Received: March 9, 2023. Revised: April 24, 2023. 

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

# An Intelligent Feature Selection Approach Based on a Novel Improve Binary Sparrow Search Algorithm for COVID-19 Classification

**Amir Yasseen Mahdi[1,2]*** **Siti Sophiayati Yuhaniz[1]**

[1]*Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia*
[2]*Computer Sciences and Mathematics College, University of Thi_Qar, Thi_Qar, Iraq*
* Corresponding author's Email: mahdi.amir@graduate.utm.my, amiryasseen@utq.edu.iq

**Abstract:** This paper proposes an improved binary sparrow search algorithm (IBSSA) as a search strategy within the feature selection (FS) methods. Its main objective is to use clinical texts to improve COVID-19 patient categorization. The constant need for an efficient FS system and the favorable outcomes of swarming behavior in numerous optimization situations drove our efforts to develop a novel FS strategy. Additionally, clinical text data are frequently highly dimensional and contain uninformative features, which have a major impact on the classifier's accuracy, making FS a key machine-learning step in data pre-processing to reduce data dimensionality. The bi-stage FS approach is used in this work to elect the features. At the initial stage, we employed a term weighting scheme (TWS) that assigned a weighted score to each feature by measuring the significance of the features obtained from the pre-processing model using a new weight calculation method called root term frequency-core-inverse exponential frequency (RTF-C-IEF). Next, finding the most relevant and almost optimal feature subset for COVID-19 illness diagnosis is done in the second stage using a freshly developed methodology that was inspired by the way sparrow's behavior. The suggested modification method for the sparrow's algorithm is composed of several stages of advancement. The main objectives are to promote the exploration of the search space and increase the algorithm's variability. In order to evaluate the proposed model, various classifiers were employed on two datasets, each of which had 1446 and 3053 cases, respectively. The experimental and statistical results demonstrate that the proposed IBSSA is significantly superior compared to other comparative optimization algorithms, and it successfully upgrades the shortcomings of the original SSA. Moreover, the IBSSA has the highest accurate performance when compared to other rivals by the SVM classifier, Where, average removed features are 77.99% and 83.5%, with improvement percentages by F1-scores: 84.95% and 95.94 % for both datasets, respectively.

**Keywords:** Natural language processing, COVID-19, Binary sparrow search algorithm, Optimization, Feature selection, Clinical text classification.

## 1. Introduction

The catastrophic spread of COVID-19 is the greatest danger to humanity since the Second World War. Virtually everyone has been affected worldwide by the coronavirus disease 2019 pandemic, including the government, medical personnel, and the general public [1]. Although extensive research is being done to create a vaccine, the virus persists and shows several patterns that have a high potential for spreading and vaccine resistance, such as Delta, Omicron, and Ihu [2]. These mutations pose a challenge in dealing with the COVID-19 epidemic [3]. According to the WHO weekly report from the second of January 2022, there have been more than 41000 new deaths and about 9.5 million new cases. Combating the early stages of COVID-19 proliferation is crucial in light of this pandemic explosion. Therefore, there is a crucial need for COVID-19 diagnosis techniques to improve patient care and strategic planning for treatment.

Artificial intelligence (AI) has recently been seen as a potentially strong tool in the fight against many evolving pandemics [4]. One of the main subfields of AI known as text mining works with the

analyzing of various forms of unstructured texts such as clinical texts in order to extract usable information, and it is receiving increased attention in several industrial domains, especially the field of medicine, to overcome the challenges which they face in clinical decision making [5]. In addition to being very helpful for analyzing, diagnosing, and forecasting illnesses, text mining techniques can also aid in the prevention of viral infections [6]. Clinical texts are a primary resource for disease-related information, the clinical narrative generally contains more thorough and accurate information for COVID-19 diagnosis, symptom description, and clinical decision-making compared to structured data that initially has a poor detection sensitivity [7]. However, the clinical texts are high-dimensional data and contain redundant and pointless features, there are tens of thousands or even hundreds of thousands of distinct terms or tokens. Even after preprocessing such as removing stop words and stemming, the feature set continues to be enormous, which is a typical problem that increases computational costs, and negatively affects the performance of classification algorithms [8, 9]. Furthermore, without a suitable set of features, a robust classification system with high predicted accuracy cannot be established. In order to significantly minimize the dimensionality curse, shorten training times, and simultaneously identify a new optimal subset of useful features for use in the classification process, feature selection is crucial [10]. The optimization challenge of feature selection calls for an efficient global approach, particularly when dealing with clinical texts that have several data dimensions. Swarm intelligence algorithms have demonstrated their suitability and efficiency for feature selection problems due in particular to their particular in overcoming the curse of dimensionality by optimizing the efficiency of classification, and the quantity of features, and their provide practical solutions in a timely manner [11]. These methods are frequently used to solve different optimization issues [12, 13]. Nonetheless, it was also noted that there is room for development, one explanation for this is because many of the suggested metaheuristics experience suffer from an imbalance between exploration and exploitation and stagnation in the local optimum [14].

Thus, the following formulation can be used to express the research question this study addresses: is it possible to improve classification accuracy and/or minimise the number of features from clinical texts by using SSA for feature selection compared to other contemporary methods that are available?

The SSA is a new kind of swarm intelligence optimization algorithms was put forth by Jiankai [15] in 2020. In order to iteratively optimize, it makes advantage of foraging behavior and anti-predation behavior in sparrow populations.

Several motivations led to the decision to use the SSA algorithm in this study to solve the FS problem. First, the SSA has been demonstrated that it has the advantages of high searching precision, faster convergence, good stability, strong competitiveness, and strong robustness [16, 17]. It also offers a brand-new way of resolving complex global optimization issues over the most recent algorithms. Second, the SSA was studied and contrasted with new swarm intelligence optimizations by Ahmed and others. The sparrow search algorithm performed far better than conventional optimization algorithms, according to a thorough comparison of experimental findings [18]. Finally, recent studies showed after a comprehensive survey [19] that SSA algorithm can run on most optimization problems due to its ease of implementation and rapid increase in the spread of agents in the problem space. In addition, SSA uses the concept of exploratory research, which makes it possible to track the characteristics of the population in the optimization process.

The goal of this study is to create a two-stage method for extracting and choosing relevant features for the COVID-19 clinical text categorization. The first stage uses the clinical texts' significant terms and concepts to identify features. In order to decrease the quantity of extracted characteristics, we employed a spaciy tool to extract concepts. The RTF-C-IEF approach is then used to order each term in the text according to how significant it is in the datasets. In the second stage, we introduce a novel improved binary sparrow search algorithm (IBSSA) based FS method to choose an ideal subset of significant features to enhance the performance of the classifier methodology. Moreover, the literature review also leads to the conclusion that the sparrow search algorithm (SSA), one of the most effective swarm intelligence methods, has not had its potential for feature selection fully explored. This encouraged authors to create a novel SSA method and modify it for use in resolving this problem. This work's primary contributions of note are:

- Proposing a new SSA approach that can effectively deal with feature selection in terms of categorisation precision, and number of elected feature and overcoming shortcomings of original SSA.
- For the first time, a feature transformation approach-based enhanced binary version of (IBSSA), developed using the new modified

initialization approach (MIA), the local search algorithm (LSA) for improving exploitation, and levy flight strategy to broaden the variety of potential solutions and provide a top level of randomization.
- Combining TWS (RTF-C-IEF), IBSSA, and SVM to offer a novel text categorization approach.
- Comparing the suggested approach to the seven of common wrapper-based feature selection techniques.

The remainder of the paper is structured as follows: section 2 discusses the FS process' related works. Section 3 provides an overview of SSA standards. The suggested methodology is described in section 4. The parameters setting and testing results of several algorithms are shown and discussed in section 5 to confirm the viability and efficiency of the IBSSA. The paper is concluded in section 6, which also outlines the work that will be done in the future.

## 2. Related work

In recent works, a variety of mixed approaches and multi-stage feature selection processes have been suggested for categorization using machine learning. Additionally, metaheuristic optimization via swarms intelligence is gaining popularity as a method for dealing with complex problems that are challenging to solve using conventional methods [20]. There has been a lot of nature-inspired optimizers published recently to replicate the evolutionary concepts and natural mechanisms for solving optimization issues.

Authors in [21] formulated feature subsets with Chi-square, Gini index, and PSO algorithms to solve FS problems in machine learning. Authors in [22] proposed the algorithm FS two-stage to enhance the effectiveness of Arabic text classification by combining the term frequency-inverse document frequency in the first phase and particle swarm optimization in the second phase. The ant lion optimizer, which was utilized in a dataset for COVID-19, was introduced in [23] as a hybrid technique for addressing the feature selection difficulty. Authors in [24] presented the hybrid MMPSO method as a proposed approach to feature engineering, it has been successfully utilized to extract features from a high-dimensional dataset by combining the feature ranking approach and the heuristic search method. When choosing features for COVID-19 patients, authors in [25] used a two-step strategy. In the first step, a filter measure was used to rank the features according to their relevance, and in the second step, a genetic algorithm and decision tree classifier were combined to find the best feature subset. This study [26], explored a two-stage feature selection pipeline that incorporates evolutionary algorithms and traditional filter approaches. Authors in [27] proposed a method based on mixing the algorithm of PSO with the butterfly optimization algorithm as a search methodology for feature selection from a COVID-19 dataset. The new crow learning algorithm has been introduced in this study [28], which uses feature selection methods as its first stage to identify the best attributes associated with COVID-19 disease. In this article [29], a novel hyper learning binary dragonfly algorithm is proposed to identify the best subset of features for a particular classification problem. An innovative two-stage technique is put forth in [30]. In the first stage, significant features from the most important concepts, such as diseases or symptoms, are extracted using a domain-specific lexicon, which that is, the unified medical language system. PSO is used in the second stage to choose additional related features from the first stage's retrieved features. The authors of this study [31] presented a two-step strategy to choose practical features for text classification. Four widely used filter ranking techniques are employed in the first stage to limit the second stage's search for PSO. To choose features in a text classification experiment for big data, term frequency-inverse document frequency (TF-IDF) and cat swarm optimization (CSO) have been proposed in [32]. The findings demonstrate that feature selection is more precise when TF-IDF and CSO are used together than when TF-IDF is used alone. A moth flame optimization (MFO) technique is put forth as a search method FS framework in this work [33] to increase the classification tasks in medical applications.

In the relevant works, all multi-stage algorithms produce accurate classification results. The chosen features subset, furthermore, is also less, thus, the classifier processing is more quickly and produces results that are more accurate when using the first stage's filtered features [31]. As a result, there is growing interested in creating frameworks and various modification techniques for the automatic discovery and elimination of pointless features.

## 3. Sparrow search algorithm (SSA)

Under the aegis of meta-heuristics and computational intelligence, Jiankai Xue introduced the SSA in 2020. It is a new swarm intelligence optimization method that draws inspiration from the

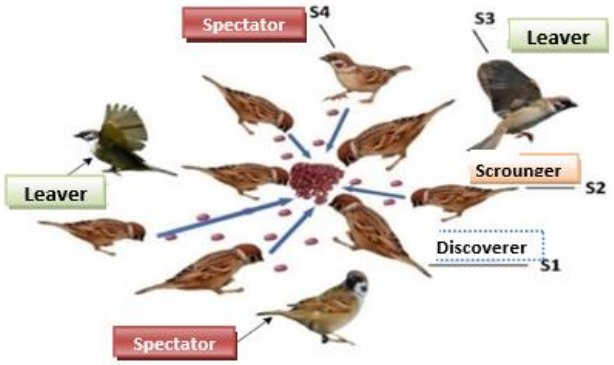


Figure. 1 Schematic diagram of Interrelationships between individuals in the sparrow's family

foraging habits of sparrows [15]. SSA can be used to resolve optimization issues in a variety of domains because of its attributes of simplicity, sparsity of parameters, and good expansibility [34]. According to their activities when looking for food, sparrows are typically divided into producersو and scroungersو as seen in Fig. 1. Scroungers S2 follow the producers in foraging, discovering, and collecting the food, whereas producers S1 often have higher energy reserves and are in charge of looking for potential food sources in the population and giving instructions to the entire flock. However, the type of identity of sparrows typically changes at any time between producer and explorer as the order of fitness changes, to find food [17]. Fig. 1 shows sparrow S4 observing its surroundings as the other sparrows continue to eat and keep a watch on him. Therefore, when the S4 chirps as a warning signal, the flock will fly away from the source of danger to another safe region for food, so, they are constantly shifting their position toward the center in quest of a better (safer) place. A sparrow in the foraging area's most hazardous boundary with the highest likelihood of flying elsewhere is represented by S3 in Fig. 1.

The mathematical model of the SSA can be developed in light of the previous sparrow description. Assume that there are $N$ sparrows in a $D$ -dimensional search space and that the $i$ - $th$ sparrow's position in the search space equals $X_i = [x_{i1}, x_{i2}, \dots, x_{id}, \dots x_{iD}]$. So, the location of the flock represented by $X$ as a vector and which contains $N$ of sparrows will be formed as a multidimensional matrix.

In SSA, the objective function of each sparrow is represented by the value of each row in $F(X_i) = f[x_{i1}, x_{i2}, \dots, x_{id}, \dots x_{iD}]$. Producers with higher fitness will have priority in obtaining food during the search process. Thus, producers should continuously update their position in relation to each dimension j during each iteration t using the following expressions:

$$X_{id}^{t+1} = \begin{cases} X_{id}^{t}.exp\left(\frac{-i}{\alpha.T}\right), & R_2 < ST, \\ X_{id}^{t} + Q.L\,, & R_2 \geq ST, \end{cases} \quad (1)$$

Where $T$ denotes the maximum number of iterations and $d \in \{1,2,\dots,D\}$ ; $t$ represents the current number of iterations. $Q$ is a random number with a normal distribution, and $\alpha \in (0,1]$ is a uniform random number; $R_2 \in [0,1]$ And $ST \in [0.5,1]$ indicate the warning(alarm) value and the safety threshold, respectively. $L$ is a matrix of size $1 \times D$ with all members being 1. $R_2 < ST$ implies that there are no predators near the foraging region and that producers can conduct greater search operations. On the contrary, $R_2 \geq ST$, the flock's detecting sparrow has identified the presence of predators and immediately warns the other sparrows. Therefore, all sparrows must swiftly leave for safer locations while changing their search method.At the same time, some scroungers may constantly observe the producers and compete to discover a suitable food source in an effort to boost their predation rate. Scroungers use the formula below to modify their position:

$$X_{i,d}^{t+1} = \begin{cases} Q.\exp\left(\frac{x_{worst,d}^{t} - x_{i,d}^{t}}{i^2}\right) & if\ \ i > \frac{N}{2}, \\ X_{p,d}^{t+1} + \left|x_{i,d}^{t} - x_{p,d}^{t+1}\right|.A_{0,d}^{+}\,.L\ if\ \ i \leq \frac{N}{2}, \end{cases} \quad (2)$$

where, $x_{worst,d}^{t}$ indicates a sparrow's current global worst position in the $d$-th dimension in the flock's $t$-th iteration, and in $t+1$-th iteration of the flock, $X_{p,d}^{t+1}$ represents the producers' current best

Table 1. SSA parameters summary

| Notation | Parameter description | Type |
|---|---|---|
| $\boldsymbol{T}$ | Maximum number of iterations | static |
| $\boldsymbol{t}$ | Current number of iterations | dynamic |
| $\boldsymbol{Q}$ | Random number | dynamic |
| $\boldsymbol{\alpha}$ | Uniform random number | dynamic |
| $\boldsymbol{R_2}$ | Warning value | dynamic |
| $\boldsymbol{ST}$ | Safety threshold | dynamic |
| $\boldsymbol{L}$ | 1×d vector with a value of 1 | static |
| $\boldsymbol{X_{worst}}$ | Current worst location | dynamic |
| $\boldsymbol{X_p}$ | Current best location | dynamic |
| $\boldsymbol{A}$ | A vector of 1×d, and the elements in A are 1 or -1. | static |
| $\boldsymbol{\beta}$ | Random number β~N(0,1) | dynamic |
| $\boldsymbol{f_i}$ | Current fitness | dynamic |
| $\boldsymbol{K}$ | Random number | dynamic |
| $\boldsymbol{\varepsilon}$ | Small constant | static |

position at the $d$-th dimension. $A$ displays a matrix of $1 \times D$, in which each component is randomly allocated to either 1 or -1, and $A^{+} = A^{T}(AA^{T})^{-1}$. $L$ symbolizes a $1 \times D$ matrix, where each entry is 1. When $i > \frac{N}{2}$, the i-th starving scrounger sparrow has a low fitness value and is in a condition of starvation; otherwise, it would fly to another site to compete for food and raise its fitness value. There are some sparrows in the population iteration process are more vigilant than others, which normally comprise up 10% to 20% of the whole swarm. The movement location of these sparrows is updated at random by Eq. (3):

$$x_{i,j}^{t+1} = \begin{cases} x_{worst,j}^{t} + \beta.\left|x_{i,j}^{t} + x_{best,j}^{t}\right| & if\ f_i > f_g, \\ x_{i,j}^{t} + K.\left(\frac{\left|x_{i,j}^{t} + x_{worst,j}^{t}\right|}{(f_i - f_w) + \varepsilon}\right) & if\ f_i = f_j, \\ x_{i,j}^{t} & if\ f_i < f_j. \end{cases} \quad (3)$$

Where, $X_{best}$ is the present global optimal location. The step control parameter is represented by $\beta$ the, is a random number $\beta \sim N(0,1)$, it adheres to a normal distribution and has a mean of 0 and a variance of 1. $K \in [-1,1]$. $f_i$ is the $t$-th generation sparrow's present fitness. $f_g$ and $f_w$ are the greatest fitness values and the worst fitness values for the present sparrow flock, respectively. When, the $f_i = f_g$, the sparrow is in the best location, it still flies around due to competition for food; when $f_i \neq f_g$ appears that the sparrow has leave the flock center and thus becomes susceptible to predators. The steps of implementation of the proposed SSA are exhibited in algorithm 1. Additionally, Table 1 lists all of the parameters of SSA.

## 4. Materials and methods

The methodology of this study passes through several stages as displayed in Fig. 2. Data collection, data pre-processing, feature extraction, feature selection, classification, and performance evaluation are the six processes that are examined in this research study.

### 4.1 Data collection

Two datasets relevant to the COVID-19 coronavirus were gathered, documented, and examined in order to be used in the investigations. Even though the strategy used several experiments to generate outstanding findings, it was often limited in the case of COVID-19 disease due to a lack of datasets [35]. The first dataset (DS1), which was acquired from numerous hospitals in Iraq, had patients with COVID-19 who tested positive by throat swab utilizing real-time reverse transcription-polymerase chain reaction (RT-PCR) testing. The sample consisted of 3053 cases, which were chosen at random from patients admitted to the referenced hospital between the end of June 2020 and the middle of December 2020. Whereas the second dataset (DS2) was gathered from a variety of sources, including GitHub, the “Italian society of medical and interventional radiology” (SIRM), and other case reports gathered from COVID-19-related medical articles on various websites like Hindawi. Lastly, the DS2 has consisted of 1446 case reports. Patient "demographic" data, including age, sex, and comorbidities, are included in both datasets (DS1 and DS2). In addition to other necessary diagnostic data and associated tests, such as symptoms, lab findings, vital signs, values from regular blood tests, and results from chest CT imaging, and others. Table 2 provides a description of each dataset. This dataset has been made publicly available at https://github.com/AmirYasseen/Clinical-Textual-Datasets-Of-Coronavirus.

Table 2. Information on datasets

| Name | Type | Label | No of samples | Rate |
|---|---|---|---|---|
| **DS1** | Textual Data | Severe<br>Non-Severe | 3053 | 55%<br>45% |
| **DS2** | Textual Data | Positive<br>Negative | 1446 | 62%<br>38% |

### 4.2 Clinical text pre-processing

The COVID-19 datasets collected were not written in standard language and especially dataset of Iraq was poorly structured. A number of pre-processing processes were performed to enhance the quality of the data and construct the feature vector because the clinical language was unstructured and complex.

The difficulties with Arabic slang words were handled in this study, and changed to English, especially with the first database that was gathered from Iraqi hospitals, then, case transformation, normalization, tokenization, stop words being removed, Pos's tagging, stemming, and lemmatization. After conducting the above-mentioned stages of pre-processing, the next step was to filter the clinical text from words that lack meaning and do not meet the requirements and create tokens from the word collection, meaning the filter of words by length. This stage also decreased

Received: March 9, 2023. Revised: April 24, 2023. 

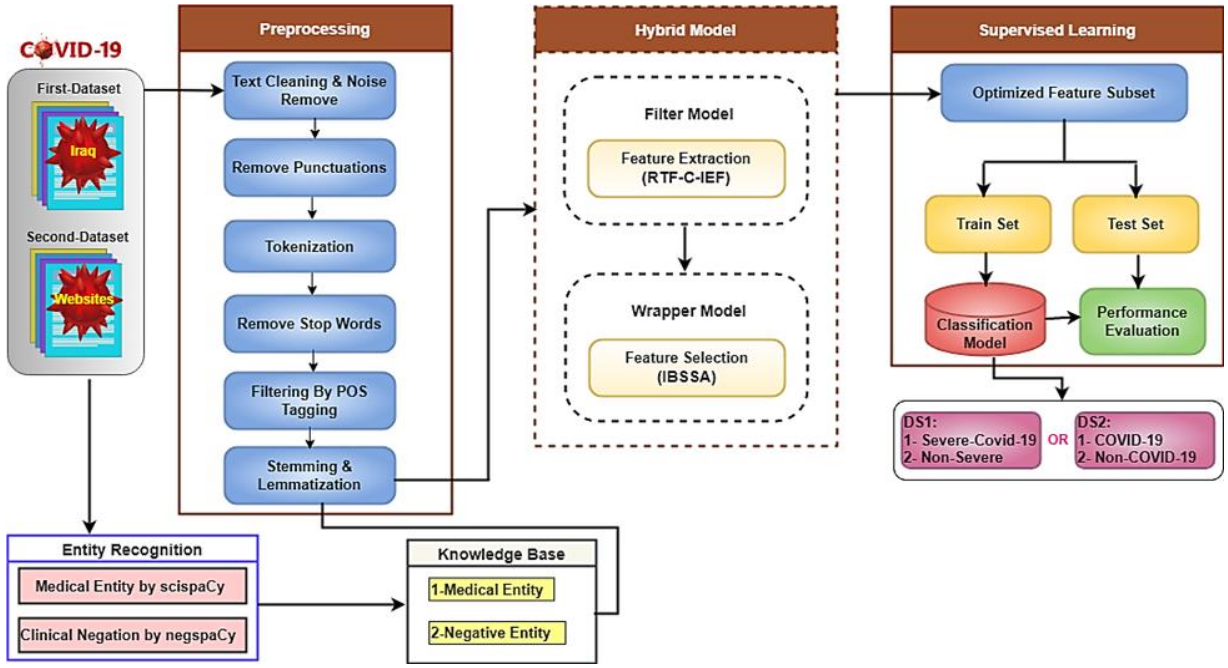


Figure. 2 An illustration of the study's workflow

the dimensions of the clinical text. For example, words with an arrangement of fewer than 3 letters were disqualified.

### 4.3 Feature extraction

From the clinical texts that have been processed, different features are extracted according to the semantics and are converted into likelihood values to be prepared for the feature choice model. In this work, feature engineering process makes use of several steps.

#### 4.3.1. SPACY and SCIPACY

In the stage step, medical terms from clinical text were extracted using the SpaCy and ScispaCy models. It is fast models for Biomedical Natural Language Processing [36], a powerful rule-matching engine, and an ontology framework containing medical vocabulary concepts are all provided. In this work, the clinical texts serve as ScispaCy's input, and the program's output are concepts from meaningful phrases that were found. Additionally, negation detection was targeted to identify the pertaining conditions for a valuable clinical decision support. In some instances, symptoms comprising many words were reduced to a single expression, such as "shortness of breath". Finally, the RTF-C-IEF measure will be used to convert the extracted terms, which are features, into a vector.

#### 4.3.2. RTF-C-IEF method

The RTF-C-IEF statistical weighting approach is used as the first step in the feature selection process for text mining to assess the importance of a term in a document. Since bag of words (BoW) and TF-IDF are less accurate than RTF-C-IEF, it was chosen as the feature extraction method [37]. RTF-C-IEF converts texts into vectors so that machine learning can process the COVID-19 clinical content properly. The feature with the highest RTF-C-IEF score throughout the entire document collection is identified. The more significant a feature is for a specific text document, the higher its RTF-C-IEF score. The RTF-C-IEF equation is as follows:

$$\mathrm{RTF-C-IEF} = \left(tf_{ij}\right)^{r_{tf}} \times \left(1 + \frac{t_x}{N}\right) \times e^{-\frac{dt(t_j)}{N}} \quad (4)$$

Where $tf_{ij}$ is the term frequency, $r_{tf}$ describes the TF distortion restriction, their predetermined default value is 0.8, $N$ is the sum of a patients records, $t_x$ is counts of frequency of the word $x$ there, and $dt(t_j)$ is the frequency of records of patients where the term $t_j$ show up in the collection.

### 4.4 Feature selection based on IBSSA

FS is an essential step before conducting the categorization, as was already indicated. The feature selection technique is designed to eliminate the least affected features and select important features before the classification stage to avoid the problem of overfitting and improve the accuracy of the diagnostic model [38]. The key problem with the suggested methodology is choosing the best features

**Algorithm 1**: Standard sparrow search algorithm

**Input:**
$\boldsymbol{T}$: $maximum\ number\ of\ iterations$;
$\boldsymbol{N}$: $number\ of\ sparrows\ overall$;
$\boldsymbol{PD}$: $number\ of\ producers$;
$\boldsymbol{SD}$: $number\ of\ scroungers$;
$\boldsymbol{ST}$: $safety\ threshold$
**Output:**
$X_{best} - Global\ optimal\ position$
$f_{best}$
$- Fitness\ of\ global\ optimal\ position$

**Start**
Initialize N sparrows plus its parameters.
$n \leftarrow 1$;
**While** $n < T$ **do**
  Rank the fitness values f(x);
  Find the current best individual and worst individual $X_p$ and $X_{worst}$;
  $R_2 \leftarrow rand(1)$, /* Randomly choose an alert value between [0, 1] */;
  **for** producer $i = 1,2, \dots, PD$ **do**
    Update the producer's position using Eq. (1);
  **end-for**
  **for** sparrow $i = \boldsymbol{PD} + 1, \boldsymbol{PD} + 2, \dots, \boldsymbol{N}$ **do**
    Update the sparrow's location using Eq. (2);
  **end-for**
  **for** scrounger $i = 1,2, \dots, \boldsymbol{SD}$ **do**
    Change the scrounger's location by Eq. (3);
  **end-for**
  Discover the most recent location(current) $X_i^{n+1}$
  Update the current position when is superior to the previous one ($X_i$, $f_i$).
  Re-rank the entire swarm according to the fitness values f(x) in ascending order.
  Search the current global optimal position $X_{best}^{n+1}$
  $X_{best} \leftarrow X_{best}^{n+1}$;
  $f_g \leftarrow f(X_{best})$
  $n \leftarrow n + 1$
**end-while**
**End**
**Return** $f_g$, $X_{best}$ /*$X_{best}$: Optimal outcome */

from clinical texts for COVID-19 diagnosis. In order to enhance the diagnostic model performance and make it a faster, more efficient model, feature selection should be done before learning the model [39]. Due to the canonical SSA algorithm's meticulous design and great performance in balancing the capabilities of exploration and exploitation, this work uses SSA as a search approach to resolving FS problems from clinical texts. It is noteworthy that several modification tactics were employed to alleviate the shortcomings of the algorithm and increase its performance in solving the FS problem. This section describes the improvements built into the default SSA algorithm. The first is a novel initialization modification technique called MIA that was introduced to the conventional SSA algorithm in order to start with high-quality individuals and therefore raise the probability of finding the optimum solution, which may improve the effectiveness of the optimization. Second, to boost diversity and the optimizer's capacity to explore additional areas of the search space, each sparrow is updated via integration with the Levy flight operator. The final enhancement, the usage of the LSA algorithm, helps the SSA exploitation phase avoid becoming stuck in local optima. In this part, we discuss these encouraging improvements. The suggested feature election technique framework is depicted in Fig. 5, and Algorithm 4 displays the IBSSA pseudo-code.

#### 4.4.1. Modified initialization approach

Population initialization is a critical factor in evolutionary algorithms, which considerably influences the diversity and convergence during the process of searching. The goal of this step is to provide a preliminary guess at probable solutions. Then, during the optimization process, these initially hypothesized solutions will be iteratively improved up until a stopping requirement is fulfilled. In general, individuals from the initial population with high-quality can discover the optimum location and hasten the convergence of the algorithm. On the other hand, using poor guesses at the outset can hinder the algorithm from discovering the optima [40]. Recent studies have shown that initialization strategies can increase the likelihood of discovering global optimums and decrease the variation of search results [41]. Moreover, the sparrow search method has a limited number of population roles and requires extensive initial optimization. In this paper, a novel initialization approach called MIA is introduced, which improves the efficiency of SSA to make it suitable for the optimization problem. Its fundamental concept is to generate a population on the basis of the first population in a simple mathematically manner without using complicated

**Algorithm 2**: A suggested MIA algorithm

$\boldsymbol{X_{ij}}$ = Position of sparrows; /* *N positions should be generated randomly;*
$\boldsymbol{X_{bin}}$ = When binary_map is achieved ($\boldsymbol{X_{ij}}$);
$\boldsymbol{Fit_{old}}$ = The fitness of all members of the population(sparrows);
$D_{max}$= Maximum number of iterations performed locally;
$M_{max}$= Maximum number of iterations performed locally;
N (populace size).

```
for d = 1 To D_max do
    Find X_bin−best /* ("Global optimal position")
    for i= 1 To N
        X_new ← (X_bin−best + X_ij) * rand; /* create a new location;
        X_bin−best ← binary_map (X_new)
        Compute the values of each Sparrow's fitness function Fi
        if F_i < Fit_old then
            Fit_old ← F_i
            X_ij ← X_new
            X_bin ← X_bin−best
        end-if
        for m= 1 To M_max do
            randomfeat = rand( ) ; / * features chosen at random, ∈ {0,1} */
            Compute the values of each Sparrow's fitness function;
            F_m (randomfeat)
            if F_m < Fit_old then
                Fit_old ← F_m
                X_bin ← randomfeat
            end-if
        end-for
    end-for
end-for
Return X_bin, X_ij, Fit_old
```

equations or significantly altering the structure of the original SSA algorithm. Then, the best individuals from the initial population will be found, and as a result, a new initial population consisting of excellent individuals is obtained. Thus, the MIA was able to control a portion of this algorithm and accurately cover the potential area. Algorithm 2 displays the entire MIA pseudo code. The suggested initialization approach also has a big effect on how good the solution, finds the best solution extremely effectively, and has contributed to increasing the chance of an initial global optimum.

#### 4.4.2. Local search based on levy flight strategy

Levy flight(LF), which is shown in Fig. 3, is a mathematical model of a random movement that complies with a distribution of possibilities [42]. Newly, it was proposed as an alternative solution to address optimization issues and has been combined into the architecture of several swarms methods to improve their ability in the fast of convergence, hop from local minima, and exploration and exploitation balancing [33, 43]. LF is suggested in this study as a way to improve the SSA optimizer's performance by integrating it into the SSA structure, preventing the algorithm from entering a local optimum when faced with complex problems of high dimensions. Thereby, improving the process of selecting features from clinical texts for diagnosing COVID-19. Therefore, in order to expand the search area, the producer's location is adjusted in Eq. (3) to update the location of the sparrows based on the Levy flight improvement expressed by Eq. (5). As a result, each modified sparrow is scheduled to employ LF once to increase search space diversity. Hence, more randomness will be obtained, and effectively lower the likelihood of the algorithm falling into the local optimum, resulting in a deeper level of exploration.

$$\boldsymbol{X_{i,j}^{t+1}} = X_{i,j}^{t} + \alpha \oplus levy(\beta) \tag{5}$$

$$\boldsymbol{Levy(\beta)} \sim \boldsymbol{\mu} = t^{-1-\beta} \qquad 0 \le \beta \le 2 \tag{6}$$

$$\boldsymbol{levy(\beta)} \sim \frac{\phi \times \mu}{\left|\nu^{1/\beta}\right|} \tag{7}$$

$$\boldsymbol{\phi} = \left[\frac{\Gamma(1+\beta)\times sin(\pi\times\beta/2)}{\Gamma(((1+\beta)/2)\times\beta\times 2^{(\beta-1)/2})}\right]^{1/\beta} \tag{8}$$

Where $\boldsymbol{X_i^t}$ denotes the $i^{th}$ sparrow at repetition $t$, rand denotes a a number chosen at random between [0, 1], the dot product is represented by $\oplus$, and $\alpha$ is parameter of control at step. Levy flight is a type of random walk that, as was already mentioned, supports a Levy distribution according to the formula in Eq. (6). Levy is computed using Eq. (7) as stochastic numbers, ν and μ are common distributions at random. Equation (8) demonstrates how to compute φ, where β = 1.5, indicated in [44], and Γ represents a common Gamma function.

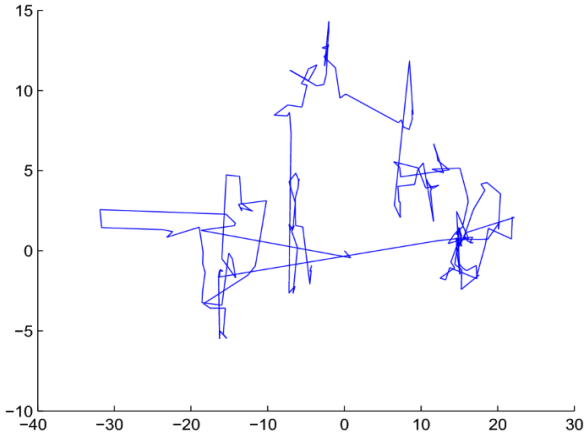

Figure. 3 Random walk using levy flight

**Algorithm 3**: A suggested LSA algorithm

$\boldsymbol{LT}$ − *Maximum number of iterations locally;*
$X_{best}^{t+1}$ /* the better location yet at the conclusion of this iteration $t+1$; $Temp \leftarrow X_{best}^{t+1}$; $Ln = 0$.

**While** $Ln < LT$ **do**
  $randfeat \leftarrow four$; /*choice of a feature at random from Temp*/.
  **for** feature $\alpha \in randfeat$ **do** /* $\alpha \in \{0,1\}$
    $\alpha \leftarrow \neg\alpha$;
  **End-for**
  $f(Temp) \leftarrow calcualte\ the\ fitness\ value$;
    **if** $f(Temp) < f(X_{best}^{t+1})$ **then**
      $X_{best}^{t+1} \leftarrow Temp$
      $\boldsymbol{X_{best}} \leftarrow X_{best}^{t+1}$
      $f_g \leftarrow f(\boldsymbol{X_{best}})$
    **end-if**
  $Ln \leftarrow Ln + 1$
**end-while**
**Return** $\boldsymbol{X_{best}}$ , $f_g$

#### 4.4.3. Improving the exploitation based on local search algorithm (LSA)

The transition of sparrows from initial random locations to better locations using the initialization algorithm MIA, results in the identification of the current best location, which is the sparrow with the current best fitness value. Then, optimization is done on the location by calling the LSA algorithm more than once. LSA is a new algorithm presented and developed by [18], and as described in Algorithm 3. The aim of this algorithm is to eliminate any remaining potentially irrelevant features.

The first call is after the initialization process; and in each current iteration of the sparrow t+1, LSA is called once more to improve the current better solution $X_{ij}^{t+1}$ obtained. Initially, the LSA algorithm creates a temporary variable called $Temp$ which stores the value of $X_{best}^{t+1}$ which is generated at the conclusion of each IBSSA repetition. LSA iteratively runs $LT$ times, to improve $Temp$. At each iteration $L_t$ of LSA, four features' $rand-feat$ are randomly chosen from $Temp$. LSA reverses the value of each variable in $rand-fea$. Subsequently, the value of fitness $f(Temp)$ of the new solution (the new *Temp*) is assessed; if it is better than $f(X_{best}^{t+1})$, then $X_{best}^{t+1}$ is set to $Temp$; else, $X_{best}^{t+1}$ and $f_g$ are maintained as-is.

#### 4.4.4. Binary discrete mapping

SSA cannot be utilized to directly solve an FS problem since the sparrow search algorithm's search domain is the real number domain in the continuous space. Therefore, in the application of feature selection and selection, a transfer function ought to utilize to convert the continuous values into binary 1 or 0 to signal whether the feature is used or not; 0 implies discard, and 1 signifies use. In this work, one of the S-shaped family functions is adopted as in Fig. 4, which has been widely used and is ideal for the solution mappings, because it generates outputs in the range [0,1], the details of this function are as follows [43, 45]:

$$TF\left(x_i^d(t)\right) = \frac{1}{1+e^{-2x_i^d(t)}} \tag{9}$$

where $x_i^d$ indicates the $i^{th}$ sparrow's position in the $d^{th}$ dimension at the $t^{th}$ iteration, $x_i$ is calculated by Eq. (1,2,3). The output of the S-shaped function is still displayed in a continuous manner in Eq. (11). Thus, to get the binary value the $i^{th}$ position is modified in the following way:

$$x_i^d(t+1) = \begin{cases} 0 & rand < TF\left(x_i^d(t)\right) \\ 1 & rand \geq TF\left(x_i^d(t)\right) \end{cases} \tag{10}$$

Where $x_i^d(t+1)$indicates the $ith$ feature in the $X$ solution at Dim $d$ in repetition "$t+1$", and rand is value between [0,1].

### 4.5 Fitness function: A two-stage approach

A function to assess the effectiveness of the optimization technique is the fitness function. This feature selection seeks to select a subset of features to increase prediction accuracy and reduce the

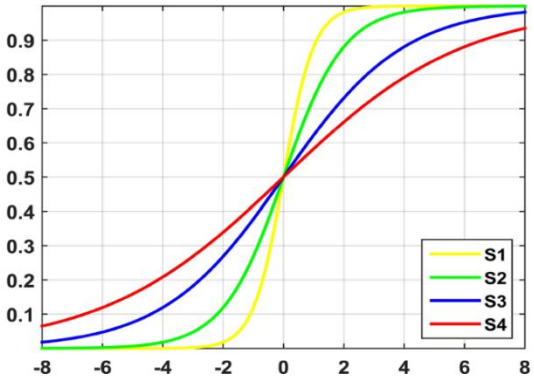

Figure. 4 Transfer function families S-shaped

number of features without noticeably lowering the prediction accuracy of the classifier constructed using only the selected features. For this reason, we employed the composite fitness function, which designs a fitness function based on two criteria: classification accuracy and a number of features elected. Thus, an individual with high classification accuracy and a limited number of features will produce a high fitness value, which should address the issue of choosing redundant and pointless features for the feature subset [46]. The evolutionary process is separated into two equal stages in our two-stage feature selection technique. The algorithm's first step focuses on reducing the classification error rate. The fitness function in the second step takes into account the number of chosen features. In the first direction, the goodness of an individual is assessed under the fitness function through the classification accuracy of a predictive model that utilizes the representation produced from feature extraction by the TWS. Therefore, we must seek out an effective classification model that can also handle the high-dimensionality of the data in a natural manner. Supportive Vector Machines (SVM) is a kind of machine learning algorithms that has proven very effective for text classification [47]. We chose SVM over alternative methods because it naturally deals with the sparseness and high dimensionality of data, as seen, for example, in [48, 49]. Therefore, the resulting values of the scale F1 by the SVM classifier represent the values of the fitness function in the optimization process for the first stage. When dealing with unbalanced data sets, it is recognized that this method of estimating the F1 measure is especially helpful[50]. The best results of the first stage are used to start the second step, which guarantees that feature minimization is based on feature subsets that leads to effective classification.

This two-stage feature selection approach's fitness function is demonstrated by developing a single objective fitness function that unifies the several objectives into one. As defined by formula (11). Eq. (11) takes both features count and classification performance into account.

$$\boldsymbol{fitness_i} = (\gamma \times ErrRate_i) + \beta \times \frac{|n|}{|N|} \qquad (11)$$

Where $n$ and $N$ represent the size of features in the selected feature subset and all features in the dataset, respectively, whereas $\gamma$ is constant values and $\gamma \in [0, 1]$. $\gamma$ shows weigh the importance of classification accuracy, and $\beta = (1 - \gamma)$ shows the weight for selected features. As for $\gamma$ and $\beta$, it has been prescribed that $\gamma$ = 0:99 and $\beta$ =0:01, based on comprehensive tests from earlier studies[51, 52]. $ErrRate_i$ is the error rate of the SVM classifier on the dataset of training using a set of features elected by IBSSA.

### 4.6 Representation of FS-IBSSA solution

Following the extraction of the features from clinical texts for COVID-19 patients and those without COVID-19, the gathered dataset must be passed to FS-IBSSA for electing the best features on cases of COVID-19. The FS-IBSSA relies on using SSA since it is an optimization method and adaptive search heuristic algorithm that mimics the intelligence of swarms.

Initially, the IBSSA creates a swarm of N sparrows (search agents) randomly in feature space. Each search agent represents a potential solution (i.e., a sub-set of informative features) in a D-dimensional search space that, in the feature selection paradigm, equalizes the initial number of features displayed in the COVID-19 dataset, hence; It is possible to express feature set (F) with 'n' features as $F = \{f_1, f_2, \dots, f_n\}$. Prior to starting the fitness evaluation procedure, the initial location of every sparrow in the swarm is discretized at each dimension, taking either 0 (elimination of the feature) or 1 (selection of the feature), in accordance with Eq. (10), to produce random binary values (which may be zero or one). Assume that ("e.g., n = 20; the number of extracted features from the clinical texts"), thus, a single agent(sparrow) can be represented as; $\{f_1, f_2, \dots, f_{20}\}$, as shown in Table 3.

Inclusively, the locations of sparrows within IBSSA are updated in accordance with Eqs. (1), (2), or (3). The components of this matrix are altered one at a time using for-loops based on their value in the preceding iteration and some random sample numbers. Thus, the entire matrix is updated simultaneously (updating the entire swarm). Fig. 5 shows the flowchart of the SSA algorithm.

Table 3. The representation of a single solution

| $S_i/f_i$ | $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ | $f_6$ | … | $f_{D-2}$ | $f_{n-1}$ | $f_n$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 0 | 1 | 1 | 0 | 1 | 0 | … | 1 | 0 | 1 |

It is important to note that, following each iteration in which the location is changed, the continuous values of the location vector are maintained for use in subsequent iterations in which the position is updated continuously. These values are likewise discretized utilizing Eq. (10), allowing us to assess the binary solution's fitness value based on the categorization error rate attained by the involved classifier utilizing the features chosen by IBSSA. Following, this process iterates until it satisfies a stopping requirement. The idea is that more and more sparrows will eventually move towards areas where better solutions are found and that the population will eventually converge to the ideal solution in accordance with an optimization problem's fitness function.

### 4.7 COVID-19 patients categorization based IBSSA-FS

In this stage, the performance of categorizing COVID-19 patients is assessed using a resultant subset of features that was chosen. These techniques are separately used to categorize datasets in which the dimension acquired at the conclusion of the application of swarm algorithms is lowered. The use of a classifier is necessary to compare the effectiveness of the proposed swarming algorithms in categorizing COVID-19 patients from clinical text. Each dataset was randomly split into a training set (80%) and a test set (20%). Using the "sklearn" Python module, the following classifiers were applied to each training set: random forest (RF), logistic regression (LR), and support vector machines (SVM). Each algorithm is thoroughly described in in [53-55].

### 4.8 Evaluation

In this study, several evaluation criteria are employed to validate the effectiveness of the suggested strategy including accuracy (Acc), precision (P), recall (R), F-measure (F1), Macro-F1, and Macro-recall. Are defined as follows:

$$\text{Acc} = \frac{\text{Num.of Correct Predictions}}{\text{Total Num.of Predictions performed}} \tag{12}$$

$$\text{P} = \frac{TP}{TP+FP} \quad , \quad \text{R} = \frac{TP}{TP+FN} \tag{13}$$

$$\text{F1\_score} = 2 \times \frac{\text{P} \times \text{R}}{\text{P}+\text{R}} \tag{14}$$

$$MacroF = \frac{1}{T}\sum_{j=1}^{T} F_j \tag{15}$$

$$MacroR = \frac{1}{T}\sum_{j=1}^{T} R_j \tag{16}$$

Where $T$ is the overall number of categorized classes and $F_j$, $R_j$ represent the F and R values for the $j^{th}$ class, respectively. Additionally, in order to raise the statistical significance of the empirical findings, every optimization algorithm is assessed 20 times separately for each dataset. To this purpose, the following important performance measures for the FS issue are adopted: average classification accuracy, selection ratio, average fitness, and standard deviation (STD).

$$\mu_{\text{feat}} = \frac{1}{20}\sum_{\text{k}=1}^{20} \frac{\text{d}_*^{\text{k}}}{\text{D}} \tag{17}$$

$$\mu_{\text{fit}} = \frac{1}{20}\sum_{\text{k}=1}^{20} \text{f}_*^{\text{k}} \tag{18}$$

$$\text{SD} = \sqrt{\frac{1}{19}\sum_{\text{k}=1}^{20}\left(\text{Y}_*^{\text{k}} - \mu_{\text{Y}}\right)} \tag{19}$$

## 5. Results and analysis

This section provides a complete empirical analysis of the behaviour of the IBSSA optimization algorithm based on several phases of development. In experiments, two medical data sets pertaining to COVID-19 patients are employed. Table 2 provided a description of these data sets' specifics.

### 5.1 Adjusting parameters

As is common knowledge, it is difficult for a metaheuristic algorithm to deliver excellent results on all optimization tasks, especially while utilizing the same parameter values. Therefore, it is preferable to adjust the crucial parameters for each optimization issue separately in order to achieve the greatest results. The results of parameter tuning are presented in Table 4. Every combination is individually run 20 times to avoid random bias, and the average outcomes are displayed. Furthermore, this study compared the suggested method with the most recent wrapper methods, including PSO, GWO, MVO, WOA, MFO, and FFA. All algorithms were implemented using the same computing platform to provide fair comparisons (Windows 10 OS 64bit, having a CPU of Intel(R) Core i7 processor, 2.20 GHz, and RAM of 16GB), as well as the identical values for each algorithm's parameters. The

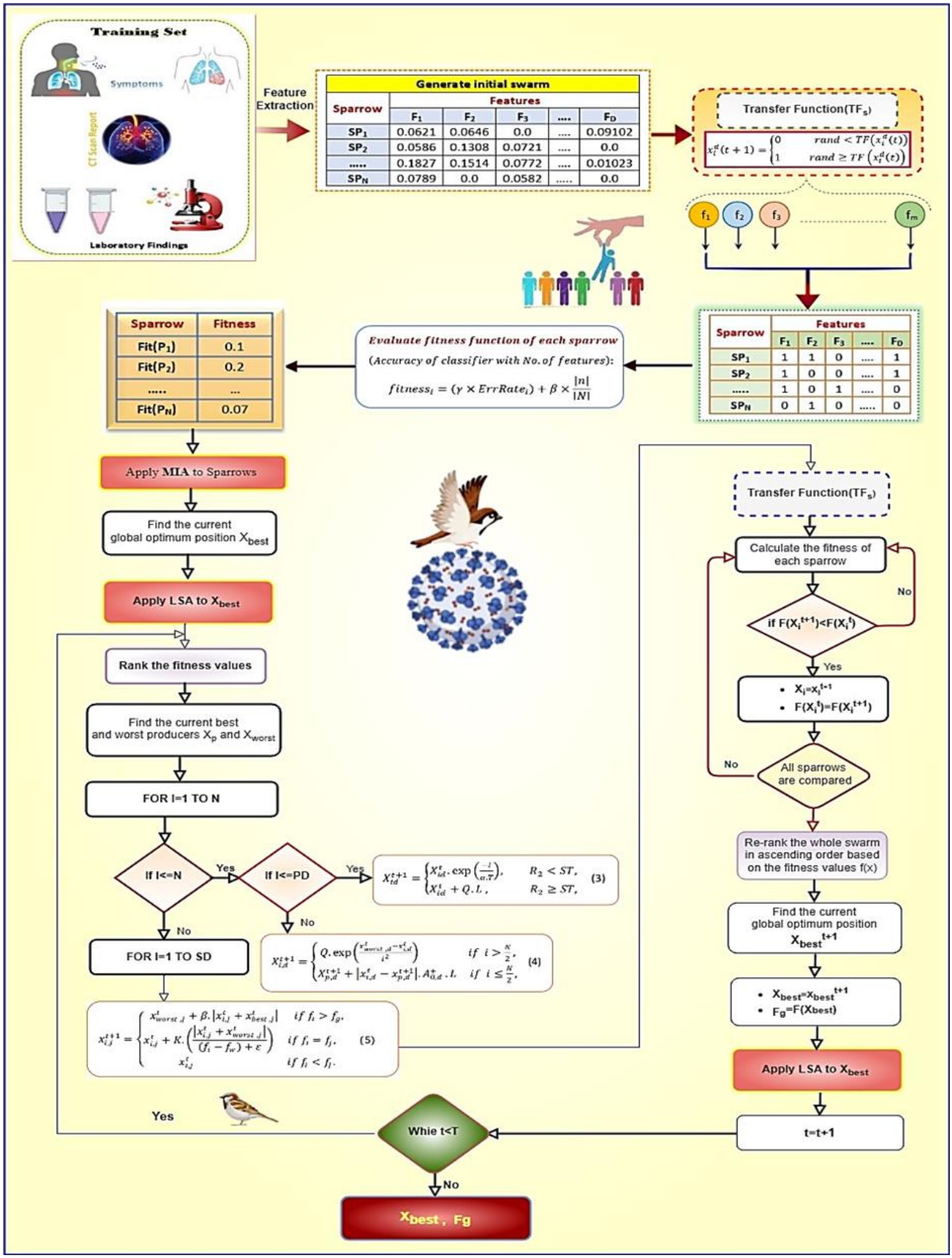


Figure. 5 The flowchart of IBSSA-FS algorithm

**Algorithm 4**: A suggested IBSSA based on levy flight, MIA, and LSA

**Start**
Initialize N sparrows and its parameters; /* Randomly generate the positions of N sparrows $X_{ij}$*/
Perform binary transformation using Eq. (10);
Determine the level of fitness for each sparrow using Eq.(11);
Apply **MIA** to $X_{ij}$ using Algorithm 2;
Find the current global optimum position $X_{best}$ and $F_{best}$ (Fitness of global optimal position)
Apply **LSA** to $X_{best}$ using Algorithm 3;
$tn \leftarrow 1$;
**While** $tn < T$ **do**
  Rank the values of fitness f(x) according to Eq.(11);
  Find the best and worst individual currently $X_p$ and $X_{worst}$;
  $R_2 \leftarrow rand(1)$; /* Randomly choose an alert value between [0, 1] */;
  **for** producer $i = 1,2, \dots, PD$ **do**
    Change the producer's location by Eq. (1);
    Levy flight is applied to modify the location of each sparrow;
  **end-for**
  **for** sparrow $i = PD + 1, PD + 2, \dots, N$ **do**
    Change the sparrow's location by Eq. (2);
  **end-for**
  **for** scrounger $i = 1,2, \dots, SD$ **do**
    Update the scrounger's location using Eq. (3);
  **end-for**
  Perform binary transformation using Eq. (10);
  Determine the level of fitness for each sparrow using Eq.(11);
  Find the current new location $X_i^{t+1}$
  **if** $f(X_i^{t+1}) < f(X_i^t)$ **then** /*if the current position is superior to the previous one, update it*/
    $X_i \leftarrow X_i^{t+1}$
    $f_i \leftarrow f(X_i^{t+1})$
  **end-if**
  Re-rank the entire swarm according to the fitness values f(x) in ascending order;
  Search for the current global optimal position $X_{best}^{t+1}$; /* First individual in the ranking*/
  $X_{best} \leftarrow X_{best}^{t+1}$;
  $f_g \leftarrow f(X_{best})$
  Apply **LSA** to $X_{best}$ using Algorithm 3;
  $tn \leftarrow tn + 1$
**end-while**
**End**
**Return** $f_g, X_{best}$ /*$X_{best}$: Optimal solution */

Table 4. Setting parameters for IBSSA

| IBSSA Parameters | Description | Setting |
|---|---|---|
| **N** | Run Time | 20 |
| **Size of the pop (N)** | Num. of sparrows (agents of search) | 50 |
| **$Iter_{max}$** | No. of iterations allowed maximum | 500 |
| **Dim** | Dimension | Size of features |
| **β** | The importance of the subset of features | 0.01 |
| **α** | classification accuracy's significance | 0.99 |
| ***PD*** | The proportion of producers | 0.2 |
| ***SD*** | The proportion of scroungers | 0.1 |

optimization algorithms are most executed in Python in the framework of EvoloPy-FS [56].

### 5.2 Experimental results

The findings of the test datasets that were related to COVID-19 are shown in this section in terms of classification performance. Each value is the average of 20 separate runs of the training/test procedure. Two stages are involved in the execution of experiments. In the first phase, the effect of TWS is studied on datasets, as we seek out the best performance through its integration into the suggested methodology. The suggested IBSSA is compared to various competing wrapper FS techniques in the second stage to demonstrate the strength of the suggested approach. The obtained result from IBSSA, which is the optimal features, is used as an input for the classifiers to use in classifying the patients into the proper groups. Note that, the feature selection stage was clearly detached from the categorization step. We assess the quality of the subsets of features using each SVM, RF, and LR. Here the SVM is used as the baseline classifier. Two crucial metrics are used in these investigations: 1) The number of elected features 2) Classification accuracy.

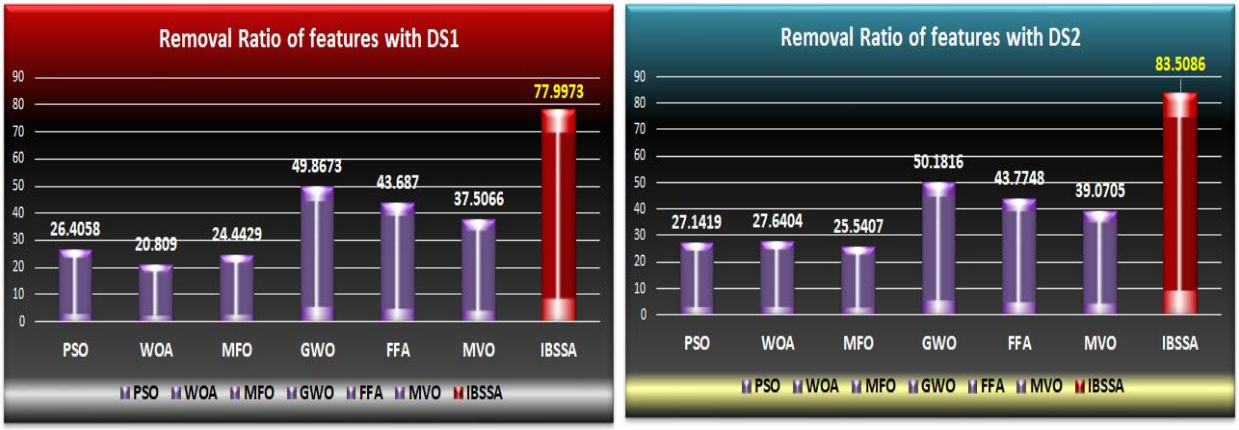


Figure. 6 Average ratio of features removed from DS1 and DS2 by IBSSA

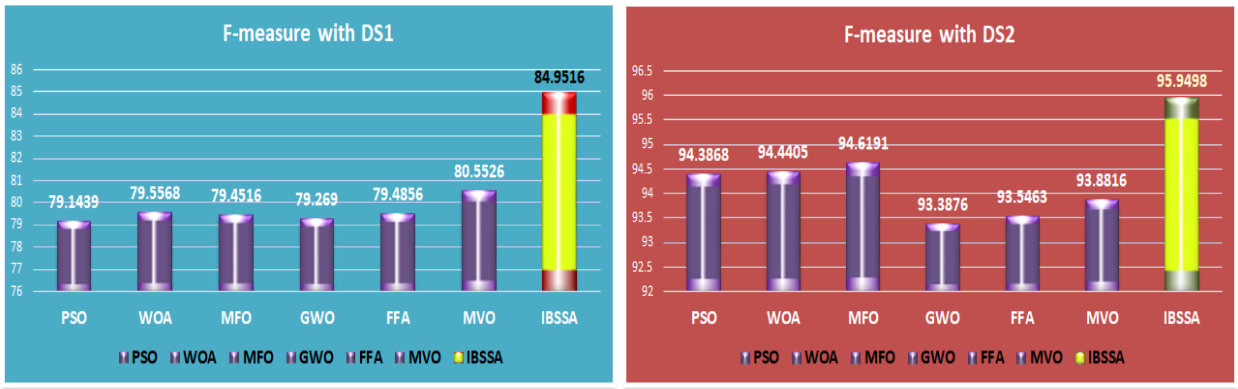


Figure. 7 Average categorization F-measure of IBSSA on DS1 and DS2 compared with other methods by SVM

Table 5. Number of featured extracted by NLP stages

| Dataset | Number of features | Categories |
|---|---|---|
| **DS1** | 377 | High Dimensionality |
| **DS2** | 2367 | High Dimensionality |

In this section, IBSSA performance on the FS problem is evaluated using a variety of metrics, including best fitness value, the average fitness score, worst fitness value, average number of features selected, the standard deviation for the average fitness values, mean accuracy value, and highest accuracy. For a clear presentation, the best results from a particular algorithm are emphasized in bold font.

Table 5 shows how many features were retrieved during pre-processing prior to feature selection, whereas Table 7 shows how many features were selected from datasets produced using various algorithms. This table also displays the numerical statistics results, which are similar to the accuracy results. The table shows that, after 20 runs, IBSSA achieves the best average number of selected features in both datasets (DS1 and DS2), which may be regarded as the greatest performance in the tests when compared to other algorithms. Note that, the accuracy and the number of chosen features are trade-offs, therefore it may be challenging to achieve the optimal outcomes to meet both of these goals for any dataset. However, we can say that the proposed IBSSA performs better than other methods in terms of in terms of choosing features in the selected datasets, as illustrated in Fig. 6.

According to accuracy, precision, and F-measure index, the performance of LR and RF with IBSSA has the best rate as shown in Tables 8 and 9; nevertheless, the difference among the average recall scores in the case of IBSSA and others is very small. Whereas Table 10 demonstrates that SVM with IBSSA has superior efficiency compared to all other classification algorithms, which is another important finding, see Fig. 7.

Classification algorithm performance on the second dataset is displayed in Tables 11, 12, and 11. Tables 9 and 10 demonstrate that, in comparison to all other methods, the classifiers had a promising performance. Whereas the IBSSA has the highest

Table 6. Values of fitness from different methods for DS1 and DS2

| Algorithm | DS1 | | | | DS2 | | | |
|---|---|---|---|---|---|---|---|---|
| | Best | Worst | SD | Mean | Best | Worst | SD | Mean |
| PSO | **11.9508** | **13.3517** | 3.6424 | **12.9468** | 4.6866 | 5.4455 | **2.067** | 5.0539 |
| WOA | 13.1452 | 14.6777 | 3.7754 | 13.7407 | 4.8834 | 5.9688 | 2.5784 | 5.6351 |
| MFO | 12.8370 | 13.7504 | **2.1992** | 13.2715 | 4.7724 | 5.5376 | 2.6126 | 5.3095 |
| GWO | 15.1563 | 16.8318 | 4.8170 | 16.1638 | 6.8914 | 9.0156 | 5.8036 | 8.0924 |
| FFA | 13.8441 | 14.8428 | 2.7810 | 14.3461 | 4.9955 | 6.1279 | 3.5989 | 5.7708 |
| MVO | 12.8401 | 14.0168 | 3.0975 | 13.5622 | 4.6566 | 5.74 | 2.6041 | 5.2127 |
| IBSSA | 12.1983 | 18.085 | 13.0699 | 15.1179 | **2.7354** | 5.9832 | 8.1708 | **4.1745** |

Table 7. The number of elected features by various methods on DS1 and DS2

| Algorithm | DS1 | | | | DS2 | | | |
|---|---|---|---|---|---|---|---|---|
| | Best | Worst | Selection Ratio | Removal Ratio | Best | Worst | Selection Ratio | Removal Ratio |
| PSO | 267 | 302 | 73.5941 | 26.4058 | 1681 | 1773 | 72.858 | 27.1419 |
| WOA | 181 | 324 | 79.1909 | 20.809 | 1156 | 1951 | 72.3595 | 27.6404 |
| MFO | 270 | 304 | 75.557 | 24.4429 | 1669 | 1830 | 74.4592 | 25.5407 |
| GWO | 175 | 208 | 50.1326 | 49.8673 | 1128 | 1245 | 49.8183 | 50.1816 |
| FFA | 197 | 225 | 56.3129 | 43.6870 | 1299 | 1377 | 56.2251 | 43.7748 |
| MVO | 214 | 256 | 62.4933 | 37.5066 | 1398 | 1500 | 60.9294 | 39.0705 |
| IBSSA | **66** | **105** | **22.0026** | **77.9973** | **356** | **440** | **16.4913** | **83.5086** |

Table 8. The classification performance comparison results that were attained by LR with DS1

| Algorithm | Accuracy | | Precision | | Recall | | F- Score | |
|---|---|---|---|---|---|---|---|---|
| | Max | Avg | Max | Avg | Max | Avg | Max | Avg |
| PSO | 79.5247 | 76.5082 | 78.9809 | 75.9353 | 85.3741 | 82.4489 | 81.5780 | 79.0423 |
| GWO | 77.6965 | 75.6307 | 77.3885 | 73.7182 | 88.7755 | **85.0850** | 80.5030 | 78.9576 |
| MFO | 77.3309 | 76.4533 | 76.3975 | 75.4018 | 84.6939 | 83.4183 | 79.8700 | 79.2028 |
| WOA | 78.7934 | 77.0658 | 77.9874 | 76.0964 | 85.034 | 83.6054 | 81.0450 | 79.6699 |
| FFA | 79.7075 | 76.1791 | 79.4212 | 74.6162 | 87.4150 | 84.4387 | 81.6520 | 79.2132 |
| MVO | 77.8793 | 76.4259 | 76.2195 | 75.016 | 86.3946 | 84.1837 | 80.7631 | 79.3323 |
| IBSSA | **83.1502** | **80.815** | **83.1715** | **81.0823** | 88.5906 | 84.6309 | **85.1613** | **82.8012** |

accurate performance when compared to other rivals when using the SVM classifier, as shown in Table 13, and see Fig.7.

In brief, to show the findings, the optimizer IBSSA with SVM has exhibited a superior classification precision in handling all chosen datasets than the other versions utilizing LR and RF classifiers. One reason is which the SVM algorithm provides the over-fitting safeguard, and is not primarily dependent on the number of features being processed. So, compared to other classifiers tested, it has a greater capability for handling the larger text feature spaces. The findings show that the SVM can perform more consistently than other models when dealing with a variety of samples. As a result, when compared to other algorithms, the IBSSA algorithm has the best performance in terms of feature selection accuracy on these chosen datasets. The included enhanced factors may be the cause since they can balance the algorithm's capacities to explore and exploit, which improves algorithm performance.

In contrast, when we employed RTF-C-IEF statistics to determine the weight of each word. These weight words were used as features to create a text dataset. As a result, the number of features decreased by the filter approach (RTF-C-IEF) is relatively small. Table 14 illustrate the results of the RTF-C-IEF and IBSSA feature selection, demonstrating how well the redundant features were eliminated. When features in two datasets are reduced to 67 and 260, respectively, for IBSSA, and

Table 9. The classification performance comparison results that were attained by RF with DS1

| Algorithm | Accuracy | | Precision | | Recall | | F- Score | |
|---|---|---|---|---|---|---|---|---|
| | Max | Avg | Max | Avg | Max | Avg | Max | Avg |
| PSO | 78.9762 | 77.1755 | 77.2871 | 75.3667 | 85.3741 | 83.1632 | 80.3226 | 79.0642 |
| GWO | 77.5137 | 76.1152 | 75.3943 | 73.6087 | 89.1156 | 83.4524 | 80.4992 | 78.1753 |
| MFO | 79.159 | 77.5502 | 76.8519 | 75.0723 | 86.3946 | 84.3027 | 80.5825 | 79.4137 |
| WOA | 79.5247 | 77.989 | 77.7429 | **75.7595** | 87.0748 | 83.8775 | 80.9135 | 79.6005 |
| FFA | 79.8903 | 77.0292 | 79.0323 | 74.6154 | 86.7347 | 83.7925 | 81.1258 | 78.9211 |
| MVO | 78.7934 | 77.3583 | 76.7081 | 75.1695 | 86.0544 | 84.0476 | 80.7018 | 79.3525 |
| IBSSA | **81.685** | **78.2509** | **80.3125** | 75.2832 | **92.2819** | **89.7147** | **84.5201** | **81.8309** |

Table 10. The classification performance comparison results that were attained by SVM with DS1

| Algorithm | Accuracy | | Precision | | Recall | | F- Score | |
|---|---|---|---|---|---|---|---|---|
| | Max | Avg | Max | Avg | Max | Avg | Max | Avg |
| PSO | 79.1590 | 76.5082 | 80.2768 | 75.8037 | 88.4354 | 82.9251 | 81.0631 | 79.1439 |
| GWO | 77.1481 | 75.5210 | 76.0125 | 72.7785 | 89.7959 | 87.1088 | 80.6107 | 79.2690 |
| MFO | 77.8793 | 76.5996 | 76.8750 | 75.2367 | 87.0748 | 84.1836 | 80.8847 | 79.4516 |
| WOA | 78.2450 | 76.9652 | 77.7070 | 76.0817 | 85.7143 | 83.3843 | 80.5873 | 79.5568 |
| FFA | 79.5247 | 76.0146 | 79.3548 | 73.6135 | 88.4354 | 86.4285 | 81.4570 | 79.4856 |
| MVO | 78.6106 | 77.2395 | 76.0479 | 74.4876 | 89.1156 | 87.7041 | 81.7473 | 80.5526 |
| **IBSSA** | **85.8974** | **82.6007** | **83.0671** | **80.4849** | **93.6242** | **89.9664** | **87.874** | **84.9516** |

Table 11. The classification performance comparison results that were attained by LR with DS2

| Algorithm | Accuracy | | Precision | | Recall | | F- Score | |
|---|---|---|---|---|---|---|---|---|
| | Max | Avg | Max | Avg | Max | Avg | Max | Avg |
| PSO | 93.8849 | 91.6187 | 94.9721 | 92.7515 | 96.0674 | 94.2977 | 95.2646 | 93.5134 |
| GWO | **94.2446** | 90.9172 | **96.0227** | 92.9304 | 96.0674 | 92.8932 | **95.4802** | 92.9027 |
| MFO | 93.5252 | **92.3201** | 94.9153 | **93.2023** | **96.6292** | **94.9438** | 95.0276 | **94.0601** |
| WOA | 93.1655 | 91.8345 | 94.3820 | 92.8152 | **96.6292** | 94.5786 | 94.7075 | 93.6820 |
| FFA | 93.5252 | 91.4388 | 93.8889 | 92.7945 | **96.6292** | 93.9325 | 95.0276 | 93.3521 |
| MVO | 92.446 | 91.4568 | 94.3503 | 92.8567 | 95.5056 | 93.9045 | 94.1504 | 93.3672 |
| IBSSA | 92.446 | 90.4676 | 94.8571 | 92.246 | 95.5056 | 92.9494 | 94.1504 | 92.5847 |

Table 12. The classification performance comparison results that were attained by RF with DS2

| Algorithm | Accuracy | | Precision | | Recall | | F- Score | |
|---|---|---|---|---|---|---|---|---|
| | Max | Avg | Max | Avg | Max | Avg | Max | Avg |
| PSO | **94.2446** | 92.2302 | 93.5484 | 90.621 | **97.7528** | 95.9269 | **95.6044** | 93.1861 |
| GWO | 93.1655 | 91.5287 | 94.7977 | 91.9839 | **97.7528** | 93.3988 | 94.7658 | 92.6582 |
| MFO | 93.8849 | **92.6978** | 93.4066 | 91.1925 | 97.191 | **96.3202** | 95.0549 | **93.6793** |
| WOA | 93.5252 | 92.5000 | 93.7500 | 91.1323 | **97.7528** | 96.0393 | 94.7368 | 93.5073 |
| FFA | 92.8058 | 91.7985 | 94.3503 | 91.8734 | 96.6292 | 94.3258 | 94.1176 | 93.0739 |
| MVO | 93.5252 | 91.9784 | 93.8889 | 91.1835 | 97.191 | 94.691 | 94.4134 | 92.8825 |
| IBSSA | 93.5252 | 91.8345 | **94.7977** | **93.15** | 95.5056 | 92.5281 | 94.1828 | 92.8265 |

Table 13. The classification performance comparison results that were attained using SVM with DS2

| Algorithm | Accuracy | | Precision | | Recall | | F- Score | |
|---|---|---|---|---|---|---|---|---|
| | Max | Avg | Max | Avg | Max | Avg | Max | Avg |
| PSO | 94.6043 | 92.7338 | 95.4802 | 93.4127 | 97.191 | 95.3932 | 95.8449 | 94.3868 |
| GWO | 93.5252 | 91.5287 | 95.9064 | 93.3714 | 96.6292 | 93.4269 | 94.9438 | 93.3876 |
| MFO | 94.2446 | 93.0395 | 95 | 93.7031 | 97.191 | 95.5617 | 95.5801 | 94.6191 |
| WOA | 93.8849 | 92.7877 | 94.4444 | 93.2464 | 97.191 | 95.6741 | 95.2909 | 94.4405 |
| FFA | 93.8849 | 91.6906 | 94.9721 | 93.0589 | 96.0674 | 94.0449 | 95.2646 | 93.5463 |
| MVO | 93.8849 | 92.0504 | 94.9721 | 92.6187 | 96.6292 | 95.1966 | 95.2381 | 93.8816 |
| IBSSA | **96.7509** | **94.9097** | **96.5714** | **94.56** | **99.4152** | **97.4031** | **97.4063** | **95.9498** |

Table 14. The comparison of the suggested IBSSA with RTF-C-IEF filter on DS1 and DS2 by macro-averaged

| Classification model | DS1 | | | | DS2 | | | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1_Score | SF | Precision | Recall | F1_Score | FS |
| RTF-C-IDF+SVM | 79.077 | 78.5681 | 78.7026 | 377 | 92.6926 | 90.8146 | 91.6259 | 2367 |
| RTF-C-IDF+IBSSA+SVM | **83.228** | **81.858** | **82.1609** | **82** | **95.084** | **94.136** | **94.548** | **390** |

Table 15. Performance comparisons between SSA, and IBSSA according to the mean fitness value, average of the features selected, and average accuracy

| Datasets | Metric | Fitness | | Feature | | Accuracy | |
|---|---|---|---|---|---|---|---|
| | | SSA | IBSSA | SSA | IBSSA | SSA | IBSSA |
| DS1 | AVE | 0.184627593 | **0.151179227** | 296 | **82.95** | 0.834433 | **0.8495162** |
| | STD | **0.0090379** | 0.0130699 | 34.564965 | **11.114120** | **0.0094387** | 0.0131858 |
| | Worst | 0.2001224 | **0.1808506** | 340 | **105** | **0.820189** | 0.819466 |
| | Best | 0.16862298 | **0.1219837** | 218 | **66** | 0.848297 | **0.87874** |
| DS2 | AVE | 0.0726331 | **0.0417456** | 1874.85 | **390.35** | 0.9519181 | **0.9594985** |
| | STD | 0.0121495 | **0.0081708** | 214.34730 | **23.698267** | 0.0093791 | **0.0082426** |
| | Worst | 0.0984971 | **0.0598327** | 2083 | **440** | 0.932584 | **0.941176** |
| | Best | 0.0530011 | **0.0273548** | 1417 | **356** | 0.966102 | **0.974063** |

only 377 and 2367, respectively, for RTF-C-IEF, it is clear that IBSSA is capable of efficient optimization. In Table 14, we can see the accuracy value obtained by a macro-average measure that “IBSSA + RTF-C-IEF + SVM” has higher classification accuracy than “RTF-C-IEF + SVM”. Additionally, “RTF-C-IEF + SVM” produced more selected features than did “IBSSA + RTF-C-IEF + SVM”. Accordingly, we conclude that using RTF-C-IEF with IBSSA for feature selection is more accurate than using RTF-C-IEF alone. It was quite interesting that the performance of the hybrid method didn’t get worse with reduced feature subsets. Also, in this study, findings proved that employing IBSSA combined with an SVM classifier provides better accuracy than others classify.

### 5.3 Comparison of IBSSA with SSA

In this section, IBSSA is investigated to quantify the extent of improvement in it and see how integrating a levy flight strategy, initialization algorithm, and a local search algorithm into SSA will affect it. So, IBSSA was compared to the original SSA based on the classifier SVM, in terms of three metrics: average fitness, the number of features selected, and average accuracy.

As shown in Table 15, the average fitness values, mean number of selected features, and mean classification accuracy are listed based on IBSSA and SSA both with the SVM classifier. In terms of mean fitness values, this table clearly shows that IBSSA outperforms the original method for both datasets. Fig. 8 displays the boxplots for datasets to evaluate the algorithms' fitness value. The boxplots, it should be mentioned, show the results of fitness values, and are exhibited following executed 20 time for every method, and 500 iterations for each run. These graphs make it possible for us to visually see the data's lower, median, and top values. As shown in this figure, IBSSA has low-fitness values (better fitness values) than the SSA approaches in both datasets. Thereby, IBSSA achieved much higher performance than the original algorithm based on fitness.

Moreover, in terms of the average number of features selected, IBSSA ranks first in the two datasets, because the largest number of features were removed by IBSSA with more than 80% than the total number of features, and selecting the fewest number of features compared to the original algorithm. Thus, IBSSA outperformed SSA in selecting a smaller number of features over all datasets. Moreover, IBSSA outperformed SSA in terms of average classification accuracy by F-score on both datasets, with 84.95% and 95.94% classification accuracies for both datasets,

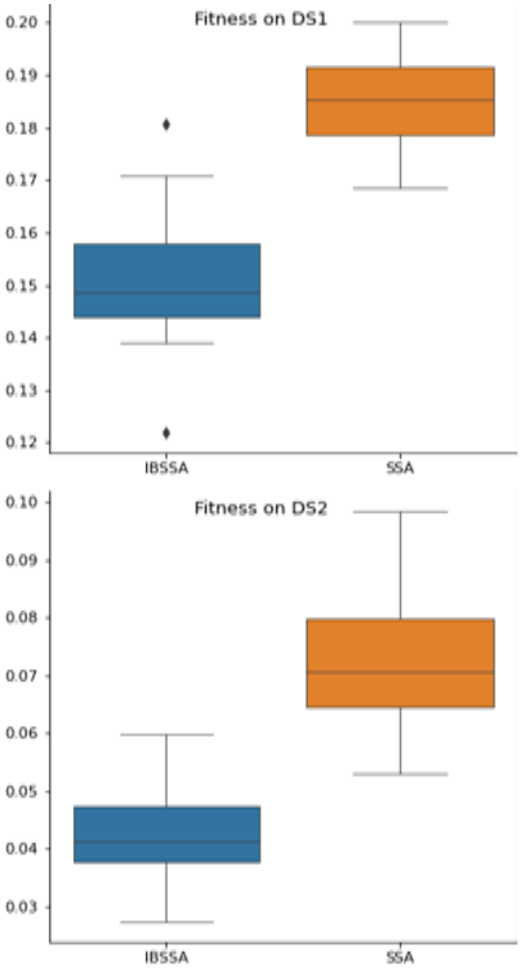


Figure. 8 Boxplots of IBSSA compared with SSA performance by fitness values for both datasets (DS1 & DS2)

respectively.

From these results, it is clearly seen that IBSSA with the SVM classifier has significantly improved both the FS and classification tasks for real-world data related to COVID-19 compared to the original SSA algorithm in terms of the overall mean number of chosen features, fitness value, and classification accuracy across all datasets.

## 6. Conclusions and future works

In this paper, we provide a precise and clever classification technique for the infection of COVID-19 patients. The suggested feature selection methodology is called RTF-C-IEF+IBSSA which combines the advantages of both the term weighting scheme and methods of wrapper selection. IBSSA selects the most useful and effective features from the extracted features from clinical texts by RTF-C-IEF which calculate the significance of the feature. The chosen features are then fed into the suggested categorization model to enable precise and informed decision-making. In IBSSA, we introduced four ways to improve both the global and local search capabilities of the algorithm. The suggested method has been compared to the most recent and well-known feature selection swarm techniques, including PSO, MFO, GWO, MVO, and FFA. The experimental findings show that the proposed approach outperformed the most recent evolutionary algorithms. According to IBSSA, the amount of diagnostic mistake in COVID-19 patients has decreased as a result of feature selection, and investigations demonstrate that the suggested strategy has a higher accuracy than other methods and is more efficient at reducing sub-features by more than 83%. Consequently, feature election allows machine learning to concentrate more on key features, lowering the risk of classifying infected people from healthy individuals. In our subsequent study, we will consider enlarging and diversifying the test datasets to more fully evaluate the suggested approach.

## Conflicts of interest

The authors declare no conflict of interest.

## Author contributions

The first author has done investigation, dataset collection, implementation, result analysis, and preparing original draft. The second author has done the supervision, review of work and validation.

## Funding statement

This work was supported in part by the Ministry of Higher Education Malaysia under Fundamental Research Grant Scheme (FRGS/1/2021/ICT06/ UTM/02/6).

## References

[1] S. L. Senanayake, “Drug repurposing strategies for COVID-19”, *Futur. Drug Discov.*, Vol. 2, No. 2, 2020.

[2] N. A. Baghdadi, A. Malki, S. F. Abdelaliem, H. M. Balaha, M. Badawy, and M. Elhosseini, “An automated diagnosis and classification of COVID-19 from chest CT images using a transfer learning-based convolutional neural network”, *Computers in Biology and Medicine*, Vol. 144, p. 105383, 2022.

[3] G. A. P. D. Souza, M. L. Bideau, C. Boschi, L. Ferreira, N. Wurtz, C. Devaux , P. Colson, and B. L. Scola, “Emerging SARS-CoV -2 genotypes show different replication patterns in human pulmonary and intestinal epithelial cells”, *Viruses*, Vol. 14, No. 1, p. 23, 2021.

[4] O. N. Oyelade and A. E. Ezugwu, "A case-based reasoning framework for early detection and diagnosis of novel coronavirus", *Informatics in Medicine Unlocked*, Vol. 20, p. 100395, 2020.

[5] F. R. Lucini, F. S. Fogliatto, G. J. C. da Silveira, J. L. Neyeloff, M. J. Anzanello, R. S. Kuchenbecker, and B. D. Schaan, "Text mining approach to predict hospital admissions using early medical records from the emergency department", *International Journal of Medical Informatics*, Vol. 100, pp. 1–8, 2017.

[6] A. A. Moammar, L. A. Henaki, and H. Kurdi, "Selecting Accurate Classifier Models for a MERS-CoV Dataset", In: *Proc. of the 2018 Conf. on Intelligent Systems and Applications*, London, UK, 2018, pp. 1070–1084.

[7] K. Lybarger, M. Ostendorf, M. Thompson, and M. Yetisgen, "Extracting COVID-19 Diagnoses and Symptoms From Clinical Text: A New Annotated Corpus and Neural Event Extraction Framework", *Journal of Biomedical Informatics*, Vol. 117, p. 103761, 2021.

[8] G. Saranya and A. Pravin, "Feature selection techniques for disease diagnosis system: A survey", *Artificial Intelligence Techniques for Advanced Computing Applications*, Singapore: Springer Singapore, pp. 249–258, 2021.

[9] B. A. I. Ji, X. Lu, G. Sun, J. Li, and Y. Xiao, "Bio-Inspired Feature Selection : An Improved Binary Particle Swarm Optimization Approach", *IEEE Access*, Vol. 8, pp. 85989–86002, 2020.

[10] J. T. Pintas, L. A. F. Fernandes, and A. C. B. Garcia, "Feature selection methods for text classification: a systematic literature review", *Artificial Intelligence Review*, Vol. 54, No. 8, pp. 6149–6200, 2021.

[11] E. O. Abiodun, A. Alabdulatif, O. I. Abiodun, M. Alawida, A. Alabdulatif, and R. S. Alkhawaldeh, "A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities", *Neural Computing and Applications*, Vol. 33, No. 22, pp. 15091–15118, 2021.

[12] F. A. Zeidabadi, S. A. Doumari, M. Dehghani, and O. P. Malik, "MLBO: Mixed Leader Based Optimizer for Solving Optimization Problems", *International Journal of Intelligent Engineering and Systems*, Vol. 14, No. 4, pp. 472–479, 2021, doi: 10.22266/ijies2021.0831.41.

[13] A. P. K, K. N. C, and R. R. K, "Pelican Optimization Algorithm for Optimal Demand Response in IslandedActive Distribution Network Considering Controllable Loads", *International Journal of Intelligent Engineering and Systems*, Vol. 15, No. 6, 2022, doi: 10.22266/ijies2022.1231.14.

[14] N. Bacanin, K. Venkatachalam, T. Bezdan, M. Zivkovic, and M. Abouhawwash, "A novel firefly algorithm approach for efficient feature selection with COVID-19 dataset", *Microprocess. Microsyst*, Vol. 98, p. 104778, 2023.

[15] J. Xue and B. Shen, "A novel swarm intelligence optimization approach: sparrow search algorithm", *Systems Science & Control Engineering*, Vol. 8, No. 1, pp. 22–34, 2020.

[16] W. Song, S. Liu, X. Wang, and W. Wu, "An Improved Sparrow Search Algorithm", In: *2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing*, Exeter, UK, 2020.

[17] W. Tuerxun, X. Chang, G. Hongyu, J. Zhijie, and Z. Huajian, "Fault Diagnosis of Wind Turbines Based on a Support Vector Machine Optimized by the Sparrow Search Algorithm", *IEEE Power & Energy Society Section*, Vol. 9, pp. 69307–69315, 2021.

[18] A. G. Gad, K. M. Sallam, R. K. Chakrabortty, M. J. Ryan, and A. A. Abohany, "An improved binary sparrow search algorithm for feature selection in data classification", *Neural Computing and Applications*, Vol. 34, No. 18, pp. 15705–15752, 2022.

[19] F. S. Gharehchopogh, M. Namazi, L. Ebrahimi, and B. Abdollahzadeh, "Advances in Sparrow Search Algorithm: A Comprehensive Survey", *Archives of Computational Methods in Engineering*, Vol. 30, No. 1, pp. 427–455, 2023.

[20] P. H. Prastyo, R. Hidayat, and I. Ardiyanto, "Enhancing sentiment classification performance using hybrid Query Expansion Ranking and Binary Particle Swarm Optimization with Adaptive Inertia Weights", *ICT Express*, Vol. 8, No. 2, pp. 189–197, 2021.

[21] R. Kamala and P. R. J. Thangaiah, "A Novel Two-Stage Selection of Feature Subsets in Machine Learning", *Engineering, Technology & Applied Science Research*, Vol. 9, No. 3, pp. 4169–4175, 2019.

[22] Y. A. Alhaj, A. Dahou, M. A. A. A. Qaness, L. Abualigah, A. A. Abbasi, N. A. O. Almaweri, M. A. Elaziz, and R. Damaševičius, "A Novel Text Classification Technique Using Improved Particle Swarm Optimization: A Case Study of Arabic Language", *Futur. Internet*, Vol. 14, No. 194, pp. 1–18, 2022.

[23] I. Strumberger, A. Rakic, S. Stanojlovic, J.

Arandjelovic, T. Bezdan, M. Zivkovic, and N. Bacanin, "Feature Selection by Hybrid Binary Ant Lion Optimizer with COVID-19 dataset", In: *Proc. of 2021 29th Telecommunications Forum (TELFOR), Belgrade, Serbia*, pp. 1–4, 2021.

[24] Y. Wang, X. Gao, X. Ru, P. Sun, and J. Wang, "A hybrid feature selection algorithm and its application in bioinformatics", *PeerJ Computer Science*, Vol. 8, pp. 1–17, 2022.

[25] S. Mustafa, A. Ali, H. Salahuddin, and M. U. Chaudhry, "Two-step Feature Selection for Predicting Mortality Risk in COVID-19 Patients", In: *Proc. of 2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, Quetta, Pakistan. pp. 1–5, 2021.

[26] V. Kumar, A. Sharma, A. Bansal, and J. S. Sandhu, "Two-Stage Feature Selection Pipeline for Text Classification", In *Computer Networks and Inventive Communication Technologies, Singapore: Springer Singapore*, pp. 795–809, 2022.

[27] I. M. E. Hasnony, M. Elhoseny, and Zahraa 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.

[28] M. A k. Alsaeedi and S. Kurnaz, "Feature selection for diagnose coronavirus (COVID-19) disease by neural network and Caledonian crow learning algorithm", *Applied nanoscience*, Vol. 13, No. 4, pp. 1–16, 2022.

[29] J. Too and S. M. Mirjalili "A Hyper Learning Binary Dragonfly Algorithm for Feature Selection: A COVID-19 Case Study", *Knowledge-Based Systems*, Vol. 212. p. 106553, 2021.

[30] M. Abdollahi, X. Gao, Y. Mei, S. Ghosh, and J. Li, "An Ontology-based Two-Stage Approach to Medical Text Classification with Feature Selection by Particle Swarm Optimisation", In: *Proc. of 2019 IEEE Congr. Evol. Comput. CEC 2019-Proc.*, Wellington, New Zealand, 2019.

[31] X. Bai, X. Gao, and B. Xue, "Particle Swarm Optimization Based Two-Stage Feature Selection in Text Mining", In: *Proc. of 2018 IEEE Congress on Evolutionary Computation (CEC)*, Rio de Janeiro, Brazil, 2018.

[32] K. L. K. Zhang, Y. H. Jason, and C. H. Neil, "Feature selection based on an improved cat swarm optimization algorithm for big data classification", *Journal of Supercomputing*, Vol. 72, No. 8, pp. 3210–3221, 2016.

[33] R. A. Khurmaa, I. Aljarah, and A. Sharieh, "An intelligent feature selection approach based on moth flame optimization for medical diagnosis", *Neural Computing & Applications*, Vol. 33, No. 12, pp. 7165–7204, 2021.

[34] Q. Liang, B. Chen, H. Wu, and M. Han, "A Novel Modified Sparrow Search Algorithm Based on Adaptive Weight and Improved Boundary Constraints", In: *Proc. of 2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*, Chengdu, China, 2021, pp. 104–109, 2021.

[35] A. E. Ezugwu, I. A. T. Hashem, O. N. Oyelade, M. Almutari, M. A. A. Garadi, I. N. Abdullahi, l. Otegbeye, A. K. Shukla, and H. Chiroma, "A Novel Smart City-Based Framework on Perspectives for Application of Machine Learning in Combating COVID-19", *BioMed Research International*, Vol. 2021, p. 5546790, 2021.

[36] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing", In: *Proc. of the 18th BioNLP Workshop and Shared Task*, Florence, Italy, pp. 319–327, 2019.

[37] A. Y. Mahdi and S. S. Yuhaniz, "Automatic Diagnosis of COVID-19 Patients from Unstructured Data Based on a Novel Weighting Scheme", *Computers, Materials & Continua.*, Vol. 74, No. 1, pp. 1375–1392, 2023.

[38] H. Lim and D. Kim, "Pairwise dependence-based unsupervised feature selection", *Pattern Recognition*, Vol. 111, p. 107663, 2021.

[39] A. H. Rabie, S. H. Ali, A. I. Saleh, and H. Arafat Ali, "A fog based load forecasting strategy based on multi-ensemble classification for smart grids", *Journal of Ambient Intelligence and Humanized Computing*, Vol. 11. pp. 209–236, 2020.

[40] P. A. Digehsara, S. N. Chegini, A. Bagheri, and M. P. Roknsaraei, "An improved particle swarm optimization based on the reinforcement of the population initialization phase by scrambled Halton sequence", *Cogent Engineering*, Vol. 7, No. 1, p. 1737383, 2020.

[41] W. H. Bangyal, A. Hameed, W. Alosaimi, and H. Alyami, "A New Initialization Approach in Particle Swarm Optimization for Global Optimization Problems", *Computational Intelligence and Neuroscience*, Vol. 2021, pp. 1–17, 2021.

[42] W. H. Bangyal, A. Hameed, W. Alosaimi, and H. Alyami, "Modified cuckoo search algorithm with rough sets for feature selection", *Neural Computing & Applications*, Vol. 29, No. 4, pp.

925–934, 2018.

[43] D. A. Elmanakhly, M. Saleh, E. A. Rashed, and S. Member, "BinHOA: Efficient Binary Horse Herd Optimization Method for Feature Selection: Analysis and Validations", *IEEE Access*, Vol. 10, pp. 26795–26816, 2022.

[44] Z. Li, Y. Zhou, S. Zhang, and J. Song, "Lévy-Flight Moth-Flame Algorithm for Function Optimization and Engineering Design Problems", *Mathematical Problems in Engineering*, Vol. 2016, pp. 1–22, 2016.

[45] P. Agrawal, T. Ganesh, D. Oliva, and A. W. Mohamed, "S-shaped and V-shaped gaining-sharing knowledge-based algorithm for feature selection", *Applied Intelligence*, Vol. 52, No. 1, pp. 81–112, 2022.

[46] Kicska and A. Kiss, "Comparing swarm intelligence algorithms for dimension reduction in machine learning", *Big Data and Cognitive Computing*, Vol. 5, No. 36, pp. 1–15, 2021.

[47] M. Afif, A. S. Ghareb, A. Saif, A. Abu Bakar, and O. Bazighifan, "Genetic algorithm rule based categorization method for textual data mining", *Decision Science Letters*, Vol. 9, No. 1, pp. 37–50, 2020.

[48] C. Wan, Y. Wang, Y. Liu, J. Ji, and G. Feng, "Composite Feature Extraction and Selection for Text Classification", *IEEE Access*, Vol. 7, pp. 35208–35219, 2019.

[49] M. Qaraad, S. Amjad, I. I. M. Manhrawy, H. Fathi, B. A. Hassan, and P. E. Kafrawy, "A Hybrid Feature Selection Optimization Model for High Dimension Data Classification", *IEEE Access*, Vol. 9, pp. 42884–42895, 2021.

[50] H. J. Escalante, M. A. G. Limón , A. M. Reyes, M. Graff, M. M. Gómez, E. F. Morales, and J. M. Carranza, "Term-Weighting Learning via Genetic Programming for Text Classification", *Knowledge-Based Systems*, Vol. 83, pp. 176–189, 2015.

[51] Y. Gao, Y. Zhou, and Q. Luo, "An Efficient Binary Equilibrium Optimizer Algorithm for Feature Selection", *IEEE Access*, Vol. 8, pp. 140936–140963, 2020.

[52] M. M. Mafarja, I. Aljarah, A. A. Heidari, H. Faris, P. F. Viger , X. Li, and S. M. Mirjalili, "Binary dragonfly optimization for feature selection using time-varying transfer functions", *Knowledge-Based Systems*, Vol. 161, pp. 185–204, 2018.

[53] N. S. M. Nafis and S. Awang, "An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification", *IEEE Access*, Vol. 9, pp. 52177–52192, 2021.

[54] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey", *Information*, Vol. 10, No. 4, pp. 1–68, 2019.

[55] A. Y. Mahdi and S. S. Yuhaniz, "Automatic extraction of knowledge for diagnosing COVID-19 disease based on text mining techniques: A systematic review", *Periodicals of Engineering and Natural Sciences*, Vol. 9, No. 2, pp. 918–929, 2021.

[56] R. A. Khurma, I. Aljarah, A. Sharieh, and S. Mirjalili, "EvoloPy-FS: An Open-Source Nature-Inspired Optimization Framework in Python for Feature Selection", *Algorithms for Intelligent Systems*, Singapore: Springer Singapore, 2020, pp. 131–173, 2020, doi: 10.1007/978-981-32-9990-0_8.