JOURNAL OF BASIC SCIENCE AND ENGINEERING

HEART DISEASE PREDICTION MODEL - OPPOSITION BASED LEARNING AND WHALE OPTIMIZATION ALGORITHM WITH RNN CLASSIFIER

[1]Dr. S. Sivasubramaniam, [2]Dr. S. P. Balamurugan

[1]Assistant Professor/Programmer, Department of Computer and Information Science,
Annamalai University.

[2]Assistant Professor / Programmer, Department of Computer and Information Science,
Annamalai University.

Email: aucissiva@gmail.com, spbcdm@gmail.com

Corresponding Author: Dr. S. P. Balamurugan, Email: spbcdm@gmail.com

ABSTRACT

Cardiovascular diseases (CVDs) remain a leading cause of mortality worldwide, emphasizing the importance of accurate prediction and early detection. This study aims to leverage cutting-edge machine learning algorithms to develop a robust predictive model for heart disease using a wide-ranging dataset obtained from the public UCI heart-disease dataset comprising 919 patients' data and 14 attributes. The dataset encompasses diverse patient characteristics, including demographic information, clinical attributes, and diagnostic tests, facilitating a holistic approach to prediction. Subsequently, comprehensive feature selection techniques were applied to refine the datasets, capturing nuanced relationships and enhancing the predictive capability. Feature Selection can improve the performance of machine learning models by creating relevant and informative features from raw data. By selecting features, Machine Learning models can make more accurate predictions, handle complex and distributed data, reduce overfitting, and extract valuable insights from categorical and numerical data. This proposed model focuses on the design of automated heart disease diagnosis model using Optimum Recurrent Neural Network (ORNN). The proposed model involves a feature selection approach using Whale Optimization algorithm and Opposition Based Learning scheme (OB-WOA) for identifying the optimal features. In addition, several classification models such as Support Vector Machine (SVM), Naïve Bayes (NB), k-Nearest Neighbours (k-NN), Decision Tree (DT), Extreme Gradient Boost (XG-Boost), and Recurrent Neural Networks (RNN). The Performance evaluation metrics including accuracy, precision, recall, and F1-score were employed to assess the models' predictive capabilities on UCI heart-disease dataset and the experimental results show that the proposed heart disease detection model OBL-WOA with RNN (ORNN) achieves the best accuracy for predicting the heart disease.

Keywords: Cardiovascular diseases, Feature Selection, Opposition Based Learning, Whale Optimization, and Recurrent Neural Networks.

1. Introduction

Heart disease (HD) reigns as the leading cause of death worldwide, surpassing even the most concerning transmissible diseases. Unlike infections like COVID-19 or swine flu, which spread rapidly, HD often develops gradually. While some infectious diseases cause significant mortality, HD poses a more substantial threat, ranking alongside other non-communicable diseases like diabetes, liver cancer, and breast cancer. Every year, a staggering 17.9 million people lose their lives to HD [1].

Cardiovascular disease (CVD) is an umbrella term that encompasses various conditions affecting the heart and blood vessels. Heart disease (HD) is a specific type of CVD that focuses on problems with the heart itself, such as coronary heart disease (CHD) and heart failure.Cardiovascular disease (CVD) is a broader term encompassing a variety of heart conditions, including heart attack and heart failure. Heart attacks are particularly critical events, often caused by blockages that prevent blood flow to the heart. These blockages typically arise from fatty deposits accumulating on the inner walls of blood vessels. CVD itself encompasses a range of heart and blood vessel ailments. This includes coronary heart disease (CHD), which affects the blood vessels supplying the heart muscle; cerebrovascular disease, targeting blood vessels supplying the brain; peripheral arterial disease, impacting blood vessels in the arms and legs; rheumatic heart disease, where heart muscle and valves are damaged by rheumatic fever caused by streptococcal bacteria; congenital heart disease, birth defects affecting the heart's structure and function; and deep vein thrombosis and pulmonary embolism, involving blood clots that travel from leg veins to the heart and lungs [2].

CVD also includes issues with blood vessels throughout the body, like cerebrovascular disease affecting the brain, peripheral arterial disease impacting blood flow in the arms and legs, and deep vein thrombosis and pulmonary embolism involving blood clots.In simpler terms, CVD is the broader category, and HD is a specific type of CVD that focuses on the heart.

Machine learning (ML) is rapidly transforming various sectors, including healthcare. Its applications in healthcare are vast, offering solutions to a multitude of challenges. ML serves as a powerful tool for medical analysis and intelligent decision-making processes [3].The medical field's technological advancements, coupled with the increasing availability of powerful computing resources and open-source datasets, have significantly boosted the use of ML. This has led to the development of techniques for predicting various diseases, such as skin cancer, brain tumors, heart disease (HD), and breast cancer [4].The electronic healthcare sector is a rich source of data. Patient information, including blood tests, medical history, electronic device readings, and image data, is continuously generated [5]. This vast amount of data provides valuable fuel for ML algorithms, enabling them to identify patterns and trends that might otherwise be missed.

Heart disease (HD) manifests differently in each patient, making it highly individualistic. Early detection is crucial, and ML-based prediction systems offer significant advantages. In some countries, HD prevalence has demonstrably increased, highlighting the need for such proactive approaches. These ML systems assist doctors in making faster and more accurate predictions of a patient's risk for HD. The approach involves training an ML model on a large dataset of patient information. Once trained, the model can analyze individual patient data and identify patterns or trends that may indicate the presence of HD. This empowers doctors with valuable insights to guide their decisions and potentially lead to earlier interventions and improved patient outcomes [6].

Structure of the Paper:

- **Section 2: Related Work:** This section reviews existing research on heart disease (HD) prediction systems.
- **Section 3: Proposed Technique:** This section introduces our proposed HD detection system, which utilizes the Cleveland dataset for prediction.
- **Section 4: Optimization and Feature Selection:** This section details how we combine opposition-based learning and the whale optimizer to select optimal features during the training process of various classifiers.
- **Section 5: Experimental Results:** This section presents the outcomes of our experiments.
- **Section 6: Conclusion:** This section summarizes our findings and discusses the potential future directions for this research.

2. Review of Literature:

In a study by Balamurugan et al. [7], an Artificial Neural Network (ANN) technique was investigated for classifying kidney disease. The researchers proposed a novel approach that leverages an optimization algorithm called Oppositional Based Grasshopper Optimization Algorithm (OBL-GOA) to select the most relevant features from the data. At this point, optimum features were selected utilizing OBL-GOA approach and chosen features were fed as to ANN technique for disease classification. The model has simply explained the role of OBL approach in increasing the efficiency of the Grasshopper Optimization algorithm in the process of optimal features selection. The results shown that the proposed OBL-GOA based ANN model achieved better accuracy than the other models.

Kumar et al. [8] proposed a machine learning model called 'An enhanced grey wolf optimization (GWO)-based learning of ANN to medicinal data classifications.' This model tackles the diagnosis of various diseases, including heart disease, breast cancer, hepatitis, and Parkinson's disease.The key innovation lies in an improved metaheuristic approach named Improved Metaheuristic Grey Wolf Optimization (IMGWO). This technique aims to refine the classification performance of an Artificial Neural Network (ANN). The researchers compared the performance of the IMGWO-ANN classifier with a standard ANN classifier using three other well-established metaheuristic techniques: Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and the original GWO algorithm. The study employed a heart disease dataset containing 303 data points, 13 attributes (features), and one class label (presence or absence of heart disease). Their findings suggest that the IMGWO-ANN achieved a superior accuracy rate compared to the models utilizing the standard GWO, GA, and PSO algorithms for optimization.

Atimbire, S.A., Appati, J.K. &Owusu, E. [11] proposed a machine learning framework for feature selection in heart disease classification. Their approach leverages awhale

Optimization algorithm (WOA) algorithm in conjunction with a Support Vector Machine (SVM).The study addresses a common challenge in WOA, which is selecting the optimal features to guide the update of bestwhales' bubble-net feeding technique during the optimization process. The authors propose a strategy that incorporates a population diversity function and a tuning function to effectively identify the most relevant features. The researchers applied the optimized feature set obtained using WOA to various classification algorithms, including SVM, Ada-Boost, etc. Their results indicated that the combination of WOA with Ada-Boost achieved the best classification accuracy for heart disease prediction.

Alwateer, M.; Almars, A.M.; Areed, K.N.; Elhosseini, M.A.; Haikal, A.Y.; Badawy, M. Ambient[12] investigated various prominent machine learning techniques in the context of building interpretable and understandable models for heart disease analysis. The study focused on achieving a balance between model accuracy and transparency.Alwateer, M.explored different approaches, including Naïve Bias with WOA attribute wrapping. This combination achieved the highest accuracy compared to other models evaluated in the study.

Lamiaa M. El Bakrawy [13] proposed a heart disease classification model that combines Grey Wolf Optimization (GWO) with a Naive Bayes (NB) classifier. The study aimed to improve the accuracy of the NB classifier through a technique called Class-Attribute Interdependence Maximization (CAIM). The approach utilizes CAIM to identify the most informative features within the dataset that are highly relevant for predicting heart disease. Grey Wolf Optimization (GWO) is then employed to automatically determine the optimal weights for these selected features within the NB classifier. Assigning appropriate weights to the features helps the NB classifier prioritize the most impactful factors during the classification process, ultimately leading to improved accuracy. The model's performance was evaluated using the Cleveland Heart Disease Database obtained from the UCI Machine Learning Repository. A 5-fold cross-validation (CV) technique was employed to ensure the robustness of the findings.

In [14] explores the potential of ensemble classification methods to improve the accuracy of heart disease risk prediction. Ensemble methods combine multiple classification algorithms, aiming to achieve better performance than any single algorithm alone. The core concept lies in leveraging the strengths of various algorithms to create a more robust model. This approach can be particularly beneficial for enhancing the accuracy of algorithms that might perform poorly on their own. The authors investigated the effectiveness of different ensemble techniques, including bagging, boosting, stacking, and majority voting, using the Cleveland Heart Disease dataset. These techniques were evaluated based on their ability to improve the prediction accuracy for heart disease detection.

Reddy et al. [15] proposed a hybrid classification model for heart disease analysis that combines a Genetic Algorithm (GA) with a Fuzzy Logic (FL) classifier. This approach leverages the strengths of both techniques to achieve accurate heart disease prediction. In this paper, different methods like Rough Set and Fuzzy rule-based classifier with adaptive GA were used as HD classification. The UCI HD datasets were used and initial feature decrease was performed by rough set model and then the hybrid the adaptive GA with FL classifier (AGAFL) was processed. The Adaptive GA gives the rules for optimization and employed to choosing the features in the dataset. The experimental study expresses that AGAFL achieved superior to other hybrid combination.

Researchers are dedicated to developing optimal techniques for HD analysis using various methods. The primary goal for all researchers is to achieve maximum classification accuracy. These techniques range from simple to complex features, and they can be categorized as supervised, unsupervised, semi-supervised FS, and advanced approaches.

3. Heart Disease Detection System (HDDS)

Heart disease prediction models rely on clinical data, also known as physician parameters. This data can include various sources, such as blood tests, patient history, interviews with patients or relatives, and treadmill test results.Before using this data for prediction, it often undergoes preprocessing. This might involve handling missing values in certain data points or normalizing the data to ensure consistency.A crucial step in the process is feature selection (FS). FS aims to identify the most relevant features from the entire dataset. Irrelevant features can potentially hinder the accuracy of classification algorithms used for prediction. By eliminating these features, the model focuses on the most impactful information for diagnosing heart disease.

3.1. Methodology used:

Including an excessive number of features in a heart disease prediction model can be counterproductive. Some features might not contribute meaningfully to the classification process, potentially hindering the model's performance. This is why feature selection (FS) is crucial.FS aims to identify the most relevant subset of features from the available data. By focusing on these key features, we can achieve accurate classification without unnecessary data. FS offers several benefits:

- **Improved Classifier Performance:** Selecting the most informative features helps classifiers learn more effectively, leading to better prediction accuracy.
- **Reduced Training Time:** With fewer features to analyze, training a model becomes faster and more efficient.
- **Noise Reduction:** FS removes irrelevant or redundant data, reducing the impact of noise on the prediction process.

This study utilizes the Whale

Optimization Algorithm (WOA) for optimal feature selection. WOA draws inspiration from the hunting behavior of humpback whales, allowing for a robust search across the entire feature space. It is further combined with the Opposition-Based Learning (OBL) method to enhance the search capability and avoid getting stuck in local optima (suboptimal solutions).

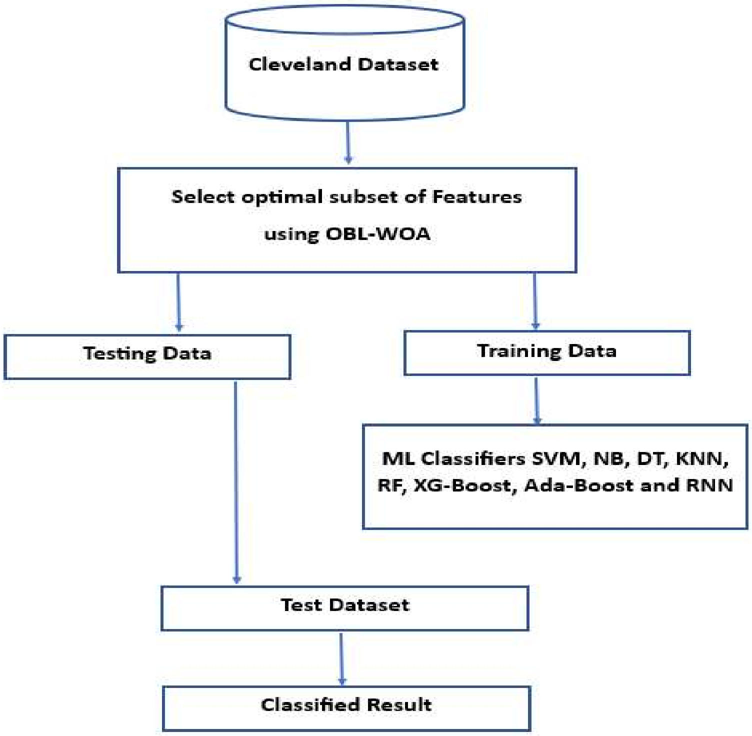

Cleveland Dataset
Select optimal subset of Features using OBL-WOA
Testing Data
Training Data
ML Classifiers SVM, NB, DT, KNN, RF, XG-Boost, Ada-Boost and RNN
Test Dataset
Classified Result

Fig-1: ProposedHeart Disease Detection System.

3.2. Heart Disease Dataset:

There are two main approaches for diagnosing heart disease: invasive and non-invasive. Coronary angiography is an example of an invasive procedure. While effective, it can be expensive, complex, and uncomfortable for patients.Fortunately, various non-invasive methods offer valuable tools for diagnosing heart disease before resorting to invasive procedures. These methods include:

- **Blood tests:**Analyzing blood composition can reveal signs of potential heart problems.
- **Patient history:** A detailed medical history can provide important clues about risk factors for HD.
- **Electrocardiogram (ECG):** This test measures the electrical activity of the heart, helping to identify abnormalities like arrhythmias.
- **Stress test (treadmill test):** This test evaluates how the heart responds to physical exertion, revealing potential blood flow issues.

Figure-2 illustrate these different diagnostic approaches used for HD detection.

This study aims to assess the effectiveness of the proposed algorithm in classifying heart disease. To achieve this, we'll utilize a dataset containing real-world patient information relevant to non-invasive diagnostic methods.Several reliable sources like Kaggle and UCI Machine Learning Repository are popular among researchers for obtaining data. In this case, we'll be using the Cleveland Heart Disease dataset from the UCI Machine Learning Repository. This dataset, originally collected by the Cleveland Clinic Foundation [16], comprises 303 instances (data points) and 76 features (characteristics) for each patient.

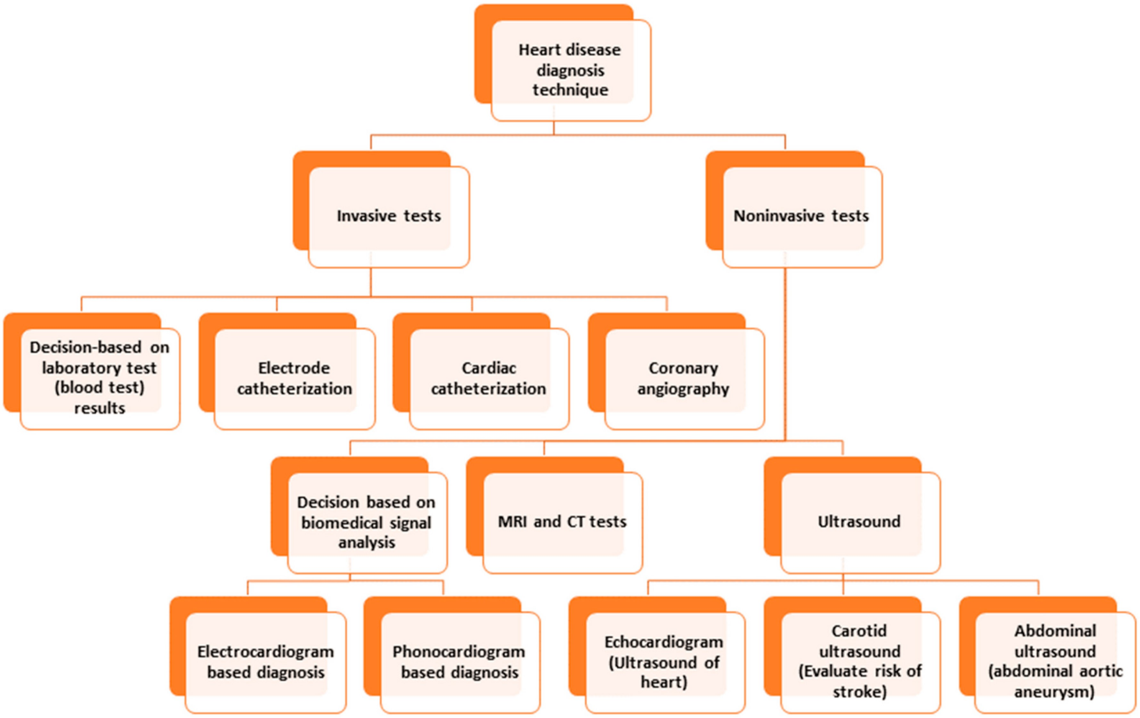

Heart disease diagnosis technique
Invasive tests
Noninvasive tests
Decision-based on laboratory test (blood test) results
Electrode catheterization
Cardiac catheterization
Coronary angiography
Decision based on biomedical signal analysis
MRI and CT tests
Ultrasound
Electrocardiogram based diagnosis
Phonocardiogram based diagnosis
Echocardiogram (Ultrasound of heart)
Carotid ultrasound (Evaluate risk of stroke)
Abdominal ultrasound (abdominal aortic aneurysm)

Fig-2: Different diagnosis methods for heart disease.

This study utilizes the Cleveland heart disease dataset, a well-known resource for evaluating machine learning algorithms in heart disease prediction. The dataset consists of 303 samples, categorized as either "normal" (139 instances) or "sick" (164 instances). It includes 14 key attributes:

- **Eight categorical features:** These features represent qualities that fall into distinct categories, such as gender (male or female) or chest pain type (typical angina, atypical angina, etc.).
- **Six numeric features:** These features represent measurable values, such as age, blood pressure, or cholesterol levels.

One of these 14 features serves as the class label, indicating the presence or absence of heart disease for each patient.

Table-1. List of features and its description in the Cleveland HD dataset

UCI ML repository's Cleveland heart disease dataset—feature subset [24].

Attribute name	Attribute description
Age	Age in years
Sex	1 denotes male and 0 denotes female
CP	Chest pain type 1, typical angina; type 2, atypical angina; type 3, nonanginal pain; and type 4, asymptomatic
trestbps	Resting blood pressure (in mmHg at entry to the health center)
chol	Serum lipid level in mg/dL
fbs	1 denotes true, i.e., the fasting blood sugar level > 120 mg/dL; 0 denotes false
restecg	Resting ECG results: null, normal; 1, ST-T wave abnormality; and 2, probable or definite left ventricular hypertrophy
thalach	Maximum heart rate achieved
exang	Exercise induced angina (1 = yes; null = no)
oldpeak	ST depression induced by exercise relative to rest
slope	The slope of the peak exercise ST segment (1, 2, and 3): 1, upsloping; 2, flat; and 3, downsloping
ca	Number of major vessels (0-3) colored by fluoroscopy
thal	Thalassemia: 3 = normal, 6 = fixed defect, and 7 = reversible defect
Num	Response: diagnosis of HD - Class (0= healthy, 1= have heart disease

3.3 Feature Selection Methods

Feature selection (FS) is a crucial step in data mining and machine learning. It helps us identify the most relevant features from a dataset, allowing classification algorithms to focus their attention on the information that matters most for predicting the target variable (class).

There are two main types of FS methods: supervised and unsupervised. The key distinction lies in how they utilize the target variable: [18]

- **Supervised FS:** These methods explicitly consider the target variable (e.g., presence or absence of heart disease) during the selection process. This allows them to identify features that are most directly related to predicting the target variable, potentially improving model accuracy. Supervised methods aim to eliminate irrelevant features that don't contribute to prediction.

- **Unsupervised FS:** These methods don't rely on the target variable. Instead, they focus on identifying inherent patterns or relationships within the data itself. This can be helpful for reducing redundancy and eliminating features that are highly correlated with each other. Unsupervised methods primarily focus on removing redundant features that don't provide unique information.

Supervised feature selection offers two main approaches: filter methods and wrapper methods. Both aim to identify the most relevant features for prediction, but they differ in their strategies:**Filter methods:** These methods rely on statistical measures to assess the relationship between individual features and the target variable (e.g., presence of heart disease). Features with strong correlations to the target variable are considered more informative and are chosen for the prediction model. Filter methods are generally fast and computationally efficient because they don't involve building and evaluating multiple prediction models. However, they might not always capture complex relationships between features.**Wrapper methods:** These methods take a more iterative approach. They evaluate different combinations of features by building and comparing multiple prediction models. The subset of features that leads to the best performance on a chosen metric (e.g., accuracy) is selected. Wrapper methods can be more flexible and potentially identify more complex feature interactions. However, they can be computationally expensive, especially for datasets with many features [19].

3.3.1. Optimization:

Supervised feature selection (FS) aims to identify the most relevant features from a dataset for building accurate prediction models. This process can be framed as an optimization problem, where the goal is to find the optimal subset of features that maximizes a specific performance metric (e.g., classification accuracy).Optimization algorithms play a crucial role in solving this problem. These algorithms work by iteratively evaluating different feature combinations and selecting the one that leads to the best performance based on the chosen metric. Here's a breakdown of two main optimization categories:**Deterministic optimization:** These algorithms follow a fixed set of rules and always converge to the same solution for a given problem with the same starting point. They are reliable but might get stuck in local optima (suboptimal solutions) depending on the initial feature selection.**Stochastic optimization:** These algorithms incorporate randomness into their search process. This allows them to explore the feature space more broadly and potentially escape local optima. However, they might not always converge to the same solution on different runs, and may require more computation time.Choosing the right optimization algorithm depends on the specific problem characteristics, such as the size and complexity of the dataset, and the desired balance between accuracy and computational efficiency [20].

A heuristic is a guiding principle that helps the algorithm make decisions during its search. In the context of feature selection, the heuristic function might evaluate the performance of different feature combinations, directing the algorithm towards combinations that seem more promising based on the chosen metric (e.g., accuracy). Here, the results may not be guaranteed to be the absolute best every time, but they aim to be consistently good approximations. Stochastic optimization algorithms introduce randomness into their search process. Unlike deterministic algorithms that follow a rigid set of rules, stochastic algorithms utilize random elements to explore the space of possible feature combinations more broadly. This offers several advantages:**Reduced Local Optima:** Local optima are suboptimal solutions that an algorithm might get stuck on. Stochastic methods, due to their randomness, have a higher chance of escaping these traps and finding a better solution (potentially the global optimum, which is the best solution across the entire search space).**Exploring Complex Feature Spaces:** For datasets with complex relationships between features, stochastic optimization can be more effective in navigating this complexity and identifying optimal feature combinations [21].

3.3.2 Metaheuristics Algorithm

Stochastic algorithms draw inspiration from biological or natural behavior and are often referred to as "metaheuristics". These algorithms play a crucial role in solving optimization problems that arise in various aspects of day-to-day life, such as medical, agriculture, and engineering. In situations,where finding an optimal solution for a given model is incredibly difficult or impractical, heuristics can be applied to enhance the process and achieve the best possible solution for predicting classification models. The term "metaheuristic" is derived from the combination of "meta" and "heuristic". "Meta" signifies a higher-level methodology, while "heuristic" refers to the art of discovering new strategies to solve problems. Metaheuristic algorithms can be categorized into two types: population-based (random search) and single-solution-based (local search). Population-based metaheuristics execute optimization by considering a collection of solutions. The search algorithm starts at a random initial position from the available solutions, and the population improves their positions at each iteration [22].

Furthermore, every metaheuristic algorithm leverages a combination of local search and randomization. Nevertheless, each stochastic algorithm incorporates arbitration and local search as part of its metaheuristic approach. Randomization serves as a powerful tool to eliminate the limitations of local search and enable global-scale exploration. Consequently, the vast majority of metaheuristic methods are well-suited for global optimization.

3.4Proposed OB-WOA model

In the proposed OB-WOA optimization algorithm, the solutions are represented as positions of whales in the search space. The algorithm iteratively updates the positions of the whales based on their current positions and the positions of the best whale found so far. This update process

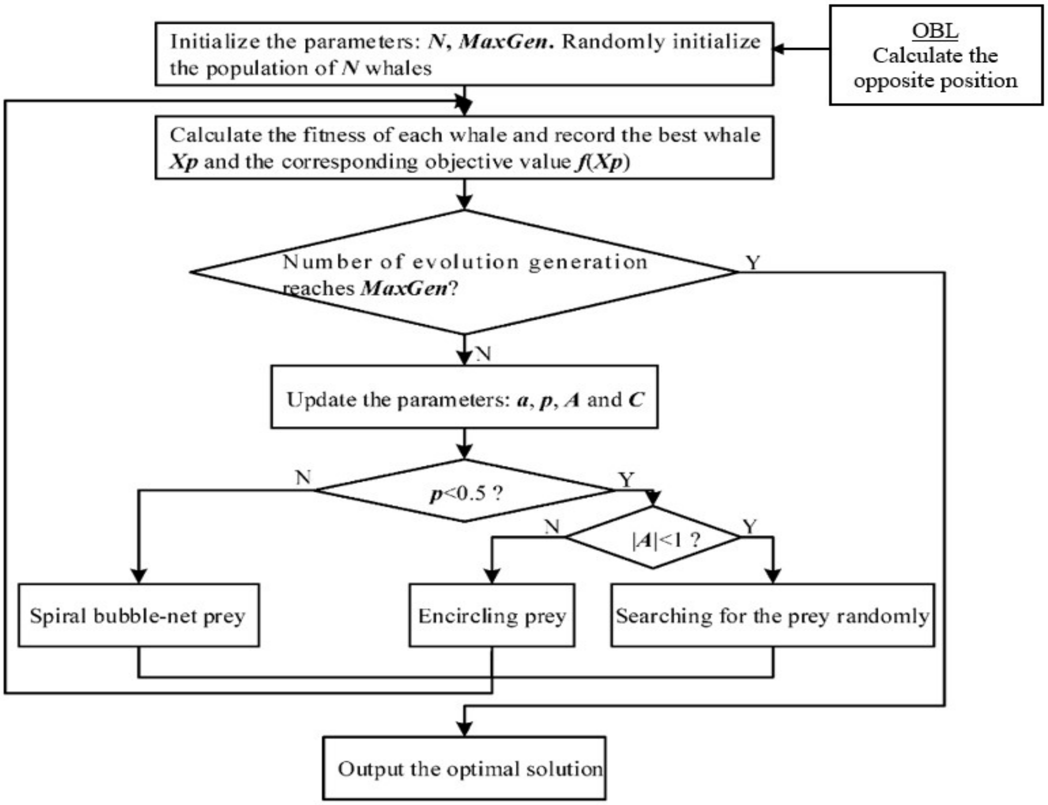

Initialize the parameters: N, MaxGen. Randomly initialize the population of N whales
OBL
Calculate the opposite position
Calculate the fitness of each whale and record the best whale Xp and the corresponding objective value f(Xp)
Number of evolution generation reaches MaxGen?
Y
N
Update the parameters: a, p, A and C
N
p<0.5 ?
Y
N
|A|<1 ?
Y
Spiral bubble-net prey
Encircling prey
Searching for the prey randomly
Output the optimal solution

Figure-4 shows the Proposed OBL-WOA HD detection model.

involves exploration to search for new solutions and exploitation to refine promising solutions.WOA has been applied to various optimization problems, including feature selection in machine learningin optimization algorithms. Its simplicity and effectiveness make it a popular choice for solving optimization problems in heart disease diagnosis domain.

3.4.1 Opposition Based Learning

The prime opposition theory was initially stated in the Yin-Yang symbol in the ancient Chinese philosophy [23].In 2005, the basic notion of OBL was developed by Hamid R. Tizhoosh[24], which consider the estimate or search or guess, and its corresponding opposite concurrently to detect the solution effectively. The primary aim of the algorithms is to find the optimum solution for the given function, consider its search or guess and its opposition simultaneously helps to enhance the accuracy[25].

The computation opposition theory was stimulatedby the opposition notion in the real-time and the opposite number wasdetermined in the following: [26]When we search for Z, and we agree that searching in opposite direction can be useful, calculate the opposite number $\vec{Z}$ is the initialphase.

Definition – 1: where Z represent a real number determined on a specific range: Z∈ [a, b]. The opposite number$\vec{Z}$can be determined by:

$$\vec{Z} = a + b - Z \qquad (1)$$

Definition – 2: (Opposite point in the D space): Where$Z$ (Z_1, ...,Z_D) represent a point in Ddimension space and $\vec{\boldsymbol{Z}}_{\mathrm{i}}$∈ [$a_i$, b_i], then i= 1, 2, ..., D. The opposite of Z is determined as$\vec{Z}_{\mathrm{i}}$(Z_1, ..., Z_D) as follow:

$$\vec{Z}_{\mathrm{i}} = a_{\mathrm{i}} + b_{\mathrm{i}} - Z_{\mathrm{i}} \qquad (2)$$

3.4.2 Whale Optimization Algorithm (WOA)

The Whale Optimization Algorithm (WOA) is a metaheuristic optimization algorithm inspired by the social behavior of humpback whales. It was proposed by SeyedaliMirjalili et al. in 2016. WOA is designed to solve optimization problems by simulating the hunting behavior of whales, where whales collaborate to locate and encircle prey.

The WOA is a recent advancement in optimization techniques, inspired by the hunting strategies of humpback whales. Unlike other hunting-based algorithms, WOA offers a unique approach. It can utilize either a randomly chosen agent or the best-performing agent within the search space to track down the optimal solution. This flexibility is further enhanced by its ability to mimic the whales' bubble-net feeding technique through the use of spirals.The core of the WOA algorithm is built on three mathematical operators that simulate different stages of a humpback whale's hunt:

1. **Search for Prey (Exploration Phase):** This initial phase allows the algorithm to broadly explore the search space for potential solutions.
2. **Encircling Prey:** Once promising areas are identified, the algorithm focuses its search by encircling the most favorable solutions.
3. **Bubble-Net Foraging (Exploitation Phase):** This final stage resembles the whales' bubble-net technique, where the algorithm refines its search around the best solutions found so far to achieve the optimal outcome.

By mimicking these natural behaviors, WOA effectively balances exploration and exploitation, leading to efficient problem-solving.The mathematical formulation is modelled and explained as follows:

1) **Prey encircling:** In this phase, the whale algorithm starts with an initial best search agent. It assumes that the current solutions are the best and it is the location of the target or almost close

to it. The rest of the agents subsequently update their locations toward the best search agent. This can be stated as the following:

$$\overrightarrow{D} = |\overrightarrow{C}.\overrightarrow{X^*}(t) - \overrightarrow{X}(t)|, \qquad (1)$$

$$\overrightarrow{X}(t+1) = \overrightarrow{X^*}(t) - \overrightarrow{A}.\overrightarrow{D}, \qquad (2)$$

where t specifies the current iteration $\vec{A}$ and $\vec{C}$ are coefficient vectors. The $\overrightarrow{X_*}$ is the location vector of the best solution attained so far and $\vec{X}$ is the location vector. In case of the presence of better solution, the $\overrightarrow{X_*}$ should be updated iteratively.

The vectors $\vec{A}$ and $\vec{C}$ are considered as follows:

$$\overrightarrow{A} = 2\overrightarrow{a}.\overrightarrow{r} - \overrightarrow{a}, \qquad (3)$$

$$\overrightarrow{C} = 2.\overrightarrow{r}, \qquad (4)$$

where $\vec{a}$ is linearly decreased from 2 to 0 over the number of iterations and r is random vector in [0, 1].

2) **Exploitation phase:** This phase is also called the bubblenet attacking and it works by two methods as follows:

- **Shrinking encircling mechanism:** in this step, the value of $\vec{a}$ in equation (3) is decreased and consequently the variation range of $\vec{A}$ is also decreased by $\vec{a}$. This implies that →−a is randomly placed in $[-\vec{a}\ ,\ \vec{a}\ ]$. where a is decreased from 2 to 0 over the optimization time.
The randomness of $\vec{A}$ in [-1, 1], the new location of the search agent can be determined anywhere in between the agent past location and the current best location. Fig. 1 shows the

possible locations from (X, Y) towards (X^*, Y^*) that can be accomplished by 0 A 1 in a 2-D space.

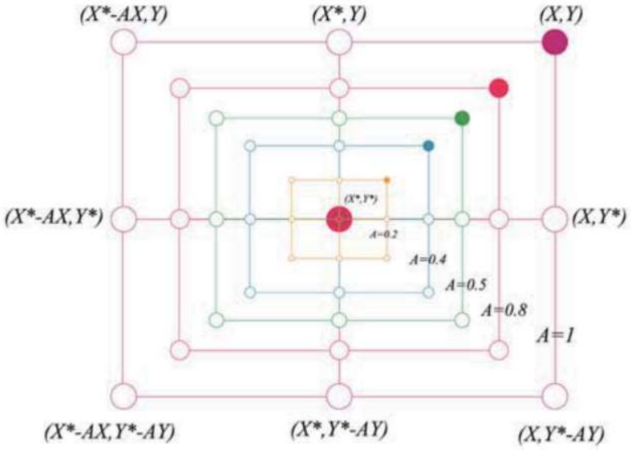


Fig. 1: WOA shrinking encircling mechanism

• **Spiral updating position:** in this step, as illustrated in Fig. 2 the distance between the positions of whale and its prey is calculated, and then an equation of spiral is created between whale and prey locations to simulate the movement of helix shape by humpback whales. This can be expressed as the following:

$$\overrightarrow{X}(t+1) = \overrightarrow{D'}.e^{bl}.\cos(2\Pi l) + \overrightarrow{X^*}(t), \qquad (5)$$

$$\overrightarrow{D'} = |\overrightarrow{X^*}(t) - \overrightarrow{X}(t)|. \qquad (6)$$

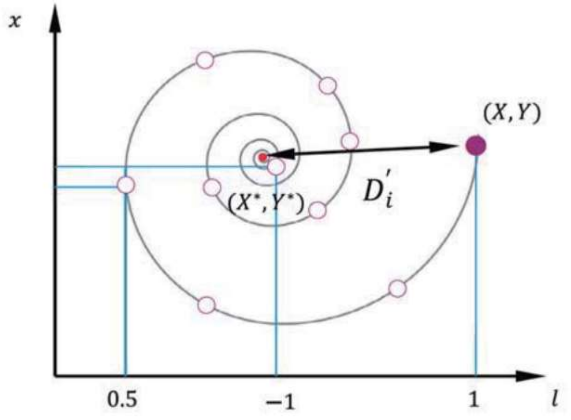


Fig. 2: WOA spiral updating position

Equation (6) shows the distance of the i-th whale to the prey (best solution attained so far), b is a constant for defining the logarithmic spiral shape and l is a random number in the region of [-1, 1]. The whale movement toward its prey is concurrently applying the shrinking circling in a spiral shaped path. Therefore, a 50% assumption of the possibility to switch between the two modes is applied to update the whale's next position as follows:

$$\overrightarrow{X}(t+1) = \begin{cases} \overrightarrow{X^*}(t) - \overrightarrow{A}.\overrightarrow{D} & \text{if } p < 0.5 \\ \overrightarrow{D'}.e^{bl}.\cos(2\Pi l) + \overrightarrow{X^*}(t) & \text{if } p \geq 0.5, \end{cases} \tag{7}$$

where p is a random number in [0, 1].

3) **Exploration phase:** In this phase, WOA achieves a global optimization. In Fig.3, whales search for its prey according to their locations to each other randomly. The $\vec{A}$ is set randomly from [-1,1] to accommodate the search agent to move away from the reference whale. This means that $\vec{A}$ has to be either greater than 1 or less than the -1. Moreover, the updated position of a search agent here is done by randomly chose an agent that allows the WOA to perform global search.

The modelling of this exploration machine is mathematically expressed as follows:

$$\overrightarrow{D} = |\overrightarrow{C}.\overrightarrow{X_{rand}} - \overrightarrow{X}|, \tag{8}$$

$$\overrightarrow{X}(t+1) = \overrightarrow{X_{rand}} - \overrightarrow{A}.\overrightarrow{D}, \tag{9}$$

where $\overrightarrow{Xrand}$ *is* a random location for random whale that chosen from the present population.

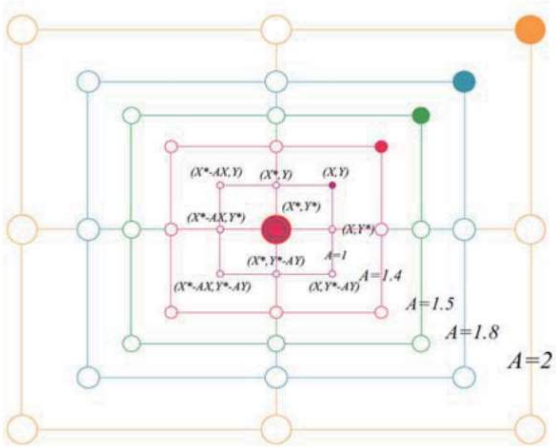


Fig. 3: WOA exploration mechanism

4. Classification Algorithms used for HD prediction

This study utilizes a novel method called OBL-WOA to identify the most relevant features within the Cleveland dataset. These selected features are then fed into various classification algorithms, including SVM, NB, KNN, DT, XGBoost, AdaBoost, and ANN, to predict the presence of heart disease (HD).

4.1. Support Vector Machine (SVM)

SVMs are one of the most popular machine learning methods used for classification and recognition tasks in supervised learning. The core concept behind SVMs lies in the assumption that there exists a non-linear relationship (often represented as a function, y = f(x)) between the input data (x) and the desired output (y), which can often be high-dimensional. In supervised learning, the only information available to the algorithm is the training data, which includes labelled examples. SVMs aim to find an optimal hyperplane in this higher-dimensional space that best separates the data points belonging to different classes. This hyperplane is constructed by maximizing the margin, which is the distance between the hyperplane and the closest data points from each class. These closest data points are called support vectors, and they play a crucial role in defining the decision boundary for future classifications [31].

4.2. Naïve Bayse (NB)

Naive Bayes is a popular classification algorithm that leverages Bayes' theorem to predict class labels. It operates under the assumption that the features (inputs) are independent of each other given the class label (output). This simplifying assumption allows for a straightforward and computationally efficient model creation process, making it particularly suitable for applications in the medical field, like heart disease (HD) diagnosis. Despite its simplicity, Naive Bayes can often achieve competitive performance compared to more complex classification models. In essence, Bayes' theorem allows the calculation of the posterior probability (P(c|x)), which represents the probability of a particular class (c) occurring given a specific set of features (x). This calculation is achieved by considering the prior probability (P(c)), the likelihood (P(x|c)), and the evidence (P(x)). The key aspect of Naive Bayes lies in the conditional independence assumption, which states that the influence of any individual feature (x) on the predicted class (c) is independent of the other features, given the class itself [32].

4.3. Decision tree (DT)

Decision Tree learning is a classification technique that utilizes a tree-like model for making predictions. This tree structure represents a series of questions based on the characteristics (features) of an item, ultimately leading to a prediction about its target value (presence or absence of heart disease in this case). The tree visually depicts the decision-making process, with the starting point being the root node and the final classifications represented by the terminal nodes (leaves) of the tree. These terminal nodes classify the data points based on the series of decisions (branches) traversed through the tree. This study investigates the effectiveness of the Decision Tree algorithm in predicting the presence of heart disease using the 14 features available in the dataset [33].

4.4. K- Nearest Neighbor (KNN)

KNN is a classification method that utilizes the concept of similarity to categorize unknown data points. It achieves this by calculating the distance between the unknown data point and the existing data points in the training set. The data point is then assigned the class label that is most frequent among its K nearest neighbors. The value of K, which represents the number of neighbors considered, plays a crucial role in the KNN algorithm's performance. Choosing an appropriate K value helps achieve a balance between accuracy and overfitting. Distance metrics are essential for KNN as they determine how similarity is measured between data points. Common distance metrics employed in KNN classification include Euclidean distance, Manhattan distance, Chebyshev distance, Canberra distance, Mahalanobis distance, and Sorensen distance. These metrics help assess the similarity between the unknown data point and the training data points during the classification process [34].

4.5. Random Forest (RF)

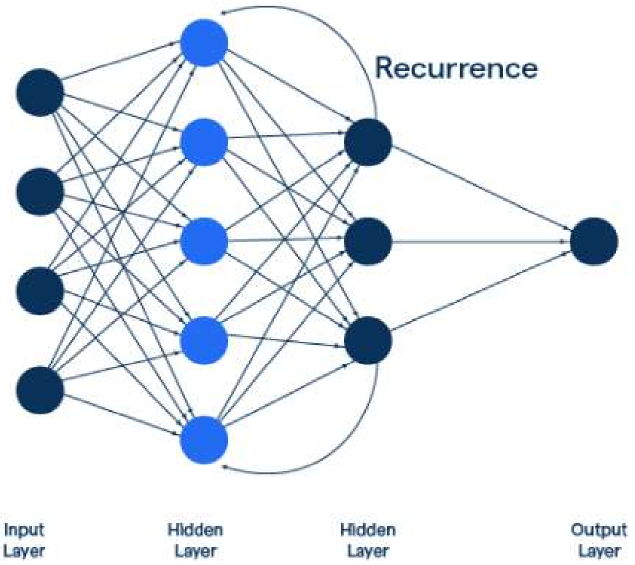
Random Forests is a supervised learning technique that leverages an ensemble of Decision Trees (DTs) to enhance prediction accuracy. These DTs are typically trained using a method called bagging. Bagging works by creating multiple training datasets from the original data by randomly sampling with replacement. Each DT in the Random Forest is then trained on a unique bagging sample. This approach introduces diversity into the ensemble, ultimately leading to improved overall performance compared to a single decision tree. While Random Forests share some hyperparameter considerations with individual DTs, they are fundamentally different models. A single DT cannot be directly integrated into a Random Forest classifier because Random Forests make predictions based on the combined output of all the trees in the ensemble[35].

4.6. Ada-Boost

AdaBoost, short for Adaptive Boosting, is a machine learning algorithm used in conjunction with other learning algorithms to improve their classification performance. It was originally developed for binary classification tasks, but the core concept can be extended to other scenarios. AdaBoost works by strategically combining multiple "weak learners" (relatively simple models) into a single "strong learner" (a more accurate model). These weak learners are trained sequentially, with each new learner focusing on the data points that the previous learners misclassified. This approach allows AdaBoost to iteratively improve the overall accuracy of the ensemble. In simpler terms, AdaBoost takes advantage of a voting system where less accurate models (weak learners) are given less weight in the final prediction. By focusing on previously misclassified data points and progressively upweighting the correct classifications, AdaBoost guides the learning process to achieve a more robust final model [36].

4.7. Recurrent Neural Network (RNN)

RNNs, or recurrent neural networks, belong to a category of neural networks that prove to be valuable in modelling sequence data. These networks, derived from feedforward networks, demonstrate a behavior akin to the functioning of the human brain. In essence, recurrent neural networks excel at generating predictive outcomes for sequential data, surpassing the capabilities of other algorithms. A Recurrent Neural Network (RNN) Classifier is a type of neural network architecture specifically designed to handle sequential data. Unlike traditional feedforward neural networks, RNNs have connections that form directed cycles, allowing them to retain information about previous inputs. This makes them particularly well-suited for tasks such as time series prediction, natural language processing, and speech recognition.



In the context of a classifier, an RNN takes a sequence of input data and processes it step by step, updating its internal state at each time step. This allows the network to capture temporal dependencies in the data and make predictions based on the entire sequence.

5. Results and discussions

This section presents the results obtained from the proposed heart disease (HD) classification model. The model was implemented in Python using Spyder 5.0 and evaluated on the publicly available Cleveland dataset in CSV format. The development environment consisted of a Windows machine equipped with an Intel Core i5 processor at 3.6 GHz and 8 GB of RAM.

5.1. Cross-validation

Cross-validation is a valuable technique used to assess the performance of machine learning models, particularly for classification tasks like heart disease prediction. It helps reduce the variability of the results and provides a more robust estimate of the model's generalizability. In cross-validation, the dataset is strategically divided into multiple folds (subsets). The model is then trained on a subset of the data (training fold) and evaluated on a different, unseen subset (testing fold). This process is repeated for all folds, ensuring that each data point is used for both training and testing. A common approach is to split the data into a ratio of 80% for training and 20% for testing. By iteratively training and testing on different portions of the data, cross-validation provides a more comprehensive understanding of the model's ability to perform well on unseen data. This helps avoid overfitting, which can occur when a model performs well on the training data but fails to generalize to new data.

5.2. Performance Evaluation Measures

The calculations of the statistical metrics on the results of classification models were defined.It can be expressed by:

(Number of correct calculations)/Number of all calculations).

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

(Number of true positive calculation)/(Number of all positive calculation)

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

(Number of true negative calculation)/(Number of all negative calculation)

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

F1-score as follows, and it deals with the mean value of precision and recall:

$$\text{F1} - \text{Score} = \frac{(2 * \text{precision} * \text{recall})}{(\text{precision} + \text{recall})}$$

Whereas TN denotes true negatives, TPsignifies true positives, FN represent false negatives, and FP means false positives.

5.3.Discussion

The basic idea of the proposed methodology is HD detection. The proposed model consists of 2models namelyclassification, and FS. For experimentation evaluation, the Cleveland datasetwas utilized. The performance wascalculated by thespecificity, F1-score, accuracy and sensitivity. Here, the prediction was made using different classification algorithms such as SVM, NB, DT, KNN and ANN.

The proposed OBL-WOA basedFS algorithm applied to select anoptimum subset of feature and increases the classification accuracy and the outcomes are shown in Table-2. It demonstrates that the proposed approach OBL-WOA with RNN attains the 96.78%of accuracy,i.e., 86.83% for employingWOA + RNN classifier, 82.33% for using all 14 features + RNN classifiers.

Table-2:Performance evaluation metrics for each classifier with optimization technique for Cleveland HD dataset

70:30 ratio (where 70% is training set and 30% is testing)												
Classifiers	**All 14features**				**WOA**				**OBL-WOA**			
	Accuracy	**Sensitivity**	**Specificity**	**F1-Score**	**Accuracy**	**Sensitivity**	**Specificity**	**F1-Score**	**Accuracy**	**Sensitivity**	**Specificity**	**F1-Score**
SVM	70%	67%	66%	64%	81%	82%	78%	81%	84%	83%	88%	86%
NB	82%	81%	83%	78%	80%	82%	82%	82%	82%	84%	87%	85%
KNN	64%	67%	72%	69%	79%	84%	91%	87%	82%	85%	82%	84%
DT	79%	79%	81%	80%	80%	81%	85%	83%	86%	83%	89%	86%
RF	82%	82%	85%	84%	81%	84%	82%	85%	84%	79%	82%	85%
XG-Boost	80%	81%	82%	78%	79%	79%	82%	85%	84%	85%	88%	81%
Ada-Boost	76%	79%	74%	76%	78%	83%	88%	86%	82%	75%	89%	81%
RNN	**82.33%**	**81%**	**86%**	**83%**	**86.83%**	**85%**	**87%**	**86%**	**96.78%**	**87%**	**96%**	**92%**

Figure-8shows the boxplot representation of the proposed OBL-WOA model. The selected optimal subset of featuresis experimented with various classification algorithms and the results displayed in boxplot diagram. The graph representation of accuracies achieved by the various classification algorithms withOBL-WOAare shown in figure-9.

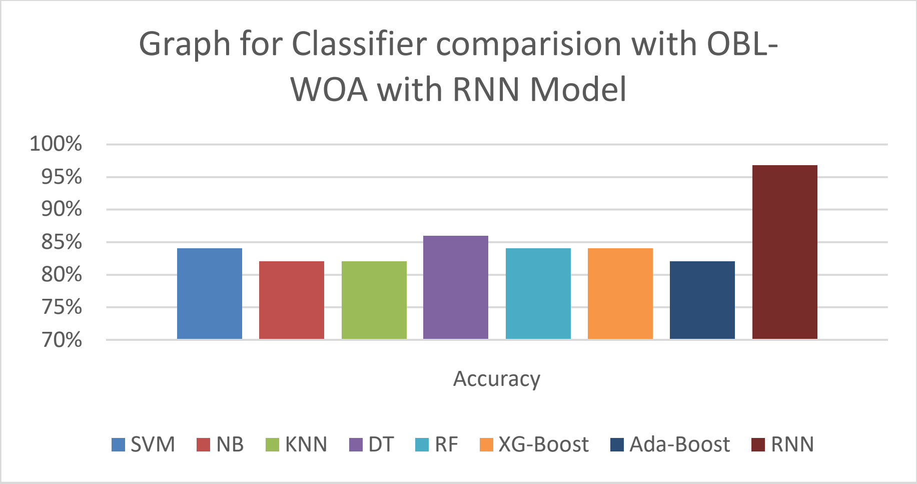


Fig-8: Boxplot for proposed OBL-WOA model

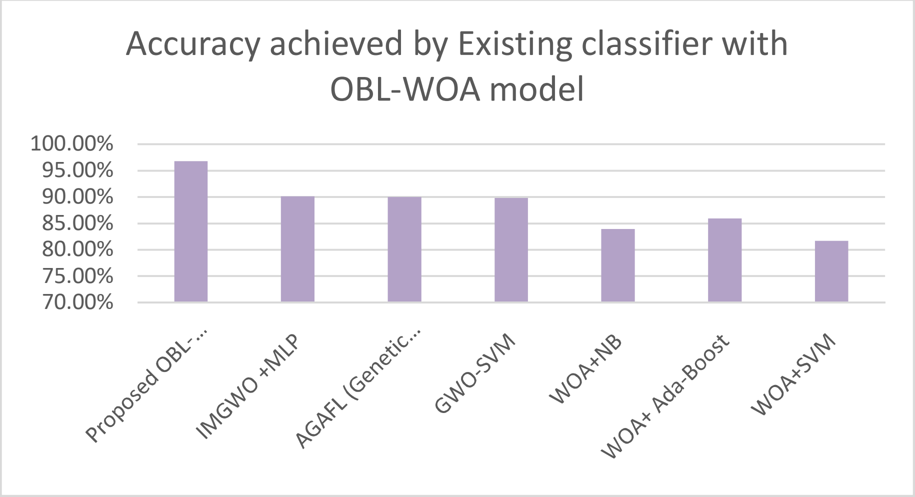


Fig-9: Accuracy achieved by existing classifier with OBL-WOA model.

From Table-3, clearly understood that the accuracy obtained by OBL-WOA+RNN outperforms IMGWO+MLPmodel presented by [8], the GWO+SVM method presented by [9], the WOA+ Ada-Boostand WOA+SVMmethod offered by [11], the WOA+NBmethod proposed by [12], and AGAFL (Genetic Algorithm with Fuzzy Logic) method offered by [15].

Table3: Comparison between the existing models with proposed OBL-WOA model

References	Models	Accuracy	Sensitivity	Specificity
Proposed method	**OBL-WOA + RNN**	**96.78%**	**87%**	**96%**
[8]2021	IMGWO +MLP	90.11%	95.45%	85.11%
[15] 2019	AGAFL (Genetic Algorithm with Fuzzy Logic)	90.0%	91.0%	90.0%
[9] 2019	GWO-SVM	89.83%	93.0%	91.0%
[12] 2021	WOA+NB	83.91%	N/A	N/A
[11] 2024	WOA+ Ada-Boost	85.91%	93.75	85.71
[11] 2024	WOA+SVM	81.67%	78.95%	84.51

6. Conclusion

The proposed model combines the Whale Optimization algorithm and Opposition Based Learning to enhance the WOA and obtain the optimal subset of features. The behavior of WOA is also demonstrated and explained. One of the major strengths of this model is the adaptive changes in the values of the best whales' bubble-net feeding technique,which make OBL-WOA an effective and efficient optimization method capable of avoiding local optima and detecting global optima consistently. The experimentation is conducted on the Cleveland (UCI) HD dataset. After identifying the optimal subset, various classification algorithms such as SVM, Decision Tree, Naïve Bayes, K-Nearest Neighbors, XG-Boost, and Recurrent Neural Networks are employed to predict HD. The results are evaluated using different measures. The overall experimental analysis reveals that OBL-WOA with RNN outperforms other hybrid combinations in terms of sensitivity, accuracy, specificity, and F1-score. Specifically, the RNN achieves a high accuracy of 96.78% compared to other methods.

In the future, the focus will be on developing a HD detection system using a variety of data, including ECG signal data and image data. Incorporating non-numerical data related to HD will help enhance the accuracy of the predictive model by utilizing different classification methods.

References:

[1] https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, YilongWang,Qiang Dong, Haipeng Shen, YongjunWang,"Artificial intelligence in healthcare: past, present and future",
Stroke and Vascular Neurology, 2:e000101, 2017.

[3] Jenni A. M. Sidey-Gibbons1 and Chris J. Sidey-Gibbons, 'Machine learning in medicine: a practical introduction', BMC Medical Research Methodology, vol. 19:64, 2019.

[4] T. Davenport and R. Kalakota, "e potential for artificial intelligence in healthcare," Future Healthcare Journal, vol. 6, no. 2, pp. 94–98, 2019.

[5] D. A. A. G. Singh, S. A. A. Balamurugan, and E. J. Leavline, "An unsupervised feature selection algorithm with feature ranking for maximizing performance of the classifiers", International Journal of Automation and Computing, vol. 12, no. 5, pp. 511–517, October. 2015.

[6] A. K. Garate-Escamila, A. Hajjam El Hassani, and E. Andres, "Classification models for heart disease prediction using feature selection and PCA", Informatics in Medicine Unlocked, vol. 19, January 2020.

[7] Balamurugan, S.P. and Arumugam, G., 'A novel method for predicting kidney diseases using
optimal artificial neural network in ultrasound images', International Journal of Intelligent
Enterprise, Vol. 7, Nos. 1/2/3, pp.37–55, 2020.

[8] Kumar, N., and Kumar, D., 'An Improved Grey Wolf Optimization-based Learning of Artificial Neural Network for Medical Data Classification', Journal of Information and Communication Technology, 20(2), 213-248, 2021.

[9] Qasem Al-Tashi, Helmi Rais1, and Said Jadid, 'Feature Selection Method Based on Grey Wolf Optimization for Coronary Artery Disease Classification', Springer Nature

Switzerland
AG 2019, IRICT 2018, AISC 843, pp. 257–266, 2019.

[10] Balamurugan, S.P. and Arumugam, 'Optimal Spatial Fuzzy Clustering Algorithm Based ROI Segmentation in Ultrasound Kidney Images',Journal of Computational and Theoretical Nanoscience, Vol. 15(9/10), pp. 2794 - 2804, 2018.

[11] Atimbire, S.A., Appati, J.K. &Owusu, E. Empirical exploration of whale optimisation algorithm for heart disease prediction. *Sci Rep* 14, 4530 (2024). https://doi.org/10.1038/s41598-024-54990-1.

[12] Alwateer, M.; Almars, A.M.; Areed, K.N.; Elhosseini, M.A.; Haikal, A.Y.; Badawy, M. Ambient Healthcare Approach with Hybrid Whale Optimization Algorithm and Naïve Bayes Classifier. Sensors 2021, 21, 4579. https://doi.org/10.3390/ s21134579.

[13] Lamiaa M. El Bakrawy, presented a 'Grey Wolf Optimization and Naive Bayes classifier Incorporation for Heart Disease Diagnosis', Australian Journal of Basic and Applied Sciences, Pages: 64-70, 11(7) May 2017.

[14] ChristalinLatha, C. B., &Jeeva, S. C. (2019), 'Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques', Informatics in Medicine Unlocked, 00203, 2019.

[15] G. Thippa Reddy,M. Praveen Kumar Reddy,
Kuruva Lakshmanna,Dharmendra Singh Rajput,
Rajesh Kaluri,Gautam Srivastava, 'Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis', International Journal of Advanced Science and Technology Vol. 29, No. 6, pp. 4225 - 4234,2020.

[16] Narender Kumar and Dharmender Kumar, 'Machine Learning based Heart Disease Diagnosis
using Non-Invasive Methods: A Review', IOP Publishing, Journal of Physics: Conference Series 1950_012081,2021.

[17] RamyaPerumal and Kaladevi AC, 'Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques'**,** International Journal of Advanced Science and Technology, pp. 4225 - 4234 Vol. 29, No. 6, 2020.

[18] Anouk Suppers, Alain J. van Gool and Hans J. C. T. Wessels, "Integrated Chemometrics and Statistics to Drive Successful Proteomics Biomarker Discovery", Proteomes, Vol:6, pp.20, April 2018.

[19] F. Kamalov and F. Thabtah, "A feature selection method based on ranked vector scores of features for classification", Annals of Data Science, 2017.

[20] Oliva, D., & Hinojosa, S. (Eds.)'Applications of hybrid metaheuristic algorithms for image processing', Springer International Publishing Vol. 890, 3-030-40977-7,2020.

[21] Suvarna, Chaitanya; Sali, Abhishek; Salmani, Sakina, 'Efficient heart disease prediction system using optimization technique', IEEE International Conference on Computing Methodologies and Communication (ICCMC) - Erode,374 –379, 2017.

[22] P. Suganya1and C. P. Sumathi, 'A Novel Metaheuristic Data Mining Algorithm for the Detection and Classification of Parkinson Disease', Indian Journal of Science and Technology, Vol 8(14), July 2015.

[23] S. Rahnamayan, "Opposition-based differential evolution", WSEAS TRANSACTIONS on COMPUTERS, ISSN: 1109-2750, Issue 10, Volume 7, October 2008 2007.

[24] H.R.Tizhoosh**, "Opposition-Based Learning: A New Scheme for Machine Intelligence",**
Proceedings of International Conference on Computational Intelligence for Modelling Control and Automation - CIMCA'2005, Vienna, Austria, vol. I, pp. 695-701, 2005.

[25] Mahdavi, S., Rahnamayan, S., & Deb, K. (2018). Opposition based learning: A literature review. Swarm and Evolutionary Computation, 39, 1–23,Published by Elsevier, 2017

[26] S.Rahnamayan, H.R.Tizhoosh, M.M. Salama, "**Opposition-Based Differential Evolution Algorithms"**, IEEE Congress on Evolutionary Computation, to be held as part of IEEE World Congress on Computational Intelligence, Vancouver, July 16-21, 2006.

[27] Mirjalili, Seyedali, 'Grey Wolf Optimizer', Advances in Engineering Software, Elsevier, 2014.

[28] FehmiBurcinOzsoydan, 'Effects of dominant wolves in grey wolf optimization algorithm', Applied Soft Computing Journal 83 Elsevier, 1568-4946, 2019.

[29] Siva Shankar G., 'Diagnosis of diabetes diseases using optimized fuzzy rule set by grey wolf
optimization Pattern Recognition', Elsevier, Letters 125 (2019) 432–438 0167-8655, 2019.

[30] Zheng-Ming Gao and Juan Zhao, 'An Improved Grey Wolf Optimization Algorithm with Variable Weights', Hindawi Computational Intelligence and Neuroscience, Article ID 2981282, 2019.

[31] Divyansh Khanna, Rohan Sahu, Veeky Baths, and Bharat Deshpande, 'Comparative Study of
Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease', International Journal of Machine Learning and Computing, Vol. 5, No. 5, October 2015.

[32] K.Vembandasamy, *R* R. Sasipriya*P*and E. Deepa, 'Heart Diseases Detection Using Naive Bayes Algorithm', IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 9, September 2015.

[33] R. Vijaya Kumar Reddy, K. Prudvi Raju, M. Jogendra Kumar, CH. Sujatha, P. Ravi Prakash,
'Prediction of Heart Disease Using Decision Tree Approach', Volume 6, Issue 3, March 2016.

[34] I KetutAgungEnriko, Muhammad Suryanegara, DadangGunawan,'Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health
Parameters', Journal of Telecommunication, Electronic and Computer Engineering, Vol. 8 No. 12, 09 February 2019.

[35] M.A. Jabbar, B.L. Deekshatulu and Priti Chandra, 'Intelligent heart disease prediction system
using random forest and evolutionary approach', Journal of Network and Innovative Computing ISSN 2160-2174 Volume 4 (2016) pp. 175-184, 30 April 2016.