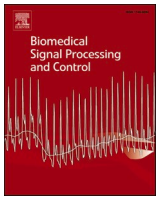
Biomedical Signal Processing and Control 102 (2025) 107417

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



A snake optimization algorithm-based feature selection framework for rapid detection of cardiovascular disease in its early stages

Zahraa Tarek [a,b], Amel Ali Alhussan [c], Doaa Sami Khafaga [c], El-Sayed M. El-Kenawy [d,e,f,g,*], Ahmed M. Elshewey [h]

[a] Department of Computer Engineering and Information, College of Engineering, Wadi Ad Dwaser, Prince Sattam Bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia
[b] Computer Science Department, Faculty of Computers and Information, Mansoura University, Mansoura 35561, Egypt
[c] Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[d] School of ICT, Faculty of Engineering, Design and Information & Communications Technology (EDICT), Bahrain Polytechnic, PO Box 33349, Isa Town, Bahrain
[e] Applied Science Research Center, Applied Science Private University, Amman, Jordan
[f] Jadara University Research Center, Jadara University, Jordan
[g] Department of Communications and Electronics, Delta Higher Institute of Engineering and Technology, Mansoura 35111, Egypt
[h] Department of Computer Science, Faculty of Computers and Information, Suez University, P.O.BOX:43221, Suez, Egypt

ARTICLE INFO

Keywords:
Cardiovascular disease
Machine learning
Feature selection
Analysis of medical data
Snake optimization

ABSTRACT

Cardiovascular disease (CVD) is a disorder that negatively affects the heart and blood vessels. Manual screening is time-consuming and prone to human error. As a result, assessors might rely on machine learning (ML) methods and automated feature selection. Feature selection aims to zero in on a dataset's most relevant and valuable subset of features. High dimensionality is addressed, complexity is reduced, and model efficiency is improved with this approach. Black-box optimization methods that take cues from nature have gained popularity to solve complex problems without resorting to formal mathematical formulations. More recently, algorithms that mimic the hunting behavior of snakes have evolved as a method for finding optimal or almost optimal solutions to difficult situations. The fundamental purpose of this research was to offer a novel framework for the analysis of cardiovascular disease (CVD) data called CVD-SO, which uses snake optimization (SO). Five machine learning methods are applied to pick and classify valid medical data efficiently. By incorporating machine learning and the SO algorithm into a single framework, we can create a CVD diagnostic model with unprecedented accuracy. The final output is a model that detects CVD with a remarkable 99.9% accuracy. These results considerably improve our comprehension of the medical data preparation procedure. The widespread burden of CVD-related disorders and the associated death rates can be alleviated, and mortality rates reduced if healthcare systems adopt this paradigm and actively combat CVD by facilitating early interventions.

1. Introduction

Cardiovascular disease (CVD) is now a public health problem and continues to be the leading cause of mortality globally. Atherosclerosis, diabetes, hypertension, inflammation, dyslipidemia, and specific genetic abnormalities are the recognized risk factors for CVD [1]. The statistics show that 31 % of all fatalities worldwide occur as a result of CVD, which claims 17.5 million lives annually [2]. An estimated $37.3 billion is spent annually on cardiovascular-related expenses connected to diabetes, as CVD is the primary cause of morbidity and death in individuals with diabetes. Dyslipidemia and Hypertension, two common diseases that coexist with type 2 diabetes, are recognized risk factors for CVD, and diabetes itself presents an independent risk. Another significant factor contributing to cardiovascular disease-related morbidity and death is heart failure. According to recent research, those with diabetes had twice as many hospitalizations for incident heart failure as people without the disease. Hospitalization rates for heart failure have decreased in recent trials using sodium-glucose cotransporter two

* Corresponding author at: School of ICT, Faculty of Engineering, Design and Information & Communications Technology (EDICT), Bahrain Polytechnic, PO Box 33349, Isa Town, Bahrain.
E-mail addresses: z.elmana@psau.edu.sa (Z. Tarek), aalhussan@pnu.edu.sa (A.A. Alhussan), dskhafga@pnu.edu.sa (D.S. Khafaga), skenawy@ieee.org (E.-S.M. El-Kenawy), ahmed.elshewey@fci.suezuni.edu.eg (A.M. Elshewey).

https://doi.org/10.1016/j.bspc.2024.107417
Received 8 January 2024; Received in revised form 22 November 2024; Accepted 17 December 2024
Available online 24 December 2024
1746-8094/

inhibitors in patients with type 2 diabetes, the majority of whom also had cardiovascular disease. Cardiovascular risk factors should be thoroughly examined in all people with diabetes at least once a year for the prevention and treatment of both ASCVD and heart failure. These risk factors include diabetes duration, overweight/obesity, smoking, dyslipidemia, hypertension, a family history of early coronary disease, chronic kidney disease, and a diagnosis of albuminuria [3]. Precision is crucial when diagnosing cardiac disease in healthy individuals since a mistake might result in a healthy patient receiving unneeded therapy. Medical issues and data collection influence the system's ability to balance the performance of various groups fairly. Not only that, but in most countries, the death rate from multiple ailments has increased due to a lack of medical specialists. The leading cause of mortality in both urban and rural environments in many of these countries is heart disease, surpassing all other causes of death. An extensive understanding of the disease's symptoms and forms must exist to diagnose the illness correctly. The area of artificial intelligence (AI), machine learning (ML), deep learning (DL), and optimizations are widely used for medical diagnosis [4–6]. In the realm of medicine, machine learning is essential. We can detect, diagnose, and forecast several diseases using machine learning. The application of machine learning and data mining techniques to predict the risk of contracting specific diseases has gained popularity recently. Data mining techniques have been applied in previous studies to forecast illness. Accurate findings have not yet been found despite several research' attempts to estimate the future risk of the illness's spread. CT scans and electrocardiograms, for example, are vital for identifying coronary heart disease, but they are frequently too costly and impractical for many low- and middle-income nations. Early detection of the disease is essential to lessen heart disease's physical and financial impact on people and organizations [7]. In medical domains, machine learning and feature selection techniques have been employed for real life human disease prediction [8–12]. Mathematical optimization, which refers to the numerical computation of parameters for a system intended to make decisions based on as-yet-unseen data, is one of the basic principles of machine learning. These parameters are selected to be optimum regarding a particular learning issue based on the available data. Many scholars across other fields have been motivated by the success of some machine learning optimization techniques to take on more challenging problems in machine learning and to develop new, more broadly applicable techniques [13]. Many heuristic algorithms that use randomized search techniques are animal instinct-based algorithms. These algorithms mimic animals' natural behaviors, such as gathering, foraging, returning home, and wooing. Particle swarm, slap swarm, chimp, grey wolf, whale, ant colony, bald eagle search, black widow, marine predators, coati optimization, etc. Additionally, based on the No Free Lunch Theorem [14], no single algorithm can handle every issue. With the aid of the learning model for a particular task, meta-heuristic algorithms (MAs) are intelligent algorithms that perform mathematical operations and attempt several times to find an optimal answer from a set of random solutions [15]. For MAs, achieving a balance between exploration and exploitation is their biggest challenge. The inquiry covers a range of contexts in the exploration stage to discover additional locations where high-quality solutions may be found. Resources are concentrated on a particular search area during exploitation [16]. SO is a meta-heuristic optimization algorithm, meaning it doesn't require specific information about the problem's gradient or other explicit mathematical characteristics. It can search through the solution space randomly, guided by heuristic rules based on snake behavior. SO needs to balance exploration and exploitation as:

- **Exploration**: This phase involves searching for potential solutions across the entire search space, akin to a snake searching for food in a wide area. The algorithm explores different regions to avoid getting stuck in a local optimum.
- **Exploitation**: Once a promising region (a solution with good potential) is found, the algorithm focuses resources on improving that specific solution, similar to how a snake would exploit a known food source. This process continues until an optimal or near-optimal solution is identified.

Using Feature Selection (FS) approaches, classification algorithms for medical applications can perform better by extracting the most recognized features. Building on the benefits of MAs in feature selection issues, Snake Optimization (SO) is a recently developed, continuous, nature-inspired technique that resembles snakes' behaviors. In the Snake Optimization algorithm, behaviors such as mating and fighting are used as metaphors for exploration and exploitation strategies. The mating process is modeled to occur when certain conditions are favorable, such as lower environmental or algorithmic 'temperature,' which refers to a situation where the search space is more constrained, and the algorithm is refining its search (exploitation). In the case of fighting, the snakes compete (like candidate solutions being evaluated against one another), and the best-performing solution (the 'winning male') is selected to continue in the optimization process. If no optimal solution is found in the current search, the algorithm enters an exploration phase, expanding its search across a broader range of possible solutions. However, once a good solution is located ('the snakes find food'), the algorithm focuses on exploiting that solution to refine it further [17]. In 2030, there will be 23.6 million deaths worldwide from CVDs, mainly from heart disease and stroke, according to a WHO report [7]. Therefore, to save lives and lessen the financial burden on society, it is imperative to employ machine learning techniques to detect CVD illness. This study aims to provide a snake-optimized framework for CVD detection called CVD-SO. As, SO is applied as a feature selection tool, selecting the best combination of features (e.g., patient data such as blood pressure, cholesterol levels) from a large set. This improves the performance of the classification models used to predict CVD risk, ensuring a lower false positive rate and better diagnostic accuracy. The present study has led to the following innovative contributions:

1. Employ a novel CVD-SO framework, which enables the analysis of CVD data with five different ML techniques. The framework was trained to diagnose cardiovascular heart disease precisely.
2. A dependable and effective workflow for preparing medical data is provided, which raises the model's learnability and predictability.
3. Following the pre-processing phase, machine learning algorithms are employed to predict CVD from the gathered features, assuring high accuracy and a decreased probability of false diagnoses. The snake optimization algorithm automatically determines the best combination of CVD features.

The rest of the paper is organized as follows: Section 2 summarizes previous research on machine learning techniques for cardiovascular diagnostics. Section 3 describes the tools and techniques used in the present investigation. The analysis and results of the experiment are reported in Section 4. Section 5 discusses the limitations of the work. Section 6 demonstrates the conclusion and future research directions.

2. Related works

In the realm of related work, various approaches have been proposed for the prediction and diagnosis of cardiovascular diseases. Malik Sh. Braik et al. [18] introduced three enhanced adaptive snake optimizer variants, demonstrating superior search performance. These optimizers were employed to develop mathematical models (power, exponential, and delayed S-shaped), showcasing more than 90 % performance scores in sensitivity, accuracy, and specificity across 12, 6, and 8 datasets. Senthilkumar Mohan et al. [19] presented a hybrid random forest with a linear model (HRFLM), achieving an accuracy level of 88.7 % in predicting heart illness. K. Saikumar et al. [20] delved into deep learning-based applications for cardiac diagnosis using IoT sensor data, utilizing the K-means algorithm for noise elimination and Linear Quadratic

Table 1
Comparison of Cardiovascular Disease Prediction Approaches.

Reference	Approach	Model/Algorithm	Performance Metrics
Ref [18]	Enhanced Adaptive Snake Optimizer	Power, Exponential, Delayed S-shaped models	>90 % Sensitivity, Accuracy, Specificity
Ref [19]	Hybrid Random Forest with Linear Model (HRFLM)	HRFLM	88.7 % Accuracy
Ref [20]	Deep Learning using IoT sensor data	DG_ConvoNet	96 % Accuracy, 80 % Sensitivity, 73 % Specificity
Ref [21]	XGBoost	XGBoost	95.9 % Accuracy, 97.1 % Precision, 94.67 % Recall, 95.35 % F1-Measure
Ref [22]	Multiple ML Models	XGBoost	91.30 % Accuracy
Ref [23]	Ensemble-based Method	Random Forest (RF)	88.70 % Accuracy
Ref [24]	Modified LeNet-5 Transfer Model	LeNet-5	9.14 % Higher Accuracy
Ref [25]	Unsupervised Density-Based Method	DBSCAN	95 % Accuracy
Ref [26]	Various ML Techniques	DT, SVM, NB, 1D CNN-LSTM	99.5 % Accuracy
Ref [27]	Deep Learning Algorithms	SqueezeNet, AlexNet, CNN	98.23 % Accuracy, 98.22 % Recall, 98.31 % Precision, 98.21 % F1-Score
Ref [28]	Multi-class Classification	Feature Selection, Multimodal Fusion	Increased Diagnosis Accuracy for CAN
Ref [29]	Catboost Model	Catboost	90.94 % Accuracy, 92.3 % F1-Score

Discriminant Analysis for feature extraction. Their DG_ConvoNet achieved 96 % accuracy, 80 % sensitivity, 73 % specificity, 90 % precision, 79 % F-Score, and a 75 % area under the ROC curve. Umarani Nagavelli et al. [21] improved heart disease diagnosis accuracy using XGBoost, outperforming Naïve Bayes, SVM, and DO in terms of accuracy, precision, recall, and f1-measure. Ranjit Chandra Das et al. [22] explored and evaluated six ML models for heart disease prediction, with Xgboost demonstrating optimum outcomes at a 91.30 % accuracy rate. Abdullah Alqahtani et al. [23] proposed an ensemble-based method employing random forest (RF) for predicting cardiovascular diseases, achieving the highest accuracy of 88.70 %. Shaimaa Mahmoud et al. [24] suggested a modified LeNet-5 transfer model for individuals with cardiovascular disease, exhibiting enhanced accuracy compared to the original model. Y. A. Nanehkaran et al. [25] presented an unsupervised density-based method using DBSCAN for identifying abnormalities in cardiac patients, showcasing an accuracy of about 95 %. Tariq Sadad et al. [26] proposed a system utilizing various ML techniques and a 1D CNN-LSTM deep learning model, achieving a remarkable accuracy of 99.5 %. Mohammed B. Abubaker and Bilal Babayiğit [27] utilized deep learning algorithms for cardiac abnormality recognition, with a suggested CNN model outperforming existing works in terms of accuracy, recall, precision, and F1-score. Md Rafiul Hassan et al. [28] introduced a multi-class classification strategy for timely identification of cardiovascular autonomic neuropathy (CAN), significantly increasing diagnosis accuracy. Nadiah A. Baghdadi et al. [29] emphasized the development of new machine learning algorithms, particularly the Catboost model, for early recognition of heart illness, achieving an average accuracy of 90.94 % and an F1-score of around 92.3 %. These diverse approaches collectively contribute to the ongoing efforts in maximizing early diagnosis and treatment of cardiovascular diseases. As shown in Table 1, the comparison provides a comprehensive overview of the diverse methodologies and their respective performance metrics in the domain of cardiovascular disease prediction.

3. Methodology

This section provides a detailed overview of the suggested approach and an extensive illustration of the dataset utilized. The procedure starts with data gathering to prepare the data for further analysis. Under the proposed methodology, medical data related to cardiovascular disease (CVD) are immediately included in a machine learning modeling

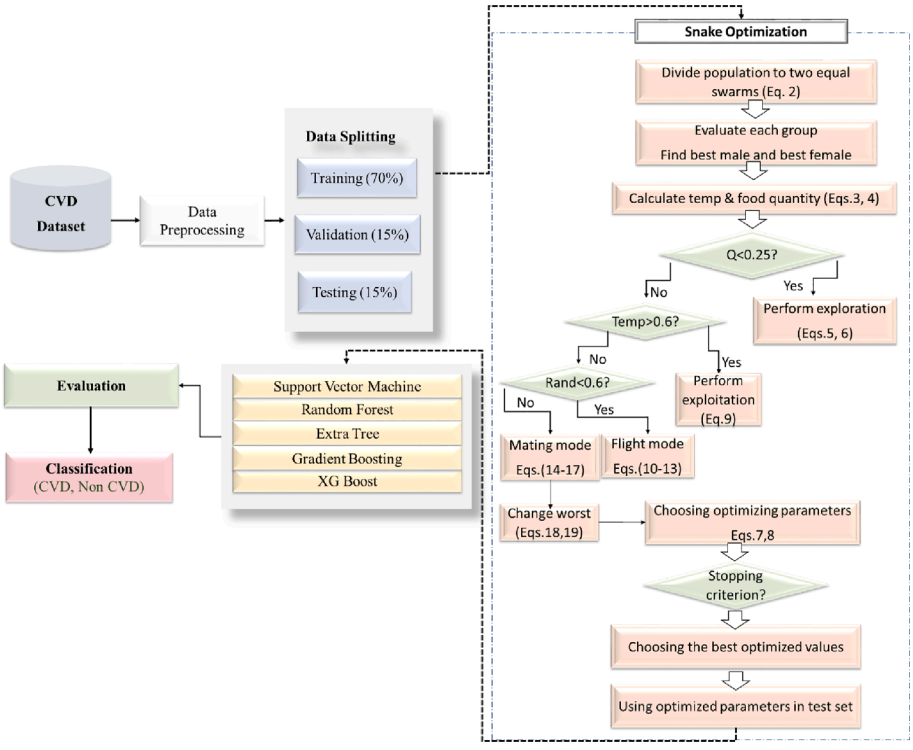


Fig. 1. The proposed CVD-SO framework.

framework to provide the ultimate decision. As cardiovascular disease continues to be the leading cause of death worldwide, efficient and accurate prediction models are critical. The use of SO for feature selection enhances the ability of machine learning models to diagnose CVD early, potentially reducing the mortality rate. This makes the integration of SO into medical diagnostics an innovative and practical approach to addressing a global health issue. With CVD-SO, an entire target system is trained using a single model, possibly handling a complicated system. The designed system is shown graphically in Fig. 1. First, data preparation methods are only used on the training set; they are also applied to the test set to maintain consistency. There is an outlier removal process as well as data class balance. After that, definite characteristics are encoded, and a standardization procedure is used to guarantee data consistency. The Snake Optimizer (SO) algorithm was introduced by Hashim and Hussien in 2022. It mimics snakes' mating habits to carry out various optimization tasks. The SO method is highly exploitable locally, is quickly realized, and is not dependent on gradient knowledge. It is established that SO performs exceptionally well when compared to LSHADE, MFO, HHO, TEO, GOA, WOA, and LSHADE-EpSin [30]. Key Parameters of SO are population size which is the number of solutions (feature sets) maintained in the algorithm at any time, max iterations that is the number of times the algorithm runs through the exploration–exploitation phases before terminating, crossover and mutation rates as the parameters that determine how much variation is introduced into new solutions during the mating phase, and stopping criteria which is a predefined threshold where the fitness score does not improve significantly over several iterations. The suggested approach uses SO techniques to comprehend characteristics and choose the most relevant features. The training and testing sets are the two separate groups into which the data is subsequently divided. This division prevents overload and data leakage. For model training, a variety of machine learning models are employed, such as Random Forest (RF), Extra Tree, Support Vector Machine (SVM), Gradient Boosting (GB), and XG Boost (XGB). Finally, the built models are extensively tested and evaluated at the final stage, and their performance is meticulously assessed.

As shown in Fig. 1, SO algorithm is employed in the feature selection process and integrated with machine learning models using the following steps:

- **Initial Setup**: The feature set for CVD data (e.g., patient medical records, test results) is first divided into multiple subsets.
- **Exploration Phase**: The SO algorithm begins by randomly selecting combinations of features from these subsets. Each set of features is tested by applying a machine learning classifier (such as support vector machines, decision trees, or neural networks) to evaluate its predictive accuracy for CVD diagnosis.
- **Evaluation and Fitness Function**: The algorithm uses a fitness function to evaluate each combination of features. This function measures how well the selected features contribute to accurately predicting whether a patient is at risk for CVD. Metrics such as accuracy, sensitivity, specificity, precision, AUC or F1-score can be used as part of this fitness function.
- **Fighting Phase (Exploitation)**: The solutions are ranked based on their performance, and weaker solutions are discarded. Stronger solutions are further refined through local searches—exploration of neighbouring solutions—helping the algorithm to narrow down the most promising feature subsets.
- **Mating Phase**: The top-performing solutions are combined to create new feature sets, which are then evaluated again. This simulates genetic combination and mutation seen in biological evolution, which leads to the generation of new, potentially better solutions.
- **Convergence**: The algorithm continues to iterate through this process, gradually refining the selected features until no significant improvement in performance is observed, indicating convergence to the optimal feature set.

While the Snake Optimization (SO) algorithm has demonstrated promising results, it still faces certain challenges. These include delayed convergence, which refers to the algorithm taking a long time to find an optimal solution. Poor solution accuracy means that the final solution may not be as precise or close to the true optimal value as desired. Additionally, the algorithm can have an easy tendency to settle into a local optimum, which occurs when the algorithm prematurely converges on a suboptimal solution rather than continuing to search for a better global optimum. These limitations stem from several factors, including the algorithm's use of fixed parameter values that do not adapt dynamically during the search process, insufficient population diversity (meaning that the candidate solutions do not vary enough from one another), and an imbalance between exploration and exploitation of the search space. This imbalance makes it more likely that the algorithm will focus too narrowly on certain areas, reducing its ability to explore new, potentially better regions. Moreover, the algorithm has a low likelihood of making significant spatial jumps—or exploring distant regions of the solution space which limits its ability to escape local optima and find the global best solution. Natural snake behavior served as the model for the SO algorithm [23]. The primary SO stages are illustrated below: the SO uses Eq. (1) to initialize a collection of random solutions in the search space. The search space is the set of all possible solutions the SO algorithm can explore. The population consists of the current candidate solutions (snakes) that are being evaluated and refined within that search space to find the optimal solution.

$$Sn_i = Sn_{min} + rand^*(Sn_{max} - Sn_{min}) \tag{1}$$

where Sn_i denotes the location in the represents of the i_{th} solution in the swarm. *rand* represents a random number between 0, 1. Sn_{min}, Sn_{max} refer to the maximum, the minimum, and values for the presented problem, respectively.

The population is split into 50 % male and 50 % female. Eq. (2) can be used to determine the number of female and male individuals.

$$N_{male} = N_{female} = \frac{N_{all}}{2} \tag{2}$$

where N_{all} represents the size of the population as a whole, N_m and N_f are the number of male and female snakes, respectively.

Two critical parameters influencing snake mating in SO are temperature and food quantity, specified by Eq. (3) and Eq. (4).

$$Temp = e^{\left(\frac{-Citer}{Titer}\right)} \tag{3}$$

$$Quan = C1^*Exp\left(\frac{Citer - Titer}{Titer}\right) \tag{4}$$

where *Citer* denotes the current iteration, *Titer* indicates the number of all iterations and $C1$ is a constant value equal to 0.5.

The snakes choose a random position and adjust their position concerning it in their quest for food if Q < Threshold (Threshold = 0.25). The following Eq (5), and Eq(6) are applied to the model exploration phase:

$$Sn_{i,male}(t+1) = Sn_{rand,male}(t) \pm c2^*A_{male}{}^*((Sn_{max} - Sn_{min})^*rand + Sn_{min}) \tag{5}$$

$$Sn_{i,female}(t+1) = Sn_{rand,female}(t) \pm c2^*A_{female}{}^*((Sn_{max} - Sn_{min})^*rand + Sn_{min}) \tag{6}$$

where $Sn_{i,male}$ denotes i_{th} male position, $Sn_{rand,male}$ represents the random male position, $Sn_{i,female}$ refers to female solution, $Sn_{rand,female}$ refers to the random female location, *rand* is a number $\in$ [0,1], and $c2$ is a constant equals 0.05. The male and female ability to locate food are denoted by A_{male} and A_{female}, which are expressed as Eq. (7) and Eq. (8):

Table 2
Fight and Mating Modes in Snake Optimizer Algorithm.

Fight Mode Eqs.(10–13)	Mating mode Eqs.(14–17)
$Sn_{i,male}(t+1) = Sn_{i,male}(t) + c3*Fmod_{male}*rand*(Quan*Sn_{best,female} - Sn_{i,male}(t))$ (10)	$Sn_{i,male}(t+1) = Sn_{i,male}(t) + c3*Mmod_{male}*rand*(Quan*Sn_{i,female}(t) - Sn_{i,male}(t))$ (14)
$Sn_{i,female}(t+1) = Sn_{i,female}(t+1) + c3*Fmod_{female}*rand*(Quan*Sn_{best,male} - Sn_{i,female}(t+1))$ (11)	$Sn_{i,female}(t+1) = Sn_{i,female}(t) + c3*Mmod_{female}*rand*(Quan*Sn_{i,male}(t) - Sn_{i,female}(t))$ (15)
$Fmod_{male}$, $Fmod_{female}$ are the fighting ability of male and female individuals, and can be computed using Eqs.(12, 13):	$Mmod_{male}$, $Mmod_{female}$ are the mating ability of male and female agents which can be represented as following Eqs.(16,17):
$Fmod_{male} = \exp(-\frac{fit_{best,female}}{f_i})$ (12)	$Mmod_{male} = \exp(-\frac{fit_{i,female}}{f_{i,male}})$ (16)
$Fmod_{female} = \exp(-\frac{fit_{best,male}}{f_i})$ (13)	$Mmod_{female} = \exp(-\frac{fit_{i,male}}{f_{i,female}})$ (17)
Where $Sn_{i,male}$ indicates i_{th} agent position in male group, $Sn_{i,female}$ refers to the position of i_{th} agent in female group and $Sn_{best,male}$, $Sn_{best,female}$ refer to the best individual position in male and female agents. $fit_{best,female}$ denotes the best agent fitness of female group, $fit_{best,male}$ represents the best agent fitness of male group, and f_i is the individual fitness.	

Table 3
CVD dataset statistical description.

	mean	std	min	50%	max
Age	54.43414	9.072290	29.0	56.0	77.0
Sex	0.695610	0.460373	0.0	1.0	1.0
Cp	0.942439	1.029641	0.0	1.0	3.0
Trestbps	131.6117	17.51671	94	130	200
Chol	246.0000	51.59251	126	240	564
Fbs	0.149268	0.356527	0.0	0.0	1.0
Restecg	0.529756	0.527878	0.0	1.0	2.0
Thalach	149.1141	23.00572	71	152	202
Exang	0.336585	0.472772	0.0	0.0	1.0
Oldpeak	1.071512	1.175053	0.0	0.8	6.2
Slope	1.385366	0.617755	0.0	1.0	2.0
Ca	0.754146	1.030798	0.0	0.0	4.0
Thal	2.323902	0.620660	0.0	2.0	3.0
Target	0.513171	0.500070	0.0	1.0	1.0

$$A_{male} = \exp(-\frac{Fitmale_{rand}}{Fitmale_i}) \quad (7)$$

$$A_{female} = \exp(-\frac{Fitfemale_{rand}}{Fitfemale_i}) \quad (8)$$

where $Fitmale_{rand}$ represents the fitness of $Sn_{rand,male}$, and $Fitmale_i$ refers to the fitness of i_{th} solution in the male group. Additionally, $Fitfemale_{rand}$ is the fitness of $Sn_{rand,female}$ and $Fitfemale_i$ denotes the fitness of i_{th} solution the in male group.

Exploiting the search space (Food discovered): If there is more food than a certain amount, and when the quantity is more than 0.25, the temperature is examined. Only utilizing Eq. (9), will the solutions move to the food if the temperature is higher than 0.25 (hot).

$$Sn_{i,j}(t+1) = Sn_{food} \pm c3*Temp*rand*(Sn_{food} - Sn_{i,j}(t)) \quad (9)$$

Where $Sn_{i,j}$ denotes the position of male individual or female, Sn_{food} represents the best individual position, and $c3$ is constant equals 2.

The snake will be in either the fight or mating mode if the temperature is below 0.6 (cold). Table 2 shows the differences between the two modes.

In the scenario that the egg hatches, change the worst male and female solutions by applying Eqs. (18), 19):

$$Sn_{worst,male} = Sn_{min} + rand*(Sn_{max} - Sn_{min}) \quad (18)$$

$$Sn_{worst,female} = Sn_{min} + rand*(Sn_{max} - Sn_{min}) \quad (19)$$

where $Sn_{worst,male}$ denotes the worst solution in the male group, $Sn_{worst,female}$ refers to the worst solution in the female group.

For the detection of CVD from health records, five ML models including Random Forest (RF), Extra Tree (ET), Support Vector Machine (SVM), Gradient Boosting (GB), and XG Boost are utilized. The pseudocode of the proposed Snake Optimizer-based feature selection is presented in algorithm 1.

Algorithm 1: Pseudocode of the Proposed Snake Optimizer

```
Define population parameters: lower and upper bounds (Sn_max, Sn_min), Total iterations
  (Titer), and population size (N_all).
Initialize snake population randomly.
Divide population into 2 equal groups: male and female (N_male, N_female) using Eq. (2).
Perform iterations of the Snake Optimization Algorithm to select features.
  for iteration in range (Titer):
    Find fit_best,female and fit_best,male
    Calculate Temp and Quan using Eqs. (3),4).
    If (Quan < 0.25) then
      Perform exploration phase for males and females using Eqs.(5–8).
    Else If (temp > 0.25) then
      Perform exploitation phase using Eq. (9).
      Else if (rand < 0.6) then
        Identify male and female snakes, respectively, in Fighting mode in Eqs.
  (10–13)
    Else
        Identify male and female snakes, respectively, in mating mode in Eqs.
  (14–17)
        Change the worst male and worst female snakes using Eqs.(18,19).
    End if
  End if
End while
Get the best feature subset from the final snake population.
Split the data into training and testing sets.
Train RF, SVM, GB, Extra Tree, and XG Boost on the training set.
Evaluate the model's performance on the testing set.
Return classification measurements based on the models' performance.
```

3.1. Dataset description

The dataset used in this study is from the UCI machine learning repository. It comprises four distinct databases: Cleveland, Hungary,

Table 4
Description of the attributes.

No.	Attribute	Description
1	Age	Age of the patient measured in years
2	Sex	Type of gender (1 = male; 0 = female)
3	Cp	Chest pain type (1 = angina; 2 = atypical form of angina; 3 = non-angina; 4 = no symptoms of angina)
4	Trestbps	Blood pressure at rest measured in millimeters of mercury (mm Hg)
5	Chol	Cholesterol level measured in milligrams per deciliter (mg/dL)
6	Fbs	Fasting blood sugar level exceeding 120 (mg/dL) (1 = true; 0 = false)
7	Restecg	Resting ECG reading (0 = normal; 1 = abnormal; 2 = definite ventricular)
8	Thalach	Recorded maximum heart rate
9	Exang	Exertion-induced angina (1 = yes; 0 = no)
10	Oldpeak	ST depression occurring during exercise
11	Slope	T segment slope during peak exercise (1 = up sloping; 2 = flat; 3 = down sloping)
12	Ca	Number of significant vessels colored during fluoroscopy
13	Thal	Types of detect (3 = normal; 6 = fixed defect; 7 = reversable defect)
14	Target	Label column (1 = disease; 0 = no disease)

Switzerland, and Long Beach V. The dataset consists of 14 attributes and is available [31]. The "Target" field is used to denote the presence or absence of heart disease in the patient and is represented as an integer: 0 signifies no disease, and 1 indicates the presence of the disease. Table 3 depicts the statistical description of the attributes. Table 4 demonstrates the description of the attributes.

The feature extraction and selection process identify the most pertinent attributes from the initial dataset, discarding those that lack relevance or are redundant. A machine learning model can enhance its performance by employing this process while streamlining computational complexity. Exploring associations between the features and the target variable is beneficial for a deeper comprehension of the data. Although the correlation coefficient is not the definitive method for assessing a feature's "relevance," it does offer valuable insights into potential relationships within the dataset. As depicted in Fig. 2, the heatmap reveals notable patterns within the data. It highlights features with high correlation coefficients and those exhibiting a substantial decline in correlation. To identify and select the most advantageous combinations of features from the cardiovascular disease (CVD) dataset, the Snake Optimization (SO) algorithm is harnessed within the framework of CVD-SO. This model draws inspiration from the behaviors of snakes, including their hunting strategies and movement patterns, to optimize feature selection.

3.2. Remove outliers

A dataset's outlier is an individual or group data points that differ significantly from the rest of dataset. The data points, or data points, in that case, are not found within the wider range of data values in a dataset, a bit outside its boarder [32]. We used Python to filter the dataset for outliers and extreme values based on interquartile ranges.

3.3. Handle class balance

By generating synthetic samples, a SMOTE approach boosts the representation of minority classes in a dataset [33]. The classifier is, therefore, given a more balanced dataset for learning purposes, mitigating the problem of imbalance. To better capture underlying patterns and enhance classification, generated samples resemble the minority class rather than replicating existing ones. Additionally, by using this technique, the overfitting issue brought on by random oversampling will be resolved.

3.4. Encoding Categorical Variables

This stage aims to transform category categories into numerical values so that machine learning algorithms can appropriately identify and comprehend them. A hot encoding approach is used to encode the values of these categories as 0 or 1 [34].

3.5. Standardization

Using data standardization, the final step before developing a predictive model is to ensure that all attributes are assessed at the same level [35]. This method's primary benefit is enhancing model learning and classification performance. Eq. (20) is used by Scikit-learn to standardize the features. Z-scores are computed by dividing a given value by the number of standard deviations that deviate from the mean. The negative z-score indicates a value below the mean, but the positive z-score indicates a value above the mean [36].

$$z = \frac{x - \mu}{\sigma} \tag{20}$$

where x is the sample with a standard value, *themeanofthetrainingsample*,

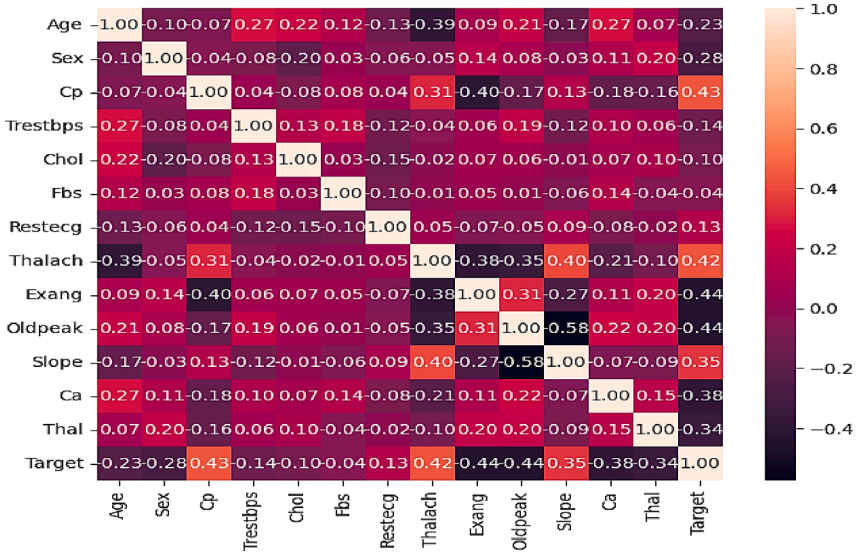


Fig. 2. Heatmap of CVD Dataset.

Table 5
Hyperparameter settings for classification models.

Models	Hyperparameters
RF	n_estimators = 50, criterion = entropy.
ET	n_estimators = 150, criterion = gini.
SVM	kernel = rbf, regularization parameter (C) = 0.2.
GB	n_estimators = 100, max_features = auto, max_depth = 1.
XGBoost	n_estimators = 50, learning_rate = 0.1.

and σ is the standard deviation.

3.6. Evaluation metrics

Machine learning models were assessed using different sets of essential features, evaluating their performance on a test dataset. A confusion matrix is employed to gauge a classifier's performance, wherein the actual values are known. The key terms in this context are described as follows:

- True Positive (TP): These indicate instances where the prediction correctly identifies the presence of cardiovascular disease (CVD).
- True Negative (TN): These indicate instances where the prediction correctly identifies the absence of CVD.
- False Positive (FP): These indicate instances where the prediction erroneously suggests the presence of CVD when it is not present.
- False Negative (FN): In these cases, the prediction fails to identify CVD when it is present.

Classification reports are generated using the following measures: accuracy, sensitivity, specificity, precision, F1 score, and area under curve (AUC) [37,38]. The receiver operating characteristic (ROC) curve serves as a widely employed metric for evaluating the effectiveness of binary classification models. It visually depicts how the actual positive rate and false positive rate trade off against each other. AUC is a metric that assesses a model's performance by analyzing its ROC curve. It summarizes the model's capacity to differentiate between positive and negative classes, with a more excellent AUC value signifying superior discrimination capability. The performance measures are represented in Eqs. (21–26).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{21}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{22}$$

$$Specitivity = \frac{TN}{TN + FP} \tag{23}$$

$$Precision = \frac{TP}{TP + FP} \tag{24}$$

$$F1score = \frac{2{*}TP}{2{*}TP + FP + FN} \tag{25}$$

$$AUC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{26}$$

4. Results and Discussion

The experimental research was carried out using Jupyter Notebook version 6.4.6, a potent tool for running and creating Python programs. Jupyter Notebook is extensively employed in artificial intelligence (AI) and machine learning as an open-source platform for implementing and executing code. The primary goal of this research is to leverage machine learning techniques for early cardiovascular disease (CVD) detection while addressing the intricacies inherent in medical data analysis. In addition, we have devised an innovative feature selection algorithm optimized through a snake-inspired approach. Different machine learning methods are used to compare the outcomes to confirm SO's efficacy. The SO algorithm greatly enhances practical problem-solving skills as well as algorithm performance. We have used five distinct machine learning models, namely, random forest (RF), extra trees (ET), support vector machine (SVM), gradient boosting (GB), and XGBoost, specifically for the task of detecting CVD. The feature selection (FS) problem revolves around the initialization of a solution utilizing a binary vector. The length of this vector is directly proportional to the dimensionality of the problem, where each bit within it corresponds to a feature present in the dataset. These bit values are binary, with '0′ denoting the exclusion of a feature and '1′ signifying its inclusion in the selected features. The snake optimization algorithm selects seven adequate and relevant features. The performance of the models was evaluated using six evaluation metrics: accuracy, sensitivity, specificity, precision, F1 score, and area under curve (AUC). Table 5 provides the hyperparameter settings for the classification models utilized in this study. The RF model is configured with 50 estimators for decision trees, and the criterion for splitting nodes is set to "entropy". For ET, the model utilizes 150 estimators for decision trees, and the criterion for splitting nodes is set to "gini". The SVM model employs a radial basis function (rbf) kernel with a regularization parameter (C) value of 0.2. In GB, this model is configured with 100 estimators and employs "auto" to determine the maximum number of features, and the full depth of the trees is set to 1. The XGBoost model has 50 boosting rounds (n_estimators) and a learning rate 0.1.

Table 6 presents the experimental results encompassing accuracy, sensitivity, specificity, precision, F1 score, AUC, and fitted time in seconds for five distinct machine learning models: random forest (RF), extra trees (ET), support vector machine (SVM), gradient boosting (GB), and XGBoost using the selected features by SO algorithm. In this table, the most outstanding results among these evaluation metrics are highlighted and emphasized by being presented in bold text.

As shown in Table 6, the performance of the classification models using RF, ET, SVM, GB, and XGBoost are demonstrated using the selected features by SO algorithm. SO-RF model presents the best results among the other models, namely, SO-ET, SO-SVM, SO-GB, and SO-XGBoost. The SO-RF model exhibited exceptional performance with an accuracy of 99.9 %, sensitivity of 99.9 %, perfect specificity of 100 %, precision of 99.9 %, and an F1 score of 100 %. It achieved an ideal AUC of 1.0, indicating the highest level of discrimination capability and the best fitted time of 1.251. The SO-ET model achieved an accuracy of 97.1 %, sensitivity of 97.1 %, specificity of 97.2 %, precision of 97.1 %, an F1 score of 97.1 %, AUC of 0.995, and fitted time of 1.324. The SO-SVM model demonstrated an accuracy of 91.7 %, sensitivity of 91.6 %, specificity of 91.7 %, precision of 91.7 %, an F1 score of 91.7 %, an AUC of 0.981, and fitted time of 1.502. The SO-GB model achieved an accuracy of 86.3 %, sensitivity of 86.3 %, specificity of 86.3 %, precision of 86.3 %, an F1 score of 86.4 %, an AUC of 0.959, and fitted time of 1.853. The worst results were achieved by the SO-XGBoost model with an accuracy of 83.4 %, sensitivity of 83.3 %, specificity of 83.4 %, precision of 83.6 %, an F1 score of 83.4 %, AUC of 0.919, and fitted time of 2.631.

Table 7 presents experimental results encompassing accuracy, sensitivity, specificity, precision, F1 score, AUC, and fitted time in seconds for five distinct machine learning models: RF, ET, SVM, GB, and XGBoost, without using the selected features by the SO algorithm.

As shown in Table 7, the performance of the classification models using RF, ET, SVM, GB, and XGBoost is demonstrated without using the selected features by the SO algorithm. RF model presents the best results among the other models, namely, ET, SVM, GB, and XGBoost. The RF model exhibited excellent performance with an accuracy of 92.4 %, sensitivity of 92.4 %, specificity of 92.4 %, precision of 92.4 %, and an F1 score of 92.5 %. It achieved a high AUC of 0.974, indicating strong discrimination capability and the best fitted time of 3.152. The ET model achieved an accuracy of 90.5 %, sensitivity of 90.6 %, specificity of 90.5

Table 6
Performance of the classification models using the selected features by SO algorithm.

Models	Accuracy	Sensitivity	Specificity	Precision	F1 score	AUC	Fitted Time
SO-RF	**99.9 %**	**99.9 %**	**100 %**	**99.9 %**	**100 %**	**1.0**	1.251
SO-ET	97.1 %	97.1 %	97.2 %	97.1 %	97.1 %	0.995	1.324
SO-SVM	91.7 %	91.6 %	91.7 %	91.7 %	91.7 %	0.981	1.502
SO-GB	86.3 %	86.3 %	86.3 %	86.3 %	86.4 %	0.959	1.853
SO-XGBoost	83.4 %	83.3 %	83.4 %	83.6 %	83.4 %	0.919	2.631

Table 7
Performance of the classification models without using the selected features by SO algorithm.

Models	Accuracy	Sensitivity	Specificity	Precision	F1 score	AUC	Fitted Time
RF	**92.4 %**	**92.4 %**	**92.4 %**	**92.4 %**	**92.5 %**	**0.974**	**3.152**
ET	90.5 %	90.6 %	90.5 %	90.6 %	90.5 %	0.952	3.582
SVM	87.2 %	87.2 %	87.3 %	87.2 %	87.2 %	0.923	3.725
GB	82.6 %	82.6 %	82.6 %	82.6 %	82.6 %	0.907	4.018
XGBoost	79.7 %	79.7 %	79.7 %	79.6 %	79.7 %	0.886	4.546

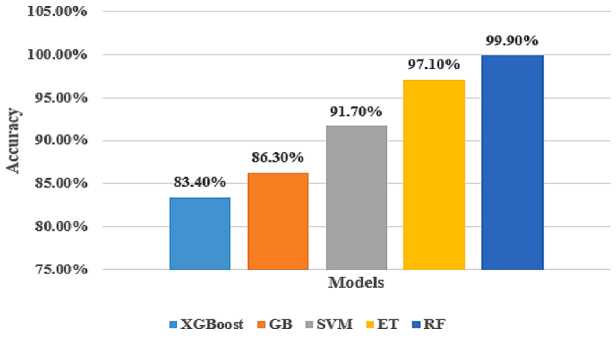


Fig. 3. Comparison between the Five Classification Models Regarding Accuracy using the Selected Features by SO Algorithm.

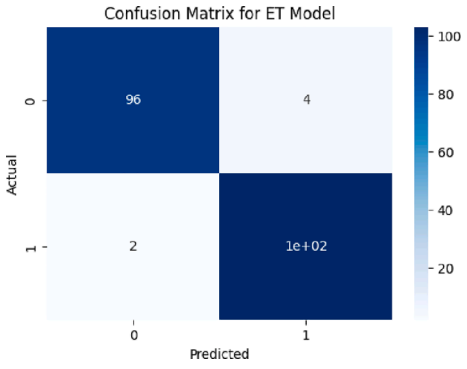


Fig. 5. Confusion Matrix for ET model.

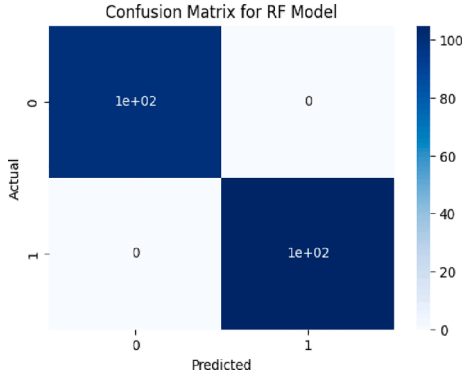


Fig. 4. Confusion Matrix for RF model.

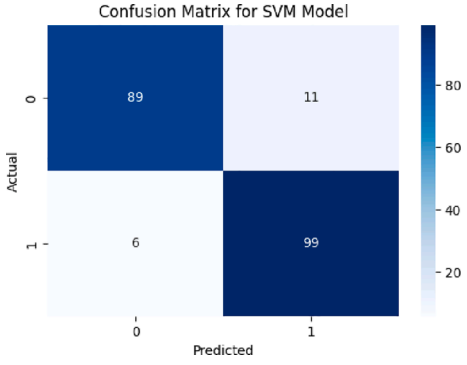


Fig. 6. Confusion Matrix for SVM model.

%, precision of 90.6 %, an F1 score of 90.5 %, an AUC of 0.952, and fitted time of 3.582. The SVM model demonstrated an accuracy of 87.2 %, sensitivity of 87.2 %, specificity of 87.3 %, precision of 87.2 %, an F1 score of 87.2 %, an AUC of 0.923, and fitted time of 3.725. The GB model achieved an accuracy of 82.6 %, sensitivity of 82.6 %, specificity of 82.6 %, precision of 82.6 %, an F1 score of 82.6 %, an AUC of 0.907, and fitted time of 4.018. The worst results were achieved by the XGBoost model with an accuracy of 79.7 %, sensitivity of 79.7 %, specificity of 79.7 %, precision of 79.6 %, an F1 score of 79.7 %, an AUC of 0.886, and fitted time of 4.546. Fig. 3 displays the accuracy of the five classification models using the selected features by the SO algorithm. Figs. 4 – 8 demonstrate the confusion matrix of RF, ET, SVM, GB, and XGBoost models, respectively using the selected features by SO algorithm.

Table 8 presents the performance of four optimization algorithms for feature selection, namely, Particle Swarm Optimization (PSO) algorithm, Genetic Algorithm (GA), Grey Wolf Optimization (GWO), and Whale Optimization Algorithm (WOA) compared to SO algorithm in terms of accuracy, sensitivity, specificity, precision, F1 score, AUC, and fitted time in seconds using RF model.

As demonstrated in Table 8, the performance of SO with RF presents the best results with an accuracy of 99.9 %, sensitivity of 99.9 %, perfect specificity of 100 %, precision of 99.9 %, and an F1 score of 100 %. It achieved an ideal AUC of 1.0, indicating the highest level of discrimination capability, and the best fitted time of 1.251. The worst results were achieved by the WOA with RF with an accuracy of 92.52 %, sensitivity of 92.53 %, specificity of 92.52 %, precision of 92.54 %, an F1 score of 92.52 %, AUC of 0.927, and fitted time of 2.172.

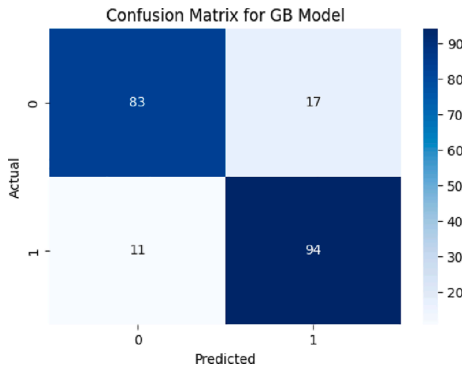


Fig. 7. Confusion Matrix for GB model.

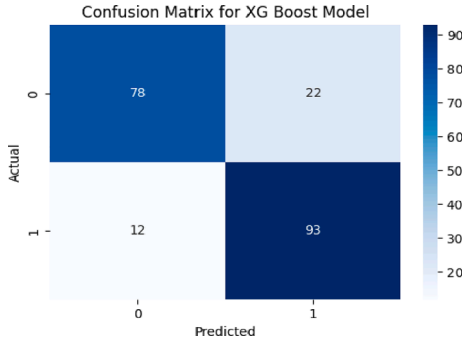


Fig. 8. Confusion Matrix for XGBoost model.

Table 9 demonstrates the hyperparameter settings for the optimization algorithms used in this study.

Table 10 presents a comparative analysis of the classification accuracy between the state-of-the-art models utilizing the UCI heart disease dataset and the proposed CVD-SO model. The comparison is organized in ascending order of accuracy.

Fig. 9 displays the sensitivity of the five classification models using the selected features by the SO algorithm.

Fig. 10 illustrates the AUC values for the models, namely, RF, ET, SVM, GB, and XGBoost, using the selected features by the SO algorithm.

4.1. Statistical analysis

The examination of statistical differentials and significance concerning the efficacy of classification models, utilizing features systematically chosen by the Sequential Optimization (SO) algorithm, was undertaken through the employment of two distinctive statistical methodologies, specifically, the Analysis of Variance (ANOVA) test and the Wilcoxon signed-rank test [47–49]. The ANOVA analysis functioned as a conduit to scrutinize the statistical import of observed performance distinctions across diverse groups of classification models. Elaborate delineations of outcomes, accompanied by pertinent statistical metrics, are meticulously documented within Table 11, affording a meticulous exposition of the ANOVA-derived insights.

Concurrent with the ANOVA scrutiny, applying the Wilcoxon signed-

Table 9
Hyperparameter settings for the optimization algorithms.

Algorithm	Hyperparameters	Value
SO	Population size	100
	Iterations	100
	Threshold values for exploration and exploitation	0.5
	Crossover and mutation rates	0.3
PSO	Population size	150
	Iterations	100
	Cognitive coefficient	2
	0000	0.4
	Social coefficient	
	0000	
GA	Crossover	0.8
	Mutation ratio	0.2
	Agents	20
GWO	Iterations	100
	Wolves	30
WOA	Population size	200
	Iterations	100

Table 10
Comparative analysis of the classification accuracy.

Studies	Model	Accuracy
Ref [39], **2022**	LR model	85.9 %
Ref [40], **2020**	Kernel discriminant analysis and adapted self-adaptive Bayesian model	90 %
Ref [41], **2017**	Feature selection using relief and rough set techniques, combined with an ensemble classifier model	92.6 %
Ref [42], **2022**	Enhanced deep CNN model	93.3 %
Ref [43], **2022**	FIS with adapted Bi-LSTM model	97.3 %
Ref [44], **2018**	SVM model.	97.5 %
Ref [45], **2021**	Deep dynamic ensemble model	98.1 %
Ref [46], **2018**	SAX with LSTM	98.4 %
Proposed CVD-SO	SO-RF model	**99.9 %**

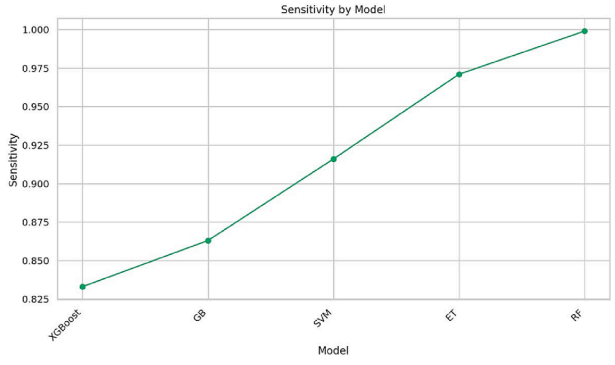


Fig. 9. Comparison between the Five Classification Models Regarding Sensitivity using the Selected Features by SO Algorithm.

Table 8
Performance of the optimization algorithms compared to SO algorithm.

Models	Accuracy	Sensitivity	Specificity	Precision	F1 score	AUC	Fitted Time
SO-RF	**99.90 %**	**99.90 %**	**100 %**	**99.90 %**	**100 %**	**1**	1.251
PSO-RF	97.81 %	97.83 %	97.01 %	97.85 %	98.84 %	0.977	1.683
GA-RF	96.65 %	96.66 %	96.65 %	96.65 %	96.66 %	0.968	1.727
GWO-RF	93.37 %	93.36 %	93.37 %	93.38 %	93.37 %	0.936	1.979
WOA-RF	92.52 %	92.53 %	92.52 %	92.54 %	92.52 %	0.927	2.172

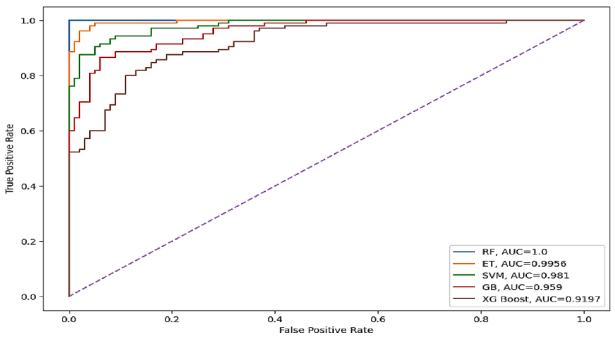


Fig. 10. AUC values for the Models are RF, ET, SVM, GB, and XGBoost.

Table 11
ANOVA for Performance of the classification models using the selected features by SO Algorithm.

ANOVA table	SS	DF	MS	F (DFn, DFd)	P value
Treatment (between columns)	0.1945	4	0.04862	F (4, 45) = 2349	P < 0.0001
Residual (within columns)	0.0009313	45	2.069e-005		
Total	0.1954	49			

Table 12
Wilcoxon for Performance of The Classification Models using the Selected Features by SO Algorithm.

	XGBoost	GB	SVM	ET	RF
Theoretical median	0.000	0.000	0.000	0.000	0.000
Actual median	0.8340	0.8630	0.9170	0.9710	0.9990
Number of values	10	10	10	10	10
Wilcoxon Signed Rank Test					
Sum of signed ranks (W)	55.00	55.00	55.00	55.00	55.00
Sum of positive ranks	55.00	55.00	55.00	55.00	55.00
Sum of negative ranks	0.000	0.000	0.000	0.000	0.000
P value (two tailed)	0.0020	0.0020	0.0020	0.0020	0.0020
Exact or estimate?	Exact	Exact	Exact	Exact	Exact
P value summary	**	**	**	**	**
Significant (alpha = 0.05)?	Yes	Yes	Yes	Yes	Yes
How significant is the discrepancy?					
Discrepancy	0.8340	0.8630	0.9170	0.9710	0.9990

rank test facilitated nuanced pairwise comparisons between classification models. This non-parametric evaluation, well-suited for scenarios marked by non-normal data distributions or modest sample sizes, yielded discerning insights. The specific findings from these comparative analyses are systematically presented in Table 12, proffering a nuanced comprehension of the statistical significance characterizing distinctions between designated pairs of classification models.

In the interest of concise and visually comprehensible representation, Fig. 11 encapsulates pivotal outcomes, enabling a reasonable grasp of the observed variabilities in classification model performance. Cumulatively, these empirical investigations and illustrative depictions enrich our scholarly understanding of the consequentiality and efficacy of features sourced from the SO algorithm and their impact on the performance landscape of classification models.

5. Limitation of the proposed work

Publicly accessible datasets from the UCI and Kaggle repositories were used to validate the proposed framework. Despite their widespread use, these datasets might not accurately reflect the variety of patient populations or clinical variation found in the real world. Snake Optimization (SO) is a meta-heuristic algorithm useful for feature selection but may show delayed convergence for very large or complex datasets. Adjusting optimization for particular use cases might be necessary. The study focuses on traditional machine learning models and does not incorporate state-of-the-art deep learning models or ensemble models that could further enhance prediction performance.

6. Conclusion and future work

This paper introduces a novel optimized learning approach called CVD-SO to detect cardiovascular disease (CVD). The primary aim of this study is to propose an effective system that surpasses the accuracy of existing CVD detection techniques by incorporating a feature selection algorithm inspired by snake optimization. The dataset utilized in this research comprises 14 distinct attributes and originates from the UCI machine learning repository. It encompasses four separate databases: Cleveland, Hungary, Switzerland, and Long Beach V. The dataset is splitted into to 70 % training, 15 % validation, and 15 % testing. To identify and categorize the most appropriate data features, a combination of five machine learning models, namely, random forest (RF), extra trees (ET), support vector machine (SVM), gradient boosting (GB), and XGBoost are employed in conjunction with the snake optimization (SO) algorithm. This integrated approach aims to generate highly accurate predictions for CVD. The classification models were evaluated using six metrics: accuracy, sensitivity, specificity, precision, F1 score, and area under curve (AUC) before and after applying the SO algorithm for feature selection. The proposed CVD-SO model achieved the highest accuracy, with 99.9 %, compared with the state-of-the-art models utilizing the UCI heart disease dataset. In the future, there is the potential for investigating the effectiveness of sophisticated nonlinear neural networks enhanced by artificial intelligence, such as GK-ARFNN and SOFNN-HPS, in predicting treatment strategies for CVD. The proposed Snake Optimization Algorithm (SO) supports scalability due to its meta-heuristic nature and ability to optimize a subset of features. The framework was successfully applied to datasets of varying sizes, demonstrating its adaptability. Moreover, the modular design of the framework allows it to integrate distributed computing techniques in future work, ensuring scalability to larger datasets and more complex real-world systems. The framework is designed to work with multiple machine learning models (e.g., RF, ET, SVM, GB, XGBoost) and diverse datasets. These capabilities underscore their flexibility and interoperability. However, we acknowledge the need to deal with healthcare standards such as Health Level 7 (HL7) and Fast Healthcare Interoperability Resources (FHIR) for seamless integration with Electronic Health Records (EHRs) and other clinical systems. While this research focuses on performance and accuracy, we recognize the critical importance of regulatory compliance for real-world applications. The framework aligns with ethical AI principles including reliability and transparency. In future work, we will ensure compliance with relevant healthcare regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR). This will include anonymizing patient data and implementing secure data handling protocols.

Our study demonstrated the efficacy of the CVD-SO framework using datasets from the UCI repository and Kaggle, these datasets are controlled and may not fully represent the diversity of real-world clinical data, so in the future we can validate the framework on larger datasets, including multi-center clinical datasets and real-time data from wearable health devices and investigating the framework's performance in dealing with data imbalances, noise, and missing values more robustly. This study focuses on traditional machine learning models (e.g., RF, ET, SVM, GB, XGBoost), integrating the Snake Optimization Algorithm (SO). Deep learning models could yield further improvements,

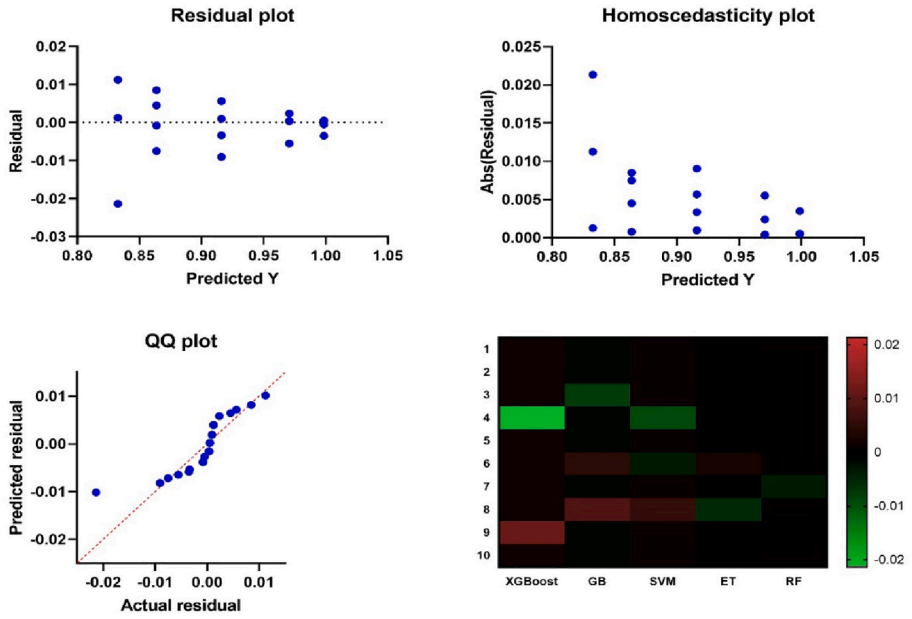


Fig. 11. Visualizing the results of CVD classification models using the selected features by SO Algorithm.

Table 13
Performance of the classification models using the selected features by SO algorithm for the first dataset.

Models	Accuracy	Sensitivity	Specificity	Precision	F1 score	AUC	Fitted Time
SO-RF	**97.45 %**	**97.45 %**	**97.47 %**	**97.45 %**	**97.46 %**	**0.978**	**2.362**
SO-ET	96.79 %	96.77 %	96.79 %	96.78 %	96.79 %	0.965	2.384
SO-SVM	93.95 %	93.94 %	93.94 %	93.95 %	93.93 %	0.937	2.588
SO-GB	91.38 %	91.36 %	91.38 %	91.37 %	91.38 %	0.914	2.694
SO-XGBoost	88.72 %	88.73 %	88.72 %	88.72 %	88.74 %	0.889	2.953

Table 14
Performance of the classification models using the selected features by SO algorithm for the second dataset.

Models	Accuracy	Sensitivity	Specificity	Precision	F1 score	AUC	Fitted Time
SO-RF	**98.07 %**	**98.08 %**	**98.09 %**	**98.07 %**	**98.08 %**	**0.984**	**1.153**
SO-ET	96.50 %	96.48 %	96.50 %	96.48 %	96.50 %	0.962	1.184
SO-SVM	95.62 %	95.61 %	95.61 %	95.63 %	95.60 %	0.953	1.203
SO-GB	92.22 %	92.23 %	92.22 %	92.23 %	92.22 %	0.922	1.266
SO-XGBoost	89.47 %	89.46 %	89.46 %	89.47 %	89.49 %	0.896	1.293

particularly in capturing complex, nonlinear relationships in high-dimensional datasets. Real-time monitoring and prediction of cardiovascular disease risk using wearable devices and Internet of Things (IoT) remains an unexplored area. Our framework can be extended to integrate streaming data from wearable devices to enable early warning systems and develop adaptive models that can dynamically update with new patient data in real time.

Code availability

Code is available on request.

CRediT authorship contribution statement

Zahraa Tarek: Writing – original draft, Resources. **Amel Ali Alhussan:** Writing – review & editing, Formal analysis, Conceptualization. **Doaa Sami Khafaga:** Writing – original draft, Visualization, Validation. **El-Sayed M. Elkenawy:** Supervision, Project administration, Methodology. **Ahmed M. Elshewey:** Writing – original draft, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R308), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Appendix A

To test the performance of the proposed approach, we applied our proposed approach on two public datasets available at https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease and https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data. The first dataset consists of 18 features and 319,795 instances. The second dataset consists of 13 features and 299 instances. Table 13 presents experimental results encompassing

accuracy, sensitivity, specificity, precision, F1 score, AUC, and fitted time in seconds for five distinct machine learning models: random forest (RF), extra trees (ET), support vector machine (SVM), gradient boosting (GB), and XGBoost using the selected features by SO algorithm for the first dataset. In this table, the most outstanding results among these evaluation metrics are highlighted and emphasized by being presented in bold text.

As demonstrated in Table 13, the performance of SO with RF presents the best results with an accuracy of 97.45 %, sensitivity of 97.45 %, perfect specificity of 97.47 %, precision of 97.45 %, F1 score of 97.46 %, AUC of 0.978, and fitted time of 2.362. The worst results were achieved by the SO with XGBoost with an accuracy of 88.72 %, sensitivity of 88.73 %, specificity of 88.72 %, precision of 88.72 %, an F1 score of 88.74 %, an AUC of 0.889, and fitted time of 2.953.

Table 14 presents experimental results encompassing accuracy, sensitivity, specificity, precision, F1 score, AUC, and fitted time in seconds for five distinct machine learning models: random forest (RF), extra trees (ET), support vector machine (SVM), gradient boosting (GB), and XGBoost using the selected features by SO algorithm for the second dataset. In this table, the most outstanding results among these evaluation metrics are highlighted and emphasized by being presented in bold text.

As demonstrated in Table 14, the performance of SO with RF presents the best results with an accuracy of 98.07 %, sensitivity of 98.08 %, perfect specificity of 98.09 %, precision of 98.07 %, F1 score of 98.08 %, AUC of 0.984, and fitted time of 1.153. The worst results were achieved by the SO with XGBoost with an accuracy of 89.47 %, sensitivity of 89.46 %, specificity of 89.46 %, precision of 89.47 %, an F1 score of 89.49 %, an AUC of 0.896, and fitted time of 1.293.

Data availability

Data availability: at link 1. https://archive.ics.uci.edu/dataset/45/heart+disease. 2. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease. 3. https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data

References

[1] M. Hemmati, S. Kashanipoor, P. Mazaheri, F. Alibabaei, A. Babaeizad, S. Asli, S. Mohammadi, A.H. Gorgin, K. Ghods, B. Yousefi, M. Eslami, Importance of gut microbiota metabolites in the development of cardiovascular diseases (CVD), Life Sci. 329 (2023) 121947.

[2] M.M. Rahman, F. Islam, M.H. -Or-Rashid, A.A. Mamun, M.S. Rahaman, M.M. Islam, A.F. Meem, P.R. Sutradhar, S. Mitra, A.A. Mimi, T.B. Emran, The gut microbiota (microbiome) in cardiovascular disease and its therapeutic regulation, Front. Cell. Infect. Microbiol. 12 (2022) 903570.

[3] N.A. ElSayed, G. Aleppo, V.R. Aroda, R.R. Bannuru, F.M. Brown, D. Bruemmer, B. S. Collins, S.R. Das, M.E. Hilliard, D. Isaacs, E.L. Johnson, 10. Cardiovascular disease and risk management: standards of care in diabetes—2023, Diabetes Care 46 (Supplement_1) (2023) S158–S190.

[4] A.M. Elshewey, A.M. Osman, Orthopedic disease classification based on breadth-first search algorithm, Sci. Rep. 14 (1) (2024) 23368.

[5] E.S. Elkenawy, A.A. Alhussan, D.S. Khafaga, Z. Tarek, A.M. Elshewey, Greylag goose optimization and multilayer perceptron for enhancing lung cancer classification, Sci. Rep. 14 (1) (2024) 23784.

[6] A.M. Elshewey, A.A. Alhussan, D.S. Khafaga, E.S. Elkenawy, Z. Tarek, EEG-based optimization of eye state classification using modified-BER metaheuristic algorithm, Sci. Rep. 14 (1) (2024) 24489.

[7] D. Waigi, D.S. Choudhary, D.P. Fulzele, D. Mishra, Predicting the risk of heart disease using advanced machine learning approach, Eur. J. Mol. Clin. Med. 7 (7) (2020) 1638–1645.

[8] L.K. Singh, M. Khanna, H. Monga, G. Pandey, Nature-inspired algorithms-based optimal features selection strategy for COVID-19 detection using medical images, N. Gener. Comput. 42 (2024) 761–824.

[9] L.K. Singh, K. Shrivastava, An enhanced and efficient approach for feature selection for chronic human disease prediction: A breast cancer study, Heliyon. 10 (5) (2024) e26799.

[10] L.K. Singh, M. Khanna, H. Garg, R. Singh, Efficient feature selection based novel clinical decision support system for glaucoma prediction from retinal fundus images, Med. Eng. Phys. 123 (2024) 104077.

[11] L.K. Singh, M. Khanna, A novel enhanced hybrid clinical decision support system for accurate breast cancer prediction, Measurement 221 (2023) 113525.

[12] L.K. Singh, M. Khanna, H. Garg, R. Singh, M. Iqbal, A three-stage novel framework for efficient and automatic glaucoma classification from retinal fundus images, Multimed. Tools Appl. 83 (2024) 85421–85481.

[13] D. Khafaga, Meta-heuristics for feature selection and classification in diagnostic breast cancer, Comput., Mater. Continua. 73 (1) (2022) 749–765.

[14] Y.C. Ho, D.L. Pepyne, Simple explanation of the no-free-lunch theorem and its implications, J. Optim. Theory Appl. 115 (2002) 549–570.

[15] M. Abdel-Basset, L. Abdel-Fatah, A.K. Sangaiah, Metaheuristic algorithms: A comprehensive review, in: Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications, 2018, pp. 185–231.

[16] D. Albashish, A.I. Hammouri, M. Braik, J. Atwan, S. Sahran, Binary biogeography-based optimization based SVM-RFE for feature selection, Appl. Soft Comput. 101 (2021) 107026.

[17] R.A. Khurma, D. Albashish, M. Braik, A. Alzaqebah, A. Qasem, O. Adwan, An augmented Snake Optimizer for diseases and COVID-19 diagnosis, Biomed. Signal Process. Control 84 (2023) 104718.

[18] M.S. Braik, A.I. Hammouri, M.A. Awadallah, M.A. Al-Betar, O.A. Alzubi, Improved versions of snake optimizer for feature selection in medical diagnosis: a real case COVID-19, Soft. Comput. 27 (23) (2023) 17833–17865.

[19] S. Mohan, C. Thirumalai, G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques, IEEE Access 7 (2019) 81542–81554.

[20] K. Saikumar, V. Rajesh, G. Srivastava, J.C. Lin, Heart disease detection based on internet of things data using linear quadratic discriminant analysis and a deep graph convolutional neural network, Front. Comput. Neurosci. 16 (2022) 964686.

[21] U. Nagavelli, D. Samanta, P. Chakraborty, Machine learning technology-based heart disease detection models, J. Healthcare Eng. 2022 (1) (2022) 7351061.

[22] R.C. Das, M.C. Das, M.A. Hossain, M.A. Rahman, M.H. Hossen, R. Hasan, Heart disease detection using ml. 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 0983-0987). IEEE, 2023.

[23] A. Alqahtani, S. Alsubai, M. Sha, L. Vilcekova, T. Javed, Cardiovascular disease detection using ensemble learning, Comput. Intell. Neurosci. 2022 (1) (2022) 5267498.

[24] S. Mahmoud, M. Gaber, G. Farouk, A. Keshk, Heart disease prediction using modified version of LeNet-5 model, Int. J. Intell. Syst. Appl. 14 (6) (2022) 1–2.

[25] Y.A. Nanehkaran, Z. Licai, J. Chen, A.A. Jamel, Z. Shengnan, Y.D. Navaei, M. A. Aghbolagh, Anomaly Detection in Heart Disease Using a Density-Based Unsupervised Approach, Wirel. Commun. Mob. Comput. 2022 (1) (2022) 6913043.

[26] T. Sadad, S.A. Bukhari, A. Munir, A. Ghani, A.M. El-Sherbeeny, H.T. Rauf, Detection of cardiovascular disease based on PPG signals using machine learning with cloud computing, Comput. Intell. Neurosci. 2022 (1) (2022) 1672677.

[27] M.B. Abubaker, B. Babayiğit, Detection of cardiovascular diseases in ECG images using machine learning and deep learning methods, IEEE Trans. Artif. Intell. 4 (2) (2022) 373–382.

[28] M.R. Hassan, S. Huda, M.M. Hassan, J. Abawajy, A. Alsanad, G. Fortino, Early detection of cardiovascular autonomic neuropathy: A multi-class classification model based on feature selection and deep learning feature fusion, Inf. Fusion 77 (2022) 70–80.

[29] N.A. Baghdadi, S.M. Farghaly Abdelaliem, A. Malki, I. Gad, A. Ewis, E. Atlam, Advanced machine learning techniques for cardiovascular disease early detection and diagnosis, Journal of Big Data. 10 (1) (2023) 144.

[30] F.A. Hashim, A.G. Hussien, Snake Optimizer: A novel meta-heuristic optimization algorithm, Knowl.-Based Syst. 242 (2022) 108320.

[31] https://archive.ics.uci.edu/dataset/45/heart+disease (accessed on 15 september 2023).

[32] W. Li, W. Mo, X. Zhang, J.J. Squiers, Y. Lu, E.W. Sellke, W. Fan, J.M. DiMaio, J. E. Thatcher, Outlier detection and removal improves accuracy of machine learning approach to multispectral burn diagnostic imaging, J. Biomed. Opt. 20 (12) (2015) 121305.

[33] A.J. Mohammed, M. Muhammed Hassan, K.D. Hussein, Improving classification performance for a novel imbalanced medical dataset using SMOTE method, Int. J. Adv. Trends Comput. Sci. Eng.. 9 (3) (2020) 3161–3172.

[34] E.H. Alkhammash, M. Hadjouni, A.M. Elshewey, A hybrid ensemble stacking model for gender voice recognition approach, Electronics 11 (11) (2022) 1750.

[35] S.A. Alzakari, A.A. Alhussan, A.S. Qenawy, A.M. Elshewey, Early detection of Potato Disease using an enhanced convolutional neural network-long short-term memory Deep Learning Model, Potato Res. 1–9 (2024).

[36] E.H. Alkhammash, S.A. Assiri, D.M. Nemenqani, R.M. Althaqafi, M. Hadjouni, F. Saeed, A.M. Elshewey, Application of Machine Learning to Predict COVID-19 Spread via an Optimized BPSO Model, Biomimetics. 8 (6) (2023) 457.

[37] Z. Tarek, A.M. Elshewey, S.M. Shohieb, A.M. Elhady, N.E. El-Attar, S. Elseuofi, M. Y. Shams, Soil erosion status prediction using a novel random forest model optimized by random search method, Sustainability. 15 (9) (2023) 7114.

[38] A.M. Elshewey, M.Y. Shams, N. El-Rashidy, A.M. Elhady, S.M. Shohieb, Z. Tarek, Bayesian optimization with support vector machine model for parkinson disease classification, Sensors 23 (4) (2023) 2085.

[39] F. Desai, D. Chowdhury, R. Kaur, M. Peeters, R.C. Arya, G.S. Wander, S.S. Gill, R. Buyya, HealthCloud: A system for monitoring health status of heart patients using machine learning and cloud computing, Internet Things 17 (2022) 100485.

[40] M.A. Khan, An IoT framework for heart disease prediction based on MDCNN classifier, IEEE Access 8 (2020) 34717–34727.

[41] X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang, Q. Wang, A hybrid classification system for heart disease diagnosis based on the RFRS method, Comput. Math. Methods Med. 2017 (1) (2017) 8272091.

[42] C. Chakraborty, A. Kishor, Real-time cloud-based patient-centric monitoring using computational health systems, IEEE Trans. Comput. Social Syst. 9 (6) (2022) 1613–1623.

[43] A.A. Nancy, D. Ravindran, P.D. Raj Vincent, K. Srinivasan, R.D. Gutierrez, Iot-cloud-based smart healthcare monitoring system for heart disease prediction via deep learning, Electronics 11 (15) (2022) 2292.
[44] S. Nashif, M.R. Raihan, M.R. Islam, M.H. Imam, Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system, World J. Eng. Technol. 6 (4) (2018) 854–873.
[45] J.N. Rao, D.R. Prasad, An Ensemble Deep Dynamic Algorithm (EDDA) to predict the heart disease, Int. J. Sci. Res. Sci. Eng. Technol. 8 (2021) 105–111.
[46] M. Liu, Y. Kim, Classification of heart diseases based on ECG signals using long short-term memory. In2018 40th Annual international conference of the IEEE engineering in medicine and biology society (EMBC) 2018 (pp. 2707-2710). IEEE.
[47] M.Y. Shams, E.S. El-Kenawy, A. Ibrahim, A.M. Elshewey, A hybrid dipper throated optimization algorithm and particle swarm optimization (DTPSO) model for hepatocellular carcinoma (HCC) prediction, Biomed. Signal Process. Control 85 (2023) 104908.
[48] A.M. Elshewey, S.M. Tawfeek, A.A. Alhussan, M. Radwan, A.H. Abed, Optimized Deep Learning for Potato Blight Detection Using the Waterwheel Plant Algorithm and Sine Cosine Algorithm, Potato Res. 1–25 (2024).
[49] A.M. Elshewey, M.Y. Shams, S.M. Tawfeek, A.H. Alharbi, A. Ibrahim, A. A. Abdelhamid, M.M. Eid, N. Khodadadi, L. Abualigah, D.S. Khafaga, Z. Tarek, Optimizing HCV Disease Prediction in Egypt: The hyOPTGB Framework, Diagnostics. 13 (22) (2023) 3439.