Wild Horse Optimizer and Support Vector Machine (SVM) Classifier Predicts the Heart Disease Converging Nature-Motivated Optimization and Machine Learning

Vishwanadham Mandala [1*], Srinivas Naveen Dolu Surabhi [2], V. R. Balaji [3], Dipak Raghunath Patil [4], Saiyed Faiayaz Waris [5], G. Shobana [6]

Abstract

Background: Heart disease is one of the most known and deadly diseases in the world and many people lose their lives from this disease every year. Early detection of this disease is vital to save people's lives. Machine Learning (ML), an artificial intelligence technology, is one of the most convenient, fastest, and low-cost ways to detect disease. Methods: This research work, presented a Wild Horse Optimizer (WHO) based feature selection and Support Vector Machine (SVM) developed a classifier for the forecasting of data related to heart diseases. The WHO algorithm that draws inspiration from the social behaviours of wild horses is presented in this work. Horses typically reside in groups consisting of a stallion, numerous mares, and young foals. Horses can be seen engaging in a variety of behaviours, including leading, grazing, chasing, and mating. The interesting quality that sets horses apart from other animals is their kindness. When a horse is decent, before they reach maturity, its foals break away from the herd and join different groups. To avoid the father mating with the siblings or daughter, this separation occurred. The horse's decent behavior served as the primary source of inspiration for the suggested algorithm. Discussion: The models were created by using several ML techniques to train the feature-selected Cleveland heart disease dataset were evaluated and their results were compared. The parameters like Sensitivity, Accuracy, Specificity, and Area under Curve of the SVM classifier model are trained on the dataset utilizing the WHO approach which yields better results when compared with the other existing approaches. Conclusion: According to the findings, the wild horse optimization algorithm and SVM classifier combo performs best when used to forecast heart disease.

Keywords: Machine Learning, Wild Horse Optimization Algorithm (WHO), Support Vector Machine (SVM).

Significance | Heart disease, responsible for millions of deaths annually, necessitates early detection. Machine Learning offers efficient, cost-effective diagnostic tools. The proposed Wild Horse Optimizer (WHO) and SVM classifier enhance heart disease forecasting, addressing critical healthcare needs.

*Correspondence. Vishwanadham Mandala, Enterprise Data Architect, IU Bloomington,107 S. Indiana Avenue, Bloomington, IN 47405, USA.
E-mail- vishwanadh.mandala@gmail.com

Editor Surendar Aravindhan, And accepted by the Editorial Board Mar 03, 2024 (received for review Jan 08, 2024)

1. Introduction

According to estimates from according to the World Health Organization, heart disease deaths 12 million people yearly. Cardiovascular illnesses account for half of all deaths in wealthy nations like the US (Soni et al., 2011). In many underdeveloped

Author Affiliation.
[1] Enterprise Data Architect, IU Bloomington,107 S. Indiana Avenue, Bloomington, IN 47405, USA.
[2] General Motors LLC, Milford, Michigan, 48380, USA
[3] Department of ECE, Sri Krishna College of Engineering and Technology, Coimbatore, Tamil Nadu, India.
[4] Computer Engineering, Amrutvahini College of Engineering Sangamner, Maharashtra, India
[5] Department of Computer Science and Engineering, Vignan's Foundation for Science, Technology & Research, Vadlamudi, Guntur, Andhrapradesh, 522213.
[6] Information Technology, Sri Krishna College of Engineering and Technology, Kuniamuthur, Coimbatore, Tamil Nadu, 641008, India

Please cite this article.
Vishwanadham Mandala et al. (2024). Wild Horse Optimizer and Support Vector Machine (SVM) Classifier Predicts the Heart Disease Converging Nature-Motivated Optimization and Machine Learning, Journal of Angiotherapy, 8(3), 1-12, 9535

2207-8843/© 2019 ANGIOTHERAPY, a publication of Eman Research, USA.
This is an open access article under the CC BY-NC-ND license.
(http.//creativecommons.org/licenses/by-nc-nd/4.0/).
(https./publishing.emanresearch.org).

nations, it is also the main cause of fatalities. It is generally thought to be the main cause of adult deaths. All types of disorders affecting the heart are included under the umbrella phrase "heart disease." In India as well as other nations, the leading cause of death was heart disease. At 34 seconds, heart disease dies. Among the several heart disease classes are coronary cardiac disease, cardiomyopathy, and cardiovascular disease. The collective term for a wide range of conditions that impact the heart, blood vessels, and the circulation of blood throughout the body is "cardiovascular diseases (CVD)." Heart disease causes a number of illnesses, disabilities, and fatalities (Shah et al., 2020). The most important and complex tasks in medicine is disease diagnosis. A medical evaluation is considered a significant but challenging undertaking that requires precise and effective execution. This method would be very beneficial if it were automated. Regrettably, not all doctors possess expertise in every discipline, and many areas suffer from a scarcity of knowledgeable individuals. Hence, the integration of these components would highly facilitate the development of an automated medical diagnostic system (Palaniappan & Awang, 2020). Clinical testing may be cheaper with the correct computer-aided data and/or decision support technology. For automated systems to be implemented accurately and efficiently, a comparative analysis of the many approaches available is essential.

Discovering hidden patterns in the clinical domain's data sets can be greatly facilitated by medical data mining. A medical diagnosis can be made with the help of these patterns. The raw medical data that are now available are extensively dispersed, diverse, and substantial (Saxena et al., 2016). It is necessary to gather these data in an organized manner. A database for hospitals can be created by integrating the data that has been gathered. A way to finding new and concealed patterns in data that is user-oriented is offered by data mining technologies. The effective method of testing that is predicated on training and testing is machine learning (ML). ML is a subset of artificial intelligence (AI), a large field of study that focuses on creating systems that mimic human capacities. However, ML algorithms are trained to analyze and utilize data; for this reason, the combination of these two technologies is often known as machine intelligence (Sai Shekhar et al., 2020). Computational statistics and ML are closely linked fields which employ mathematical optimization to provide techniques and application domains to address real-world corporate, industrial, and medical challenges. The primary types into which it can be separated are supervised learning and unsupervised learning. When learning under supervision, a method creates an equation from a set of data that includes the inputs and the intended results. Unsupervised learning involves the creation of a theoretical framework by a method utilizing a collection of data that simply consists of inputs and no desired output labels (Katarya & Meena, 2021). The goal is to employ physical body functions to forecast the likelihood of having heart disease. And when predicted inputs and desired outputs are present, supervised learning is unquestionably an excellent option.

For heart disease diagnosis, data mining and neural network algorithms are used. Numerous techniques, including Decision Trees (DT), K-Nearest Neighbor Algorithm (KNN), Genetic Algorithm (GA), and Naive Bayes (NB), are employed to classify the extent of the condition (Kumar et al., 2018). Because heart illness has a complicated character, it needs to be treated properly. Failure to do so may lead to heart disease or premature mortality. Data mining and clinical investigations are used to discern various metabolic conditions. Data mining and classification considerably enhance the accuracy of heart disease forecasts and facilitate data interpretation. The reliability of events connected to heart disease can be estimated by decision trees (Rani et al., 2021). Many techniques were employed to the established data mining approaches for the prediction of heart disease in order to abstract knowledge. However, the accuracy of the classifiers is less while the dataset has complex. So this research work introduced Wild Horse Optimizer (WHO) based feature selection which is a novel optimizer algorithm that draws inspiration from the social behaviors of wild horses. And then the SVM based classifier is proposed for the detection of heart disease.

2. Literature Review

Here, the role and effectiveness of different feature extraction and nature-inspired techniques employed for diagnosis of the given heart disease data were accessed and presented.

Maheswari et al (Maheswari & Pitchai, 2019) proposed the intelligent prediction method provides the user with instance-specific assistance for heart disease. An array of cardiac indications is given into the program. The user checks the specifics and indications for heart illness before beginning any procedures. The information related to every patient is retrieved using the ID3 and navie Bayes approaches in data mining. System efficiency is examined depending on the precise outcome prediction.

Pattekari et al (Pattekari & Parveen, 2012) developed A smart system employing the Naive Bayes data mining modeling method. The user responds to the pre-established questions through the implementation of a web-based application. It analyzes the user values with the training data set and returns concealed data from the database that was actually saved. It can diagnose cardiac disease by providing complicated answers to questions, which conventional decision-support technologies are unable to, enabling medical professionals to make more informed clinical decisions. It contributes to lower treatment costs by offering efficient treatments.

Mythili et al (Mythili et al., 2013) the accuracy of adding rules to the individual results of logistic regression, decision trees, and SVM on

the Cleveland Heart Disease Database using a rules-driven methodology suggests a reliable heart disease prediction method.

Mahmoodi et al (Mahmoodi, 2017) introduced a fuzzy method and SVM method were effective at finding the condition quickly, when it came to diagnosing cardiac illness. Data from 270 individuals with 13 attributes were utilized in this qualitative and quantitative investigation. To identify patients with heart disease, the fuzzy system and SVM classifier were integrated utilizing the capabilities of the MATLAB program and were emulated by an equipment with a core i5 processor and Windows 7. The platform's assessment standards were sensitivity and classification rates; the system's efficiency was 85.8% and 85% for these two metrics, accordingly.

Zulkiflee et al (Zulkiflee & Rusiman, 2021) presented three approaches (BLR simulations, BLR systems with LQD, and BLR systems with MA) were applied to the heart disease data. Following a comparison of the three approaches, it was discovered that the binary logistical framework with the implemented MAD approach tended to be the most accurate model with the highest accuracy %. Only thalassemia, the quantity of main arteries, and the type of chest pain are important and strongly correlated with heart disorders. The purpose of this type of research is to educate the public about the key risk factors for heart disease so they may avoid or postpone the onset of the condition.

Reddy et al (Reddy et al., 2019) proposed a classification model, to determine which particular traits, utilizing the Cleveland and Statlog Project Heart datasets, are most important for predicting heart disease. Utilizing three separate percentage splits, the random forest method's accuracy across the feature selection and classification models was shown to be 90–95%. It appears that the eight and six features that were chosen are the very minimum needed to create a more accurate efficiency model. However, the prediction system may not perform any better if the remaining 8 or 6 features are dropped.

Anbarasi et al (Anbarasi et al., 2010) introduced a Genetic algorithm which is employed to identify the characteristics that are helpful in diagnosing cardiac conditions, hence lowering the quantity of tests that a patient requires. Genetic search reduces thirteen variables to six attributes. 3 classifiers: Naive Bayes, Classification by Clustering, and Decision Tree are then employed to forecast patient diagnoses with an accuracy comparable to that which was achieved prior to the reduction of variable value. Additionally, results show that, after including feature subset selection with comparatively high model construction time, the Decision Tree data mining method works better than the other two data mining strategies. When comparing the model construction time prior to and following attribute reduction, Naïve Bayes works adequately.

Bharti et al. (Bharti et al., 2021) presented three methods for predicting cardiac disease (without FS and outlier detection, with FS and no outlier detection, and with FS and outlier detection). According to them, FS along with outlier identification is superior to the other two approaches. The UCI heart dataset was trained with LR, KNN, SVM, RF, and DT classifiers, and the least absolute shrinkage and selection operation (LASSO) technique was used to pick the most important characteristics for heart illness. While deep learning achieved 94.2% accuracy, the KNN approach only achieved 84.8% accuracy.

Ghosh et al. (Ghosh et al., 2021) worked on UCI Dataset and employed RELIEF (Chikhi & Benhammada, 2009) and LASSO (Zhou & Wieser, 2018) feature selection mechanisms with classification. With a selection of 13 essential features from the dataset, they achieve the highest accuracy of 92.65% using Random Forest Bagging Method algorithm. With a selection of 11 features using the LASSO method they achieve the highest accuracy of 97.85% using the Gradient Boosting Method algorithm, and with a selection of 10 features using the RELIEF method they achieve the highest accuracy of 99.05% using RFBM algorithm.

Nitant et al. (Nitanta & Priyab, 2021) have suggested a decision tree-based artificial neural network model for forecasting cardiac disease. In order to extract the key features from the input in each layer and utilize them as the input for the subsequent layer of the ANN model, the authors used the decision tree as the activation function in the layers of the ANN. By the integration of heart datasets; Cleveland, Hungary, Long Beach, and Swiss, to understand heart disease, Kanagarathinam et al. (Kanagarathinam et al., 2022) produced a dataset titled "Sathvi" consisting of 531 instances, 12 attributes, and no missing values. For the purpose of predicting cardiovascular problems, they applied Pearson's correlation coefficient approach. In this study, the CatBoost algorithm was used for classification and obtained an accuracy of 87.85%.

Gupta et al. (Gupta et al., 2022) employed the Cleveland heart dataset to identify high-risk individuals with heart disease. The model learned several algorithms and standardized the data to the conventional scales. Using Logistic Regression, it achieved the greatest accuracy of 92.30%. The KNN classifier was also tweaked with k values ranging from 2 to 20; at k = 14, the classification accuracy was 90.11%.

Saboor et al. (Saboor et al., 2022) predicted heart disease using the UCI heart disease dataset and standardized heart dataset features for optimal prediction results, then using the GridSearchCV technique to fine-tune the hyperparameters of machine learning classifiers. An accuracy of 96.72% was attained using a support vector machine (SVM) classifier with a sigmoid kernel and a complexity value of C = 0.5.

Using exploratory analysis of features from a dataset pertaining to heart disease, Chang et al. (Chang et al., 2022) created a python-based model of the Random Forest technique. A correlation matrix

plot was used to assess the significance of the features in the heart dataset, and an accuracy of 82.18% was attained using a random forest classifier.

Reddy et al. (Reddy et al., 2022) combined the Cleveland and Statlog cardiac datasets and extracted the finest core features using the CFS method (Correlation-based Feature Selection). Second, utilizing three individual and three ensemble classifiers on datasets yielded the appropriate hyperparameters which depict the best prediction outputs. The author claims an accuracy of 97.91% with the Random Forest ensemble classifier.

Ozcan et al. (Ozcan & Peker, 2023) used an extensive dataset including data from five heart disease datasets of 1190 individual patients. To understand the connections between input and output data, the Classification and Regression Tree (CART) algorithm had been used to make predictions about cardiac disease. The author ranked the selected features based on their importance and achieved an accuracy of 87%.

Ogundepo et al. (Ogundepo & Yahya, 2023) used two datasets, the Cleveland dataset was considered for classification and the Statlog dataset was considered to validate the model. The authors performed an in-depth exploratory analysis of the Cleveland data using the Chi-square test of independence and followed by training ten classification models for prediction. The author stated support vector machine provided the most accurate predictions with a classification accuracy of 85% and validation accuracy of 87.04% with the Statlog dataset.

Using the BRFSS 2015 heart disease dataset, Fernando et al. (Fernando et al., 2022) assessed the performance of various supervised classification models in terms of accuracy. These models included Naive Bayes, LightGBM, Decision Trees, Random Forest, XGBoost, K Nearest Neighbours, and ADABoost. Smoteen and SmoteTomek outperformed the other sample methods and were used by the authors to address the class imbalance problem in the dataset, achieving 97.10% classification accuracy with random forest model.

Using data from the 2015 BRFSS survey of US citizens, Das et al. (Das et al., 2023) developed and compared six machine learning models for predicting cardiovascular disease. These models included Xgboost, Bagging, Random Forest, Decision Tree, K-Nearest Neighbour, and Naive Bayes. The six machine learning models were compared in terms of their accuracy, sensitivity, F1-score, and area under the curve (AUC). According to the authors, the Xgboost model produced the most optimal outcomes, with an accuracy rate of 91.30%.

In order to conclude the literature, there are certain limitations in the work of previous authors, some authors analysed a small amount of individual data for classification, and the models implemented with small data may not validate the model for bulk data. The authors mainly used machine learning approaches for classification, while deep learning and optimization approaches may yield better results. The evaluation metrics in most of the literature are accuracy, sensitivity, specificity, and f1 score but some validation curve might be helpful for better understanding.

3. Methodology

This research work, presented a wild horse optimizer (WHO) SVM-based classifier and feature selection for heart disease prediction. The WHO, a novel optimizer algorithm that draws inspiration from the social behaviors of wild horses, is presented in this article. Horses typically reside in groups consisting of a stallion, numerous mares, and young foals. Horses can be seen engaging in a variety of behaviors, including leading, grazing, chasing, and mating. The interesting quality that sets horses apart from other animals is their kindness. When a horse is decent, before they reach maturity, its foals break away from the herd and join different groups. The reason for the father's absence was to avoid the siblings or daughter from mating. The horse's decent behavior served as the primary source of inspiration for the suggested algorithm. The models that were created by using several ML techniques to train the feature-selected Cleveland heart disease dataset were evaluated and their results were compared. The suggested methodology's entire procedure is provided in figure 1.

3.1. Dataset Description and Statistics

Out of the 303 occurrences with 76 attributes in the Cleveland Heart dataset, only 14 features are thought to be more appropriate to be utilized for study experiments (Reddy et al., 2021). Table 1 shows the Cleveland Heart dataset attribute descriptions from the UCI machine-learning repository.

Nominal or categorical types are qualities with fewer than ten classes. Gender-based classes comprise the attribute'sex': 0 = female and 1 = male. Four classifications of chest pain types are included in the term "cp": 1 is normal angina, 2 is atypical angina, 3 is non-angina pain, and 4 is asymptomatic. Considering if the fasting blood sugar is greater than 120 mg/dL, the characteristic "fbs" has two classes: 1 = true and 0 = false. Three classifications of resting electrocardiographic results make up the feature "restecg": 0 represents normal, 1 indicates an aberrant ST-T wave, and 2 indicates substantial left ventricular hypertrophy. According to exercise-induced angina, the attribute "exang" is divided into two classes: 1 = yes and 0 = no. Three peak exercise kinds The attribute "slope" includes ST segment slope: Upslope = 1, flat = 2, downslope = 3. Calculating the number of fluoroscopy-colored major vessels (0–3) divides "ca" into four groups. The attribute "thal" might be 3 for normal, 6 for fixed, or 7 for reversible heart status. There are five prediction classes for the attribute "target": 0 denotes no risk of heart disease, whereas 1-4 denotes many stages. Given that determining a patient's risk of getting heart disease was the primary goal of this research initiative, values between 1 and 4 were transformed to 1. Thus, the classes 0 and 1 were the only ones contained in the "target"

attribute. Numeric/integer types are assigned to the attributes "oldpeak," "trestbps," "chol," "thalach," and "age".

Table 2(a) provides the statistical properties of the numerical qualities, including the minimum, standard deviation, maximum, mean, distinct, missing, and unique values. The Cleveland dataset's numerical properties don't contain any missing values.

Table 2(b) displays the statistical properties of the nominal attributes, including label, count, missing, and distinct values. Six (6) instances, or 2% of the dataset, were found to have missing values two (2) from the "thal" variable and four (4) from the "ca" attribute out of 303. 164 instances of the target class labels 0 (no risk) and 139 instances of label 1 (risk) made up 54% and 46% of the sample, accordingly.

3.2. Data preprocessing: normalization

Each feature in a dataset that is employed to train an algorithm often has a distinct distribution. A SVM finds it extremely challenging to fit the data in these situations. Numerous methods exist to address this issue, all attempting to modify each feature to achieve a comparable range inside the real number set. Several common normalizers include:

- • After applying Eq. 1, MaxMin Normalization considers the maximum and minimum values needed to fix the data into the range [0,1].

$$\hat{X} = \frac{X - X_{Min}}{X_{Max} - X_{Min}} \quad (1)$$

3.3. Feature Selection using Wild Horse Optimization (WHO)

The WHO adopt a wild horse identity. Non-terrestrial horses are referred to as wild horses. They reside in two groups: a family group for mares, or female horses, and a separate group for stallions, or male horses. Among the family group and the single group, mating takes place. Foals, or young horses, are concerned about grazing when they are first born (Naruei & Keynia, 2022). After leaving their home group, female foals join other groups. Once a male colt reaches maturity, they are referred to as stallions. Stallion, and join the "single group" respectfully. Decency, in the sense that grouping the stallions prevents incest. The behavior of powerful leaders, who can reach water holes while other less dominant members must wait for hours, was emphasized by their search for water during dry seasons. Family groups are led by mares, but as subordinates, they have to submit to a leader chosen by the stallions. The following are the WHO's primary steps.

1) Population Initialization and Leadership Selection

N individuals make to the initial population $(\vec{x})$, which is chosen at random. $(\vec{x}) = \{\vec{x}_1, \vec{x}_2, \dots \dots \dots \vec{x}_n\}$and To create the following vector, the goal function of every population is computed.

$$(\vec{O}) = \{\vec{O}_1, \vec{O}_2, \dots \dots \dots \vec{O}_n\} \quad (2)$$

Groups G are created from the population, wherein G = NXPS and PS represents the proportion of stallions among the overall population. At the beginning of the method, each group has a randomly chosen stallions leader; however, as the method progresses, the highest fitness value determines which leaders are elected.

2) Grazing Behavior

Equation 3 depicts the grazing pattern.

$$\bar{X}_{i,G}^{j} = 2Zcos(2\pi RZ) \times \left(Stallion^{j} - X_{i,G}^{j}\right) + Stallion^{j} \quad (3)$$

where $X_{i,G}^{j}$ is the member's group present position, $Stallion^{j}$ indicates leader's group position, Equation 5 shows that the Z variable is given as follows: R is a randomly generated number in the interval [−2, 2] causing horses to graze at various angles (360 degrees) of the group leader; π is taken as 3.14; movement across various radii is caused by the cosine function of R and π; and the final location of a member $\bar{X}_{i,G}^{j}$ is the updated position of a member.

$$P = \vec{R}_1 < TDR: \qquad IDX = (P == 0); \quad (4)$$

$$Z = R_2 \ominus IDX + \vec{R}_3 \ominus (\sim IDX) \quad (5)$$

here P is a vector ∈ [0, 1], R_2 states a random number ∈ [0, 1], $\vec{R}_1$ and $\vec{R}_3$ are random vectors ∈ [0, 1], the $\vec{R}_1$ yields' IDX indices that meet the requirement (P == 0). TDR falls to zero, as seen in Equation 6.

$$TDR = 1 - iter \times \left(\frac{1}{maxiter}\right) \quad (6)$$

3) Horse Mating Behavior

Equations 7, 8, 9, and 10 illustrate decency and mating behavior.

$$X_{G,K}^{P} = Crossover(X_{G,i}^{q}, X_{G,j}^{z}) \quad (7)$$

$$i \neq j \neq k \; p = q = end, \quad (8)$$

$$Crossover = Mean \quad (9)$$

where $X_{G,K}^{P}$ denotes horse p's position as it departs group k is replaced by a horse whose parents depart groups i and j as a result of puberty. They have mated to generate $X_{G,i}^{q}$., yet they are unrelated to one another. The horse z and the foal q, who belongs to the i group, mated, whose position $X_{G,j}^{z}$ is in the j group, when it reached adulthood.

B. Group leadership

The leader's group is responsible for leading the group to the appropriate section of the water. This water is fought over by leaders for the use of the dominating group; other members are not allowed to utilize it until the dominating group has departed. Equation 9 depicts this action in the same way as in equation (10), where WH is the water position, Stallion Gi is the group i's present leader's position,

$$\overline{Stallion}_{G_i} = \begin{cases} 2Ccos(2\pi RZ) \times \left(WH - Stallion_{G_i}\right) + WH & if\ R_3 > 0.5 \\ 2Ccos(2\pi RZ) \times \left(WH - Stallion_{G_i}\right) - WH & if\ R_3 \leq 0.5 \end{cases} \quad (10)$$

1) Exchange and Leadership Selection

The leaders are initially chosen at random. At an additional step of the process, the population that is the fittest is chosen to be the leader. The roles of the chosen member and the leader are displayed in equation (11),

Box 1. The pseducode of WHO is represented in algorithm 1.

Algorithm 1: WHO
Input: Raw data
Output: optimized features
1: Initialization: set the parameters PC, PS.
2: Set populations.
3: Determine each population's fitness value
4: Assemble groups and choose leaders.
5: while (tier <= maxiter) do
6: compute TDR using Equation 5.
7: for each stallion do
8: compute Z using Equation 4.
9: for each foal inside the group do
10: if rand > PC then
11: update position by Equation 3
12: else
13: update position by Equation 7
14: end if
15: end for
16: if rand > 0.5 then
17: update position of $\overline{Stallion\ G_i}$ by Equation 10 first part
18: else
19: update position of $\overline{Stallion\ G_i}$ by Equation 10 second part part
20: end if
21: if fitness ($\overline{Stallion\ G_i}$) >fitness (Stallion) then
22: Stallion = $\overline{Stallion\ G_i}$
23: end if
24: Sort group foals based on fitness levels
25: Choose the foal that is least fit
26: if fitness(foal)<fitness (Stallion) then
27: exchange foal and stallion position according to eq 11
28: end if
29: end for
30: m=m+1
31: end while
32: Return the solution with best fitness

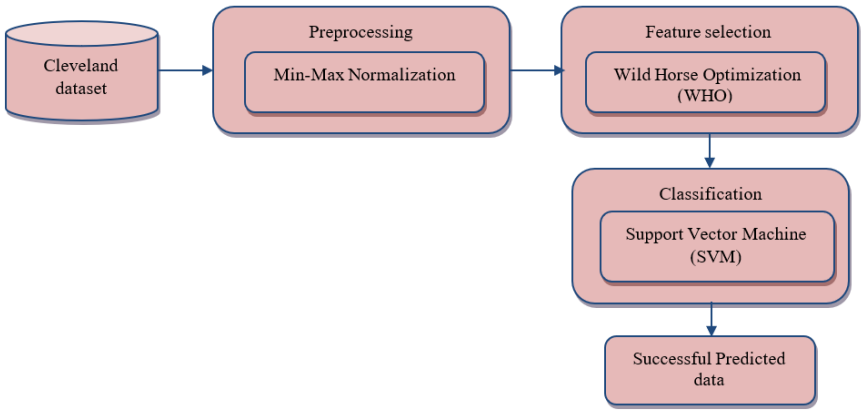

Figure 1. The suggested methodology's entire procedure

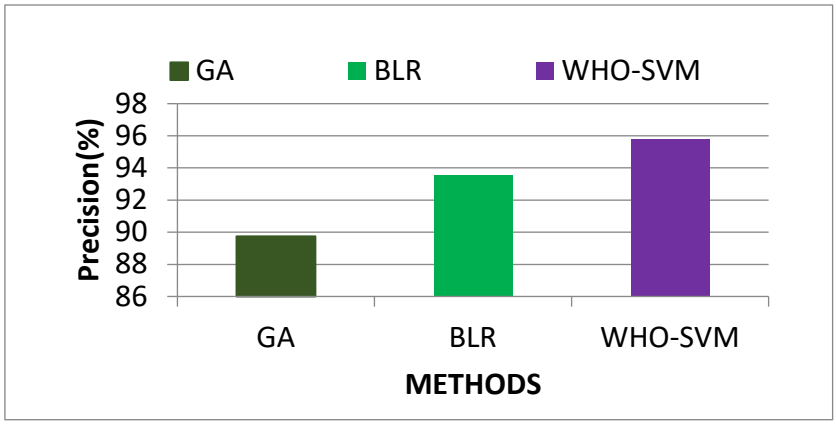

Figure 2. Precision comparison results of the proposed WHO-SVM and existing classifiers

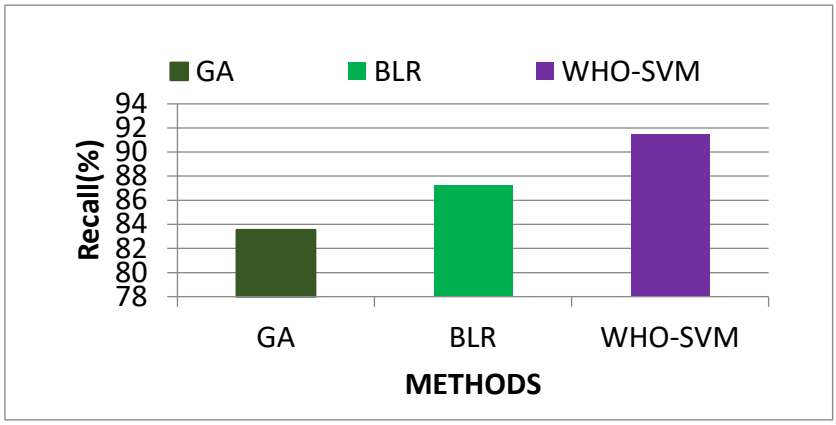

Figure 3. Recall comparison results of the proposed WHO-SVM and existing classifiers

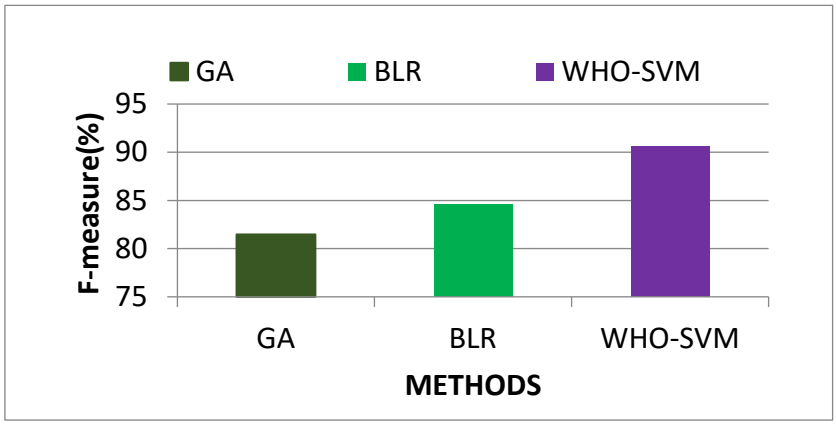

Figure 4. F-measure comparison results of the proposed WHO-SVM and existing classifiers

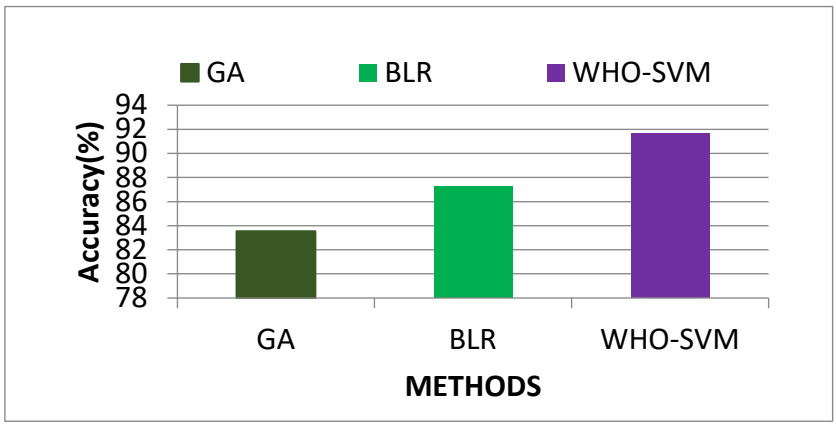

Figure 5. Accuracy comparison results of the proposed WHO-SVM and existing classifiers

Table 1. Cleveland Heart dataset attribute descriptions from the UCI machine-learning repository.

Attribute	Description	Type of Attribute	Attribute value Range
Age	Age in years	Numeric	29 to 77
Sex	Gender	Nominal	0=female, 1=male
cp	chest pain type	Nominal	1=typical angina, 2= a typical angina, 3= non- angina pain, 4= asymptomatic
trestbps	resting blood pressure in mm Hg on admission to the hospital	Numeric	94-200
chol	serum cholesterol in mg/dl	Numeric	126-564
fbs	fasting blood sugar >120 mg/dl	Nominal	0=false 1= true
restecg	Resting electrocardiographic results	Nominal	0=normal 1= ST-T wave abnormality 2= definite left ventricular hypertrophy by Estes' criteria
thalach	Maximum heart rate achieved	Numeric	71 to 202
exang	Exercise induces angina	Nominal	0=no 1-yes
oldpeak	ST depression induced by exercise related to rest	Numeric	0 to 6.2
slope	the slope of the peak exercise ST segment	Nominal	1= upsloping 2= flat 3= downsloping
ca	number of major vessels colored by fluoroscopy	Nominal	0-3
thal	the heart status	Nominal	3= Normal, 6=fixed defect, 7=reversible defect
target	Prediction attribute	Nominal	0 =no risk of heart disease, 1 to 4 = risk of heart disease

Table 2. (a) The numerical properties' statistical framework. (b) The nominal attributes' statistical framework.

Attribute	Min.	Max.	Mean	StdDev	Missing	Distinct	Unique
age	29	77	54.439	9.039	0	41	4(1%)
trestbps	94	200	31.69	17.6	0	50	17(6%)
chol	126	564	246.693	51.777	0	152	61(20%)
thalach	71	202	149.607	22.875	0	91	28(9%)
oldpeak	0	6.2	1.04	1.161	0	40	10 (3%)

$$Stallion_{G_i} = \begin{Bmatrix} X_{G,i} & if\ \cos\ t(X_{G,i}) < \cos t\ (Stallion_{G_i}) \\ Stallion_{G_i} & if\ \cos\ t(X_{G,i}) > \cos t\ (Stallion_{G_i}) \end{Bmatrix} \quad (11)$$

The pseudocode of WHO is represented in algorithm 1 (Box 1).

3.3. Classification using Support Vector Machine (SVM)

The capacity of SVM to enable non-linear classification employing a kernel function has long made them appealing for anomaly detection (Suthaharan & Suthaharan, 2016). After giving a brief overview of SVM fundamentals, concentrate on the EOC-SVM utilized in this study.

Utilize the conventional two-class SVM, where a set of n training cases, $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. $x_i \in R^d$, the class label of each instance, y_i is associated with x_i, and $y_i \in [-1, +1]$. The linear SVM classifier finds the best separation hyperplane by optimizing the classifier's "margin" utilizing the following equation: $w^T x + b = 0$, where w∈F and b∈R are two factors that define the decision hyperplane's location in feature space F (w determines the decision hyperplane's orientation, while b determines its movement). Thus, a general representation of the decision function is as

$$f(x, w, b) = sign(w^T x + b) \in \{-1, +1\} \quad (12)$$

There are few overruns that will significantly impact the classifier profile determined by the decision function.

where,

$$sign\ (w^T x + b) = \begin{cases} +1, & if\ (w^T x + b) \geq 0 \\ -1, & otherwise \end{cases} \quad (13)$$

SVMs work by locating (w,b) that, to minimize the generalization error, the hyperplane is placed at the greatest distance of the closest training samples of the two classes. The "margin" is defined by this distance. The first applications of SVMs were in linearly distinct classification tasks. They were expanded to include non-linearly distinct classification issues, though. A non-linear function Φ(x) allows certain samples to exceed the margin (soft-margin SVMs), and extending the data into greater dimension space yields a non-linear decision boundary. Although data points are "lifted" into a feature space F wherein a hyperplane may divide them, they might not be linearly distinct in their original space. The form of that hyperplane is non-linear when it is reflected back into the input space. Slack variables ξ are added to enable certain data points to reside within the margin to avoid the SVM classifier from over-fitting noisy data. A value of C > 0 adjusts the trade-off among the classification error on the training data and margin maximization. The following minimization technique applies to the objective function of SVM classifiers:

$$\min_{w,b,\xi_i} \frac{\|w\|^2}{2} + C \sum_{i=1}^{n} \xi_i \quad (14)$$

subject to $y_i(w^T \phi(x_i) + b \geq 1 - \xi_i$

$$\xi_i \geq 0, \quad i = 1, \ldots \ldots n$$

Lagrange multipliers $\alpha_i, i = 1, \ldots \ldots n$ are utilized to resolve the minimization issue. For a given data point x, a novel decision function rule is specified as

$$f(x) = sign(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b) \quad (15)$$

Each α_i>0 contributes to the machine's support because it is weighted in the decision function. Because SVMs are thought to be sparse, there aren't many Lagrange multipliers that have a value other than zero. The kernel function is the name given to the function $K(x, xi) = \Phi(x)\ T\ \Phi(xi)$. It is not required to carry out an explicit projection because the decision function's result simply depends on the dot-product of the vectors in the feature space F. A function K can be substituted as long as it yields the same outcomes. It is the kernel trick. Three common options for the kernel function are sigmoidal, polynomial, and linear. Applied the Gaussian Radial Base Function in this work.

$$K(x, x_i) = \exp\left(\frac{-\|x - x_i\|^2}{2\sigma^2}\right) \quad (16)$$

here the dissimilarity parameter is (x, x_i) and the kernel parameter is σ∈R. A non-linear decision function can be utilized for splitting a set of data points into two classes utilizing this collection of formulas and concepts. The technique's power stems from its utilization of kernel functions, which allow it to work in an implicit, high-dimensional feature space without ever requiring the computation of the coordinates of the data. Instead, feature space inner products are calculated from images of every pair of data. Compared to the explicit computation of the coordinates, this process is frequently less expensive analytically.

4. Results and Discussion

The experimental analysis is implemented in MATLAB. The Parameter setting for WHO used in testing is PS is 0.2, PC is 0.13, Population Size is about 25 - 100 and the No. of Iterations is 75 – 250.

The efficiency of the suggested Wild Horse Optimization with Support Vector Machine (WHO-SVM) model is compared with the existing classifier namely Genetic Algorithm (GA) and Binary Logistic Regression (BLR). Along with classification accuracy, the classifier is evaluated utilizing the statistical metrics provided in equations (17)–(20) and the average performance of each classifier. Precision is the proportion of correctly found positive results to predicted positive observations.

Precision = TP/TP+FP (17)

Sensitivity is the ratio of effectively-recognized results to all authentic class observations.

Recall = TP/TP+FN (18)

The F1 score is the Precision and Recall weighted average. There must be false positives and negatives.

F1 Score = 2*(Recall * Precision) / (Recal

Utilizing both positive and negative values, accuracy is computed as follows:

$$\text{Accuracy} = (TP+FP)/(TP+TN+FP+FN) \quad (20)$$

here True Positive, False Positive, True Negative, and False Negative are the definitions of TP, FP, TN, and FN, accordingly. These variables are computed for each of the five benchmark datasets (UCI, 2010) and a dataset on heart disease for all standard and suggested classifiers.

Figure 2 displays the effectiveness of the suggested WHO-SVM's precision comparative findings. The findings confirm that the features selection technique centered on WHO guidelines may accurately predict the categorization of heart disease. The quantity of helpful characteristics in the suggested WHO does not significantly impact linear transformation efficiency. This is a desirable feature since it eliminates the need for tedious regularization parameter tuning in the classifier. The proposed WHO has highly effective technique for solving the classification problem.

The performance outcomes of the suggested WHO-SVM based classifier are presented in figure.3. The findings demonstrate that, in comparison to the current approach, which yields lower recall results for example, the BLR approach metric has 89.68% and the GA method metric has 87.25%—the suggested approach yields high recall outcomes of 91.74%.

Figure 4 indicates that, in comparison to the BLR and GA, the suggested WHO-SVM performs remarkably well with respect to of the disease prediction rate. The ML-based qualitative analysis and the quantitative study's F-measure findings converge. The suggested WHO-SVM is evaluated in terms of accuracy for the heart disease dataset against various cutting-edge classification methods.

Figure .5. shows the proposed WHO-SVM give more accuracy 98.68% which is higher than the existing classifier. In a comparable way, when applied to static data, all of the previously described classifiers execute poorly in comparison to the WHO-SVM classifier, demonstrating the technique's efficacy in all crucial scenarios for the categorization of heart disease. As a result, the classifiers' accuracy will be greater than that of another classifier created from a prior framework.

5. Conclusion

Early detection of irregularities in heart diseases and long-term life preservation will be facilitated by learning the way raw healthcare data related to heart data is processed. In this study, raw data was processed utilizing ML methods to produce a unique and unique diagnosis of heart disease. This research work, presented a wild horse optimizer (WHO) based feature selection and SVM based classifier for the prediction of heart disease data. The wild horse optimizer (WHO), a novel optimizer algorithm that draws inspiration from the social behaviors of wild horses, is presented in this article. Horses typically reside in groups consisting of a stallion, numerous mares, and young foals. Horses can be seen engaging in a variety of behaviors, including leading, grazing, chasing, and mating. The interesting quality that sets horses apart from other animals is their kindness. When a horse is decent, before they reach maturity, its foals break away from the herd and join different groups. The reason for the father's absence was to keep the siblings or daughter from mating. The horse's decent behavior served as the primary source of inspiration for the suggested algorithm. The models that were created by using several ML techniques to train the feature-selected Cleveland heart disease dataset were evaluated and their results were compared. Sensitivity, Accuracy, Specificity, and Area Under Curve of the SVM classifier model trained on the dataset utilizing the wild horse optimization approach yielded the best results. The proposed WHO-SVM gives more accuracy 98.68% which is higher than the existing classifiers. Expand on this study by utilizing more hybrid swarm based optimization methods including association rules, time series, and clustering.

Author contributions

V.M. collected data and wrote the original draft. S.N.D.S. conceptualized the study. V. R. B. supervised the project, reviewed and edited the writing, and administered the project. D.R.P. developed the methodology. S.F.W. interpreted the data, and G. S. analyzed the data.

Acknowledgment

The authors expressed gratitude to their department.

Competing financial interests

The authors have no conflict of interest.

References

Anbarasi, M., Anupriya, E., & Iyengar, N.C.S.N. (2010). Enhanced prediction of heart disease with feature subset selection using genetic algorithm. International Journal of Engineering Science and Technology, 2(10), pp. 5370-5376.

Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., & Singh, P. (2021). Prediction of heart disease using a combination of machine learning and deep learning. Computational intelligence and neuroscience, pp. 1-11. https://doi.org/10.1155/2021/8387680.

Chang, V., Bhavani, V.R., Xu, A.Q., & Hossain, M.A. (2022). An artificial intelligence model for heart disease detection using machine learning algorithms. Healthcare Analytics, 2, pp. 100016. https://doi.org/10.1016/j.health.2022.100016.

Chikhi, S., & Benhammada, S. (2009). ReliefMSS: a variation on a feature ranking ReliefF algorithm. International Journal of Business Intelligence and Data Mining, 4(3-4), pp. 375-390. https://doi.org/10.1504/IJBIDM.2009.029085.

Das, R.C., Das, M.C., Hossain, M.A., Rahman, M.A., Hossen, M.H., & Hasan, R. (2023). Heart disease detection using ml. In IEEE 13th Annual Computing and

Communication Workshop and Conference (CCWC), pp. 0983-0987. https://doi.org/10.1109/CCWC57344.2023.10099294.

Fernando, C.D., Weerasinghe, P.T., & Walgampaya, C.K. (2022). Heart Disease Risk Identification using Machine Learning Techniques for a Highly Imbalanced Dataset: a Comparative Study, pp. 43-55. http://doi.org/10.4038/kjms.v4i2.50.

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F.J.M., Ignatious, E., Shultana, S., Beeravolu, A.R., & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access, 9, pp. 19304-19326. https://doi.org/10.1109/ACCESS.2021.3053759.

Gupta, C., Saha, A., Reddy, N.S., & Acharya, U.D. (2022). Cardiac Disease Prediction using Supervised Machine Learning Techniques. In Journal of Physics: Conference Series, 2161(1), pp. 1-11. https://iopscience.iop.org/article/10.1088/1742-6596/2161/1/012013#:~:text=DOI%2010.1088/1742%2D6596/2161/1/012013.

Kanagarathinam, K., Sankaran, D., & Manikandan, R. (2022). Machine learning-based risk prediction model for cardiovascular disease using a hybrid dataset. Data & Knowledge Engineering, 140, pp. 102042. https://doi.org/10.1016/j.datak.2022.102042.

Katarya, R., & Meena, S.K. (2021). Machine learning techniques for heart disease prediction: a comparative study and analysis. Health and Technology, 11(1), pp. 87-97. https://doi.org/10.1007/s12553-020-00505-7.

Kumar, M.N., Koushik, K.V.S., & Deepak, K. (2018). Prediction of heart diseases using data mining and machine learning algorithms and tools. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 3(3), pp. 887-898. http://dx.doi.org/10.13140/RG.2.2.28488.83203.

Maheswari, S., & Pitchai, R. (2019). Heart disease prediction system using decision tree and naive Bayes algorithm. Current Medical Imaging, 15(8), pp. 712-717. https://doi.org/10.2174/1573405614666180322141259.

Mahmoodi, M.S. (2017). Designing a heart disease prediction system using support vector machine. Journal of Health and Biomedical Informatics, 4(1), pp. 1-10.

Mythili, T., Mukherji, D., Padalia, N., & Naidu, A. (2013). A heart disease prediction model using SVM-decision trees-logistic regression (SDL). International Journal of Computer Applications, 68(16), pp. 11-15.

Naruei, I., & Keynia, F. (2022). Wild horse optimizer: A new meta-heuristic algorithm for solving engineering optimization problems. Engineering with computers, 38(4), pp. 3025-3056. https://doi.org/10.1007/s00366-021-01438-z.

Nitanta, D.R.P., & Priyab, R. (2021). Predicting Heart disease using Machine Learning. Turkish Journal of Computer and Mathematics Education, 12(13), pp. 370-376.

Ogundepo, E.A., & Yahya, W.B. (2023). Performance analysis of supervised classification models on heart disease prediction. Innovations in Systems and Software Engineering, 19(1), pp. 129-144. https://doi.org/10.1007/s11334-022-00524-9.

Ozcan, M., & Peker, S. (2023). A classification and regression tree algorithm for heart disease modeling and prediction. Healthcare Analytics, 3, pp. 100130. https://doi.org/10.1016/j.health.2022.100130.

Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. In IEEE/ACS international conference on computer systems and applications, 8(8), pp. 343-350. https://doi.org/10.1109/AICCSA.2008.4493524.

Pattekari, S.A., & Parveen, A. (2012). Prediction system for heart disease using Naïve Bayes. International Journal of Advanced Computer and Mathematical Sciences, 3(3), pp. 290-294.

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

Reddy, K.V.V., Elamvazuthi, I., Aziz, A.A., Paramasivam, S., Chua, H.N., & Pranavanand, S. (2022). An efficient prediction system for coronary heart disease risk using selected principal components and hyperparameter optimization. Applied Sciences, 13(1), pp.1-28. https://doi.org/10.3390/app13010118.

Reddy, K.V.V., Elamvazuthi, I., Aziz, A.A., Paramasivam, S., Chua, H.N., & Pranavanand, S. (2021). Heart disease risk prediction using machine learning classifiers with attribute evaluators. Applied Sciences, 11(18), pp. 1-22. https://doi.org/10.3390/app11188352.

Reddy, N.S.C., Nee, S.S., Min, L.Z., & Ying, C.X. (2019). Classification and feature selection approaches by machine learning techniques: Heart disease prediction. International Journal of Innovative Computing, 9(1), pp. 39-46.

REFERENCES

Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M.F., & Ullah, N. (2022). A method for improving prediction of human heart disease using machine learning algorithms. Mobile Information Systems, pp. 1-9. https://doi.org/10.1155/2022/1410169.

Sai Shekhar, M., Mani Chand, Y., & Mary Gladence, L. (2020). Heart Disease Prediction Using Machine Learning. In International Conference on Emerging Trends and Advances in Electrical Engineering and Renewable Energy, pp. 603-609. https://doi.org/10.1007/978-981-15-8685-9_63.

Saxena, K., & Sharma, R. (2016). Efficient heart disease prediction system. Procedia Computer Science, 85, pp. 962-969. https://doi.org/10.1016/j.procs.2016.05.288.

Shah, D., Patel, S., & Bharti, S.K. (2020). Heart disease prediction using machine learning techniques. SN Computer Science, 1(6), pp. 1-6. https://doi.org/10.1007/s42979-020-00365-y.

Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), pp. 43-48.

Suthaharan, S., & Suthaharan, S. (2016). Support vector machine. Machine learning models and algorithms for big data classification: thinking with examples for effective learning, pp. 207-235. https://doi.org/10.1007/978-1-4899-7641-3_9.

Zhou, C., & Wieser, A. (2018). Jaccard analysis and LASSO-based feature selection for location fingerprinting with limited computational complexity. In Progress in

Location Based Services, 14, pp. 71-87. https://doi.org/10.1007/978-3-319-71470-7_4.

Zulkiflee, N.F., & Rusiman, M.S. (2021). Heart Disease Prediction Using Logistic Regression. Enhanced Knowledge in Sciences and Technology, 1(2), pp. 177-184.