INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

An optimized Model for Heart Disease Prediction with Customized Ensemble Voting Classifier and Nature Inspired Optimization

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

1 Assistant Professor, Department of Information Technology, JIS College of Engineering, India
2 Associate Professor, Department of Computer Science and Engineering, Swami Vivekananda University, India
3 Associate Professor, Department of Computer Science and Engineering, JIS College of Engineering
4 Manager, Capgemini, India

OPEN ACCESS

Received: 10-05-2023

Accepted: 03-08-2023

Published: 14-09-2023

Citation: Majumder AB, Gupta S, Singh D, Majumder S (2023) An optimized Model for Heart Disease Prediction with Customized Ensemble Voting Classifier and Nature Inspired Optimization. Indian Journal of Science and Technology 16(34): 2719-2729. https://doi.org/10.17485/IJST/v16i34.1004

* **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

Objective: To develop an optimized model for prediction of heart disease. **Methods/findings:** The model has been built applying Customized Ensemble Voting Classifier where the weights of each base classifier have been calculated considering accuracy, specificity and sensitivity. For feature selection Chi Square method has been applied. The performance of the model has been enhanced using Firefly algorithm. The proposed model has achieved 85.52% accuracy. **Novelty:** The novelty of our research paper on heart disease prediction lies in the integration of machine learning algorithms to develop an accurate predictive model. Multiple classifiers have grouped together through a customized ensemble voting classifier. For assigning the weight not only accuracy is considered but also specificity and sensitivity scores have also been considered and then performance has been optimized with Firefly algorithm which makes the method comprehensive and reliable for predicting heart disease risk and could ultimately lead to more effective prevention and treatment strategies.

Keywords: Chi Square; Ensemble Voting Classifier; Firefly Algorithm

1 Introduction

Heart disease is one of the leading causes of death globally, affecting millions of individuals every year. Identifying the risk factors associated with heart disease ,developing accurate prediction models is crucial for effective prevention and treatment of this condition. In recent years, machine learning algorithms have shown great potential in predicting heart disease risk by analysing large and complex datasets.

The objective of this research paper is to develop a heart disease prediction model using machine learning algorithms and to evaluate its accuracy and efficacy.

The proposed model has been built using a customized weighted ensemble voting classifier trained on the dataset collected from UCI data repository. For feature selection Chi Square method has been applied. And finally, the model's performance has been optimized using nature inspired Firefly Algorithm. The Chi-square test has benefits such as ease of computation, robustness with regard to data distribution, and the ability to derive precise information. The experiment has used the Chi square test in light of these characteristics. As different machine learning models has their advantages and disadvantages, ensemble classifier overcomes these by combining multiple classifiers. In this method Decision Tree, Adaboost and Gradient Boosting have been utilized together as ensemble voting classifier with customized weight values calculated considering accuracy, specificity and sensitivity. Over that for performance enhancement Firefly algorithm has been applied. Firefly is highly nonlinear, can handle multi class model. Its convergence rate is very high and it does not require a good initial solution to start. Considering all these Firefly optimization has been considered for this work. To evaluate the performance of the proposed model accuracy, precision, recall, and F1-score have been measured. Area under the receiver operating characteristic (ROC) curve (AUC) to assess the discriminatory power of our models has been used.

The major contribution of our research paper on heart disease prediction is the development of an optimized machine learning model that integrates clinical data and machine learning methodologies to accurately predict the occurrence of heart disease.

In section 2 few existing works have been discussed and analysed to get the insight of their working principles. In section 3 proposed methodology has been represented followed by outcome analysis of the work in section 4. Conclusion has put in section 5.

1.1 Literature Review

In this section several recent works have been analysed which have worked as motivation of our proposed work.

The study by Rohit Bharti et al. presented a novel approach to predict heart disease using machine learning and deep learning techniques. The authors applied Isolation Forest and Robust Scaler algorithms to perform feature selection and handle outliers in the dataset, respectively. Different machine learning classifiers used over this work are Random Forest. Logistic Regression K Neighbor, Support Vector Machine, Decision Tree and XGBoost with 80.3%, 83.3%, 84.26%, 83.29% .82.33% and 71.4% accuracy respectively. Authors also applied and tested deep learning for achieving better performance in (1).

Applying Logistic Regression, LightGBM, XGBoost, Gaussian Naive Bayes, Support Vector Machine (SVM), authors proposed a model of heart disease prediction where 80.32%, 78.68%, 80.32%, 77.04%, 73.77%, and 88.5% accuracy were respectively achieved by each method. For feature selection Chi Square test was used. Authors justified the performance measured applying Confusion Matrix and AUC-ROC curve in (2).

Using machine learning techniques, Sneha Grampurohit and Chetan Sagarnal presented their research where a dataset with 4920 patients' records diagnosed with 41 diseases was analyzed, with 95 optimized symptoms selected as independent variables. The study compares the performance of Decision Tree, Random Forest, and Naïve Bayes classifiers. Early disease prediction can improve patient care and community services, making this research valuable in addressing health-related issues in (3).

An estimation model for heart disease prediction using bagging techniques was suggested by Annwesha Banerjee et al. This suggested model used Naive Bayes, K Nearest Neighbour, and Logistic Regression as its base learner. The proposed model was trained by data collected from UCI data repository. The motivation of the work was to achieve a better prediction result applying multiple bagged classifiers. Bagged Logistic Regression, Gaussian Naïve Bayes and K nearest neighbour had achieved accuracy of 82.8%, 82.5% and 83.2% respectively in (4).

K. Polaraju et al. suggested a multi-linear regression analysis-based model, and the significance of each individual coefficient was examined using the T test and the F test. Authors worked on a dataset consisting of 3000 instances, description of which was presented in the paper. The major claim of authors through the work was that regression model outperformed other machine learning model in (5).

Enriko, I Ketut, et al. produced a model applying Naive Bayes, Decision Tree, and K-Nearest Neighbor (KNN) for heart disease prediction that took into account 8 different features and had an accuracy of 81.85%. In the initial phase the authors applied 10 Fold cross validation mechanism also. For performing the KNN analysis, Microsoft Excel and Macro Visual Basic (Excel Macro) were used in (6) .

Vanitha Guda, Shalini K., and Shivani C. sought to predict heart disease by employing an innovative hybrid approach that combined the power of a Random Forest and the interpretability of a Linear Model. Leveraging the Framingham and Cleveland datasets, both well-established sources of heart disease-related information, the researchers devised a method that harnessed the non-linear modelling capabilities of the Random Forest in (7).

A model for heart disease prediction utilising Random Forest classifier, Naïve Bayes , J48 algorithm was proposed by Sanchayita Dhar et al. They used ERIC to get data. The proposed model was implemented through WEKA. It was observed that Random Forest outperformed the performance of other two classifiers in[8].

A machine learning-based model developed by Riddhi Kasabe to forecast cardiac illness was 87% accurate. In this work author claimed that Naive Bayes and Random Forest are able to identify the detail insight of data compare to other techniques. A detail description of the classification algorithm was represented in the paper in[9].

S. Shylaja et al. in[10] created a hybrid model for heart disease prediction where SVM and ANN were combined. The details of system architecture were described in the paper. The data set was collected form UCI data repository. For performance evaluation accuracy, specificity and sensitivity were used and it was observed. The significance of the proposed model was justified by comparison the performance of RIPPER , Naïve Baye, SVM and ANN.

A method of k-modes clustering with Huang beginning was put up by M. Chintan et al., and it can increase classification accuracy. Models like the Decision Tree classifier (DT), Multilayer Perceptron (MP), Random Forest (RF), and XGBoost (XGB) were employed. GridSearchCV was used to fine-tune the model's parameters in[11].

Sadiq Jaffer et al. proposed a work of heart disease prediction using different machine learning techniques. Decision Tree, KNN, Logistic Regression and Random Forest were used and their comparative analysis was also presented. The model was trained with UCI heart disease dataset. The major observation out of the work was that Random Forest had achieved greater accuracy over the other applied algorithm in[12].

V. Mane et al. represented a work of heart disease prediction using multiple classifiers. Different basic machine learning methods, ensemble learning mechanisms along with deep learning methods were applied in their proposed work. Information Gain, Fisher Score, and correlation coefficient are the methodologies that were applied for feature selection in[13].

Applying different boosting algorithm and SVM authors proposed a work of heart disease prediction where the issues out of missing data and imbalanced data were addressed for better performance. For data normalization MinMax normalization technique was applied. Along with that to address the issue of missing data KNN, Random Forest and Multiple Imputation by Chained Equation were applied in[14].

Emran Kabir Hashi et al. proposed a method of heart disease prediction which was based on Hyper parameter tuning. The dataset was collected from UCI data repository on which the model was trained and tested. Redundant value and missing value issues were been addressed by author. The classifiers used were Logistic Regression, K Nearest Neighbour, Support Vector Machine, Decision Tree and Random Forest in[15].

Amin et al. used a layered neural network in their study to get an incredibly low error rate when performing analysis for the incidences of heart conditions. The proposed neural network was consisting of 12 input nodes and 10 hidden nodes. Levenberg-Marquard algorithm was applied for training the model. Genetic Algorithm was utilized in the proposed work for adjustment of weights where mean square function was used as fitness function in[16].

Machine learning was used by Khoudrifi et al. to compare algorithms with various performance metrics. In some circumstances, each algorithm performed better than it did in others. The models most likely to perform well on the data set utilised in this study include K-NN, RF, and Multilayer Perceptron (MLP) with hybrid Particle Swarm Optimisation (PSO) and Ant Colony Optimisation (ACO) in[17].

Sajja, T.K., and Kalluri, H.K. proposed a model for predicting cardiac disease using Convolutional Neural Networks, and they also compared it to other models such as Logistic Regression, K-Nearest Neighbours (KNN), Naive Bayes (NB), Support Vector Machines (SVM), and Neural Networks (NN). The detail description of dataset was represented through visualization in the paper. The performance measured was evaluated through accuracy and AUC-ROC curve in[18].

Subramani S et al. proposed a model for heart disease prediction applying various machine learning method. Gradient Boostinging Decision Tree was applied for feature selection in this work. The various methods applied over were Naïve Bayes, Decision Tree, Support Vector Machine, Logistic Regression, K Nearest Neighbour and Random Forest. The average AUC score achieved was 0.88 in[19].

Majumder et al. proposed an explainable hybrid method for heart disease prediction applying Logistic Regression, Naïve Bayes, K Nearest Neighbour, Support Vector Machine, Kernel SVM, Random Forest and Artificial Neural Network. Based on accuracy, sensitivity, specificity best classifier was being chosen as final classifier in[20].

2 Proposed Methodology

This paper represents a model for heart disease prediction using an optimized novel weighted ensemble mechanism. The model has been trained using the dataset collected from UCI data repository. After selection of competent features, a weighted ensemble voting classifier has been used then performance of the model has been optimized applying Firefly algorithm

2.1 Data set Description

The collected dataset from UCI data repository has all total 14 features[21] which are age, sex,cp, trestbps, chol,fbs,restecg,thalach, exang ,oldpeak, slope, ca, thal and target. All the independent features are not equally impactful in decision making so in the next step feature selection method has been applied. The description of dataset has been put below.

Age: Age is a risk factor for cardiovascular disease, and the risk increases with older age.

Sex: This is a categorical field where 0 represents female and 1 represents male.

Cp : Chest Pain is a categorical variable represents different types of chest pain - 0 for asymptomatic, 1 for atypical angina, 2 for non-angina pain, and 3 for typical angina. Angina is chest pain caused by reduced oxygen flow to the heart.

Tresbps: Resting Blood Pressure is a continuous field represents the blood pressure of patients at rest. High resting blood pressure may indicate the presence of heart diseases.

Chol: Cholesterol is a continuous field represents the level of cholesterol in the blood. High cholesterol levels are associated with an increased risk of heart diseases.

Fbs: Fasting Blood Sugar is a categorical field indicates whether the patient's blood sugar level is above 120mg/dl (1 for above 120 mg/dl, 0 for not above 120 mg/dl).

Restecg : Resting Electrocardiogram is a categorical field represents the results of an echocardiogram when the patient is at rest. It provides information about the heart's health.

Thalach: Maximum Heart Rate is a continuous field represents the maximum heart rate observed during stress.

Exang: Exercise-Induced Angina is a categorical field indicates whether the patient experiences angina during a stress test (1 for feeling angina, 0 for no angina).

Old peak: Oldpeak (ST Depression) indicates the ST segment's decrease during exercise, which can be relevant for diagnosing heart diseases.

Slope: This categorical field indicates the slope of the ST segment during exercise - 0 for descending, 1 for flat, and 2 for ascending.

Ca: Number of colored Blood Vessels field represents the number of blood vessels colored by the radioactive dye during the thallium stress test.

Target: This is the dependent variable in the dataset, identifying whether a patient has heart disease or not (0 for heart disease, 1 for no heart disease).

The histogram of the used dataset has shown in the Figure 1 .

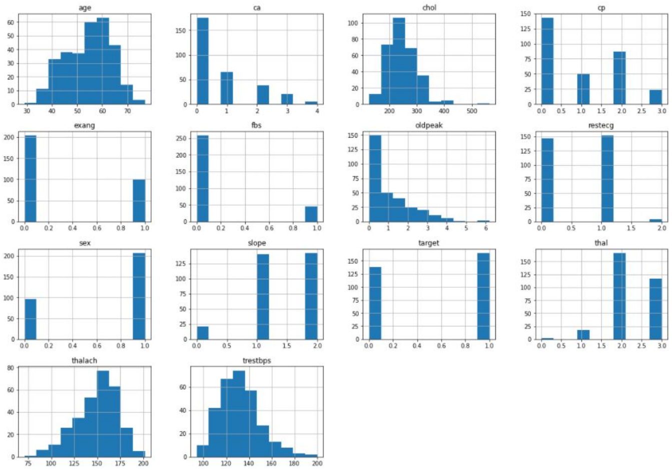
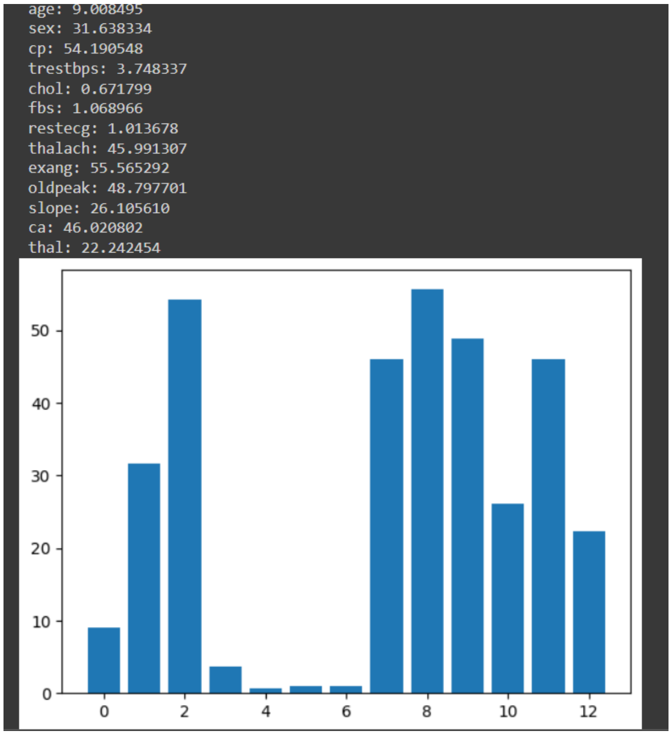


Fig 1. Dataset Histogram

2.2 Feature Selection

In this phase Chi Square Test has been applied for feature selection. The literature with data analysis with Chi Square Testing has been widely circulated in versatile domain including disease prediction. It has been observed that Chi Square Tests are ordinarily used to examine the order correspondence between observed result and expected outcome. The Chi Square Test is proven most appropriate test when the data is from a random sample. The Chi Square Test helps to identify which categorical features are more informative and have a significant association with the binary target variable. Features with a high Chi Square statistic and low p-value indicate stronger associations and are more likely to contribute to the classification outcome. When dealing with binary classification the Chi Square Test can be used to assess the relationship between each categorical feature and the binary target variable. Models built using selected features from the Chi Square Test are more interpretable since they focus on the most relevant and meaningful predictors. Applying this mechanism total ten competent independent features have been selected which are age, sex,cp, trestbps, thalach, exang ,oldpeak, slope, ca, and thal. The feature importance applying Chi Square Test has been shown in the Figure 2 .



```
age: 9.008495
sex: 31.638334
cp: 54.190548
trestbps: 3.748337
chol: 0.671799
fbs: 1.068966
restecg: 1.013678
thalach: 45.991307
exang: 55.565292
oldpeak: 48.797701
slope: 26.105610
ca: 46.020802
thal: 22.242454
```

Fig 2. Feature importance applying Chi Square Test

2.3 Model development Applying Voting Classifier

In this phase a weighted Voting Classifier has been applied for classification.

The Voting Classifier is an ensemble machine learning technique that combines the predictions from multiple individual classifiers to make a final prediction. It is a type of model averaging, where each classifier contributes its predictions, and the majority vote or the average is used to make the final decision. The Voting Classifier can be applied to both classification and regression problems, but we'll focus on its usage for classification tasks. The Voting Classifier typically combines two or more diverse classifiers (e.g., Decision Trees, Random Forest, Logistic Regression, Support Vector Machines, etc.). These individual classifiers can be different in their learning algorithms, hyper parameters, or feature representations.

There are two main voting methods employed in the Voting Classifier. In our proposed method Soft voting technique has been applied.

2.4 Soft Voting

In soft voting, the individual classifiers provide probability scores (confidence levels) for each class label. The final prediction is determined by averaging these probabilities and selecting the class with the highest average probability.

Let p_i(c) be the probability of classifier $C_i$predicting class c for input X, and let y_{vote}be the final prediction of the Voting Classifier. This can be represented through equation no 1.

Then, for each class c∈C

$$y_{vote}(X) = argmax(\frac{1}{N}\sum_{k=i}^{N} p_i\,(c)) \tag{1}$$

The Voting Classifier selects the class label with the highest average probability.

The learners used in this ensemble voting classifiers are Decision Tree, Adaboost and Gradient Booting. Weights of this proposed classifier has been set based on the performance of each base classifier. Accuracy, Sensitivity and Specificity have been considered for weight calculation which has been represented in below equation no 2. Accuracy can be used to measure a model's overall performance. A great statistic is accuracy, but only when datasets are symmetric. Recall or sensitivity refers to the percentage of positives that were accurately categorised as positives. Specificity, which measures the proportion of true negatives to all other negatives in the data, is another metric. Sensitivity is a crucial factor in the model's explanation since proper diagnosis of patients is essential for treatment to start, which could have catastrophic repercussions if done incorrectly.

$$W_i = Ac_i + Sp_i + 2 * Sn_i \tag{2}$$

Where Wi is the weight for base classifier i.

Ac_i is the Accuracy of base classifier i

Sp_i is the Specificity of classifier i

Sn_i is the Sensitivity of classifier i

In this experiment the classifiers considered are Decision Tree, Adaboost and Gradient Booting. As the weight for the voting ensemble classifier can only be 0 to 1 scale so the calculated weight in equation no 2 has been scaled to 0 to 1 range.

2.5 Model Optimization applying Firefly Algorithm

In this phase the performance of weighted ensemble voting classifier has been optimized applying Firefly optimization. Nature serves as the inspiration for Firefly. Yang, X. S.'s FA is a population-based optimisation technique that imitates the attraction of a firefly to a flashing light[22].

The Firefly Algorithm is a nature-inspired optimization algorithm, and its mathematical representation involves defining equations for the attraction between fireflies and their movement in the search space. It consists of following main steps.

1. Initialization: Fireflies are randomly scattered in the search space, and each firefly represents a potential solution to the optimization problem.

2. Attractiveness: The attractiveness of a firefly is determined by the objective function value of the corresponding solution. A better solution has a higher attractiveness, and the goal is to maximize this value.

3. Movement: Fireflies move towards other fireflies in the search space, and their movement is influenced by two main factors:

- Attraction: Fireflies are attracted to brighter fireflies (i.e., solutions with higher objective function values).
- Distance: Fireflies are repelled by other fireflies that are too far away.

4. Intensity of Light: The intensity of light produced by a firefly corresponds to its objective function value. As the algorithm progresses, the fireflies' positions are updated, and they move towards brighter fireflies.

Convergence: As the iterations progress, fireflies converge towards better solutions, and the algorithm tries to find the optimal or near-optimal solution.

3 Result and Discussion

In this section details observations through the experiments have been represented and analysed.

3.1 Observation: Impact of feature selection

Initially the voting classifier has been applied on the raw dataset (without applying any feature selection) and accuracy achieved through this is 73.684%. The confusion matrix and Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve have been shown in the Figures 3 and 4 below.

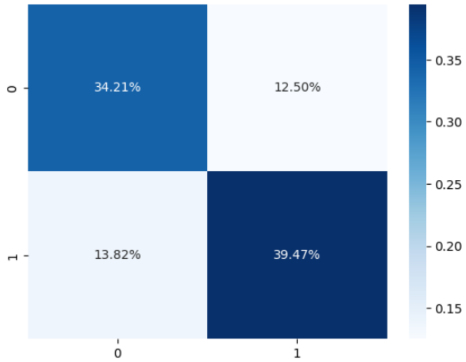


Fig 3. Confusion Matrixgenerated through the proposed model using raw dataset (without featureselection)

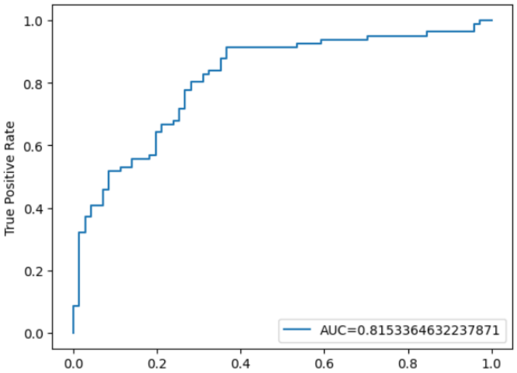


Fig 4. Area Under the Curve (AUC)of the Receiver Operating Characteristic (ROC) curve of the proposed classifiertrained with raw dataset(without feature selection)

In the next phase of experiment Chi Square method has applied for selection of competent features. The proposed voting classifier trained with new feature set has achieved higher accuracy which has discussed in below section.

3.2 Observation: Application of Weighted Ensemble Voting Classifier (with selected features through Chi Square)

In this phase the outcome of application of Weighted Ensemble Classifier on the selected feature set has been observed.

Individual accuracy achieved applying Decision Tree, Adaboost and Gradient Boosting is 78.94%, 82.89% and 78.94% respectively. With the achieved accuracy, sensitivity and specificity weights have been calculated for each classifier following the equation no 2 to build the Ensemble Voting Classifier.

The calculated weight= [0.8,1,0.9]

Applying these weight values an ensemble classifier has been built which achieved an enhanced accuracy of 84.21%. In the Figure 5 below the generated confusion matrix out the experiment of this section has shown.

In the below Table 1 the performance of individual classifiers and our proposed ensemble voting classifier have been shown.

Table 1. Performance Analysis of individual and our proposedCustomized voting classifier

Classifiers	**Accuracy Score**
Decision Tree	78.94%
Adaboost	82.89%
Gradient Boosting	78.94%
Customized Ensemble Voting Classifier	84.21%

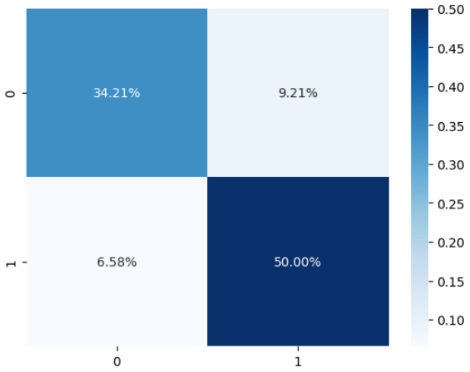


Fig 5. Generated Confusion Matrix applying Weighted Ensemble Classifier over selected dataset

For analyzing the performance of the proposed model Precision Recall and F1 Score has also been measured. In the Table 2 below the performance matrices considered- accuracy, precision, recall and F1 scores have shown.

Precision is the proportion of true positive predictions (correctly predicted positive instances) out of all positive predictions made by the model. In other words, it measures how many of the instances predicted as positive are actually positive. Precision can be represented with below mentioned equation no 3.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{3}$$

Recall is the proportion of true positive predictions out of all actual positive instances in the dataset which can be represented by equation no 4. It measures how well the model identifies positive instances.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{4}$$

The F1 score is the harmonic mean of precision and recall. It is a balanced measure that takes into accounts both precision and recall. The F1 score is especially useful when the data is imbalanced, meaning one class is much more prevalent than the other. F1 score can be represented using equation no 5.

$$F1\ score = \frac{(2 * Precision * Recall)}{(Precision + Recall)} \tag{5}$$

Table 2. Overall performance of the Customized Weighted Ensemble Voting Classifier

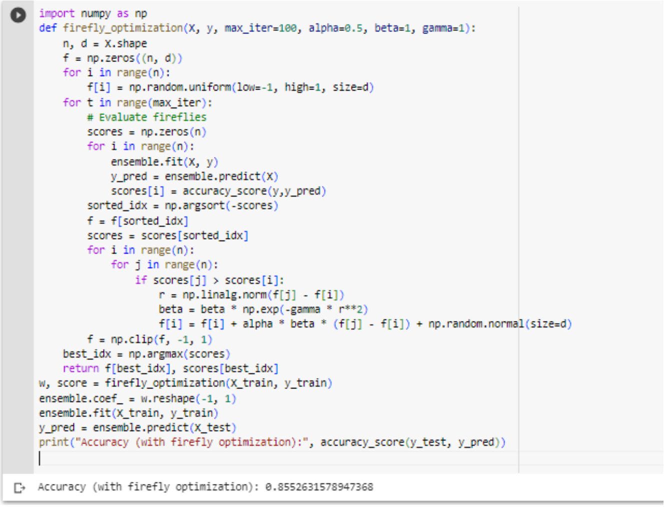
Matrices	**Score**
Accuracy	84.21%
Precision	78.78%
Recall	83.86%
F1 Score	81.24%

3.3 Observation: Application of Firefly Optimization

In this phase of work Firefly optimization algorithm has been applied to achieve better performance.100 iterations all together has been used where accuracy score has been used as objective function. The optimized accuracy by applying this nature inspired algorithm is 85.52%. The sample run of the used FireFly algorithm has shown in the Figure 6 below.

The Area Under the Curve have been shown in the Figure 7 below.

Comparative analysis of our proposed model with few recent works in this field has shown in the Table 3 .



```
import numpy as np
def firefly_optimization(X, y, max_iter=100, alpha=0.5, beta=1, gamma=1):
    n, d = X.shape
    f = np.zeros((n, d))
    for i in range(n):
        f[i] = np.random.uniform(low=-1, high=1, size=d)
    for t in range(max_iter):
        # Evaluate fireflies
        scores = np.zeros(n)
        for i in range(n):
            ensemble.fit(X, y)
            y_pred = ensemble.predict(X)
            scores[i] = accuracy_score(y,y_pred)
        sorted_idx = np.argsort(-scores)
        f = f[sorted_idx]
        scores = scores[sorted_idx]
        for i in range(n):
            for j in range(n):
                if scores[j] > scores[i]:
                    r = np.linalg.norm(f[j] - f[i])
                    beta = beta * np.exp(-gamma * r**2)
                    f[i] = f[i] + alpha * beta * (f[j] - f[i]) + np.random.normal(size=d)
        f = np.clip(f, -1, 1)
    best_idx = np.argmax(scores)
    return f[best_idx], scores[best_idx]
w, score = firefly_optimization(X_train, y_train)
ensemble.coef_ = w.reshape(-1, 1)
ensemble.fit(X_train, y_train)
y_pred = ensemble.predict(X_test)
print("Accuracy (with firefly optimization):", accuracy_score(y_test, y_pred))
```

```
Accuracy (with firefly optimization): 0.8552631578947368
```

Fig 6. Sample run of FireFly Algorithm applied

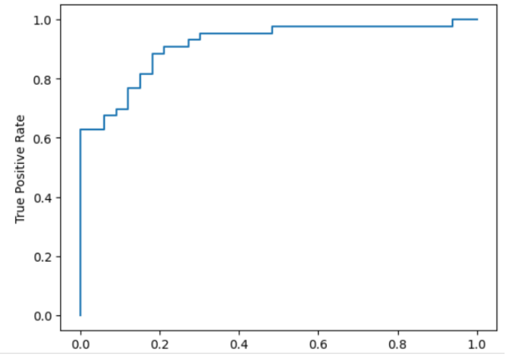


Fig 7. Area under the Curve of the Optimized classifier

Table 3. Comparative analysis of our proposed model with few recent works

Proposed Work	Year of Publication	Observation
(1)	2021	Applied Random Forest. Logistic Regression K Neighbour, Support Vector Machine, Decision Tree and XGBoost and achieved 80.3%, 83.3%, 84.26%, 83.29% .82.33% and 71.4% respectively.
(2)	2022	Applied Logistic Regression, LightGBM, XGBoost, Gaussian Naïve Bayes, Support Vector Machine (SVM), and achieved 80.32%, 78.68%, 80.32%, 77.04%, 73.77%, and 88.5% accuracy respectively
(4)	2022	Applied Bagged Logistic Regression, Gaussian Naïve Bayes and K Nearest Neighbour and achieved 82.8%, 82.5% and 83.2% accuracy respectively. No feature selection method applied.
(9)	2020	Applied Naïve Bayes and Random Forest for detail data insight. Accuracy achieved : 87%
(13)	2023	Information Gain, Fisher Score, and correlation coefficient applied for feature selection. For classification applied ensemble learning and deep learning classifier.
(19)	2023	Applied Naïve Bayes, Decision Tree, Support Vector Machine, Logistic Regression, K Nearest Neighbour and Random Forest. For feature selection applied: Gradient Boosting Decision Tree. Average AUC score : 0.88

Continued on next page

Table 3 continued

[20]	2023	Applied machine learning methods-Logistic Regression, Naïve Bayes, K Nearest Neighbour, Support Vector Machine, Kernel SVM, Random Forest with 83%, 83%, 82%, 86%, 86% and 83% accuracy. Applying deep learning achieved better performance No feature selection method applied
Our Proposed Model	2023	Applied Ensemble Voting Classifier. For feature selection applied Chi Square Test. AUC score=0.92 Accuracy 85.52%

4 Conclusion

In this paper, a heart disease prediction model has been proposed applying customized weighted ensemble classifier. For considering weight of base classifier not only accuracy but also specificity and sensitivity have also been considered. The achieved accuracy through the customized voting classifier is 84.21%. Over that FireFly optimization has been applied for better performace and accuracy has been increared to 85.52%. This study makes a contribution to heart disease prediction using Artificial Intelligence and Machie Learning. As future scope we would implement model collecting real time data through IoT based application.

Acknowledgement

The authors thank their Organizations for supporting the research work

References

1) Bharti R, Khamparia A, Shabaz M, Dhiman G, Pande S, Singh P. Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. . Available from: https://doi.org/10.1155/2021/8387680.
2) Karthick K, Aruna SK, Samikannu R, Kuppusamy R, Teekaraman Y, Thelkar AR. Implementation of a Heart Disease Risk Prediction Model Using Machine Learning. *Computational and Mathematical Methods in Medicine*. 2022;2022:1–14. Available from: https://doi.org/10.1155/2022/6517716.
3) Grampurohit S, Sagarnal C. Disease Prediction using Machine Learning Algorithms. *2020 International Conference for Emerging Technology (INCET)*. 2020;p. 1–7. Available from: https://doi.org/10.1109/INCET49848.2020.9154130.
4) Majumder AB, Gupta S, Singh D. An Ensemble Heart Disease Prediction Model Bagged with Logistic Regression, Naïve Bayes and K Nearest Neighbourïve Bayes and K Nearest Neighbour. *Journal of Physics: Conference Series*. 2022. Available from: https://doi.org/10.1088/1742-6596/2286/1/012017.
5) Polaraju K, Prasad DD. Prediction of Heart Disease using Multiple Linear Regression Model. 2017. Available from: https://www.ijedr.org/papers/IJEDR1704226.pdf.
6) Enriko KA, Suryanegara M, Gunawan D. Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters. 2016. Available from: https://core.ac.uk/download/229279242.pdf.
7) Vanitha Guda, Shalini K, Shivani C. Heart Disease Prediction Using Hybrid Technique. *Journal of Interdisciplinary Cycle Research*. 2020;XII.
8) Dhar S, Roy K, Dey T, Datta P, Biswas A. A Hybrid Machine Learning Approach for Prediction of Heart Diseases. In: 2018 4th International Conference on Computing Communication and Automation (ICCCA). IEEE. 2018;p. 1–6. Available from: https://doi.org/10.1109/CCAA.2018.8777531.
9) Kasabe R, Narang G. Heart Disease Prediction using Machine Learning. *International Journal Of Engineering Research Technology* . 2020;09(08).
10) Shylaja S, and RM. Classifier is used for Heart Disease Prediction System. *International Journal of Engineering Research* . 2020;9. Available from: https://www.researchgate.net/publication/334965479_Hybrid_SVM-ANN_Classifier_is_used_for_Heart_Disease_Prediction_System.
11) 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.
12) Jaffer S, Pasha S, K SL. Heart Disease Prediction using. *Machine Learning International Journal of Advanced Research in Computer and Communication Engineering*. 2021;10. Available from: https://doi.org/10.17148/IJARCCE.2021.10727.
13) Mane V, Tobre Y, Bonde S, Patil A, Sakhare P. Heart Disease Prediction Using Machine Learning and Neural Networks. In: Shukla, K P, Singh, P K, Tripathi, K A, et al., editors. Computer Vision and Robotics. Springer Nature Singapore. 2023;p. 205–228. Available from: https://doi.org/10.1007/978-981-19-7892-0_17.
14) Louridi N, Douzi S, Ouahidi BE. Machine learning-based identification of patients with a cardiovascular defect. *Journal of Big Data*. 2021;8(1):133–133. Available from: https://doi.org/10.1186/s40537-021-00524-9.
15) Hashi EK, Zaman MSU. Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction. *Journal of Applied Science & Process Engineering*. 2020;7(2):631–647. Available from: https://doi.org/10.33736/jaspe.2639.2020.
16) Amin SU, Agarwal K, Beg R. Genetic neural network based data mining in prediction of heart disease using risk factors. *Proceedings of 2013 IEEE Conference on Information and Communication Technologies*. 2013;p. 1227–1231. Available from: https://doi.org/10.1109/CICT.2013.6558288.
17) Khourdifi Y, Bahaj M. Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization. *International Journal of Intelligent Engineering and Systems*. 2019;12(1):242–252. Available from: https://doi.org/10.22266/ijies2019.0228.24.
18) Sajja TK, Kalluri HK. A Deep Learning Method for Prediction of Cardiovascular Disease Using Convolutional Neural Network. *International Information and Engineering Technology Association*. 2020;34(5):601–606. Available from: https://doi.org/10.18280/ria.340510.
19) 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. Available from: https://doi.org/10.3389/fmed.2023.1150933.

20) Majumder AB, Gupta S, Singh D, Majumder S. An Explainable Hybrid Intelligent System for Prediction of Cardiovascular Disease. *Journal of Mines, Metals and Fuels*. 2023;71(5):687–694. Available from: https://doi.org/10.18311/jmmf/2023/34171.

21) Heart Disease Data Set. . Available from: https://archive.ics.uci.edu/ml/datasets/heart+disease.

22) Yang XS. Nature-Inspired Metaheuristic Algorithm. Luniver Press. 2008. Available from: https://staff.fmi.uvt.ro/~daniela.zaharie/ma2016/projects/techniques/FireflyAlgorithm/Yang_nature_book_part.pdf.