Proceedings of the SMART–2023, IEEE Conference ID: 59791
12th International Conference on System Modeling & Advancement in Research Trends, 22nd–23rd, December, 2023
College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, India

2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART) | 979-8-3503-6988-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/SMART59791.2023.10428617

Heart Disease Prediction Using Enhanced Whale Optimization Algorithm Based Feature Selection With Machine Learning Techniques

A. Lakshmi[1] and Dr. R. Devi[2]
[1]*Research Scholar, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu 600117*
[2]*Associate Professor, Department of Computer Science, School of Computing Sciences, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu 600117*
E-mail: [1]lakshmiaacw@gmail.com, [2]devi.scs@velsuniv.ac.in

***Abstract*—Cardiovascular disease, a type of heart disease, is the leading cause of death worldwide. Early detection of heart disease can help get proper treatment and save lives. Machine Learning (ML) models are becoming increasingly popular for use in a wide range of clinical diagnostic tasks. Making accurate predictions is essential for such tasks because the results can have a big impact on patients and reduce mortality. ML algorithms for efficient identification of heart disease plays an important role in healthcare, especially in cardiology. Initially, the Framingham heart disease dataset was collected from a Kaggle website for analyzing heart disease prediction. The preprocessing stage is applied to manage and remove the inappropriate data from the dataset. Then, an Enhance Whale Optimization Algorithm - based feature selection technique applied to the dataset to select the most relevant features (best-reduced feature divisions) for the detection of heart disease. Finally, machine learning classification algorithms, both conventional and hybrid methods, were implemented on the reduced feature dataset. The trained classifiers were evaluated in terms of accuracy, precision, recall and F1-score.**

Keywords: Cardiovascular disease, Machine Learning, Enhanced Whale Optimization Algorithm (EWOA), feature, weight

I. Introduction

Cardiovascular disease is the leading cause of death worldwide today. According to the recent statistics World Health Organization (WHO), heart disease is the most dangerous disease. Billions of people worldwide suffer from heart disease, and 12 million people die from these diseases every year. Many people experience symptoms that were previously unrecognized or overlooked before death. Heart disease has several major causes. Some of these can be high cholesterol levels, blood pressure and smoking, alcohol consumption, high sugar and physical inactivity, and hypertensive heart [1].

The main reason for this high number of deaths is not detecting this problem early. Early detection of heart disease can prevent many deaths in patients. Early heart disease prediction has always been vital in its control and diagnosis. Early diagnosis of heart disease can help provide more effective and accurate treatment to patients [2]. Hence, the need to develop such early prediction and clinical diagnostic systems is increasing daily. The main objective of such systems is that they should provide high accuracy with low operating cost.

Machine learning-based early detection of cardiac disease can assist patients, physicians, and policymakers in a number of ways [3]. For patients, it can empower them to take proactive steps to prevent or delay the onset of heart disease, such as modifying their lifestyle, taking medications, or undergoing interventions. It can also help them monitor their condition and seek timely medical help if needed. For clinicians, it can assist them in making informed decisions about diagnosis and treatment based on the best available evidence and the patient's preferences. It can also help them prioritize high-risk patients and allocate resources efficiently.

II. Literature Survey

Joshi et al. [4], carried out a thorough analysis of the ML applications for disease diagnosis. The paper described the many categories of healthcare diagnosis systems that classify data using ML algorithms. The main goal of the analysis is to provide in-depth information about how ML and AI are used in disease diagnosis. Last but not least, this study provides an abundance of knowledge to the practising physician and also helps to assess the ML algorithms used in disease diagnosis by providing prediction results on disease datasets.

Saurabh et al. [5] did a thorough analysis that compares various techniques on two datasets related to medical

diagnosis, drug reviews, clinical trials and biomedical literature. The datasets, D1 takes into account information pertaining to conjunctivitis, diarrhoea, stomach ache, coughing, and nausea, whereas D2 includes the common dataset known as WebKB4. SVM, MLP, and the Random Forest (RF) were included in the machine learning algorithms known as the Radial function. It also evaluates the effect of combining multiple features and applying dimensionality reduction methods. The analysis presents a comprehensive analysis of the results and discusses the implications for developing effective and robust text classification systems for healthcare domains. Finally, their experimental findings show the best accuracy of 97%.

A novel machine learning-based approach for identifying heart diseases was proposed by Kanwal, Samina, et al.[6]. The authors classify the patients into groups that are healthy or diseased by applying a classification algorithm after using a genetic algorithm to identify the most pertinent features from a sizable dataset of medical records. The machine learning methods DL, SVM, NN, NB, and LR use the attributes chosen by the Genetic Algorithm (GA) as input. The model is implemented using the two heart disease datasets. The model is implemented using the two heart disease datasets. Accuracy, precision, and f-measure are used to gauge the evaluation of the results. In terms of accuracy, the suggested model performs at 92%. 96% of the results are obtained in terms of precision.

Using machine learning, hyper-parameter optimization, and genetic algorithms, Jinny et al. [7] presented a unique predictive model for CHD. To choose the best features from a wide range of medical and demographic parameters, genetic algorithms are employed. Next, the hyper-parameter optimization to adjust the parameters of several machine learning models, including ANN, LR, SVM, and RF. The model performance results indicate that, out of all the models, the suggested model obtains the highest levels of accuracy, sensitivity, specificity, and F1-score. Additionally, the feature significance analysis and cross-validation to show the interpretability and resilience of our model. In addition to being a trustworthy tool for early CHD prediction, this model can assist doctors in making well-informed decisions on CHD treatment and prevention.

Zamani et al. [8] devised a meta-heuristic approach for feature selection called FSWOA. It is created based on Humpback Whale hunting strategies, which consist of three key steps: encircling prey, spiral bubble-net attacking, and searching for prey. This algorithm's performance is examined using different heart disease medical datasets. The result shows with acceptable accuracy, the dimension of medical dataset also reduced.

Hegazy et al. [9] made some changes to the fundamental WOA structure and obtained an improved whale optimisation algorithm (IWOA). The IWOA is evaluated using various functions to transform continuous values into binary solutions. The proposed methodology uses the KNN classifier with a feature selection to identify feature subsets that improve classification precision while reducing the number of selected features. The optimisation results show that the IWOA improves the basic whale optimisation technique and also outperforms five other optimizers in terms of robustness and generalisation capacity. In addition, 27 datasets were used to complete this task, with IWOA being compared to WOA, PSO, GA, ALO, and GWO.

Vijayarani et al. [10] devised a ML approach to diagnose liver diseases based on clinical data. The performance of the classification algorithms, such as SVM and NB, on a dataset of 583 patients with different liver conditions. The findings showed that SVM algorithm achieves higher accuracy, sensitivity, and specificity than the NB algorithm in predicting liver diseases.

Olaniy et al. [11] obtained accuracy of 63% and 70%, respectively, for each of the proposed techniques, in a study using BPNN neural network and radial function on BUPD dataset. In future investigations, it is advised that the factors impacting the disease be investigated utilising an optimisation method. Early diagnosis of the condition would result in good outcomes for both physicians treating the disease in its early stages and for patients, as it would lower treatment expenditures.

Junejo et al. [12], developed deep learning techniques (DLTs) that were used to analyse stable CVD, providing vital information to help reduce misdiagnosis in the robust healthcare industry (RHI). The goal is to first do molecular diagnostics (MD), and then use DLTs to synthesise and characterise data from CVD patients. A machine learning approach is used to analyze and predict functional recovery in CVD patients. ANN performs better than KNN when considering the test dataset. Also, the KNN accuracy ratio performs better. Then, multiple feature selection approaches are used to rank the attributes that contribute most to the CVD classifier, therefore reducing the number of diagnostic procedures required for the patient.

III. Proposed Method

Figure 1 described the phases of proposed heart disease prediction system using EWOA feature selection technique

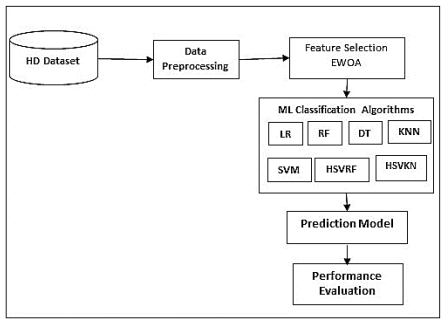

Fig. 1: Proposed heart disease prediction system using EWOA

A. Data Collection

The Framingham Cardiovascular dataset was obtained from the Kaggle website. The Framingham binary class heart disease dataset is a collection of data from a long-term study of cardiovascular health in the town of Framingham, Massachusetts [13]. The dataset contains 4238 records of individuals who were examined every two years for a period of 10 years. Each record has 16 attributes, such as age, sex, blood pressure, cholesterol, smoking status, and whether or not the individual developed coronary heart disease (CHD) during the follow-up period. The dataset is useful for exploring the risk factors and predictors of CHD, as well as for developing and evaluating machine learning models for binary classification of heart disease. These datasets are probably going to contain clinical and medical data, target variable indicating the presence or absence of heart disease. The description of Framingham dataset is given in table 1.

Table 1: Description of Features

S.No.	Attribute Name	Attribute Description
1.	male	A binary characteristic that denotes a person's gender (one for men and zero for women)
2.	age	The person's age.
3.	education	The individual's level of education (a category feature with values such as 1,2,3, and 4)
4.	currentSmoker	Indicates if the person now smokes (1 = yes, 0 = no).
5.	cigsPerDay	How many cigarettes a person smokes each day
6.	BPMeds	If the person is taking blood pressure medicine, it is indicated by a 1 or a 0.
7.	prevalentStroke	Prevalent whether the person has ever had a stroke (1 for yes, 0 for no)
8.	prevalentHyp	Prevalent Whether the person experiences hypertension on a regular basis (1 for yes, 0 for no).
9.	diabetes	A person's level of diabetes (yes or no, 1 or 0).
10.	totChol	The person's total cholesterol level.
11.	SysBP	Systolic blood pressure
12.	diaBP	Diastolic blood pressure
13.	BMI	Body Mass Index is calculated based on a person's height and weight
14.	heartrate	The person's resting heart rate
15.	glucose	The amount of glucose present in a person's blood.
16.	TenYearCHD	A binary label that states whether or not a person is at risk for coronary heart disease within the next ten years (yes or no)

B. Data Preprocessing

Data Preprocessing is the method of transforming an unstructured data into accessible format for analysis using Machine Learning (ML) algorithms [14]. Preprocessing can include tasks such as data cleaning, normalization, and encoding. The primary goal of preprocessing is to improve the quality of the data, as well as to reduce the noise, outliers, missing values, and inconsistencies that can affect the performance of the machine learning models. Preprocessing can also help to discover hidden patterns and insights from the data, as well as enhance the interpretability and explainability of the results.

C. Feature Selection

The Whale Optimisation Algorithm (WOA), a strategy for optimisation inspired by nature and the hunting habits of humpback whales [15]. The WOA excels at resolving challenging optimisation issues in a variety of industries, including engineering, finance sector, healthcare, and more. The EWOA is an improved version of the WOA, which overcomes some of its drawbacks such as slow convergence and low exploration ability. The EWOA introduces two new operators: the spiral shrinking operator and the chaotic spiral operator, which enhance the exploitation and exploration phases of the algorithm, respectively. The EWOA also adopts a dynamic parameter control strategy to balance the exploration and exploitation abilities throughout the search process.

The pseudocode of the Enhanced Whale Optimisation Algorithm (EWOA):

```
Step1: Initialize population of whales (feature subsets)
        Initialize max_iter as maximum number of iterations
        For iteration in range(max_iter):
                Calculate fitness values for each whale
        using
                 ObjF(Features)
                Rank whales based on fitness values
Step 2: Calculate T and F coefficients based on the iteration
 number
        T = 2 - 2 * iteration / max_iter
        F = 2 * iteration / max_iter
        For each whale:
                Select a random whale as the target whale
                Select another random whale as the source
 whale
Step 3: Update whale's position based on target and source
whales
        for feature in features:
         r1 = random() # Random value [0, 1]
         r2 = random() # Random value [0, 1]
         A1 = 2 * A * r1 - T
         C1 = 2 * r2
         D = abs(C1 * target_whale[feature] –
         source_whale[feature])
         new_position = target_whale[feature] - A1 * D
 Step 4: Update feature position based on importance ranking
 if new_position > 0 and new_position <= 1:
                feature_importance =
                calculate_feature_importance(feature)
                new_position *= feature_importance
                whale[feature] = new_position
                Select the whale with the highest fitness
                value as the final selected features.
```

ObjF(Features) in this pseudocode stands for the objective function that assesses the fitness of a feature subset. The exploration and exploitation trade-off during position updates is governed by the parameters A1 and C1.

The positions of the source and target whales as well as the ranking of the feature's relevance are taken into account when updating each feature's position. EWOA's inspiration and background come from the occurrence of whale group hunting, in which humpback whales cooperate and use particular hunting techniques to catch their prey.

The model employs whale movement patterns and communication methods to adaptively look for the best answers in challenging search regions. EWOA features an adaptation mechanism for a dynamic search space. As a result, the algorithm can concentrate on promising sections of the solution space while systematically probing less-explored regions. This adaptability aids in increasing convergence and escape local maxima. Greedy Selection for EWOA incorporates a technique for greedy selection that gives the best solutions priority at each iteration.

The Enhanced Whale Optimisation Algorithm (EWOA) is used for feature selection and it involves optimising feature subsets to increase predictive performance. Following the application of EWOA, the selected features are employed for further analysis and prediction of heart disease. The EWOA algorithm is applied individually on the binary class dataset. EWOA analyses multiple subsets of features during optimisation to determine the most relevant ones for heart disease prediction. Following the execution of EWOA, a subset of features is chosen depending on their significance or contribution to predicted accuracy. The binary class heart disease dataset, which includes several features that describe people's traits and medical data. The columns reflect several characteristics, and each row represents a patient. By applying an Enhanced whale optimization algorithm to the heart disease dataset, it detects heart disease and derives results for the target attribute, which indicates the presence or absence of heart disease. Table 2 shows the optimal features selected using EWOA from the heart disease dataset.

Table 2: List of Optimal Features Selected Using EWOA

Dataset	Optimal Features
Framingham Heart Disease Dataset	male, age, BPMeds, prevalentStroke, prevalentHyp, diabetes, totChol, sysBP, diaBP, heartRate

D. Classification Algorithm

Five popular machine learning algorithms that can be applied to regression or classification applications are Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), and K-Nearest Neighbor (KNN). A linear model called SVM looks for the best hyperplane to divide the data into distinct classes. RF is an ensemble technique that builds a more reliable and accurate prediction by combining several decision trees. Data is divided into branches using the hierarchical DT structure according to certain parameters. A LR method is used to calculate the likelihood of a binary result. A label is assigned using KNN, a non-parametric technique, based on the majority vote of the neighbors who are nearest to the object in the feature space. Depending on the data and problem domain, these algorithms have benefits and drawbacks. In high-dimensional and sparse data, for instance, SVM are better suited for achieving effective generalization with fewer support vectors. Due to its superior noise and outlier handling capabilities, RF is better suited for low-dimensional and dense data sets.

Tuning the parameters in machine learning classification algorithms is an important step to optimize the performance and accuracy of the system. The performance of these methods was analysed. A hybrid model for disease prediction is a combination of different methods that can leverage the strengths of each approach and overcome their limitations. It can improve the accuracy and reliability of the predictions by reducing the bias and variance of individual methods. Therefore, the most commonly used algorithms such as SVM, RF, and KNN were selected to construct the hybrid methods. The performance of Hybrid SVM-RF (HSVRF) and hybrid SVM-KNN (HSVKN) was analysed using evaluated metrics such as accuracy, precision, recall, and F1-score.

IV. Result and Evaluation

Table 3 shows the performance metrics of the classification techniques when applied on the binary dataset after using the Enhanced Whale Optimisation Algorithm.

Table 3: Performance Metrics of Proposed System

Classification Methods	Precision %	Recall %	Fl-Score %	Accuracy %
LR	66.15	66.13	65.96	67.01
DT	78.15	78.15	78.14	78.21
KNN	80.48	78.37	78.15	76.45
SVM	66.76	66.79	66.74	66.74
RF	74.06	70.01	68.24	70.01
HSVKN	79.34	78.01	77.88	78.01
HSVRF	86.70	86.44	86.38	85.79

The precision of a model is determined by dividing the total number of predicted positives by the ratio of genuine positives. This ratio indicates how effective the model is at classifying positive data. Recall gauges how sensitive the model is to identifying the positive class; it is expressed as the ratio of true positives to the total number of actual positives. The F1-score, which equalizes both measures and provides a single score that represents the overall quality of the heart disease prediction system, is the harmonic mean of precision and recall. HSVRF achieves the greatest

accuracy of 85.79%, Precision 86.70%, Recall 86.44% and F1-score 86.38% indicating that it effectively collected a high number of true positives. F1-score of HSVRF has the highest F1-score, balancing precision and recall. Based on these measures, the 'Hybrid Support Vector Machine and Random Forest (HSVRF)' technique appears to be the best-performing method for heart disease prediction on this binary class dataset, with the highest accuracy.

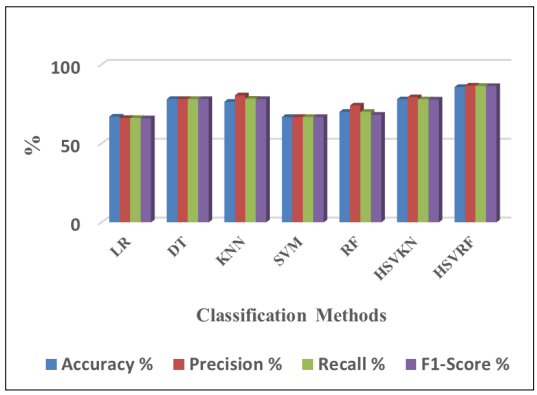

Fig. 2: Performance Analysis of the Proposed System

Figure 2 shows the performance measures for each model used to predict heart disease on the binary class dataset. This format presents the findings in a graphical format, making it easier to compare metrics across different models.

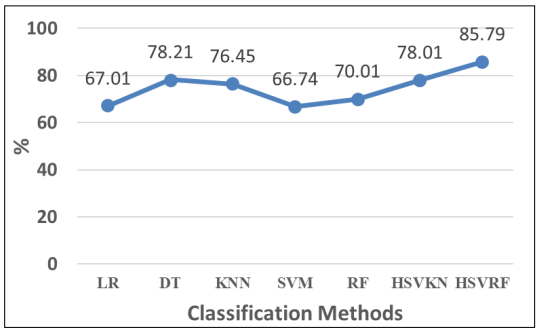

Fig. 3: Accuracy Analysis of EWOA

Figure 3 shows the accuracy of the Enhanced Whale Optimisation Algorithm (EWOA) when applied to a binary class dataset, as well as the accuracy of conventional and the proposed hybrid algorithms trained with feature selection using EWOA. The above figure shows the hybrid classification methods are giving better results when compared to the conventional classification machine learning methods.

V. Conclusion

This research work involves predicting cardiac disease using machine learning approaches on binary class dataset. To identify significant characteristics from the dataset, feature selection methods, notably the Enhanced Whale Optimisation Algorithm (EWOA) was used. Several machine learning methods were trained and evaluated for their performance in predicting heart disease, including LR, RF, DT, KNN, SVM, HSVRF, and HSVKN. The HSVRF classification method provided the best result for the Framingham dataset with an accuracy of 85.79%, and F1-score of 86.38%. The proposed prediction system acquired excellent accuracy as well as high precision, recall, and F1-score values, making them appropriate for their individual dataset.

References

[1] Ramalingam VV, Dandapath A, Raja MK, " Heart disease prediction using machine learning techniques: a survey", Int J Eng Technol. 2018;7(2.8):684–7.

[2] Islam, Md Touhidul, Sanjida Reza Rafa, and Md Golam Kibria. "Early prediction of heart disease using PCA and hybrid genetic algorithm with k-means." 2020 23rd International Conference on Computer and Information Technology (ICCIT). IEEE, 2020.

[3] Chandrasekhar, Nadikatla, and Samineni Peddakrishna. "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization." Processes 11, no. 4 (2023): 1210.

[4] Joshi, KK, Gupta, KK & Agrawal, J 2020, 'A Review on Application of Machine Learning in Medical Diagnosis,' 2nd International Conference on Data, Engineering and Applications (IDEA), Bhopal, India.

[5] Saurabh Kumar Srivastava, Sandeep Kumar Singh & Jasjit S Suri 2019, 'Effect of incremental feature enrichment on healthcare text classification system: A machine learning paradigm', Computer Methods and Programs in Biomedicine, vol. 172, pp. 35-51.

[6] Kanwal, Samina, et al. "An effective classification algorithm for heart disease prediction with genetic algorithm for feature selection." 2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC). IEEE, 2021.

[7] Jinny, S. Vinila, and Yash Vijay Mate. "Early prediction model for coronary heart disease using genetic algorithms, hyper-parameter optimization and machine learning techniques." Health and Technology 11 (2021): 63-73.

[8] Zamani, Hoda, and Mohammad-Hossein Nadimi-Shahraki, "Feature selection based on whale optimization algorithm for diseases diagnosis", International Journal of Computer Science and Information Security, Vol. 14, Issue 9, pp. 1243-1247, 2016

[9] Hegazy, Ah E., M. A. Makhlouf, and Gh S. El-Tawel, "Dimensionality reduction using an improved whale optimization algorithm for data classification", International Journal of Modern Education and Computer Science, Vol. 10, Issue 7, pp. 37, 2018.

[10] Vijayarani, S., and S. Dhayanand, "Liver disease prediction using SVM and Naïve Bayes algorithms", International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, Issue 4, pp. 816-820, 2015.

[11] Olaniyi, Ebenezer Obaloluwa, and Khashman Adnan, "Liver Disease Diagnosis Based on Neural Networks", Advances in Computational Intelligence, pp. 48-53, 2013.

[12] Junejo, Ar, Yin Shen, Asif Ali Laghari, Xiaobo Zhang, and Hao Luo, "Molecular Diagnostic and Using Deep Learning Techniques for Predict Functional Recovery of Patients Treated of Cardiovascular Disease", IEEE Access, Vol. 7, pp. 120315-120325, 2019.

[13] Mahmoud, Walaa Adel, Mohamed Aborizka, and Fathy Ahmed Elsayed Amer. "Heart Disease Prediction Using Machine Learning and Data Mining Techniques: Application of Framingham Dataset." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12.14 (2021): 4864-4870.

[14] Lakshmi, A., and R. Devi. "Comparative Analysis of Multiclass Heart Disease Prediction Classification Models using Preprocessing and Feature Selection." 2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC). IEEE, 2022.

[15] Gharehchopogh, Farhad Soleimanian, and Hojjat Gholizadeh. "A comprehensive survey: Whale Optimization Algorithm and its applications." Swarm and Evolutionary Computation 48 (2019): 1-24.