






Research Article

WOA-COVID-19: Whale Optimization Algorithm for Selection of Multi-Examination Features based on COVID-19 Infections

Karrar Hameed Abdulkareem^{1,2,*}, Mazin Abed Mohammed^{3,4}, Zaid Abdi Alkareem Alyasseri^{2,5}, Dawood Zahi Khutar⁶
Osama Ahmad Alomari⁷

1 College of Agriculture, Al-Muthanna University, Samawah 66001, Iraq

2 College of Engineering, University of Warith Al-Anbiyaa, Karbala, 56001, Iraq

3 Department of Artificial Intelligence, College of Computer Science and Information Technology, University of Anbar, Anbar 31001, Iraq

4 College of science, Al-Farabi University, Baghdad, Iraq

5 Information Technology Research and Development Center (ITRDC), University of Kufa, Najaf, Iraq

6 Al-Muthanna University, College of Engineering, Department of Electronic and Communications, Al-Muthanna, Iraq

7 Department of Computer Science and Information Technology, College of Engineering, Abu Dhabi University, Abu Dhabi, United Arab Emirates

ARTICLE INFO

Article History

Received 9 Jul 2025
Revised 20 Jul 2025
Accepted 25 Jul 2025
Published 6 Aug 2025

Keywords

Whale Optimization
Algorithm
COVID-19
Feature selection
K-Nearest Neighbors
Multi-examination
features

ABSTRACT

Since its emergence in late 2019, COVID-19 (Coronavirus Disease 2019) has become one of the most critical global health threats, claiming millions of lives and placing many more at serious risk. The complexity of diagnosing COVID-19 lies in the wide range of clinical and examination features involved, prompting researchers to explore various advanced diagnostic methods. However, one of the main challenges is identifying the most relevant features that can streamline and improve diagnostic accuracy. In this study, we propose a feature selection approach based on the Whale Optimization Algorithm (WOA) to identify key examination indicators associated with COVID-19. We used a dataset of 78 patients that included 25 features, covering demographics, symptoms, vital signs, laboratory findings, and chronic health conditions. The WOA was applied as a single-objective optimization technique to select the most informative features. These selected features were then used with the K-Nearest Neighbors (KNN) algorithm to classify patients into three categories of severity: mild, moderate, and severe. To evaluate the effectiveness of WOA, we compared it against six other well-established metaheuristic algorithms: Particle Swarm Optimization (PSO), Multi-Verse Optimizer (MVO), Grey Wolf Optimizer (GWO), Moth-Flame Optimization (MFO), Firefly Algorithm (FFA), and the BAT algorithm. Results showed that WOA successfully reduced the feature set from 25 to just 6 key features while achieving a high classification accuracy of 92.5%. It also demonstrated strong robustness, as reflected in its low standard deviation compared to other methods. Overall, the proposed WOA-COVID-19 framework proved to be a highly effective and efficient solution for feature selection in the context of COVID-19 diagnosis.



1 INTRODUCTION

The widespread adoption of computer and internet technologies has led to the generation of vast amounts of data with several characteristics. Extracting valuable information from large datasets is essential in data mining. Choosing pertinent and beneficial characteristics may greatly impact many applications including text mining, image processing, Bioinformatics [1], and industrial applications[2]. The Internet of Things (IoT) is an advanced technology where physical devices containing sensors are interconnected over a network to share data[3]. Challenges in IoT applications involve the collection and processing of large volumes of data obtained from IoT sensors. Another obstacle is the presence of superfluous, inconsequential, and disruptive features. One way to address these difficulties is to employ feature selection to choose the best subset of characteristics [4]. Making predictions using train data and relevant attributes is one of the main objectives of data modeling and classification. For machine learning applications, large datasets with a high dimensional features space and a comparatively small number of samples are essential. One of the most important methods for getting rid of characteristics that are redundant and unnecessary from the original feature collection is dimensionality reduction [5].

The search strategy and sub-set quality evaluation are two major elements in the feature selection process, according to the feature selection system. The search approach employs a first-stage feature subgroup selection mechanism. The following stage involves employing a classifier to evaluate the subset's quality as indicated by the search strategy module. However,

there are three categories of feature selection strategies: methods based on filters, wrappers, and embedded systems. Traditional approaches relying on exhaustive search for huge data sets are insufficient and time-consuming, leading to limits in identifying the optimal collection of characteristics. For instance, when the feature size is d , selecting the necessary subset of features among 2^d options might be challenging. The FS Wrapper approach relies on an internal categorization to pinpoint a more pertinent subset of characteristics, which can significantly affect its performance, especially when dealing with large datasets. Strategies exist for moving both backward and forward to include or remove features that do not align with a wider set of standards. Metaheuristic algorithms(MA) enhance the efficiency of FS processes based on these challenges[6].

Various approaches have been suggested to address the FS issue. The aforementioned methodologies may be categorized into three distinct types: First, filter methods which are used to determine the association between the characteristics, metrics, and output variable where a statistical procedure is applied here. The appropriate characteristics are then chosen in light of this. Examples of these techniques are Mutual Information, Fisher Score, and Relief [7]. Second, wrapper methods in this case the model training is performed through learning algorithm. For instance, binary particle swarm optimization[8] and ant colony optimization[9].Third, embedded methods: these methods combine the two previous methods. Examples of this type are LASSO and RIDGE[10].Filter techniques are often more efficient than wrapper methods. The reason for this is that wrapper approaches employ a supervised learning strategy, which is a time-intensive process. However, wrapper approaches often yield superior classification results compared to filter methods[7].For all mentioned reasons we have employed wrapper approach is a base for feature selection task. Wrapper-based method, the selection procedure incorporates a learning algorithm (such as a classification method). Optimization is the method of identifying the most optimal solution for a certain problem from a range of potential alternatives[11]. Meta-heuristic algorithms are quite effective in solving radiology issues. These algorithms are mostly inspired by the logical behavior of physical algorithms found in nature [12]. Optimization is the method of identifying the most optimal solution for a certain problem from a range of potential alternatives. Meta-heuristic algorithms are quite effective in solving radiology issues. These algorithms are mostly inspired by the logical behavior of physical algorithms found in nature [13].

There are many different types of implementations that MA may be used for, including FS. The FS issue has been effectively solved utilizing certain traditional techniques, including Differential Evolution (DE), PSO, and GA. Furthermore, FS has also made use of contemporary MA algorithms including the Gravitational Search Algorithm (GSA), Grasshopper Optimization algorithm (GOA), Competitive Swarm Optimizer (CSO), and others. Given that FS may be understood as an optimization issue, not all FS challenges can be solved via MA. This latter data is established in accordance with no-free-lunch; as a result, further research into novel alternative MA is required[14]. Also, several academics have attempted to tackle feature selection issues using stochastic approaches, such as PSO, GA, ABC, and SA [15]. Modern methods utilized to effectively address feature selection issues include the Dragonfly Algorithm (DA) [16] and the GWO [17]. The researcher's interest has been sparked by Binary butterfly optimization (BOA), a newly developed optimization algorithm, due to its dependability, simplicity, and resilience in tackling engineering and real-world efficiency. Using principles from food-finding and butterfly-matching algorithms, BOA resolves optimization issues on a global scale. The effectiveness of BOA is unparalleled by other optimization techniques [18]. This population-based metaheuristic can mitigate the issue of stagnation in local optimal to a certain degree. It can also converge effectively towards the optimal solution. The main contributions of this study are:

- To model the problem of selection for multi-examination (demographic, laboratory findings, vital signs, symptoms, and chronic conditions) features of COVID-19 as an evolutionary-based optimization task.
- To propose a single objective technique combined with the KNN classifier for feature selection and detection of COVID-19 infections.
- An experimental evaluation of the performance of the proposed WOA–KNN algorithm using standard metrics is carried out.
- Different experiments were conducted for the purpose of comparing the obtained results after running six feature selection metaheuristic algorithms which are PSO, MVO, GWO, MFO, FFA, and BAT algorithm.

The paper structure is organized as follows: Section 2 presents previous works related to our study. Section 3 provides problem definition. Section 4, concentrates on methodological steps that adopted by proposed study. Section 5, discusses the results that obtained and how are significant. Finally, Section 6 shows how the proposed study resolved the claimed challenge and discusses main constraints.

2 LITERATURE REVIEW

Computed Tomography (CT) and X-ray imaging have demonstrated high sensitivity in diagnosing COVID-19 [19]. However, due to their high cost, limited availability, and associated radiation risks, these techniques are not always ideal for widespread patient screening. As a result, effectively distinguishing between COVID-19 positive and negative cases remains a pressing challenge in efforts to control the pandemic[20]. While several studies have attempted to predict COVID-19 infection using models based on single-feature inputs, few have proposed diagnostic frameworks that incorporate multiple clinical variables to improve classification accuracy. This research addresses that gap by applying machine learning classification methods to predict COVID-19 status (positive or negative) using 14 clinical features [21]. Clinical data suggest that the incubation period for COVID-19 is typically around four days, with a range of two to seven days. Approximately 81% of infected individuals experience mild or moderate symptoms, while the remaining 19% develop severe or critical illness [22]. Evidence also shows that elderly patients and those with underlying health conditions are significantly more likely to suffer from serious complications or death [23]. The prognosis of COVID-19 varies widely among patients, making early clinical assessment particularly challenging. Current diagnostic and treatment guidelines are insufficient for identifying the factors associated with severe disease progression or for making early judgments in critical cases. Therefore, there is an urgent need to establish more effective clinical protocols for early case classification. This would help prevent disease escalation, enable early intervention for severe and critical cases, and ensure timely separation and treatment of high-risk individuals in both confirmed and suspected patient populations [24].

The illness caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), commonly known as COVID-19, has had a profound impact on global health. It remains a major international concern due to the continued rise in infections and mortality rates worldwide [25]. The standard initial diagnostic method for COVID-19 is the real-time quantitative reverse transcription polymerase chain reaction (qPCR) test, a molecular technique widely used to detect viral RNA [26]. In addition, chest X-rays and computed tomography (CT) scans have gained considerable attention in both clinical practice and research for their role in diagnosing and monitoring COVID-19 patients [19]. Numerous studies have compared these diagnostic approaches, highlighting the strengths and limitations of each. While qPCR may show variable sensitivity depending on the type of biological sample, CT imaging may fail to detect very small or early-stage lung lesions [27]. Beyond diagnosis, considerable research has also focused on predicting patient outcomes by analyzing early clinical indicators observed during the course of the disease[28]. Understanding these prognostic factors is essential for improving treatment strategies and patient management during the ongoing pandemic.

Artificial Intelligence (AI) has gained widespread acceptance in the medical field as a valuable decision-support tool for clinicians [29-31]. Among its key branches, Machine Learning (ML) has shown exceptional promise in enabling the development of models capable of automatically diagnosing various diseases with high accuracy [32]. One of ML's most common applications is classification, where data are categorized based on a defined set of features. However, challenges such as overfitting—especially when using small training datasets—and the presence of non-Gaussian noise in the samples can severely impact model generalizability and performance [33]. Moreover, large and unfiltered feature sets often introduce irrelevant or noisy attributes, further increasing uncertainty and reducing accuracy. One effective strategy to address these challenges is selecting a minimal yet highly informative subset of features, which not only improves classification accuracy but also reduces computational time and complexity [34]. Creating a robust classification model requires careful consideration of several factors, including hyperparameter tuning, which plays a crucial role in determining model performance[35]. However, the space of possible hyperparameter combinations is often vast, making manual tuning impractical, time-consuming, and dependent on expert knowledge. As a result, there is a growing need for automated hyperparameter optimization, with various methods available—each with its own strengths and weaknesses [33].

In this context, metaheuristic algorithms have shown great promise for solving hyperparameter optimization challenges. Among the most widely used are Particle Swarm Optimization (PSO) and Genetic Algorithms (GA) [36, 37]. These techniques have been applied across diverse domains and have demonstrated strong performance in identifying optimal parameter configurations [38]. Particularly in feature selection, several wrapper-based algorithms have emerged, employing three general search strategies: exponential, sequential, and random search [39]. While exponential search delivers precise results, it is computationally intensive. Sequential methods incrementally add or remove features, but once a feature is selected or discarded, it cannot be revisited—making the approach susceptible to local optima [40]. Random search methods, including simulated annealing and metaheuristic algorithms, aim to avoid such pitfalls by introducing stochastic behavior and population-based exploration[41]. Due to their strong exploratory capabilities, metaheuristic algorithms have gained considerable attention for feature selection tasks, as outlined in the related literature [42]. Well-known algorithms in this space include PSO, Ant Colony Optimization (ACO), Artificial Bee Colony (ABC), and the Whale Optimization Algorithm (WOA), all of which have been used to build effective wrapper models [40]. However, many of these algorithms still face limitations related to imbalanced search dynamics, insufficient population diversity, and poor local/global convergence strategies, often leading to suboptimal results[43]. To overcome these limitations, researchers have continued to refine and hybridize these methods. Therefore, WOA stands out for its simplicity, minimal

control parameters, and adaptive behavior, making it an attractive candidate for wrapper-based feature selection [40]. Nonetheless, both theoretical analyses and empirical studies have revealed that WOA tends to perform well on simple or low-dimensional problems but struggles with complex, high-dimensional tasks such as extracting relevant features from medical datasets[44]. These challenges often lead to premature convergence and reduced diversity within the population. Therefore, this study is motivated by the need to enhance the WOA by improving its exploration-exploitation balance and boosting its ability to extract a compact yet meaningful subset of features. With its efficient structure and flexibility, an improved version of WOA has the potential to significantly strengthen feature selection performance in healthcare applications.

3 PROBLEM FORMULATION

In COVID-19 patients' classification as mild, moderate and severe, a machine learning model is sought, mathematically: a decision function f indicates if a Covid-19 patients (CP) condition in Mild (M) or Moderate (MO) or Severe (S).

The set of Covid-19 patients' indicators can be represented by CP.

A function $f: CP \rightarrow \{M, MO, S\}$ is searched.

To acquire this function, a machine learning system is trained on a dataset of pre-classified COVID-19 patients: $\{(cp1, a1), (cp2, a2), \dots, (CPn, an)\}$, $cp_i \in CP$, $ai \in \{M, MO, S\}$.

This study included two important components: Feature selection involves employing the WOA as a feature selector to extract relevant features $v_j = f_1, f_2, f_3, \dots, f_n$ from an Covid-19 patients data and construct feature vectors $V = \langle v_1, v_2, v_3, \dots, v_i \rangle$ that was then fed and utilized in the classification phase and the classification which was carried out with use of KNN classifier.

4 RESEARCH METHODS

This section provides a detailed explanation of the proposed WOA-COVID-19 technique, which integrates the Whale Optimization Algorithm (WOA) with a K-Nearest Neighbors (KNN) classifier to address the challenge of selecting the most relevant features for COVID-19 diagnosis. These features span multiple examination categories, including demographic information, laboratory findings, vital signs, symptoms, and chronic health conditions. Furthermore, Figure 1 illustrates the classification of COVID-19 patients into three primary categories based on severity: mild, moderate, and severe.

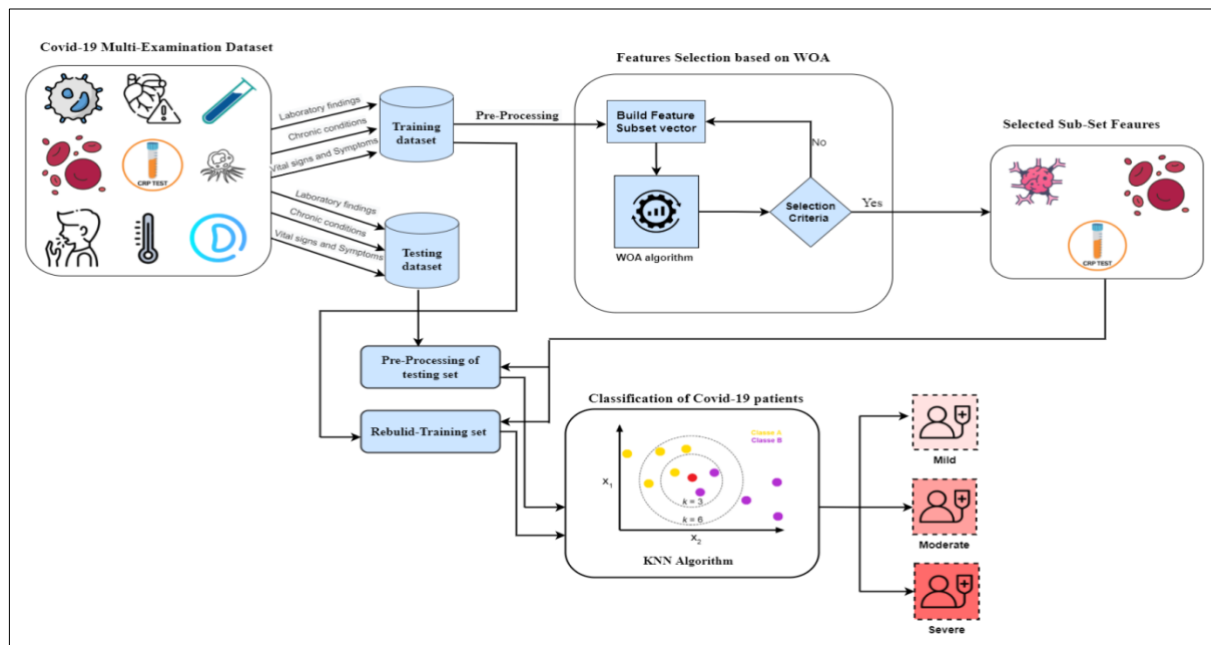


Fig.1. Proposed WOA-COVID-19 feature selection approach.

4.1 Dataset

The dataset used in this study was sourced from prior investigations [45, 46], with samples collected between April 8, 2020, and December 3, 2020. These samples were employed to develop and evaluate the proposed classification model. A total of 88 individuals were confirmed to be infected with COVID-19 by expert physicians at the Azizia Primary Healthcare Sector in the Wasit Governorate, Iraq. Among the 78 patients included in the final dataset, the most commonly reported symptoms were loss of taste (92.3%) and loss of smell (91.02%). The dataset comprises 25 clinical examination features relevant to COVID-19, categorized into five dimensions. The demographic dimension includes patient age. The laboratory findings dimension contains several biomarkers such as lymphocyte count, C-reactive protein (mg/L), urea (mmol/L), and creatinine ($\mu\text{mol/L}$), among others. The vital signs dimension consists of body temperature ($^{\circ}\text{C}$) and oxygen saturation (%). The symptoms dimension includes cough, nasal congestion, pleuritic chest pain, and binary indicators (1/0) for loss of smell and taste. Lastly, the chronic conditions dimension accounts for comorbidities such as cancer, diabetes, and heart disease. Each of these 25 features is associated with the patient's clinical condition—categorized as mild, moderate, or severe—as detailed in Table 1. These features serve as input for the proposed method's feature selection and classification process. Initially, the complete dataset was used to train a Whale Optimization Algorithm (WOA) to identify the most informative subset of features. Then, a new training model was constructed using only the selected features, with the aim of evaluating the K-Nearest Neighbors (KNN) classifier's performance. Importantly, the feature vector used during the testing phase of the KNN classifier consists solely of the final subset selected by the WOA.

TABLE I: DATASET DETAILS

Characteristic	No. of instance
No of Features	25
Data size	78 samples
Class distribution	22 Sever 37 Moderate 19 Mild

4.2 Data pre-processing

In traditional detection systems, preprocessing is a critical initial step to ensure that input data is properly formatted and standardized for subsequent processing. The collected clinical dataset contains a mixture of categorical and numerical features. For example, categorical features such as chronic disease status include discrete classes like cancer, heart disease, and diabetes. In contrast, numerical features—such as white blood cell count, C-reactive protein (mg/L), and oximetry saturation (%)—are represented as continuous values. This non-uniformity in data representation can adversely affect the performance and accuracy of classification models. To address this issue, all categorical features were converted into appropriate numerical representations to ensure consistency across the dataset. Additionally, min-max normalization was applied to all features in both the training and testing sets. This widely used technique scales each feature so that the minimum value becomes 0, the maximum value becomes 1, and all intermediate values are linearly mapped to a range between 0 and 1 [47]. This normalization step is essential for improving the stability and performance of machine learning algorithms by ensuring that all input features contribute equally to the learning process.

$$y = (x - \min) / (\max - \min) \quad (1)$$

4.3 Whale Optimization Algorithm (WOA)

WOA is a recently proposed bio-inspired optimization algorithm[48]. It replicates the sociable hunting behavior of Humpback whales in locating and capturing prey. The WOA algorithm is inspired by the spiral bubble-net hunting method used by humpback whales. This approach entails the whales diving down and creating bubbles in a spiral arrangement over their food. Then the whales swim up towards the surface, as displayed in Figure 2.

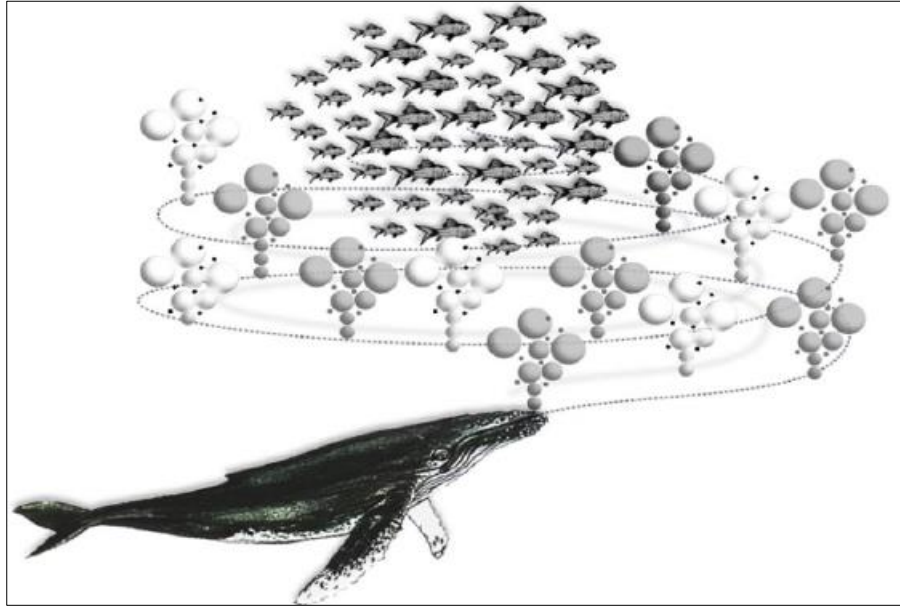


Fig.2. Approach of Humpback whales engage in bubble-net hunting [49].

Three operators model the three phases of humpback whale behavior: exploration (searching for prey), surrounding prey, and exploitation (bubble-net foraging). What follows is an explanation and model of the mathematical formulation:

- 1) Prey encircling: this stage comprises the start of the whale algorithm to the preliminary best search agent. It presupposes that the existing solutions are optimal and that it is situated near or in close proximity to the prey. Consequently, the other remaining agents update their locations toward the best search agent. The following expression illustrates this process:

$$\vec{D} = |\vec{C} \cdot \vec{X}^*(t) - \vec{X}(t)|, \quad (2)$$

$$\vec{X}(t+1) = \vec{X}^*(t) - \vec{A} \cdot \vec{D} \quad (3)$$

Where t specifies the present (iteration) \vec{A} and \vec{C} is coefficient vectors. \vec{X}^* represents the vector of location for the top solution gained yet while \vec{X} is the location vector. In the occurrence of a more effective solution, the X^* needs to be modified continuously.

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

$$\vec{A} = 2\vec{a} \cdot \vec{r} - \vec{a}, \quad (4)$$

$$\vec{C} = 2\vec{r} \quad (5)$$

Where \vec{a} is gradually reduced from 2 to 0 over the amount of iterations (repetitions) and r is random vector in range [0, 1]. This modeling allows an agent to update its location within the vicinity of the current optimal solution and mimics the behavior of surrounding the prey. The search algorithm may explore a larger search space in n dimensions, and the presence of agents near the optimal answer will aid in navigating across hypercubes [50].

- 2) Bubblenet attacking: It operates through a pair of techniques, which are as follows:

- Shrinking encircling technique: in this stage, \vec{a} value that presented in equation (4) is reduced and subsequently the variation range of \vec{A} is also lessened by \vec{a} . This infers that is randomly sited in $[-\vec{a}, \vec{a}]$. Where a is reduced from 2 to 0 over the process time for optimization.

The randomness of \vec{A} in $[-1, 1]$, the search agent's new position can be determined anywhere between its former position and the recent top position.

-Spiral updating position: This stage calculates the distance that exists between the whale's location and the target, and then a formula of spiral is generated between whale and target locations to simulate the movement of helix shape by humpback whale, and as the following expressions shows:

$$\vec{X}(t+1) = \vec{D}^t \cdot e^{bl} \cdot \cos(2\pi l) + \vec{X}^*(t), \quad (6)$$

$$\vec{D}^t = |\vec{X}^*(t) - \vec{X}(t)|. \quad (7)$$

Equation (7) represents the distance between the i th whale and the prey, which is the best solution obtained thus far. The constant b determines the shape of the logarithmic spiral, while l is a random number within the range of -1 to 1. The whale moves towards its prey by simultaneously executing a diminishing circular motion in a spiral-shaped trajectory. Hence, a 50% probability of transitioning between the two modes is utilized to calculate the whale's subsequent location in the following manner:

$$\vec{X}(t+1) = \begin{cases} \vec{X}^*(t) - \vec{A} \cdot \vec{D} & \text{if } p < 0.5 \\ \vec{D}^t \cdot e^{bl} \cdot \cos(2\pi l) + \vec{X}^*(t) & \text{if } p \geq 0.5 \end{cases} \quad (8)$$

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

3) Exploration phase: WOA achieves a global optimization in this phase. As shown in Figure 2, whales looking for its target with accordance to their location to each other in random existence. The \vec{A} value is appointed randomly between (-1) and (1) to accommodate the agent of searching to travel far from the reference whale. This means that \vec{A} should have a value that is either less than -1 or greater than 1. In addition, in order for the WOA to perform a thorough search, the new position of a search agent is decided by randomly picking one.

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

$$\vec{D} = |\vec{C} \cdot \vec{X}_{rand} - \vec{X}| \quad (9)$$

$$\vec{X}(t+1) = \vec{X}_{rand} - \vec{A} \cdot \vec{D}, \quad (10)$$

Where \vec{X}_{rand} is a random location for random whale that chosen from the current population[51].

Algorithm1: Whale Optimization Algorithm

Input:

Number of whales (Population) size (n)

Maximum number of iterations

Output:

Best position of whale X^*

1: Randomly Initialize population of whale positions where $X_i(i=0,1,2,3,\dots,n)$.

2: Initialize a, A and C.

3: Find fitness value per whale.

4: Make X^* as the best whale.

5: $k=1$

6: While $k < \text{maximum number of iterations}$ do

7: for each whale do

8: Modify the location of the present whale to a location given by equation 2

9: end for

10: Modify a, A and C

11: Compute each whale fitness value.

12: If there is a better solution modify X^*

13: $k=k+1$

14: end while

15: return X^*

TABLE II : PARAMETERS SETTING FOR EXPERIMENTS

Algorithms	Parameter	Value
FFA	alpha, gamma, beta	=0.5,=1,=0.2
	Iterations	100
	dim	50
	Population Size	20
	Number Of Runs	25
BAT	A	0.5
	r	0.5
	Iterations	100
	dim	50
	Population Size	20
	Number Of Runs	25
PSO	Maximum inertia weight=	0.90
	Minimum inertia weight=	0.20
	C1 =	2
	C2 =	2
	Iterations	100
	dim	50
	Population Size	20
MVO	Number Of Runs	25
	WEPMax	1
	WEPMin	0.2
	p	6
	Iterations	100
	dim	50
	Population Size	20
WOA	Number Of Runs	25
	r random number	[0 1]
	Iterations	100
	dim	50
	Population Size	20
MFO	Number Of Runs	25
	-	
	Iterations	100
	dim	50
	Population Size	20

4.4 Classification of Covid-19 patients based on KNN model

The KNN method is a nonparametric algorithm. Analysis for learning and prediction is conducted according to the provided issue or dataset. The KNN classification model makes predictions only based on neighboring data values without any prior assumptions about the dataset. In KNN, ' K ' denotes the quantity of closest neighboring data points. The KNN method classifies the provided dataset based on the number of closest neighbors, denoted as ' K '. The KNN model classifies the training dataset directly. The prediction of a new instance is determined by identifying the ' K ' nearest neighbor examples in the training set and categorizing it based on the majority class of those instances. Another example is identified by using the Euclidean distance formula. Euclidean distance is calculated as the square root of the sum of squared differences between a new instance (C_a) and an old instance (D_b)[52] .

$$Eu_{a,b} = \sqrt{\sum_{k=1}^n (C_{ak} - D_{bk})^2} \quad (11)$$

To mention selected subset of features by WOA algorithm is the main input for training and testing of KNN model. Patients with the severe condition were placed in class 1, patients with the moderate case were placed in class 2, and patients with the light case were placed in class 3.

4.5 Evaluation Metrics

In the training phase, the performance of each individual feature subset is assessed using 10-fold cross-validation to determine the KNN classification error rate utilized in the fitness function. A single feature subset is represented by each whale location. In order to direct the feature selection process, the training set is utilized to assess the KNN on the validation set during optimization. The chosen characteristics are next assessed on the test set to provide the final assessment of the chosen features. To assure the stability and statistical significance of the results acquired, the process of partitioning the data instances is iterated through 20 independent trials. The subsequent metrics are documented for every iteration on the unobserved test sets:

- 1- Feature_ratio: illustrates the relationship between the total number of features in the dataset and the average number of features picked.
- 2- Best_Acc and Av_Acc: show the best and the average classification accuracies.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (12)$$

The classifier's accurate prediction is mean cases of true positives (TP) and true negatives (TN). While False Positive (FP) and False Negative (FN) are represent the classifier's erroneous predictions, respectively.

- 3- Std_Acc: The testing accuracy standard deviation. It is used to gauge the stability and robustness of the optimization process; lower standard values mean that the optimization process always converges to the same solution, whilst bigger values mean that the outcomes are much more random.

5 Results and Discussion

The proposed algorithm was implemented in Python and executed on a system equipped with an Intel Core i5 CPU and 6 GB of RAM. To identify the optimal subset of features, the algorithm was run 20 times using the evaluation functions summarized in Table 2. This section presents the results obtained from the selected dataset, based on the defined evaluation criteria. All results are analyzed in the context of the training phase of the proposed system. Three main scenarios were considered for performance analysis in comparison with state-of-the-art methods. First, the best classification accuracy achieved with the fewest selected features is presented in Table 3. Second, the worst classification accuracy relative to the number of selected multi-examination COVID-19 indicators is shown in Table 4. Third, the average accuracy alongside execution time is summarized in Table 5. Additionally, the mean values from 10 independent runs were calculated and visualized in Figures 3 and 4. Figure 3 illustrates the average convergence rate per iteration, while Figure 4 depicts the average number of selected features per iteration.

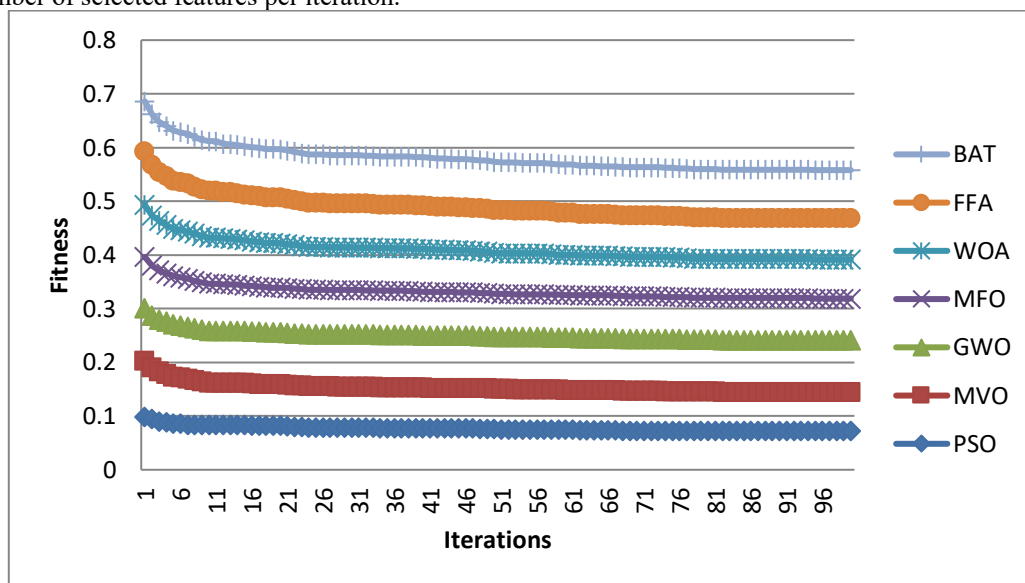


Fig.3. Convergence Rate Analysis.

Since all the algorithms employ the same initialization procedure, Figure 3 shows that most of them have similar performance at the beginning stage when comparing their average convergence rates. Convergence rates for BAT, FFA, WOA, MFO, GWO, and MVO have all dropped significantly from the beginning stage. In the process of determining the optimal solution, PSOs exhibit less variation in the value of the convergence rate at both the beginning and the end. Although there is a gradual improvement in convergence performance across all methods, the best solution is still not found until much later in the process.

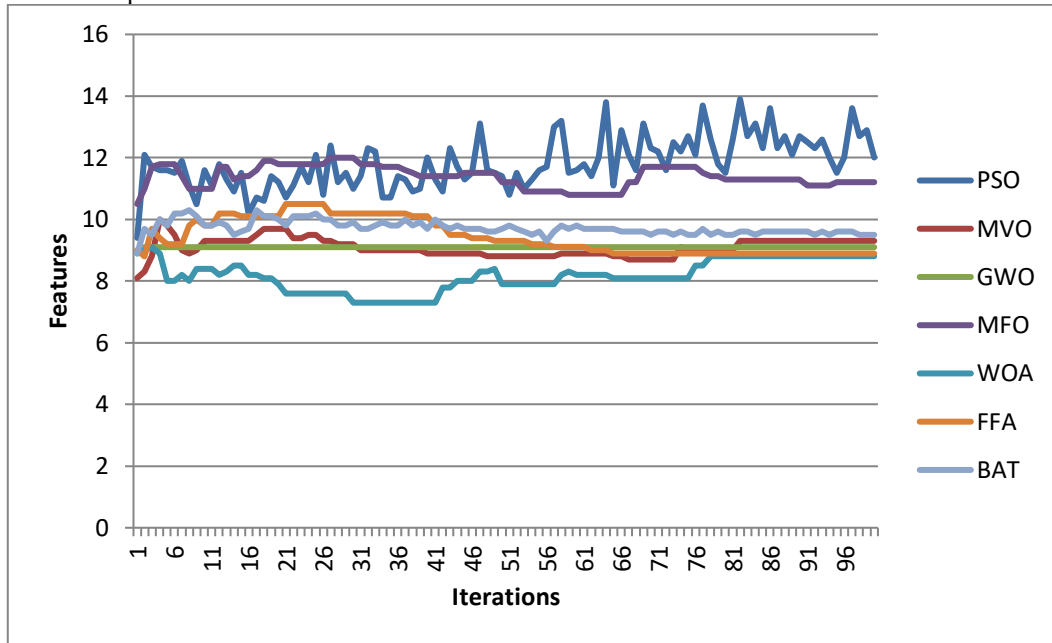


Fig.4. Selected Features Analysis.

Figure 4 shows the number of selected features per-iteration into seven tested algorithms. In general, it is observed that most algorithms have fluctuation into number of selected features. On other hand, fewer algorithms have very small change into number of selected multi-examination Covid-19 features. The maximum fluctuation in the number of selected features can be observed through the performance PSO algorithm. This is followed by MVO algorithm but with less change compared to PSO algorithm. To mention, the maximum number of selected features is obtained by PSO and MVO during all experiment iterations. The other six algorithms have selected a smaller number of features compared to PSO and MVO. However, WOA showed the best performance during all identified iterations where the minimum number of Covid-19 patients attributes is selected. Thus, WOA is considered as the best algorithm for feature selection process.

TABLE III: BEST ACCURACY WITHIN SELECTED FEATURES

Algorithm	Total features	Selected features	Best -Accuracy
PSO	25	11	91.8%
MVO	25	9	92.6
GWO	25	10	91.6%
MFO	25	10	93%
WOA-Covid-19	25	6	92.5%
FFA	25	9	93.6%
BAT	25	11	90.6%

According to the Table 3, all seven algorithms have scored significant classification performance where classification accuracy reaches above 90%. However; the maximum classification accuracy has scored by FFA algorithm. While, the minimum classification rate has presented by BAT algorithm. When considering the problem of computational load into real world application the number of features showed is as less as possible. In this direction, in our analysis the less number of selected features is given by WOA algorithm where only 6 out of 25 features are selected as base for classification process. Furthermore, the classification accuracy reaches 92.5 % that less than 1% of best scored accuracy. Therefore, the proposed WOA-Covid-19 algorithm is the best from accuracy and less number of selected features perspectives.

TABLE IV: WORST ACCURACY WITHIN SELECTED FEATURES

Algorithm	Total features	Selected features	Worst -Accuracy
PSO	25	15	88.8%
MVO	25	8	89.3%
GWO	25	9	87.5%
MFO	25	11	89.3%
WOA-Covid-19	25	7	89.6%
FFA	25	9	89.6%
BAT	25	7	87%

Table 4 shows very surprising results for instance, comparing to results of BAT algorithm into Table 3 the number of selected features was decrease to 7 with accuracy rate reach 87% in BAT algorithm. However, in the proposed WOA-Covid-19 algorithm have increased with only one more features comparing to six ones that selected by best accuracy scenario. Therefore, the performance of WOA-Covid-19 algorithm still significant even with worst accuracy scenario.

TABLE V: AVERAGE ACCURACY WITHIN EXECUTION TIME

Algorithm	Ratio of features reductions	Average-Accuracy	Execution Time	Std
PSO	44%	90%	5.431480932	0.008994371
MVO	36%	91%	5.064110494	0.009654982
GWO	40%	91.7%	5.739459682	0.011334436
MFO	40%	91%	5.567290997	0.011922351
WOA-Covid-19	24%	90.1%	6.56030283	0.009754048
FFA	36%	90.1%	6.938513851	0.012428291
BAT	44%	89%	5.013561296	0.011397092

Table 5 confirms the observations into previous two tables where the highest reduction rate have scored by WOA-Covid-19. This algorithm have selected only 24% of Covid-19 features as attributes for identification process. While the worst reduction rate are presented based on results of PSO and BAT. Moreover, the best average accuracy value have provide by GWO where this algorithm scored 91.7%. The proposed WOA-Covid-19 has scored 90.1% with only 1.6 less than GWO. However, GWO presented less reduction rate where this algorithm has selected 40% of Covid-19 features. To mention, when it comes to execution time WOA-Covid-19 algorithms have consumed more times than other algorithms while BAT algorithm is less time consuming algorithm. The PSO algorithm provides the least standard ratio. Table 5 shows that the resilience of WOA-Covid-19 is good compared to other methods, with less standard deviation on the given dataset, except for PSO. The results suggest that the proposed mechanism enhances the efficiency of WOA-Covid-19 in addressing feature selection difficulties. To mention, each of Spo2, chronic disease, lymphocyte count, and C-reactive protein are most important features for classification covid-19 cases that satisfied medical and technical requirements.

By implement the output of our proposed study in form of real application, this can be generalized to biology, medical domain, drug design and several other areas. Additionally, researchers may apply it to any dataset with two different feature selection techniques and three different learner types to produce a variety of tables and diagrams depending on the dataset's characteristics. The algorithms and approaches can also be combined to create an ensemble method.

6 CONCLUSIONS

COVID-19 is a highly complex disease that has affected billions of people across various regions of the world. This complexity arises from its diverse symptoms and numerous clinical examinations attributes, making decision-making in the medical sector particularly challenging. To address this, the present study introduces a feature selection method based on optimization principles—specifically, the Whale Optimization Algorithm (WOA), referred to as WOA-COVID-19. A total of 25 features derived from various COVID-19 examination approaches were utilized in the feature selection process using the WOA method. The best- and worst-case results demonstrate that WOA outperforms six other metaheuristic feature selection algorithms evaluated during the study. While some of the other algorithms showed competitive performance, none surpassed the results achieved by WOA-COVID-19. For the classification task, the K-Nearest Neighbors (KNN) algorithm was employed to categorize patients into three main severity levels: mild, moderate, and

severe. The classification results in both worst and best scenarios confirmed the effectiveness of using KNN in conjunction with WOA-selected features. Furthermore, the average accuracy analysis revealed that WOA-COVID-19 not only reduced the number of features required for diagnosis but also maintained high classification accuracy. Specifically, the feature reduction rate reached 24%, without compromising classification performance. Additionally, statistical analysis using standard deviation (STD) confirmed the robustness and reliability of the WOA-COVID-19 approach. Overall, the proposed method proved highly effective in solving the multi-examination feature selection problem and accurately classifying different types of COVID-19 patients. However, one noted limitation is the relatively high execution time of the WOA-COVID-19 algorithm. Future work will aim to address this limitation by introducing a more efficient strategy, especially for real-time deployment scenarios.

Additional Information and Declarations

Funding: Authors state no funding involved.

Conflict of Interests: The authors declare no conflict of interest.

Author Contributions: **Karrar Hameed Abdulkareem:** Supervision; conceptualization; methodology, formal analysis, investigation, data duration, writing –original draft, **Mazin Abed Mohammed:** Formal analysis, investigation, writing – original draft, visualization, **Zaid Abdi Alkareem Alyasseri:** methodology, data duration, review & editing, **Dawood Zahi Khutar:** conceptualization, writing – review & editing, **Osama Ahmad Alomari:** methodology, data duration, review & editing.

Data Availability: The data that support the findings of this study are available in articles [45, 46].

References

- [1] A. M. Khalid, H. M. Hamza, S. Mirjalili, and K. M. Hosny, "BCOVIDOA: A Novel Binary Coronavirus Disease Optimization Algorithm for Feature Selection," *Knowledge-Based Systems*, vol. 248, p. 108789, 2022/07/19/ 2022.
- [2] S. Egea, A. R. Mañez, B. Carro, A. Sánchez-Esguevillas, and J. Lloret, "Intelligent IoT Traffic Classification Using Novel Search Strategy for Fast-Based-Correlation Feature Selection in Industrial Environments," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1616-1624, 2018.
- [3] L. Zhao and X. Dong, "An Industrial Internet of Things Feature Selection Method Based on Potential Entropy Evaluation Criteria," *IEEE Access*, vol. 6, pp. 4608-4617, 2018.
- [4] P. Wongthongtham, J. Kaur, V. Potdar, and A. Das, "Big Data Challenges for the Internet of Things (IoT) Paradigm," in *Connected Environments for the Internet of Things: Challenges and Solutions*, Z. Mahmood, Ed. Cham: Springer International Publishing, 2017, pp. 41-62.
- [5] M. Rostami, K. Berahmand, E. Nasiri, and S. J. E. A. o. A. I. Forouzandeh, "Review of swarm intelligence-based feature selection methods," vol. 100, p. 104210, 2021.
- [6] I. M. EL-Hasnony, M. Elhoseny, and Z. Tarek, "A hybrid feature selection model based on butterfly optimization algorithm: COVID-19 as a case study," vol. 39, no. 3, p. e12786, 2022.
- [7] R. Bandyopadhyay, A. Basu, E. Cuevas, and R. Sarkar, "Harris Hawks optimisation with Simulated Annealing as a deep feature selection method for screening of COVID-19 CT-scans," *Applied Soft Computing*, vol. 111, p. 107698, 2021/11/01/ 2021.
- [8] J. Wei *et al.*, "A BPSO-SVM algorithm based on memory renewal and enhanced mutation mechanisms for feature selection," *Applied Soft Computing*, vol. 58, pp. 176-192, 2017/09/01/ 2017.
- [9] S. Kashef and H. Nezamabadi-pour, "An advanced ACO algorithm for feature subset selection," *Neurocomputing*, vol. 147, pp. 271-279, 2015/01/05/ 2015.
- [10] H. Zhang, R. Zhang, F. Nie, and X. Li, "A Generalized Uncorrelated Ridge Regression with Nonnegative Labels for Unsupervised Feature Selection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2781-2785.

- [11] M. M. Fouad, A. I. El-Desouky, R. Al-Hajj, and E. S. M. El-Kenawy, "Dynamic Group-Based Cooperative Optimization Algorithm," *IEEE Access*, vol. 8, pp. 148378-148403, 2020.
- [12] M. A. A. Al-qaness, A. A. Ewees, H. Fan, and M. Abd El Aziz, "Optimization Method for Forecasting Confirmed Cases of COVID-19 in China," vol. 9, no. 3, p. 674, 2020.
- [13] E. S. M. El-Kenawy, A. Ibrahim, S. Mirjalili, M. M. Eid, and S. E. Hussein, "Novel Feature Selection and Voting Classifier Algorithms for COVID-19 Classification in CT Images," *IEEE Access*, vol. 8, pp. 179317-179335, 2020.
- [14] D. Yousri, M. Abd Elaziz, L. Abualigah, D. Oliva, M. A. A. Al-qaness, and A. A. Ewees, "COVID-19 X-ray images classification based on enhanced fractional-order cuckoo search optimizer using heavy-tailed distributions," *Applied Soft Computing*, vol. 101, p. 107052, 2021/03/01/ 2021.
- [15] I. M. El-Hasnony, M. Elhoseny, and Z. Tarek, "A hybrid feature selection model based on butterfly optimization algorithm: COVID-19 as a case study," *Expert Systems*, vol. 39, no. 3, p. e12786, 2022/03/01 2022.
- [16] M. A. Tawhid and K. B. Dsouza, "Hybrid binary dragonfly enhanced particle swarm optimization algorithm for solving feature selection problems," *Mathematical Foundations of Computing*, vol. 1, no. 2, pp. 181-200, 2018/05/02 2018.
- [17] E. Emary, H. M. Zawbaa, and A. E. Hassanien, "Binary grey wolf optimization approaches for feature selection," *Neurocomputing*, vol. 172, pp. 371-381, 2016/01/08/ 2016.
- [18] S. Arora and P. Anand, "Binary butterfly optimization approaches for feature selection," *Expert Systems with Applications*, vol. 116, pp. 147-160, 2019/02/01/ 2019.
- [19] T. Ai *et al.*, "Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases," (in eng), *Radiology*, vol. 296, no. 2, pp. E32-e40, Aug 2020.
- [20] D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi, and F. Cabitza, "Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study," *Journal of Medical Systems*, vol. 44, no. 8, p. 135, 2020/07/01 2020.
- [21] I. Arpaci, S. Huang, M. Al-Emran, M. N. Al-Kabi, and M. Peng, "Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11943-11957, 2021/03/01 2021.
- [22] W. J. Guan *et al.*, "Clinical Characteristics of Coronavirus Disease 2019 in China," (in eng), *N Engl J Med*, vol. 382, no. 18, pp. 1708-1720, Apr 30 2020.
- [23] C. Wu *et al.*, "Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China," (in eng), *JAMA Intern Med*, vol. 180, no. 7, pp. 934-943, Jul 1 2020.
- [24] Z. Li *et al.*, "Efficient management strategy of COVID-19 patients based on cluster analysis and clinical decision tree classification," *Scientific Reports*, vol. 11, no. 1, p. 9626, 2021/05/05 2021.
- [25] N. Zhu *et al.*, "A Novel Coronavirus from Patients with Pneumonia in China, 2019," (in eng), *N Engl J Med*, vol. 382, no. 8, pp. 727-733, Feb 20 2020.
- [26] Y. Fang *et al.*, "Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR," (in eng), *Radiology*, vol. 296, no. 2, pp. E115-e117, Aug 2020.
- [27] H. X. Bai *et al.*, "Performance of Radiologists in Differentiating COVID-19 from Non-COVID-19 Viral Pneumonia at Chest CT," *Radiology*, vol. 296, no. 2, pp. E46-E54, 2020/08/01 2020.
- [28] I. Shiri *et al.*, "Machine learning-based prognostic modeling using clinical data and quantitative radiomic features from chest CT images in COVID-19 patients," *Computers in Biology and Medicine*, vol. 132, p. 104304, 2021/05/01/ 2021.
- [29] K. Sailunaz, T. Özyer, J. Rokne, and R. Alhajj, "A survey of machine learning-based methods for COVID-19 medical image analysis," *Medical & Biological Engineering & Computing*, vol. 61, no. 6, pp. 1257-1297, 2023/06/01 2023.
- [30] H. M. A. Ghanimi, F. T. Ayasrah, V. C. Jadala, T. C. Manjunath, K. Balasaranya, and B. Srinivasarao, "An Innovative Artificial Intelligence Based Decision Making System for Public Health Crisis Virtual Reality Rehabilitation," *Journal of Machine and Computing*, Article vol. 5, no. 1, pp. 561-575, 2025.
- [31] S. H. Nowfal, S. Sengan, J. S. Deol, S. G. Bhatta, V. Saravanan, and B. Veeramallu, "The Diagnosis of Heart Attacks: Ensemble Models of Data and Accurate Risk Factor Analysis Based on Machine Learning," *Journal of Machine and Computing*, Article vol. 5, no. 1, pp. 589-599, 2025.
- [32] I. I. Al Barazanchi *et al.*, "Optimizing the Clinical Decision Support System (CDSS) by Using Recurrent Neural Network (RNN) Language Models for Real-Time Medical Query Processing," *Computers, Materials and Continua*, Article vol. 81, no. 3, pp. 4787-4832, 2024.

- [33] V. Stojanovic and D. Prsic, "Robust identification for fault detection in the presence of non-Gaussian noises: application to hydraulic servo drives," *Nonlinear Dynamics*, vol. 100, no. 3, pp. 2299-2313, 2020/05/01 2020.
- [34] M. Soui, N. Mansouri, R. Alhamad, M. Kessentini, and K. Ghedira, "NSGA-II as feature selection technique and AdaBoost classifier for COVID-19 prediction using patient's symptoms," *Nonlinear Dynamics*, vol. 106, no. 2, pp. 1453-1475, 2021/10/01 2021.
- [35] R. Elshawi, M. Maher, and S. J. a. p. a. Sakr, "Automated machine learning: State-of-the-art and open challenges," 2019.
- [36] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295-316, 2020/11/20/ 2020.
- [37] H. Liu *et al.*, "Multi-objective optimization of buckling load and natural frequency in functionally graded porous nanobeams using non-dominated sorting genetic Algorithm-II," *Engineering Applications of Artificial Intelligence*, vol. 142, p. 109938, 2025/02/15/ 2025.
- [38] R. Olivares *et al.*, "An Optimized Brain-Based Algorithm for Classifying Parkinson's Disease," *Applied Sciences*, vol. 10, no. 5. doi: 10.3390/app10051827
- [39] P. Agrawal, H. F. Abutarboush, T. Ganesh, and A. W. Mohamed, "Metaheuristic Algorithms on Feature Selection: A Survey of One Decade of Research (2009-2019)," *IEEE Access*, vol. 9, pp. 26766-26791, 2021.
- [40] M. H. Nadimi-Shahraki, H. Zamani, and S. Mirjalili, "Enhanced whale optimization algorithm for medical feature selection: A COVID-19 case study," *Computers in Biology and Medicine*, vol. 148, p. 105858, 2022/09/01/ 2022.
- [41] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2015, pp. 1200-1205.
- [42] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A Survey on Evolutionary Computation Approaches to Feature Selection," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606-626, 2016.
- [43] M. H. Nadimi-Shahraki and H. Zamani, "DMDE: Diversity-maintained multi-trial vector differential evolution algorithm for non-decomposition large-scale global optimization," *Expert Systems with Applications*, vol. 198, p. 116895, 2022/07/15/ 2022.
- [44] H. Chen, Y. Xu, M. Wang, and X. Zhao, "A balanced whale optimization algorithm for constrained engineering design problems," *Applied Mathematical Modelling*, vol. 71, pp. 45-59, 2019/07/01/ 2019.
- [45] A. M. Dinar *et al.*, "Towards Automated Multiclass Severity Prediction Approach for COVID-19 Infections Based on Combinations of Clinical Data," vol. 2022, 2022.
- [46] K. H. Abdulkareem *et al.*, "MEF: Multidimensional Examination Framework for Prioritization of COVID-19 Severe Patients and Promote Precision Medicine Based on Hybrid Multi-Criteria Decision-Making Approaches," *Bioengineering*, vol. 9, no. 9. doi: 10.3390/bioengineering9090457
- [47] R. R. N. Alogaili *et al.*, "AntDroidNet Cybersecurity Model: A Hybrid Integration of Ant Colony Optimization and Deep Neural Networks for Android Malware Detection," *Mesopotamian Journal of CyberSecurity*, Article vol. 5, no. 1, pp. 104-120, 2025.
- [48] S. Mirjalili and A. Lewis, "The Whale Optimization Algorithm," *Advances in Engineering Software*, vol. 95, pp. 51-67, 2016/05/01/ 2016.
- [49] M. H. Nadimi-Shahraki, H. Zamani, Z. Asghari Varzaneh, and S. Mirjalili, "A Systematic Review of the Whale Optimization Algorithm: Theoretical Foundation, Improvements, and Hybridizations," *Archives of Computational Methods in Engineering*, vol. 30, no. 7, pp. 4113-4159, 2023/09/01 2023.
- [50] H. F. Eid, "Binary whale optimisation: an effective swarm algorithm for feature selection," *International Journal of Metaheuristics*, vol. 7, no. 1, pp. 67-79, 2018/01/01 2018.
- [51] M. Sharawi, H. M. Zawbaa, E. Emary, H. M. Zawbaa, and E. Emary, "Feature selection approach based on whale optimization algorithm," in *2017 Ninth International Conference on Advanced Computational Intelligence (ICACI)*, 2017, pp. 163-168.
- [52] P. Theerthagiri, I. J. Jacob, A. U. Ruby, and Y. Vamsidhar, "Prediction of COVID-19 possibilities using KNN classification algorithm," 2020.