S. M., Abualigah, L., & Mirjalili, S. (2024). Puma optimizer (PO): a novel metaheuristic optimization algorithm and its application in machine learning. *Cluster Computing*, *27*(4), 5235–5283.

Adler, E. D., Voors, A. A., Klein, L., Macheret, F., Braun, O. O., Urey, M. A., Zhu, W., Sama, I., Tadel, M., Campagnari, C., Greenberg, B., & Yagil, A. (2020). Improving risk prediction in heart failure using machine learning. *European Journal of Heart Failure*, *22*(1), 139–147. https://doi.org/10.1002/ejhf.1628

Ahmadi, E., Weckman, G. R., & Masel, D. T. (2018). Decision making model to predict presence of coronary artery disease using neural network and C5. 0 decision tree. *Journal of Ambient Intelligence and Humanized Computing*, *9*, 999–1011.

Akinola, S., Leelakrishna, R., & Varadarajan, V. (2024). Enhancing cardiovascular disease prediction: A hybrid machine learning approach integrating oversampling and adaptive boosting techniques. *AIMS Medical Science AIMS Medical Science*, *2*, 58–71.

Al-Mahdi, I. S., Darwish, S. M., & Madbouly, M. M. (2025). Heart Disease Prediction Model Using Feature Selection and Ensemble Deep Learning with Optimized Weight. *CMES-*

Computer Modeling in Engineering and Sciences, *143*(1), 875–909.

Alghamdi, F. A., Almanaseer, H., Jaradat, G., Jaradat, A., Alsmadi, M. K., Jawarneh, S., Almurayh, A. S., Alqurni, J., & Alfagham, H. (2024). Multilayer perceptron neural network with arithmetic optimization algorithm-based feature selection for cardiovascular disease prediction. *Machine Learning and Knowledge Extraction*, *6*(2), 987–1008.

Ambrish, G., Ganesh, B., Ganesh, A., Srinivas, C., & Mensinkal, K. (2022). Logistic regression technique for prediction of cardiovascular disease. *Global Transitions Proceedings*, *3*(1), 127–130.

Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*, *36*, 82–93.

Asif, D., Bibi, M., Arif, M. S., & Mukheimer, A. (2023). Enhancing heart disease prediction through ensemble learning techniques with hyperparameter optimization. *Algorithms*, *16*(6), 308.

Barry, K. A., Manzali, Y., Lamrini, M., Rachid, F., & Elfar, M. (2024). Heart Disease Prediction Using Weighted K-Nearest Neighbor Algorithm. *Operations Research Forum*, *5*(3), 76.

Biswas, N., Ali, M. M., Rahaman, M. A., Islam, M., Mia, M. R., Azam, S., Ahmed, K., Bui, F. M., Al-Zahrani, F. A., & Moni, M. A. (2023). Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques. *BioMed Research International*, *2023*(1), 6864343.

Cleveland Heart disease dataset. (n.d.). https://archive.ics.uci.edu/dataset/45/heart+disease

da Costa, C., da Costa Linch, G. F., & Nogueira de Souza, E. (2016). Nursing Diagnosis Based on Signs and Symptoms of Patients With Heart Disease. *International Journal of Nursing Knowledge*, *27*(4), 210–214. https://doi.org/10.1111/2047-3095.12132

Darwish, A. (2018). Bio-inspired computing: Algorithms review, deep analysis, and the scope of applications. *Future Computing and Informatics Journal*, *3*(2), 231–246. https://doi.org/10.1016/j.fcij.2018.06.001

Eberhart, R., & Kennedy, J. (1995). A new optimizer using particle swarm theory. *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, 39–43.

El-Bialy, R., Salamay, M. A., Karam, O. H., & Khalifa, M. E. (2015). Feature analysis of coronary artery heart disease data sets. *Procedia Computer Science*, *65*, 459–468.

El-Kenawy, E.-S. M., Khodadadi, N., Mirjalili, S., Abdelhamid, A. A., Eid, M. M., & Ibrahim, A. (2024). Greylag goose optimization: nature-inspired optimization algorithm. *Expert Systems with Applications*, *238*, 122147.

Elsedimy, E. I., AboHashish, S. M. M., & Algarni, F. (2024). New cardiovascular disease prediction approach using support vector machine and quantum-behaved particle swarm optimization. *Multimedia Tools and Applications*, *83*(8), 23901–23928.

Eskandari, M., & Hassani, Z. (2019). Intelligent application for Heart disease detection using Hybrid Optimization algorithm. *Journal of Algorithms and Computation*, *51*(1), 15–27.

Everything you need to know about heart disease. Medical News Toda. (n.d.). https://www.medicalnewstoday.com/articles/237191#types

Ganie, S. M., Pramanik, P. K. D., Malik, M. B., Nayyar, A., & Kwak, K. S. (2023). An Improved Ensemble Learning Approach for Heart Disease Prediction Using Boosting Algorithms. *Comput. Syst. Sci. Eng.*, *46*(3), 3993–4006.

Gazzaz, N. M., Yusoff, M. K., Ramli, M. F., Juahir, H., & Aris, A. Z. (2015). Artificial neural network modeling of the water quality index using land use areas as predictors. *Water Environment Research*, *87*(2), 99–112.

Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, *2018*(1), 3860146.

Isik, I. (2024). Heart disease prediction with feature selection based on metaheuristic optimization algorithms and electronic filter model. *Arabian Journal for Science and Engineering*, *49*(9), 11953–11966.

Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2016). Prediction of heart disease using random forest and feature subset selection. *Innovations in Bio-Inspired Computing and Applications: Proceedings of the 6th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2015) Held in Kochi, India during December 16-18, 2015*, 187–196.

Kamkar, I., Akbarzadeh-T, M.-R., & Yaghoobi, M. (2010). Intelligent water drops a new optimization algorithm for solving the vehicle routing problem. *2010 IEEE International Conference on Systems, Man and Cybernetics*, 4142–4146.

Kazerani, R. (2024). Improving breast cancer diagnosis accuracy by particle swarm optimization feature selection. *International Journal of Computational Intelligence Systems*, *17*(1), 44.

Khourdifi, Y., & Baha, M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *International Journal of Intelligent Engineering & Systems*, *12*(1).

Lakshmi, A., & Devi, R. (2023). Heart Disease Prediction Using Enhanced Whale Optimization Algorithm Based Feature Selection With Machine Learning Techniques. *2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART)*, 644–648.

Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203.

Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., & Wang, Q. (2017). A hybrid classification system for heart disease diagnosis based on the RFRS method. *Computational and Mathematical Methods in Medicine*, *2017*(1), 8272091.

Natarajan, K., Vinoth Kumar, V., Mahesh, T. R., Abbas, M., Kathamuthu, N., Mohan, E., & Annand, J. R. (2024). Efficient heart disease classification through stacked ensemble with optimized firefly feature selection. *International Journal of Computational Intelligence*

Systems, *17*(1), 174.

Ozcan, M., & Peker, S. (2023). A classification and regression tree algorithm for heart disease modeling and prediction. *Healthcare Analytics*, *3*, 100130.

Paul, A. K., Shill, P. C., Rabin, M. R. I., & Akhand, M. A. H. (2016). Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, 145–150.

Rani, P., Kumar, R., Ahmed, N. M. O. S., & Jain, A. (2021). A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments*, *7*(3), 263–275. https://doi.org/10.1007/s40860-021-00133-6

Rimal, Y., Sharma, N., & Alsadoon, A. (2024). The accuracy of machine learning models relies on hyperparameter tuning: student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms. *Multimedia Tools and Applications*, *83*(30), 74349–74364.

Sadar, U., Agarwal, P., Parveen, S., Dhand, G., & Sheoran, K. (2024). HEART DISEASE PREDICTION USING MACHINE LEARNING CLASSIFIERS WITH VARIOUS BALANCING TECHNIQUES. *Proceedings on Engineering*, *6*(4), 1871–1878.

Saqlain, S. M., Sher, M., Shah, F. A., Khan, I., Ashraf, M. U., Awais, M., & Ghani, A. (2019). Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. *Knowledge and Information Systems*, *58*, 139–167.

Shah, S. M. S., Batool, S., Khan, I., Ashraf, M. U., Abbas, S. H., & Hussain, S. A. (2017). Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis. *Physica A: Statistical Mechanics and Its Applications*, *482*, 796–807.

Sharma, A., Lysenko, A., Jia, S., Boroevich, K. A., & Tsunoda, T. (2024). Advances in AI and machine learning for predictive medicine. *Journal of Human Genetics*, *69*(10), 487–497.

Sumwiza, K., Twizere, C., Rushingabigwi, G., Bakunzibake, P., & Bamurigire, P. (2023). Enhanced cardiovascular disease prediction model using random forest algorithm. *Informatics in Medicine Unlocked*, *41*, 101316.

Suresh, P., & Ananda Raj, M. D. (2018). Study and analysis of prediction model for heart disease: an optimization approach using genetic algorithm. *Int J Pure Appl Math*, *119*(16), 5323–5336.

Tama, B. A., Im, S., & Lee, S. (2020). Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble. *BioMed Research International*, *2020*(1), 9816142.

Tharwat, A., & Hassanien, A. E. (2019). Quantum-behaved particle swarm optimization for parameter optimization of support vector machine. *Journal of Classification*, *36*(3), 576–598.

Valarmathi, R., & Sheela, T. (2021). Heart disease prediction using hyper parameter optimization (HPO) tuning. *Biomedical Signal Processing and Control*, *70*(March), 103033. https://doi.org/10.1016/j.bspc.2021.103033

Verma, L., & Mathur, M. K. (2019). Deep learning based model for decision support with case

based reasoning. *Int J Innov Technol Explor Eng*, *8*(6C), 149–153.

Vijayashree, J., & Sultana, H. P. (2018). A machine learning framework for feature selection in heart disease classification using improved particle swarm optimization with support vector machine classifier. *Programming and Computer Software*, *44*, 388–397.

Wang, S., Zhang, L., Liu, X., & Sun, J. (2024). Optimization of multidimensional feature engineering and data partitioning strategies in heart disease prediction models. *Alexandria Engineering Journal*, *107*, 932–949.

Yaqoob, M. M., Nazir, M., Yousafzai, A., Khan, M. A., Shaikh, A. A., Algarni, A. D., & Elmannai, H. (2022). Modified artificial bee colony based feature optimized federated learning for heart disease diagnosis in healthcare. *Applied Sciences*, *12*(23), 12080.