Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/120712

A Proposed Framework to Improve Diagnosis of Covid-19 Based on Patient's Symptoms using Feature Selection Optimization

Manar Elhussiny[1,*], *O. E. Emam* [2] *and Safaa.M.Azzam* [2]

[1]Department of Software Engineering, Faculty of Computers and Information, Kafr El Sheikh University, Kafr El Sheikh, Egypt
[2]Department of Information Systems, Faculty of Computers and Artificial Intelligence, Helwan University, Helwan, Egypt

Received: 2 Mar. 2023, Revised: 22 May 2023, Accepted: 12 Jun. 2023
Published online: 1 Jul. 2023

Abstract: Recently, an epidemic called COVID-19 appeared, and it was one of the largest epidemics that affected the world in all economic, educational, health, and other aspects due to its rapid spread worldwide. The surge in infection rates made traditional diagnostic methods ineffective. Systems for automatic diagnosis and detection are crucial for controlling the outbreak. Other than PCR-RT, further diagnostic and detection techniques are needed. Individuals who receive positive test results often experience a range of symptoms, ranging from mild to severe, including coughing, fever, sore throats, and body pains. In more extreme cases, infected individuals may exhibit severe symptoms that make breathing challenging, ultimately leading to catastrophic organ failure. A hybrid approach called SDO-NMR-Hill has been developed for diagnosing COVID-19 based on a patient's initial symptoms. This approach incorporates traits from three models, including two distinct feature selection optimization methods and a local search. Supply-demand optimization and the naked mole rat were preferred among metaheuristic methods because they have fewer parameters and a lower computing overhead, which can help you find superfluous and uninformative characteristics. Hill climbing was preferred among local search methods to maximize a criterion among several candidate solutions. We used decision trees, random forests, and adaptive boosting machine-learning classifiers in various experiments on three COVID-19 datasets. We carried out a natural selection of the classifier's hyper-parameters to optimize outcomes. The optimal performance was attained using the adaptive boosting classifier, with an accuracy of 88.88% and 98.98% for the first and third datasets, respectively. The optimal performance for the second dataset was attained using the random forest classifier, with an accuracy of 97.97%. The suggested SDO-NMR-Hill model is evaluated using nine benchmark UCI datasets, and 15 different optimization techniques are contrasted.

Keywords: COVID-19, Hill climbing (HC), Naked mole-rat algorithm (NMR), Supervised machine learning algorithms, Supply-demand-based optimization (SDO).

1 Introduction

The novel coronavirus-induced pneumonitis began to spread in Wuhan, China, at the end of December 2019. Epidemiological studies revealed that the illness epidemic was connected to the seafood market in Wuhan, where the first patient was admitted to the hospital on December 12, 2019. The WHO designated this illness Coronavirus Disease 2019 on February 11, 2020. By infecting more than 152 million individuals throughout more than 216 nations and territories, COVID-19 has recently turned into a pandemic. The vast rise in infections has rendered conventional diagnostic methods useless. The major method of illness transmission was direct contact with affected people. Isolation from society can therefore reduce the chance of illness. Disease transmission can happen up to 6 feet away. The breath droplets created when an infected person talks or sneezes are, therefore, one of the primary ways that illnesses spread. In certain circumstances, COVID-19 symptoms go undetected [1,2,3]. Fever, breathing, and a dry cough difficulty are the primary symptoms of the COVID-19 infection. Up to 10% of individuals may develop gastrointestinal symptoms, including diarrhea, while others may feel muscle pains, weariness, or a loss of smell or taste. The COVID-19 clinical characteristics and other silent infections are shown in Table 1 [1]. According to experts, although a medicine to

* Corresponding author e-mail: manarelhussiny2016@gmail.com

© 2023 NSP
Natural Sciences Publishing Cor.

effectively treat COVID-19 is not yet available, several antiviral and other medications have demonstrated beneficial therapy. To treat COVID-19 infection, 15 different medications have been used, including famotidine, chloroquine, nafamostat, hydroxychloroquine, omifenovir, etc. [4]. To avoid Covid-19, scientists went for the vaccination. Biological treatments called vaccinations provide a particular infectious disease with active acquired immunity. They accomplish this by triggering an immune response to an antigen, a chemical involved in the illness. The first vaccination was developed in 1796 by Edward Jenner, who is credited with creating the smallpox vaccine. The COVID-19 vaccine has evolved at a rate unheard of in the history of immunization. There are 104 candidate vaccines in the clinical stages of research and 184 candidate vaccines in the preclinical stages. According to recent statistics, there are now 18 COVID-19 vaccines that have been authorized and are being used worldwide. The COVID-19 vaccines fall into four main groups and utilize various delivery systems: whole virus vaccines, vaccines based on proteins, vaccines using viral vectors, and vaccines using nucleic acids [5].

Table 1: The Clinical Characteristics of COVID-19.

Level	Characteristics of clinical	COVID-19 RT-PCR test
Asymptomatic	chest imaging results and absence of clinical signs.	Positive
Mild	Mild clinical symptoms include headache, fever, nasal congestion, exhaustion, coughing, malaise, anorexia, sore throat, malaise, muscular soreness, and weariness. No unusual imaging results for the chest.	Positive
Moderate	mild to medium clinical characteristics. Imaging of the chest revealed a little pneumonia symptom.	Positive
Severe	signs of a respiratory infection suspected along with any of the following breathing difficulty, $respiratoryrate >= 30\ breaths/min$, oxygen saturation at rest $<= 93\%$; $Pao_2/Fio_2 <= 300$ mmHg (1 mmHg Z 0.133 kPa). The lesions dramatically advanced $> 50\%$ within 24–48 hours, according to chest imaging, indicating a serious illness	Positive
critical	The disease's fast progression combined with any of the following Shocks, along with other organ failure and the necessity for mechanical ventilation, respiratory failure, and need for intensive care unit monitoring therapy.	Positive
RT-PCR, reverse transcriptase-polymerase chain reaction; Pao_2, arterial partial pressure of oxygen; Fio_2, oxygen concentration;		

With the use of machine learning (ML), a computer may reason like a person and come to a decision all by itself. It is the method of automatically teaching computers to learn without having them be directly programmed. Building computer software that can access data and use it for learning processes is the main goal of ML. Information technology, probability, statistics, artificial intelligence, psychology, and many other fields all provide a vast field called machine learning. A subset of artificial intelligence is ML. On the other hand, artificial intelligence refers to the creation of computer systems that can do tasks that frequently require human input, such as decision-making. Choosing the appropriate course of action for each choice is crucial to achieving our goals. In this situation, a variety of ML methods are used for both classification and regression problems. When using a continuous prediction strategy, regression is the most effective course of action. Classification is used when the prediction aim is a discrete value or a class label. As an illustration of categorization, let's look at the process of sorting legitimate emails from spam ones. Depending on whether or not the email is spam, the result should be yes or no. Classification algorithms are widely used in the medical field to diagnose and forecast illnesses accurately. Diseases and health issues such as diabetes, cardiac syndrome, chronic renal disease, breast cancer, and liver cancer [6,7,8].

Feature selection (FS) acts as an operation that selects the best attributes from the datasets. The diversity of FS applications can be proven in diverse domains. These applications include biomedical problems, text mining, image analysis, etc. The FS function is used to pick out related attributes. Suppose dataset D contains N features number where dataset D equals A_1, A_2,.., A_N. The base is to pick out the better feature subsets from D. Elicit Subset d equal A_1, A_2,..,A_n and, n < N and A_1, A_2,..,A_n is the attributes of any dataset [9].

The metaheuristic algorithms that were utilized to take care of the FS issue Binary vectors Offer are considered important attributes. In the planned calculation, the arrangement vector is addressed. By [100101 ...], this means that 1 implies that an extraordinary attribute has been chosen, and a worth of 0 implies that the attribute isn't chosen in the subgroup [9].

To summarize, the following are our significant contributions:

1. Hybrid research using features obtained from two distinct metaheuristic feature selection approaches and a local search for diagnosis using COVID-19 was given.

© 2023 NSP
Natural Sciences Publishing Cor.

a. Among metaheuristic approaches, NMR and SDO were favored for identifying pointless and uninformative features since they required less computational work and fewer parameters.
b. Hill climbing was selected among local search techniques to optimize a criterion among a set of candidate solutions.

2. We employed three distinct COVID-19 datasets, varying in size and number of symptoms, to assess the proposed model.
3. Adaptive boosting, random forest, and decision tree are three machine-learning classifiers that we utilized to compare the results of several experiments.
4. Comparing the SDO-NMR-hill model to 15 different optimization techniques to prove its superiority.
5. Using a confusion matrix (CM), we contrasted the results of some recent work on the same datasets with our superior findings to assess the quality of the suggested framework.
6. Nine benchmark UCI datasets are used to test the proposed SDO-NMR-hill model, and 15 alternative optimization strategies are compared.

This study will include in the section 2 a background, the section 3 will include related works, the section 4 will talk about materials and methods, the section 5 will talk about results and discussion, and the section 6 will include the conclusion and future work.

2 Background

The technologies and algorithms that are necessary for our work are reviewed in this part. It is divided into three sections A. Supervised Machine Learning Algorithms B. Feature Selection Optimization Algorithms C. Local Search Algorithm.

A. Supervised machine learning algorithms

A simple machine learning model's learning process consists of two stages: training and testing. The learning algorithm or learner builds the learning model using samples of training data that are used as input throughout the training process. The execution engine is used by the learning model during testing to produce a prediction for test or production data. The learning model's output, tagged data, provides the final forecast or classified data. In supervised machine learning methods, the underlying algorithm is initially trained using a labeled training dataset. The unlabeled test dataset is then supplied to the trained algorithm to classify the individuals into corresponding groups. Classification and regression are the two primary forms of supervised learning, both of which need an input and an output, with the primary goal being to map the input to the outcome. Discrete output variables are used in classification tasks. This variable can be classified into a variety of groups, like "red" or "black," "diabetic" or "non-diabetic", "spam" or "not spam", and a program that determines the species of iris from measurements of its flower. In regression issues, such as the likelihood that a person will acquire cardiovascular disease, the relevant output variable is a real number [6,7,10]. An overview of supervised machine learning algorithms is provided here.

1. Logistic Regression (LR)

LR may be used to guess the chance that a new instance belongs to a specific class. Considering that the outcome is a probability, it ranges from 0 to 1. To ply the LR as a binary classifier, a threshold must be established to differentiate between the two labels. For instance, if the probability value is greater than 0.50, an input instance will be classed as "class A"; otherwise, "class B". A prediction technique, it is similar to ordinary least squares (OLS) regression. However, in LR, the output of the prediction is binary [10,11].

2. Support Vector Machine (SVM)

Support-Vector Machine models and traditional multilayer perceptron neural networks have many similarities. Vapnik initially introduced it in 1992. Regression, classification issues, and feature selection all employ SVM. This method is effective for resolving issues involving linear and nonlinear datasets. An n-dimensional feature space, where n is the total number of features, is initially created for each individual piece of data. The process of finding the hyperplane that separates the data points into two groups maximizes marginal distance for both classes and minimizes classification errors [6,8,10,11].

3. K-Nearest Neighbor (KNN)

One of the most popular machine learning techniques for classification issues, pattern recognition, and regression is the KNN, which is a straightforward model. Using the Euclidean distance between data points, KNN finds neighbors among the data. The "K" represents the number of closest neighbors that should be included in the "vote." For the same sample item, different values for "K" can provide various categorization outcomes. The K value is important and must be carefully determined since, if K is too tiny, the system may overfit [8,10].

© 2023 NSP
Natural Sciences Publishing Cor.

4. Multilayer Perception (MLP)

It is a well-known neural network-based classification system, often consisting of an input layer, an output layer, and one or more hidden layers between the input and output layers, frequently trained through backpropagation. These layers of nodes all have unidirectional connections. Numerous "neurons" that link all the layers together may be found in each layer. It is also a classifier in which, unlike conventional neural network training, the weights of the network are determined by solving a quadratic programming problem with linear constraints as opposed to a nonconvex, unconstrained minimization problem [11,12,13].

5. Decision Tree (DT)

By organizing the logic of the decision in a way that compares and evaluates the results for the classification of data items into a structure resembling a tree that categorizes instances by sorting them according to feature values, it is a hierarchical design that uses the divide-and-conquer strategy. Each node in a decision tree corresponds to a characteristic of an instance that has to be classified, and each branch to a potential value for that node. Instances are categorized and ordered starting at the root node based on the values of their characteristics. There are sometimes several layers to the nodes in a DT, with the top level acting as the root or parent node and the lesser ones acting as child nodes [6,11,12].

6. Random Forest (RF)

A data categorization method called RF is based on DT and ensemble learning. The result is a forest of decision trees and a massive number of trees. The class label for each occurrence during the testing period is predicted by every tree in the forest. Each test set's final pick is determined by a majority vote when a class label is predicted by each tree. The classification that gets the greatest support is assumed to be the one that should be used with the test data. Each item of data in the data collection goes through this process again [12].

7. Adaptive Boosting (AdaBoost)

It is a methodology that combines several weak classifiers into a single strong classifier using ensemble learning techniques. To improve recognition accuracy, it is integrated with other artificial intelligence techniques. It generates the final result by employing the majority vote approach to weigh all of these weak classifiers' judgments. According to [15], The AdaBoost algorithms have the benefits of simplified feature selection, easier implementation, and reasonable generalization, but they only provide mediocre solutions and are susceptible to outliers and noisy data [14,15,16].

B. Feature selection optimization Algorithms

Most datasets have relevant and unrelated features that are unhelpful but contribute to the overall dimensions of the problem space, and this increases computational cost, so FS plays an important role in identifying relevant features and eliminating irrelevant features. The goal of FS is to reduce the dimensionality of selecting a small subset of relevant features by removing redundant or noisy features to perform better, have higher accuracy, and lower computational costs [17,18]. We used two feature selection optimization algorithms. We will address them below.

1. The naked mole-rat algorithm (NMR)

it is a naked mole-rat natural-inspired swarm intelligence method. That closely mirrors the natural mating habits of NMRs and possesses the following distinguishing characteristics [19].

1. Eusocial NMRs have a population of 295 individuals, with 70–80 individuals being the average.
2. Breeders are the most productive NMRs in the working group, and they are only in charge of mating, whilst workers are in charge of a range of tasks. The group is directed by a female queen who divides the population into breeders and workers, are shown in Figure 1.
3. Employees do crucial tasks, and breeders replace the best of them. Instead, low-performing breeders are returned to the worker pool, while high-performing employees become breeders.
4. The most prolific breeder among breeders, who mate with the queen.

There are three phases to the algorithm. The initialization of the NMR population takes place in the first phase, followed by the worker phase and the breeder phase [19].

1. In the beginning, it creates a population of n NMR with a uniform distribution of random numbers, where every NMR in the range [1, ... n] is a D-dimensional vector. Here, D stands for the number of variables or features that will be put to the test to solve the issue. every NMR starts out as in equation 1 [19].

$$NMR_{(i,j)} = NMR_{(min,j)} + U(0,1)X(NMR_{(min,j)} - NMR_{(max,j)}) \quad (1)$$

Where $j \in [1,\ldots D]$, $i \in [1,\ldots..n]$, $NMR_{(i,j)}$ is the i^{th} solution in the j^{th} dimension, $NMR_{(min,j)}$, $NMR_{(max,j)}$ are the problem function's lower and upper limits. $U(0,1)$ is a random number with a uniform distribution.

© 2023 NSP
Natural Sciences Publishing Cor.

Fig. 1: a) Breeding mates. b) Breeders. c) Workers.

2. Then, the workers prefer to grow fitter during this phase so that they can have the opportunity to start breeding and finally mate with the queen. In this case, each NMR generates a new solution based on its own knowledge and local data. Here, the new NMR's fitness is assessed, and if it is found to be more suited for mating, the old solution is set aside, and the new one is remembered. Otherwise, the previous answer is kept. The ultimate fitness of each worker rat is recorded when they have all finished the search procedure. The NMR uses the following equation 2 to create a fresh solution from an old one [19]:

$$w_i^{(t+1)} = w_i^t + \lambda (w_k^t - w_j^t) \tag{2}$$

Where $w_i^{(t+1)}$ is the new solution to workers, w_i^t is the old solution in t^{th} iteration. The w_k^t, w_j^t are two randomly selected solutions from the pool of workers. The λ is a mating factor. A uniform distribution in the [0, 1] range is used to determine the value of λ [19].

3. The breeder NMR also improves themselves in order to continue breeding and being picked for mating. The breeder NMRs are updated based on a breeding probability (bp) in relation to the overall best d. This bp is a completely random number between [0, 1]. Some breeders could be relegated to the workers' group because they are unable to maintain their fitness. According to the following equation 3, the breeders adjust their positions [19]:

$$b_i^{(t+1)} = (1-\lambda) b_i^t + \lambda (d - b_i^t) \tag{3}$$

Where $b_i^{(t+1)}$ is the new solution to breeders, b_i^t is the old solution in t^{th} iteration. As an initial value, the value of bp has been set to 0.5 [19].

2. *The Supply-demand-based Optimization (SDO)*

The supply and demand mechanism in economics serves as the inspiration for SDO, a swarm-based optimizer. The supply and demand interactions between producers and consumers are both simulated by this algorithm. The supply-demand process is an economic theory that describes how prices are determined in a market economy. According to this theory, the producers' supply connection, $q_{(t+1)} = f(p_t)$, where f is a linear supply function, determines a commodity's quantity, $q_{(t+1)}$, which is dependent on its current market price, p_t. F is an increasing function since it assumes that the supply of a certain product will increase the next time its price increases on the market. The equation $p_{(t+1)} = g(q_{(t+1)})$, where g is a linear demand function, represents the demand relationship between consumers and sets the price of a product $p_{(t+1)}$ in the subsequent time period in dependence on its quantity $q_{(t+1)}$ in the same period. The decreasing function g states that the market price of a good will increase as supply increases. The equilibrium point with the oscillations is a location $P(x_0, y_0)$ where these two functions finally intersect. The values x_0 and y_0, respectively, indicate a commodity's equilibrium value [20]. The function of supply f might be written as in equation 4:

$$q_{(t+1)} - q_0 = a(p_t - p_0) \tag{4}$$

The function of demand g might be written as in equation 5:

$$p_{(t+1)} - p_0 = -b(q_{(t+1)} - q_0) \tag{5}$$

When t represents the time, b and a are the linear coefficients, and There are commonly two ways in which the supply-demand mechanism works. One is the stability mode. When $\mid ab \mid < 1$ and the function of supply f is steeper than

© 2023 NSP
Natural Sciences Publishing Cor.

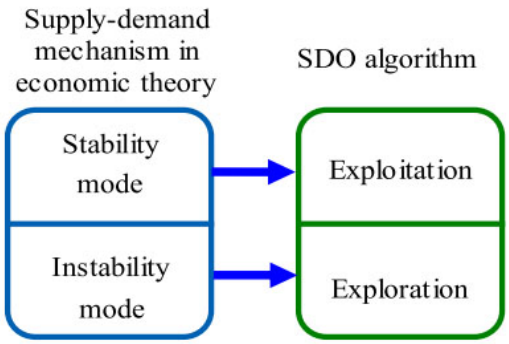

Fig. 2: The switching to the SDO algorithm from the supply-demand mechanism.

the function of demand g, the commodity price and quantity curve with respect to time tends to spiral inwards. As a result, with each period that passes, the oscillations get smaller. After certain iterations, number, quantity, and price approach equilibrium (x_0, y_0). The alternate is the instability mode. When $\mid ab \mid > 1$ and the function of demand g is steeper than the function of supply f, the oscillations become more pronounced with time, and the price-quantity curve with respect to time tends to spiral higher [20]. As the oscillations get more intense over time, the price and quantity curve with respect to time has a tendency to spiral ahead. Because of this, the quantity and price move away from the equilibrium point (x_0, y_0) over time. In a stable mode, both the quantity and the price of the commodity encourage the exploitation of the area close to the equilibrium point. Over time, the extent of this exploitation process varies. With the help of the supply and demand mechanism, it is used to conduct a neighborhood search in a profitable region. This causes the commodity's quantity and price to continuously deviate from the equilibrium point while it is in an unstable condition. A worldwide search is conducted in the demand and supply mechanism in the search space. The transition from the economic theory's supply-demand mechanism to the SDO algorithm is seen in Figure 2 [20].

We assume that there are n marketplaces, each of which has d unique categories of items, each of which has a unique price and quantity, for the SDO algorithm. Where n stands for the number of markets (possible candidate solutions) and d for the amount of the commodity in each market (also variables) as well as the number of prices of the commodity in each market (also potential candidate solutions). The candidate solution is replaced by the prospective candidate solution if the potential candidate solution is superior to the candidate solution. There are two matrices that provide the commodity's price and quantity, respectively, as shown in equations 6 and 7 [20].

$$x = \begin{bmatrix} x_1^1 & \dots & x_1^d \\ \vdots & \vdots & \vdots \\ x_1^n & \dots & x_d^n \end{bmatrix} \tag{6}$$

$$y = \begin{bmatrix} y_1^1 & \dots & y_1^d \\ \vdots & \vdots & \vdots \\ y_1^n & \dots & y_d^n \end{bmatrix} \tag{7}$$

where X represents the market's commodity price matrix and Y represents the market's commodity quantity matrix. The j^{th} price of commodity in the i^{th} market is represented by $x_i^j = (i = 1, \dots, n), (j = 1, \dots, d)$, and $x_i (i = 1, \dots, n)$ is the i^{th} commodity price vector corresponding to a potential answer. The j^{th} commodity quantity in the i^{th} market is represented by $y_i^j = (i = 1, \dots, n), (j = 1, \dots, d)$, while the i^{th} commodity quantity vector is represented by $y_i (i = 1, \dots, n)$ [20].

Each market's quantity and price vectors for commodities are assessed using the fitness function. $F_x = [f_{x_1}, f_{x_2}, \dots, f_{x_n}]^T$ is an array for keeping the commodity price vectors fitness values in n markets for whole the markets. $F_y = [f_{y_1}, f_{y_2}, \dots, f_{y_n}]^T$ is another array for keeping the commodity quantity vectors function values in n markets, where T denotes the array's transposition. Two crucial elements of SDO are the commodity equilibrium price X_0 and equilibrium quantity Y_0; however, they are not known beforehand during iterations. Thus, they must first be

© 2023 NSP
Natural Sciences Publishing Cor.

named [20].
As seen below, the equilibrium quantity vector Y_0

$$N_i = |F_{y_i} - \frac{1}{n}\sum_{i=1}^{n} F_{y_i}| \tag{8}$$

$$Q = \frac{N}{\sum_{i=1}^{n} N_i} \tag{9}$$

$$Y_0 = Y_k, k = Roulettewheelselection \tag{10}$$

As seen below, the equilibrium price vector X_0

$$M_i = |F_{x_i} - \frac{1}{n}\sum_{i=1}^{n} F_{x_i} \tag{11}$$

$$P = \frac{M}{\sum_{i=1}^{n} M_i} \tag{12}$$

$$X_0 = \begin{cases} r_1.\frac{\sum_1^n x_i}{n} & \text{If } rand < 0.5 \\ x_k, & \text{k=Roulette wheel select P If } rand >= 0.5 \end{cases} \tag{13}$$

where r_1 is a chance number between 0 and 1. Following are suggested values for the supply function and demand function, respectively [20].

$$y_i(t+1) = y_0 + \alpha.(x_i(t) - x_0) \tag{14}$$

$$x_i(t+1) = x_0 - \beta.(y_i(t+1) - y_0) \tag{15}$$

$x_i(t)$ denotes the i^{th} commodity price at the time t, $y_i(t)$ denotes the i^{th} commodity quantity at the time t, and denote the weight of supply and weight of demand, respectively. These two weights can be expressed as follows [20]:

$$\alpha = \frac{2(T-t+1)}{T}\sin(2\pi r) \tag{16}$$

$$\beta = 2\cos(2\pi r) \tag{17}$$

where T is the iteration maximum number and r is a random value between [0, 1]. Let L equal the sum of the weights and.

$$L = \alpha\beta \tag{18}$$

where, the stability mode is denoted by $|L| < 1$ and the instability mode by $|L| > 1$. The values of the variable L are depicted over time, where the maximum number of iterations T is set to 1000 [20].

C. Local Search algorithms

Local search-based techniques begin with solution x. A neighborhood approach (i.e., N (x)) is then used to iteratively modify this solution until a local optimal solution is attained. As a result, the search space S is shown as a collection of locations s where $s \subset S$ and $\cup s = S$. Local search-based approaches may go deeper into the search region and discover the local best solution, in contrast to population-based approaches. Such techniques, however, do not cover the whole search field [21]. The most basic version of the local search-based technique is hill climbing (HC). The definition of the search space S, the objective function f(x), and the neighborhood structure N(x) all obviously influence HC. For the majority of optimization problems, the results from HC frequently produce extremely poor solutions because they frequently become trapped in local optima. hill climbing pseudo-code in algorithm 1 [22].

Algorithm1: hill climbing pseudo-code

© 2023 NSP
Natural Sciences Publishing Cor.

$x_i = LB_i + (UB_i - LB_i) \cup (0,1), \forall_i = 1,2,\ldots,N\{The initial solution x\}$
$Calculate(f(x))$
$itr = 0$
$while(itr <= Max\_Itr)$ do
  $x^{'} = improve(N(x))$
  $if(f(x^{'}) <= f(x))$ then
    $x = x^{'}$
  end if
$itr = itr + 1$
end while

3 Related works

Coronavirus has claimed the lives of many around the world, and no medicine has been found so far. It poses a great threat to all countries of the world, especially as it is spreading rapidly, which makes many people infected with the disease and need medical care, whether at home or in the hospital, and here we face the problem of rapidly consuming resources, including medicines and oxygen tubes. Masks, sterilization tools, and so on, especially in developing countries or poor villages, so scientists turned to technology and used machine learning and feature selection techniques to diagnose Corona disease based on X-rays, CT scans, tests, and symptoms, and we will see some of these methods below.

3.1 COVID-19 diagnosis based on X-ray

As the symptoms of pneumonia are similar to the symptoms of COVID-19, it is necessary to distinguish COVID-19 patients from pneumonia patients. They created a study that contains a set of X-ray data that includes three categories, namely, corona patients, pneumonia patients, and healthy people. They used many trained networks to extract features from X-ray images, and these networks are ALEX Net, VGG16, and others. These features are then taken to classify the patient as COVID or not using the RESNET 50 PLUS SVM with an accuracy of 98.6% [23].

Others have created a model to discover whether the patients who were examined are healthy, Covid-19, or pneumonia, and this model performed four basic functions. Pre-processing X-ray images, extracting features from X-ray images using CNN techniques, and choosing the most useful features using principle component analysis (PCA) and Relief finally, the classification stage, and here the accuracy reached 97.33%, and 100% [24].

Other scientists have proposed a method for the detection of COVID-19 based on X-rays. This method consists of four steps. The first step is the rudimentary processing of X-ray samples. the second step is extracting attributes from the images, which is called the Residual exemplar local binary pattern (ResEXLBP). the third step is to choose the most important features for the diagnostic stage by suggesting an iterative model called iterative ReliefF (IRF). Fourth they used five classifiers. They reached an accuracy of 99.69% using SVM plus LOOCV and a precision of 100% using SVM plus 10-FOLD CV [25].

In addition, they build a model called COVIDetectionNet uses features derived using CNN-based AlexNet to diagnose Covid-19 disease. The influential features are picked up using the Relief algorithm and then assorted using a Support vector machine (SVM), and this model achieved a precision of 99.18% [26].

In addition, others used other techniques to diagnose corona disease using X-rays, as they were able to reach an accuracy of 99.43%. The proposed method applies a deep convolution neural network to X-ray images to extract features, and the useful features were selected by a binary differential meta-algorithm. These picked features up passed to the SVM to assort the selected features for one of three image categories such as Health, Pneumonia, and COVID-19 [27].

others use deep learning models to obtain deep features from x-ray images; they extracted these features using three models from CNN (InceptionResNetV2, ResNet50, ResNet101). Both ACO and PSO were applied to select the relevant and important features. They also used both SVM and KNN as classifiers. They used five metrics to evaluate the performance of the classifiers (recall, specificity, accuracy, precision, and f1-score). SVM + PSO achieved an accuracy of 99.86%, and KNN + PSO achieved an accuracy of 99.41%. SVM+ACO achieved an accuracy of 99.83% and KNN + ACO achieved an accuracy of 99.35% [28].

In this research [29], deep learning models were also used to make a methodology for diagnosing COVID-19 based on

© 2023 NSP
Natural Sciences Publishing Cor.

X-ray images. This methodology is called MH-Covid Net, and it consists of four steps; the first is to pre-process images, and the second step works to extract features from these images using four Different models of deep learning (ResNet, VGG19, AlexNet, and GoogleNet). The third step is to select the relevant and necessary features using binary gray wolf optimization and binary particle swarm optimization. finally, SVM was used as a classifier. The proposed method achieved an accuracy of 99.38% [29].

Another study suggested a way to detect COVID-19. This method consists of four stages. The first stage is the preprocessing of the X-ray images. In the second stage, they used DCNN techniques like InceptionV3, VGG16, and Xception to extract the deep and related features. This output enters the next stage, which is the selection of features using coalition game theory (Approximated Shapley values) and Nystrom Sampling; then, the selected features are entered for the MLP classifier stage. The proposed model achieved an accuracy of 93.45% [30].

3.2 COVID-19 diagnosis based on Computerized tomography

Computerized tomography of the chest has an important role in the accurate diagnosis of coronavirus, as they suggested a model for detecting coronavirus from chest CT images. Therefore, they used CNN techniques to extract features from chest CT images. Then they used a two-phase feature selection. The first phase use relief techniques and mutual information (MI), and the second phase use the Dragonfly algorithm to define the most related features finally and, after that, takes these to classify whether they are COVID-19 and non-COVID-19 using a Support vector device classifier. the results for the first dataset were with an accuracy of 98.39% and the second with an accuracy of 90% [31].

other researchers have created a strategy called (covid-19 patients detection strategy), which contributes to two methods. The first method was called hybrid feature selection methodology (HFSM) to select more effect features from features that were extracted from chest CT images through a broad and powerful method used for image analysis called gray level co-occurrence matrix (GLCM). The second method is enhanced K-Nearest Neighbor (EKNN). This strategy has an accuracy of 96% [32].

Another study diagnosed COVID-19 using CT scans. CT images were processed using Nearest Neighbour interpolation and min-max scalar. then the features were extracted from the images using CNN models like (GoogleNet, VGG16 Net, and VGG19Net). the important features were selected using the Guided whale optimization algorithm based on stochastic fractal search After this stage. They use the Voting Classifier based on particle swarm optimization - A guided whale optimization algorithm to aggregate the predictions of four individual classifiers, called SVM, NN, KNN, and DT [33].

Others also used CT scans to distinguish pneumonia resulting from Covid-19 from suspected cases, so they processed CT scan images to reduce noise and increase sensitivity using Imaging Biomarker Explorer software. They used nine feature selection techniques to recognize worthy features from these methods Pearson correlation (PC), mutual information (MI), and so on. They used ten classifiers to assort and prophesy COVID-19 cases like Bagging (BAG), k-nearest neighborhood, and so on. RFE plus BAG classifier and RFE plus DT classifier achieved the highest accuracy at 98.4% [34].

In addition, building a model for detecting covid-19 disease depends on CT images consisting of two stages. The first stage is to extract features from the images using CNN, depending on DenseNet. And the second stage is to remove repetitive and useless features using a metaheuristic algorithm called Harris Hawks optimization (HHO) that is combined with Simulated Annealing (SA) and Chaotic Initialization to overcome on HHO's Shortfalls. Accuracy reached 98.85% with SA + Chaotic Initialization and reached 98.42% without SA + Chaotic Initialization [35].

Others suggested a model for diagnosing COVID-19 using CT scan images based on two steps. The first step is to extract features from CT scan images through different methods for CNN, such as ResNet189 and VGG1910. Then take the extracted features using a hybrid algorithm consisting of (a manta ray foraging optimizer and golden ratio optimization). The proposed model was evaluated on three datasets (COVID-CT, SARS-COV-2, and MOSMED). and they reached an accuracy of 99.15%, 99.42%, and 95. 57% of the datasets, respectively. they used three classifiers (SVM, MLP, and ELM) [36].

They combined manta ray foraging optimization (MRFO) as feature selection with Opposition-based learning (OBL) as a local search to create an efficient version of MRFO for diagnosing COVID-19 using a CT scan dataset consisting of 349 images. The proposed method was evaluated using Otsu's method and compared with six metaheuristic algorithms like the sine-cosine algorithm, based on three measures, Which best fitness values, SSIM matrices, and PSNR [37].

© 2023 NSP
Natural Sciences Publishing Cor.

3.3 COVID-19 diagnosis based on laboratory results

There are other multiple techniques to detect infection with COVID-19 disease or not. One of these techniques is a real-time reverse transcriptase-polymerase chain reaction (RT-PCR) that is taken from the throat or nose to obtain DNA and use it in diagnosis, but sometimes the results were incorrect, either for false-negative results or false-positive results. So the chest CT scan was considered more accurate, but it is also not specific enough for diagnosis because it sometimes fails to detect coronary lung tissue, and also, CT scans lack details. So digital lab tests are the best accuracy, so researchers created a strategy to detect corona patients based on digital laboratory results called distance-biased Naive Bayes (DBNB). It is divided into two phases (feature selection phase to select the most important features from laboratory test findings, and classification phase that uses advantages of naive bays algorithm and overcomes its shortages) [38].

Another study diagnosed COVID-19 based on the laboratory results of patients in a hospital in São Paulo, Brazil, consisting of 5,644 patients. This method consists of three stages. The first stage is pre-processing to normalize the dataset. The second stage is feature selection. They use a new Caledonian crow learning algorithm for selecting important features. Finally, selected features are inputs for the artificial neural network. This method achieved an accuracy of 94.31% [39].

They made a model that diagnoses corona patients based on epidemiological factors, comorbidities, Clinical symptoms, and laboratory data. It consists of four steps, first, preliminary data processing, second methods of selecting features, third classification model, and fourth matrixes to evaluate performance. They relied on this study on algorithms such as Chi-square, ANNOVA, mutual information, and f score. They arrived in this study using ANNOVA with the highest accuracy and were 93.8% [40].

Others proposed a model for diagnosing corona patients depending on machine learning and IOT. This model also depends on the laboratory results of the diagnosis, extracting 18 features from the results, using feature selection to pick up the most important features, then using these features for classification and repeating this process. The diagnostic accuracy reached 95%. The model is suitable for smart hospitals [41].

3.4 COVID-19 diagnosis based on patient's symptoms

They made a prognosis model that set apart stomachic cases of Coronavirus based on symptoms and features. This model is broken into three steps. The initial step selects the features and then increases the weight of each feature, as each feature has weight, and they are based on the information value technique to select important features. The second step classifies suspected patients into two categories. While the last step evaluates the efficiency of the suggested model based on testing the data and conducting a group of experiments, where the suggested model is rated using two different datasets. The results of this model reached an accuracy of 85% for the first dataset and 95.56% for the second dataset [42].

In Table 2, we summarize the related works to the technical terms used, accuracy, and what the diagnosis depends on.

Table 2: The summary of related work.

no	Corona disease diagnosis based on	Technologies used	Accuracy	References
1	x-ray	1. ALEX Net, VGG16, google Net, and others. 2. ResNet 50 plus Support vector machine (SVM).	98.66%	[23]
2	x-ray	1.CNN techniques like ResNet50, ResNet101, GoogleNet, and ResNet18. 2. Principle component analysis (PCA) and Relief. 3. Support vector machine (SVM).	97.33%. (Separate between three categories) 100% (Separate between two categories)	[24]
3	x-ray	1. Residual exemplar local binary pattern (ResEXLBP). 2. Iterative ReliefF (IRF). 3. Subspace discriminant, support vector machine (SVM), Decision Tree (DT), K Nearest neighborhood (KNN), and linear discriminant (LD).	99.69% of SVM plus LOOCV and 100% of SVM plus 10-FOLD CV	[25]

© 2023 NSP
Natural Sciences Publishing Cor.

4	x-ray	1. CNN-based AlexNet. 2. Relief algorithm. 3. Support vector machine (SVM).	99.18%	[26]
5	x-ray	1. Deep convolution neural network. 2. Binary differential meta-algorithm. 3. Support vector machine (SVM).	99.43%	[27]
6	x-ray	1. CNN techniques like ResNet50, ResNet101, Inception ResNetV2. 2. particle swarm optimization (PSO) algorithm and ant colony algorithm (ACO). 3. support vector machines (SVM) and a k-nearest neighbor (k-NN).	SVM + PSO 99.86% KNN + PSO 99.41% SVM+ACO 99.83% KNN + ACO 99.35%	[28]
7	x-ray	1. Image contrast enhancement algorithm. 2. Deep learning models like ResNet, VGG19, AlexNet, and GoogleNet. 3. Gray wolf optimization, binary particle swarm optimization. 4. SVM.	99.38%	[29]
8	x-ray	1. Bicubic interpolation. 2. DCNN techniques like InceptionV3, VGG16, and Xception. 3. Coalition game theory (Approximated Shapley values) and Nystrom Sampling. 4. MLP	93.45%	[30]
9	Chest CT images	1. CNN. 2. Mutual information (MI). 3. Relief-F. 4. Dragonfly.	98.39% for dataset1 90% for dataset2	[31]
10	Chest CT images	1. Enhanced K-Nearest Neighbor (EKNN). 2. Hybrid feature selection methodology (HFSM). 3. Gray level co-occurrence matrix (GLCM).	96%	[32]
11	Chest CT images	1. Nearest Neighbour interpolation and min-max scalar. 2. CNN models like (AlexNet, GoogleNet, VGG16 Net, ResNet50, and VGG19Net). 3. Guided whale optimization algorithm based on stochastic fractal search. 4. SMOTE and LSH-SMOTE. 5. SVM, NN, KNN, and DT.	—	[33]
12	Chest CT images	1. Imaging Biomarker Explorer (IBEX, MD Anderson Cancer Center) software. 2. Open-source IBEX feature extraction software, GLRLM, intensity direct, and so on. 3. Pearson correlation (PC), mutual information (MI), Logistic Regression (LR), Recursive feature elimination (RFE), and so on.	98.4%	[34]
13	Chest CT images	1. CNN depends on DenseNet. 2. Metaheuristic algorithm called Harris Hawks optimization (HHO). 3. Simulated Annealing (SA) and Chaotic Initialization.	98.85% (with SA + Chaotic Initialization and reached) 98.42% (without SA + Chaotic Initialization.)	[35]
14	Chest CT images	1. CNN, such as ResNet189, GoogleNet8, and VGG1910. 2. Manta ray foraging optimizer, golden ratio optimization. 3. SVM, MLP, and ELM.	COVID-CT 99.15% SARS-COV-2 99.42% MOSMED 95. 57%	[36]
15	Chest CT images	1. Manta ray foraging optimization (MRFO). 2. Opposition-based learning (OBL).	—	[37]
16	Digital laboratory results	1. Advanced particle swarm optimization (APSO). 2. Weighted naïve Bayes module (WBNM). 3. Distance Reinforcement module (DRM).	94.271% for APSO	[38]

© 2023 NSP
Natural Sciences Publishing Cor.

17	laboratory results	1. Caledonian crow learning algorithm. 2. Artificial neural network.	94.31%	[39]
18	Epidemiological factors, comorbidities, clinical symptoms, and laboratory data	1. Chi-square. 2. ANNOVA. 3. Mutual information (MI).	93%	[40]
19	ML, IoT, and laboratory	1. Naïve Bayes (NB). 2. Random Forest (RF). 3. Support vector machine (SVM).	95%	[41]
20	Patient's symptoms	1. Non-dominated sorting genetic algorithm (NSGA-11). 2. AdaBoost classifier.	85% for dataset1 95.56% for dataset2	[42]

4 Materials and Methods

The process of the suggested method for COVID-19 detection has been detailed in this part one by one. The entire piece is broken down into three subsections, which are: (1) The description of the dataset, (2) The feature selection model, and (3) The parameters used for algorithms.

4.1 The dataset description

In our study, we used three datasets, each with different features and sizes, and they are described in Tables 3, 4, and 5. We obtained the first two datasets from a study [42], and the third dataset from Kaggle [43]. The first dataset included 1495 cases and 12 features. It included 757 people with Covid-19. In contrast, the second dataset included 99,232 samples and 8 features, including 5 features of primary symptoms. Where it includes 8,393 infected with Covid-19, the third dataset included 5434 cases and 20 features, including 7 features of primary symptoms and 6 of chronic diseases and other information. There are currently 4383 individuals who have been infected with Covid-19. Pre-processing is one of the key stages for datasets before it's used in a training model to address incompleteness, imbalance, and other issues. This does not apply to the first and third datasets, unlike the second dataset, as it suffers from imbalance, as the number of people infected with Covid-19 is much lower than the number of people without Covid. 19 To address this problem, we used the Synthetic Minority Oversampling Technique (SMOTE) [44].

Table 3: The dataset 1 description.

DataType: Numeric	
Features	**Values**
Age	[1:98]
Gender	1: Male 0: Female
Fever	1: Yes 0: No
Cough	1: Yes 0: No
Fatigue	1: Yes 0: No
Pains	1: Yes 0: No
Nasal Congestion	1: Yes 0: No
Shortness of breath	1: Yes 0: No
Runny Nose	1: Yes 0: No
Sore Throat	1: Yes 0: No
Diarrhea	1: Yes 0: No
Chills	1: Yes 0: No
Headache	1: Yes 0: No
Vomiting	1: Yes 0: No
Lives in affected area	1: Yes 0: No
Output	1: Covid-19 0: Non Covid19

Table 4: The dataset 2 description.

DataType: Numeric

© 2023 NSP
Natural Sciences Publishing Cor.

Features	**Values**
Age (age 60 years or above)	1: Yes 0: No
Sex	1: Male 0: Female
Cough	1: Yes 0: No
Fever	1: Yes 0: No
Sore throat	1: Yes 0: No
Shortness of breath	1: Yes 0: No
Headache	1: Yes 0: No
Known contact with an individual confirmed to have Covid-19	1: Yes 0: No
Output	1: Covid-19 0: Non Covid19

Table 5: The dataset 3 description.

DataType: Numeric	
Features	**Values**
Breathing Problem	1: Yes 0: No
Fever	1: Yes 0: No
Dry Cough	1: Yes 0: No
Sore throat	1: Yes 0: No
Running Nose	1: Yes 0: No
Asthma	1: Yes 0: No
Chronic Lung Disease	1: Yes 0: No
Headache	1: Yes 0: No
Heart Disease	1: Yes 0: No
Diabetes	1: Yes 0: No
Hyper Tension	1: Yes 0: No
Fatigue	1: Yes 0: No
Gastrointestinal	1: Yes 0: No
Abroad travel	1: Yes 0: No
Contact with COVID Patient	1: Yes 0: No
Attended Large Gathering	1: Yes 0: No
Visited Public Exposed Places	1: Yes 0: No
Family working in Public Exposed Places	1: Yes 0: No
Wearing Masks	1: Yes 0: No
Sanitization from Market	1: Yes 0: No
Output	1: Covid-19 0: Non Covid19

4.2 The feature selection model

Developing a method to reduce feature size is essential as it helps to decrease the measurement of feature vectors without compromising data and enhances accuracy by reducing computational load. We utilized two optimizers, SDO and NMR, and preferred hill climbing as the local search technique to optimize a criterion among multiple potential solutions. These models are addressed in a background section. Figure 3 shows a hybrid approach using characteristics generated from two separate metaheuristic models, a local search model. The proposed SDO-NMR-hill approach work in two stages.

The first stage: is the selection of features, and There are two steps in this stage. The first step is to use SDO to return the best initial solutions for the second step. We also note the SDO algorithm Pseudo-code in algorithm 2. The second step is to use NMR, and Hill climbing are used to determine which feature combination is the best for the second stage. We also note the NMR and hill-climbing algorithm Pseudo-code in algorithm 3. We have picked three different classifiers, including RF, DT, and AdaBoost, to calculate the fitness function. The transfer function's job is to turn the feature set into a string of 0s and 1s so that the sample can be trained to its final state. According to table 6 [45], there are two main categories of transfer functions. We have utilized the S4 function for binarization for this purpose.

© 2023 NSP
Natural Sciences Publishing Cor.

In the second stage: the (RF, DT, and AdaBoost) classifiers are applied to predict the accuracy based on the selected optimal features. The class distribution is balanced using the sampling approach (SMOTE) before the training procedure starts since dataset 2 appears unbalanced.

Table 6: Two different types of transfer functions.

S-Shaped functions Family		V-Shaped functions Family	
Label	**Transfer function**	**Label**	**Transfer function**
S1	$T(x)=\frac{1}{1+e^{x}}$	V1	$T(x)=\lvert\tanh(x)\rvert$
S2	$T(x)=\frac{1}{1+e^{2x}}$	V2	$T(x)=\lvert\mathrm{erf}(\frac{\sqrt{\pi}}{2}x)\rvert$
S3	$T(x)=\frac{1}{1+e^{3x}}$	V3	$T(x)=\lvert\frac{x}{\sqrt{1+x^2}}\rvert$
S4	$T(x)=\frac{1}{1+e^{\frac{x}{2}}}$	V4	$T(x)=\lvert\frac{2}{\pi}\arctan(\frac{\pi}{2}x)\rvert$

Algorithm 2: The supply-demand-based optimization (SDO) Algorithm Pseudo-code

Initialize price and quantity population
Calculate price and quantity population fitness
while $iter <= maxiter$
  for i: market_size
    $N_i = |F_{y_i} - \frac{1}{n}\sum_{i=1}^{n} F_{y_i}|$
    $Q = \frac{N}{\sum_{i=1}^{n} N_i}$
    $Y_0 = Y_k, k = Roulettewheelselection(Q)$
    $M_i = |F_{x_i} - \frac{1}{n}\sum_{i=1}^{n} F_{x_i}|$
    $P = \frac{M}{\sum_{i=1}^{n} M_i}$
    $X_0 = \begin{cases} r_1.\frac{\sum_1^n x_i}{n} & \text{If } rand < 0.5 \\ x_k, & \text{k=Roulette wheel select P If } rand >= 0.5 \end{cases}$
    $\alpha = \frac{2(T-t+1)}{T}\sin(2\pi r)$
    $\beta = 2\cos(2\pi r)$
    Find the new solution for quantity using: $y_i(t+1) = y_0 + \alpha.(x_i(t) - x_0)$
    Find the new solution for price using: $x_i(t+1) = x_0 - \beta.(y_i(t+1) - y_0)$
      if $y_i(t+1) <= y_i(t)$
        $y_i(t) = y_i(t+1)$
      End if
      if $x_i(t+1) <= x_i(t)$
        $x_i(t) = x_i(t+1)$
      End if
  End for
  Calculate new price and new quantity population fitness
  Compare between the new price and new quantity population fitness,
  update the overall best, update Iter count
End while
Return the best initial solutions

Algorithm 3: The Naked Mole Rat (NMR) and hill climbing Algorithm Pseudo-code

Declare population numbers: n,breeders number: B=n/5,workers number:w=n-B,breeding probability: bp,dimensional vector: D
take the initial population from SDO
while $iter <= maxiter$
  itr=0
  while $(itr <= n)$ do
    $x_i = randomsolution$
    Calculate(f(x))
    $x^{'} = improve(N(x))$

© 2023 NSP
Natural Sciences Publishing Cor.

if($f(x^{'}) <= f(x)$) then
$x = x^{'}$
end if
itr=itr+1
end while
for $i = 1 : w$
Find the new solution for worker using:$w_i^{(t+1)} = w_i^t + \lambda(w_k^t - w_j^t)$
End for
For $i = 1 : B$
If $U(0,1) > bp$
Find the new solution for breeder using: $b_i^{(t+1)} = (1-\lambda)b_i^t + \lambda(d - b_i^t)$
End if
End for
combines the new worker and breeder population
evaluate the population, update the overall best d, update Iter count
Save overall best
end while
return the final best (d)

4.3 The parameters used for algorithms

The following are the parameters that the feature selection optimization and classification algorithms use:

A. Common parameters:
Population Size= 100, Max iteration=30, Run num=15.

B. RF parameters:
Dataset 1: N_estimators=10, criterion= entropy.
Dataset 2: N_estimators=10, criterion= gini, max_depth=5, max_feature=1.
Dataset 3: N_estimators=5, criterion= entropy.

C. DT parameters:
Dataset 1: ccp_alpha=0.002, criterion= gini, max_depth=10, min_samples_leaf=1, min_samples_split=5.
Dataset 2: Max_depth=6, criterion= entropy, min_samples_leaf=3.
Dataset 3: max_depth=13, criterion=gini, min_samples_leaf=3.

D. AdaBoost parameters:
Dataset 1, Dataset 3: N_estimators=177, max_depth=6, learning_rate=0.002.
Dataset 2: N_estimators=99, max_depth=11, learning_rate=0.022.

E. NMR parameters:
Bp=0.5, Breeder size=20, Worker size=80.

5 Results and Discussion

The following groups make up the experimental study:
1. the three COVID-19 datasets
2. benchmarking

5.1 The three COVID-19 datasets

We submit the empirical findings from the three COVID-19 detection datasets, whose summaries were previously provided in the preceding section. The experiment covers the outcomes (accuracy after FS, mean accuracy, accuracy standard divination (SD), and #features after FS) of various classifiers (DT, RF, AdaBoost) utilized for the SDO-NMR-hill

© 2023 NSP
Natural Sciences Publishing Cor.

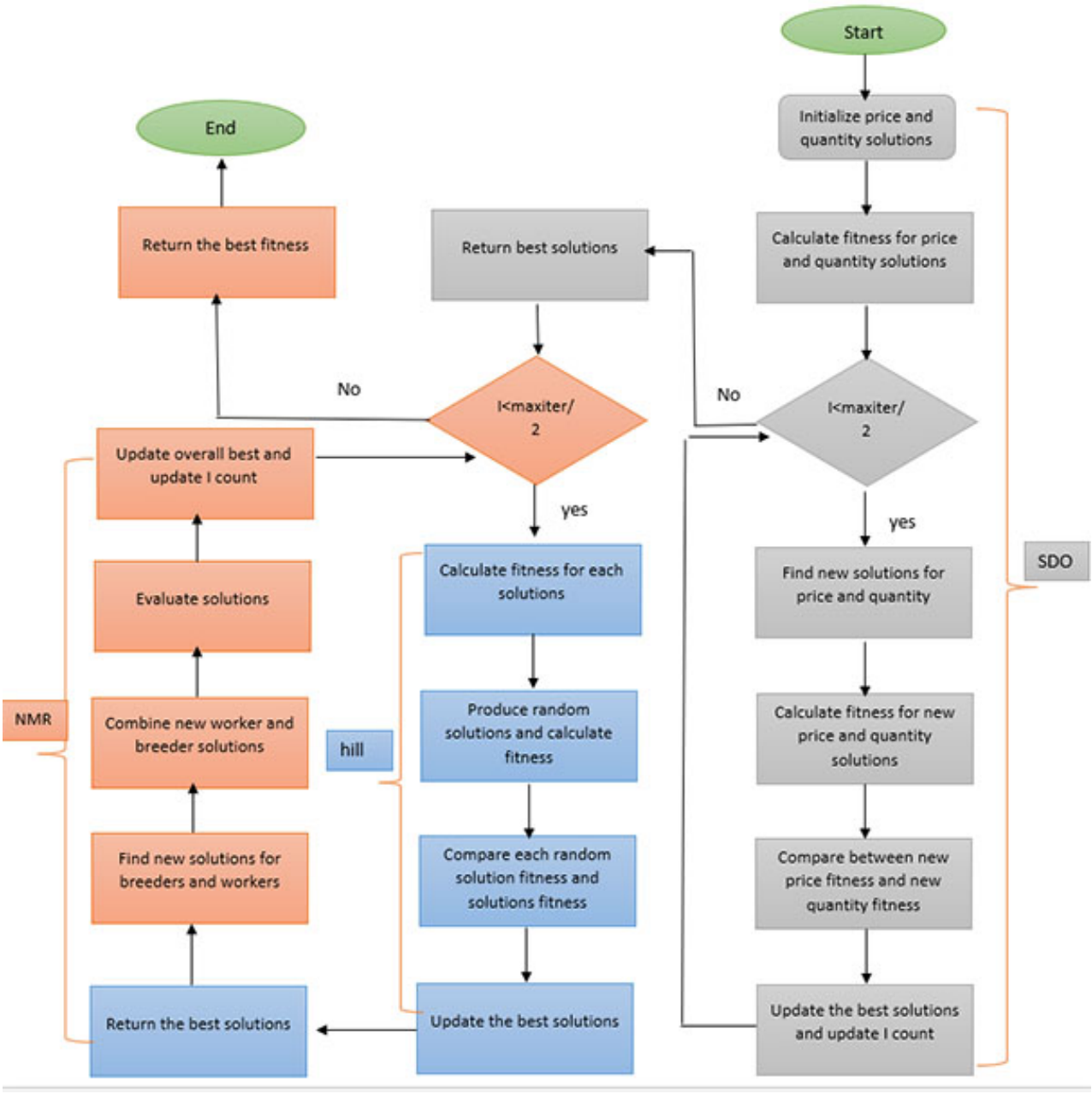

Fig. 3: The flowchart showcases the hybrid of two distinct metaheuristic models with a local search model.

algorithm's fitness computation. Comparative analysis of alternative optimization techniques. To verify the SDO-NMR-Hill algorithm's superiority, we tested some well-known optimization methods on all three datasets and compared the outcomes to those attained by the SDO-NMR-Hill algorithm. **The algorithms which we have chosen for comparison:** improved squirrel search algorithm (ISSA) [46], squirrel search algorithm (SSA) [47], artificial bee colony (ABC) [48], particle swarm optimization (PSO) [49], bat algorithm (BA) [50], grey wolf optimizer (GWO) [51], whale optimization algorithm (WOA) [52], grasshopper optimization algorithm (GOA) [53], sailfish optimizer algorithm (SFO) [54], Harris hawks optimization (HHO) [55], Atom search optimization (ASO) [56], Henry Gas Solubility Optimization (HGSO) [57], and Bird swarm optimization (BSA) [58].

From the results shown in tables 7,8, 9, and 10, it can be seen that the accuracy and optimal results of the SDO-NMR-hill algorithm running 15 times on three datasets are higher than other algorithms.

© 2023 NSP
Natural Sciences Publishing Cor.

We divided the dataset1 by 80% to 20% for training and testing, respectively, and the accuracy was extracted at 86.2%, 85.61%, and 86.2% from RF, DT, and AdaBoost, respectively. We divided the dataset1 by 88% to 12% for training and testing, respectively, and the accuracy of 87.22%, 87.22%, and 88.88% was extracted from RF, DT, and AdaBoost, respectively. As shown in the table 7. AdaBoost plus the SDO-NMR-Hill achieved the highest accuracy for the first dataset is 88.88%, but it is higher in terms of processing time. Unlike DT, it was the fastest one in extracting the outputs. In the second dataset, we did two different pre-processing, where the first pre-processing achieved accuracy 95.89%, 95.89%, and 95.89 from RF, DT, and AdaBoost, respectively. The second pre-processing achieves an accuracy of 97.97%, 97.93%, and 97.93% from RF, DT, and AdaBoost, respectively. As shown in the tables 8, and 9. RF plus the SDO-NMR-Hill achieved the highest accuracy for the second dataset is 97.97%. Finally, the third dataset achieves an accuracy of 98.52%, 98.80%, and 98.98% from RF, DT, and AdaBoost, respectively, as shown in the table 10. AdaBoost plus the SDO-NMR-Hill achieved the highest accuracy for the third dataset is 98.98%.

Table 7: The dataset1 results.

Dataset Name: dataset 1												
Model Name	**Random forest**				**Decision Tree**				**AdaBoost**			
	Accuracy after FS (%)	**Mean Accuracy**	**Accuracy SD**	**# Selected features**	**Accuracy after FS (%)**	**Mean Accuracy**	**Accuracy SD**	**# Selected features**	**Accuracy after FS (%)**	**Mean Accuracy**	**Accuracy SD**	**# Selected features**
ISSA	82.61	82.27	0.0030	10	81.61	81.40	0.0016	9	80.94	80.94	0.0000	6
SSA	82.61	82.23	0.0036	10	81.61	81.32	0.0011	9	80.94	80.94	0.0000	5
ABC	82.61	82.23	0.0024	10	81.27	81.29	0.0008	5	80.94	80.94	0.0000	5
PSO	81.94	81.34	0.0037	7	81.27	81.05	0.0034	5	80.94	80.65	0.0062	5
BA	82.27	81.63	0.0052	7	81.27	81.16	0.0032	5	80.94	80.78	0.0042	5
GWO	82.61	82.32	0.0034	10	81.61	81.38	0.0016	9	80.94	80.94	0.0000	5
WOA	82.61	82.25	0.0045	10	81.27	81.25	0.0019	5	80.94	80.91	0.0008	5
GOA	82.61	82.34	0.0022	10	81.61	81.36	0.0015	9	80.94	80.89	0.0011	5
SFO	82.61	82.19	0.0029	10	81.61	81.34	0.0013	9	80.94	80.94	0.0000	6
HHO	82.61	82.07	0.0036	10	81.27	81.40	0.0000	5	80.94	80.71	0.0043	6
BSA	82.61	82.30	0.0029	10	81.34	81.40	0.0018	9	80.94	80.91	0.0008	5
ASO	82.61	81.98	0.0027	10	81.34	81.40	0.0018	9	80.94	80.89	0.0011	6
HGSO	82.27	81.76	0.0036	7	81.61	81.16	0.0016	6	80.94	80.42	0.0067	7
SDO	81.27	81.24	0.0003	2	82.94	76.52	0.1244	4	81.27	78.50	0.0041	8
NMR	83.61	83.58	0.0003	3	83.61	83.61	0.0000	3	83.61	79.93	0.011	4
SDO-NMR-hill 0.12	**87.22**	**87.22**	**0.0000**	**5**	**87.22**	**87.11**	**0.0004**	**5**	**88.88**	**88.88**	**0.0000**	**6**
SDO-NMR-hill 0.2	**86.2**	**86.2**	**0.0000**	**6**	**85.61**	**85.61**	**0.0000**	**5**	**86.2**	**86.2**	**0.0000**	**6**

Table 8: the dataset2 results for first pre-processing.

Dataset Name: dataset 2												
Model Name	**Random forest**				**Decision Tree**				**AdaBoost**			
	Accuracy after FS (%)	**Mean Accuracy**	**Accuracy SD**	**# Selected features**	**Accuracy after FS (%)**	**Mean Accuracy**	**Accuracy SD**	**# Selected features**	**Accuracy after FS (%)**	**Mean Accuracy**	**Accuracy SD**	**# Selected features**
ISSA	95.65	95.65	0.0000	4	95.65	95.65	0.0000	4	95.19	95.19	0.0000	2
SSA	95.65	95.65	0.0000	4	95.65	95.65	0.0000	4	95.19	95.19	0.0000	2
ABC	95.65	95.65	0.0000	4	95.65	95.65	0.0000	4	95.19	95.19	0.0000	2
PSO	95.65	95.63	0.0007	4	95.65	95.68	0.0006	4	95.17	95.17	0.0000	2
BA	95.65	95.67	0.0014	4	95.65	95.69	0.0007	4	95.19	95.19	0.0000	2
GWO	95.65	95.65	0.0000	4	95.65	95.67	0.0005	4	95.19	95.19	0.0000	2
WOA	95.65	95.65	0.0000	4	95.65	95.67	0.0005	4	95.15	95.19	0.0000	2
GOA	95.65	95.65	0.0000	4	95.65	95.67	0.0005	4	95.19	95.19	0.0000	2
SFO	95.65	95.65	0.0000	4	95.65	95.65	0.0000	4	95.19	95.19	0.0000	2

© 2023 NSP
Natural Sciences Publishing Cor.

HHO	95.65	95.65	0.0000	4	95.65	95.65	0.0000	4	95.19	95.19	0.0000	2
BSA	95.65	95.65	0.0000	4	95.65	95.65	0.0000	4	95.19	95.19	0.0000	2
ASO	95.65	95.65	0.0000	4	95.65	95.65	0.0000	4	95.19	95.19	0.0000	2
HGSO	95.65	95.67	0.0000	4	95.65	95.72	0.0011	4	95.19	95.19	0.0000	2
SDO	93.44	93.44	0.0000	1	93.99	93.99	0.0000	2	93.92	93.92	0.0000	2
NMR	93.99	93.42	0.0036	2	93.99	93.47	0.0028	2	93.99	93.43	0.0030	2
SDO-NMR-hill	**95.89**	**95.89**	**0.0000**	**4**	**95.89**	**95.89**	**0.0000**	**4**	**95.89**	**95.89**	**0.0000**	**4**

Table 9: The dataset2 results for second pre-processing.

Dataset Name: dataset 2												
Model Name	**Random forest**				**Decision Tree**				**AdaBoost**			
	Accuracy after FS (%)	**Mean Accuracy**	**Accuracy SD**	**# Selected features**	**Accuracy after FS (%)**	**Mean Accuracy**	**Accuracy SD**	**# Selected features**	**Accuracy after FS (%)**	**Mean Accuracy**	**Accuracy SD**	**# Selected features**
ISSA	97.93	97.93	0.0000	7	97.93	97.93	0.0000	7	97.86	97.86	0.0000	6
SSA	97.93	97.93	0.0000	7	97.93	97.93	0.0000	7	97.86	97.86	0.0000	6
ABC	97.93	97.93	0.0000	7	97.93	97.93	0.0000	7	97.86	97.86	0.0000	6
PSO	97.93	97.82	0.0016	7	97.93	97.85	0.0012	7	97.86	97.76	0.0008	6
BA	97.93	97.92	0.0004	7	97.93	97.91	0.0006	7	97.86	97.82	0.0007	6
GWO	97.93	97.93	0.0000	7	97.93	97.93	0.0000	7	97.86	97.86	0.0000	6
WOA	97.93	97.93	0.0000	7	97.93	97.93	0.0000	7	97.86	97.86	0.0000	6
GOA	97.93	97.93	0.0000	7	97.93	97.93	0.0000	7	97.86	97.84	0.0005	6
SFO	97.93	97.93	0.0000	7	97.93	97.93	0.0000	7	97.86	97.86	0.0000	6
HHO	97.93	97.93	0.0000	7	97.93	97.93	0.0000	7	97.86	97.86	0.0000	6
BSA	97.93	97.93	0.0000	7	97.93	97.93	0.0000	7	97.86	97.84	0.0005	6
ASO	97.93	97.93	0.0001	7	97.93	97.91	0.0006	7	97.86	97.86	0.0000	6
HGSO	97.93	97.93	0.0001	7	97.93	97.93	0.0001	7	97.86	97.75	0.0012	6
SDO	96.22	96.18	0.0014	1	96.22	96.16	0.0021	1	96.22	96.22	0.0000	1
NMR	96.82	96.15	0.0033	2	96.22	96.12	0.0025	1	96.22	96.22	0.0000	1
SDO-NMR-hill	**97.97**	**97.97**	**0.0000**	**7**	**97.93**	**97.93**	**0.0000**	**7**	**97.93**	**97.93**	**0.0000**	**7**

Table 10: The dataset3 results.

Dataset Name: dataset 3												
Model Name	**Random forest**				**Decision Tree**				**AdaBoost**			
	Accuracy after FS (%)	**Mean Accuracy**	**Accuracy SD**	**# Selected features**	**Accuracy after FS (%)**	**Mean Accuracy**	**Accuracy SD**	**# Selected features**	**Accuracy after FS (%)**	**Mean Accuracy**	**Accuracy SD**	**# Selected features**
ISSA	97.33	96.85	0.0049	9	98.62	98.55	0.0005	11	97.70	97.62	0.0010	14
SSA	97.33	96.88	0.0052	6	98.62	98.55	0.0008	11	97.70	97.57	0.0011	15
ABC	97.33	97.09	0.0046	7	98.62	98.59	0.0005	10	97.52	97.45	0.0012	11
PSO	97.33	96.50	0.0045	6	98.62	98.44	0.0007	12	97.15	96.95	0.0018	12
BA	97.33	96.10	0.0058	9	98.62	98.52	0.0008	12	97.52	97.28	0.0019	13
GWO	97.33	96.92	0.0045	7	98.62	98.56	0.0006	10	97.70	97.66	0.0005	14
WOA	97.33	96.72	0.0059	6	98.53	98.54	0.0007	10	97.70	97.60	0.0010	14
GOA	97.33	97.05	0.0034	6	98.62	98.55	0.0008	11	97.70	97.42	0.0017	14
SFO	97.33	97.01	0.0044	7	98.62	98.60	0.0005	11	97.61	97.56	0.0014	13
HHO	97.33	96.67	0.0058	10	98.62	98.51	0.0009	11	97.61	97.52	0.0018	13
BSA	97.33	96.92	0.0038	9	98.53	98.53	0.0006	10	97.61	97.49	0.0017	12
ASO	97.33	96.85	0.0048	9	98.62	98.56	0.0007	11	97.70	97.49	0.0016	14
HGSO	97.15	96.06	0.0039	9	98.44	98.45	0.0005	10	97.42	97.30	0.0014	11
SDO	97.70	95.76	0.0299	12	97.33	94.79	0.0340	13	97.79	95.28	0.0488	12

© 2023 NSP
Natural Sciences Publishing Cor.

NMR	98.16	97.68	0.0028	11	98.06	97.57	0.0043	11	98.16	97.81	0.0024	11
SDO-NMR-hill	**98.52**	**98.52**	**0.0000**	**11**	**98.80**	**98.80**	**0.0000**	**10**	**98.98**	**98.97**	**0.0004**	**12**

Comparison with modern techniques. Results from recent research on the aforementioned datasets have been compared with those from the current one to assess the effectiveness of the proposed framework. We used a confusion matrix (CM) for performance measurements. The confusion matrix, depicted in table 11, is a table used to illustrate how well a classification model performs on a set of test data for which the real values are known. Accuracy, sensitivity, specificity, precision, and F-Score data were taken from CM to evaluate model performance.

Table 11: The confusion matrix.

		Predicted	
		Negative	Positive
Actual	Negative	TN(True Negative)	FP (False Positive)
	Positive	FN (False Negative)	TP (True Positive)

TP: To identify the sick as sick. FP: To identify the health of a sick. TN: To identify the healthful as healthy. FN: To identify the sick as healthful [28]. Accuracy: shows the proportion of rightly predicted instances in the dataset as a whole. The formula is as follows [42]:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FN+FP)} \tag{19}$$

Precision: is the accuracy that captures the proportion of rightly predicted members of the positive class. It is an indicator of the percentage of patients who the classifier identified as having COVID-19 who really had the virus. The formula is as follows [42]:

$$Precision = \frac{TP}{(TP+FP)} \tag{20}$$

Sensitivity (Recall): the proportion of COVID-19 instances that were rightly anticipated (Probability of a positive test given the probable COVID-19 virus infection in the patients). This is how it is calculated [42]:

$$Sensitivity = \frac{TP}{(TP+FN)} \tag{21}$$

Specificity: reflects the proportion of NON-COVID-19 instances that were mistakenly anticipated (Given that there are probable COVID-19-free patients, the likelihood of a negative test is high). The formula is as follows [42]:

$$Specificity = \frac{TN}{(FP+TN)} \tag{22}$$

F1-Score: A coordinated mean of the values for precision and recall. It achieves the ideal poise between Recall and Precision, giving a reliable assessment of the model's effectiveness in categorizing COVID-19 sick [42]:

$$F1-score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \tag{23}$$

AUC: This metric makes it possible to determine how well a model can differentiate between COVID-19-infected and uninfected patients. The formula is as follows [42]:

$$AUC = \frac{(Sensitivity + Specificity)}{2} \tag{24}$$

Table 12 presents the findings of the comparative investigations. Over all of the aforementioned datasets, the suggested technique produces the best result.

Table 12: The results of the comparative studies.

Dataset	**Reference**	**Name of model**	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**	**F1-score**	**AUC**

© 2023 NSP
Natural Sciences Publishing Cor.

1	[42]	NSGA-II + DT	84.51	82.8	88.46	80	85.54	84.23
		NSGA-II + RF	83.74	85.78	82.48	85.12	84.10	83.8
		NSGA-II + AdaBoost	85	90	89.32	85	86.01	87.16
	our model	SDO-NMR-hill +DT	**87.22**	**86.96**	**87.91**	**86.52**	**87.43**	**87.215**
		SDO-NMR-hill +RF	**87.22**	**86.96**	**87.91**	**86.52**	**87.43**	**87.215**
		SDO-NMR-hill + AdaBoost	**88.88**	**89.01**	**89.01**	**88.76**	**89.01**	**88.885**
2	[42]	NSGA-II + DT	92.52	92.52	92.52	93.91	92.52	93.22
		NSGA-II + RF	92.44	92.44	92.44	93.91	92.44	93.26
		NSGA-II + AdaBoost	95.56	95.56	95.56	98.19	95.56	96.87
	our model	SDO-NMR-hill +DT	**97.93**	**97.93**	**97.93**	**99.06**	**97.93**	**98.495**
		SDO-NMR-hill +RF	**97.97**	**97.97**	**97.97**	**99.38**	**97.97**	**99.675**
		SDO-NMR-hill + AdaBoost	**97.93**	**97.93**	**97.93**	**99.06**	**97.93**	**98.495**

5.2The benchmark

The proposed SDO-NMR-Hill model is evaluated on nine benchmark UCI datasets and compared with other optimization algorithms using the best and mean accuracy and a number of selected features. The datasets are described in terms of the feature number, sample number, class number, and dataset field in Table 13. We have studied the impacts of the performance of SDO-NMR-hill with two classifiers, RF and DT. From the results shown in tables 14 and 15, it can be seen that the accuracy and optimal results of the SDO-NMR-hill algorithm running 15 times on nine datasets are higher than those of other algorithms in some datasets and equalize the power of algorithms in others.

Table 13: The UCI datasets description.

label	Name of dataset	Attributes number	Samples number	Classes number	domain of Dataset
1	CongressEW	16	435	2	Politics
2	KrvskpEW	36	3196	2	Game
3	PenglungEW	325	73	2	Biology
4	Lymphography	18	148	2	Biology
5	M-of-n	13	1000	2	Biology
6	Tic-Tac-Toe	9	958	2	Game
7	WineEW	13	178	3	Chemistry
8	Zoo	16	101	6	Artificial
9	Exactly	13	1000	2	Biology

Table 14: The algorithm ran 15 times on nine datasets benchmarked with RF and compared with 15 other algorithms.

Dataset name ↓	Model name →	ISSA	SSA	ABC	PSO	BA	GWO	WOA	GOA	SFO	HHO	BSA	ASO	HGSO	SDO	NMR	SDO-NMR-hill
CongressEW	Best accuracy	98.85	98.85	98.85	97.70	97.70	98.85	97.70	98.85	98.85	97.70	98.85	98.85	97.70	100	100	100
	Mean accuracy	98.01	97.78	98.01	97.62	97.70	97.78	97.70	97.78	98.24	97.70	98.16	97.93	97.70	97.85	100	100
	#Selected features	7	7	7	3	3	7	3	7	7	3	7	7	4	6	3	3
KrVsKpEW	Best accuracy	95.00	95.00	95.31	94.69	95.16	95.47	94.84	96.56	95.31	94.53	95.78	95.63	94.53	96.24	99.21	99.68
	Mean accuracy	93.94	94.07	94.44	93.87	93.52	94.55	94.11	94.51	94.42	93.79	94.44	94.38	93.30	80.70	98.43	99.68
	#Selected features	19	14	18	12	16	18	12	13	14	21	14	19	14	14	23	21
PenglungEW	Best accuracy	80.00	86.67	80.00	80.00	80.00	86.67	80.00	80.00	86.67	80.00	86.67	73.33	73.33	86.66	93.33	100
	Mean accuracy	74.22	75.56	74.22	75.56	73.78	74.22	76.00	73.78	81.33	72.44	74.22	72.44	68.44	64.88	89.33	91.55
	#Selected features	56	230	147	3	16	36	3	144	8	147	161	149	139	149	315	61
Lymphography	Best accuracy	90.00	93.33	90.00	86.67	90.00	90.00	93.33	93.33	90.00	90.00	90.00	93.33	86.67	90	96.66	100
	Mean accuracy	87.56	87.33	87.11	84.00	84.22	87.78	86.89	87.56	87.56	86.22	86.89	86.89	84.00	80	96.00	95.55

© 2023 NSP
Natural Sciences Publishing Cor.

	#Selected features	10	13	13	5	10	8	13	13	7	8	8	13	7	7	7	8
M-of-N	Best accuracy	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	Mean accuracy	96.93	96.80	99.80	94.17	95.40	9850	97.57	99.03	98.27	96.57	98.70	98.03	93.97	80.40	100	100
	#Selected features	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
Tic-tac-toe	Best accuracy	86.98	86.98	86.98	86.98	86.98	86.98	86.98	86.98	86.98	86.98	86.98	86.98	86.98	88.02	88.54	91.14
	Mean accuracy	86.98	86.98	86.98	86.15	86.28	86.98	86.98	86.91	86.98	86.98	86.98	86.98	86.49	75.87	88.54	90
	#Selected features	7	7	7	7	7	7	7	7	7	7	7	7	7	8	7	8
WineEW	Best accuracy	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	Mean accuracy	100	100	100	100	100	100	100	100	100	100	100	100	100	92.22	99.81	100
	#Selected features	3	3	3	3	3	3	3	3	3	3	3	3	3	9	7	3
Zoo	Best accuracy	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	Mean accuracy	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	#Selected features	4	4	4	4	4	4	4	4	4	4	4	4	5	6	5	4
Exactly	Best accuracy	84	84	84	84	72	84	84	84	84	84	84	84	72	85	100	100
	Mean accuracy	74.13	78.27	79.45	70.10	69.60	73.17	71.45	72.08	75.37	71.28	71.48	69.17	69.77	71.46	99.93	100
	#Selected features	6	6	6	6	5	6	6	6	6	6	6	6	5	10	7	6

Table 15: The algorithm ran 15 times on nine datasets benchmarked with DT and compared with 15 other algorithms.

Dataset name ↓	**Model name →**	**ISSA**	**SSA**	**ABC**	**PSO**	**BA**	**GWO**	**WOA**	**GOA**	**SFO**	**HHO**	**BSA**	**ASO**	**HGSO**	**SDO**	**NMR**	**SDO-NMR-hill**
CongressEW	Best accuracy	97.70	97.70	97.70	97.70	97.70	97.70	97.70	97.70	97.70	97.70	97.70	97.70	97.70	97.70	100	100
	Mean accuracy	97.70	97.70	97.70	97.70	97.62	97.70	97.70	97.70	97.70	97.70	97.70	97.70	97.70	97.70	100	100
	#Selected features	3	3	4	3	3	3	3	3	3	3	4	3	3	4	5	3
KrVsKpEW	Best accuracy	99.84	99.84	99.38	98.447	99.84	99.84	99.84	99.53	99.69	99.69	99.38	99.38	98.59	97.65	99.37	99.73
	Mean accuracy	99.72	99.76	99.10	97.47	99.19	99.77	99.74	98.98	99.67	99.57	99.05	98.82	98.02	81.50	98.72	99.73
	#Selected features	31	29	25	19	32	30	26	21	29	25	21	24	18	19	29	21
PenglungEW	Best accuracy	86.67	86.67	86.67	86.67	86.67	93.33	93.33	93.33	93.33	93.33	93.33	86.67	86.67	80	86.66	93.33
	Mean accuracy	86.67	85.33	85.33	84.44	82.67	87.11	86.67	84.89	89.33	85.33	87.11	84.89	80.89	72.88	84.88	86.66
	#Selected features	40	17	145	4	37	65	61	175	7	159	151	145	155	29	67	53
Lymphography	Best accuracy	90.00	90.00	86.67	86.67	90.00	90.00	86.67	90.00	90.00	86.67	90.00	90.00	86.67	93.33	90	100
	Mean accuracy	86.00	85.56	86.00	83.11	83.56	86.67	84.89	85.56	87.11	85.78	86.44	85.56	83.33	84.88	87.33	100
	#Selected features	7	7	7	8	8	7	7	7	9	7	9	11	9	9	11	5
M-of-N	Best accuracy	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	Mean accuracy	100	100	100	100	100	100	100	100	100	100	100	100	100	80.76	100	100
	#Selected features	6	6	6	6	6	6	6	6	6	6	6	6	6	10	6	6
Tic-tac-toe	Best accuracy	92.71	92.71	92.71	92.71	92.71	92.71	92.71	92.71	92.71	92.71	92.71	92.71	92.71	88.02	87.5	92.71
	Mean accuracy	92.71	92.71	92.71	88.85	90.24	92.71	92.71	92.43	92.71	92.71	92.71	92.71	91.28	75.76	87.5	92.71
	#Selected features	8	8	8	8	8	8	8	8	8	8	8	8	8	7	6	8
	Best accuracy	100	100	100	100	100	100	100	100	100	100	100	100	100	97.22	97.22	100

© 2023 NSP
Natural Sciences Publishing Cor.

WineEW	Mean accuracy	100	100	100	100	100	100	100	100	100	100	100	100	100	90	96.66	100
	#Selected features	3	3	3	3	3	3	3	3	3	3	3	3	3	8	5	4
Zoo	Best accuracy	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	Mean accuracy	100	100	100	100	100	100	100	100	100	100	100	100	100	96	100	100
	#Selected features	4	4	4	4	4	4	4	4	4	4	4	4	5	7	5	5
Exactly	Best accuracy	100	100	100	100	100	100	100	100	100	100	100	100	100	91.5	100	100
	Mean accuracy	100	100	100	97.97	97.80	100	99.97	99.93	100	99.97	100	100	99.53	73.93	99.81	100
	#Selected features	6	6	6	6	6	6	6	6	6	6	6	6	6	10	6	6

6 Conclusion and Future work

Although the Corona epidemic has continued for more than three years, scientists have not found a drug to treat COVID-19. Therefore, they tended to make the vaccine to reduce infection and limit the spread of the disease, especially for asymptomatic people. As of January 16, 2023, a total of 13,131,550,798 vaccine doses have been administered [59]. However, a new variant has recently emerged. WHO has reported 664,097,132 confirmed cases of COVID-19 as of 5:33 p.m. CET on January 23, 2023, including 6,716,108 deaths [59]. Because of its rapid spread among people through communication, methods based on CT scans, x-rays, and laboratory tests require dealing with people, which leads to its spread, increasing the number of cases, and delaying diagnosis, which leads to the deterioration of some cases that may lead to death. Therefore, it is necessary to find an effective method for early diagnosis of COVID-19 while limiting the spread of the disease. In this research, we presented a strategy for diagnosing corona based on symptoms, which is divided into two stages. The first stage selects the necessary features and takes place in two steps. The first step is SDO to improve the initial solution for the second step, then NMR and hill climbing to choose the final features that enter the second stage, which is classification to predict accuracy. Finally, the efficiency of the proposed strategy was evaluated using three datasets for COVID-19 and nine datasets for UCI. Then compare the results with 15 other metaheuristic algorithms with different classification algorithms. Our future trends are building a web-based mobile system to help health professionals detect cases of COVID-19 infection. Using the proposed model for other diagnostic methods for COVID-19, such as X-rays, CT scans, and laboratory tests, Using different feature selection methods and optimization techniques may be helpful to get better model performance. Using the proposed model to diagnose other diseases.

References

[1] Z. Gao, Y. Xu, C. Sun, X. Wang, Y. Guo, S. Qiu, K. Ma., "A systematic review of asymptomatic infections with COVID-19," J. Microbiol. Immunol. Infect., 54 (1), 12–16, (2021).
[2] F. Wu, S. Zhao, B. Yu, Y. Chen, W. Wang, Z. Gang Song, Y. Hu, Z. Tao, J. Tian, Y. Pei, M. Yuan, Y. Zhang, F. Dai, Y. Liu, Q. Wang, J. Zheng, L. Xu, E. Holmes, and Y. Zhang., "A new coronavirus associated with human respiratory disease in China," Nature, 579 (7798), 265–269, (2020).
[3] Z. A. A. Alyasseri, M. A. Al-Betar, I. Abu Doush, M. A. Awadallah, A. K. Abasi, S. N. Makhadmeh, O. A. Alomari, K. H. Abdulkareem, A. Adam, R. Damasevicius, M. A. Mohammed, R. Abu Zitar. , "Review on COVID-19 diagnosis models based on machine learning and deep learning approaches," Expert Syst., 39 (3), 1–32, (2022).
[4] M. Khan and S. T. Khan, "COVID-19: A Global Challenge with Old History, Epidemiology and Progress So Far," Moléculas, 26 (1), 1–25, (2021).
[5] D. Ndwandwe and C. S. Wiysonge, "COVID-19 vaccines," Curr. Opin. Immunol., 71, 111–116, (2021).
[6] A. E. Mohamed, "Comparative Study of Four Supervised Machine Learning Techniques for Classification," Int. J. Appl. Sci. Technol., 7 (2), 5–18, (2017).
[7] V. Nasteski, "An overview of the supervised machine learning methods," Horizons.B, 4, 51–62, (2017).
[8] M. Fatima and M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic," J. Intell. Learn. Syst. Appl., 9 (1), 1–16, (2017).
[9] 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, 9, 26766–26791, (2021).
[10] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Med. Inform. Decis. Mak., 19, 1–16, (2019).
[11] O. F.Y, A. J.E.T, A. O, H. J. O, O. O, and A. J, "Supervised Machine Learning Algorithms: Classification and Comparison," Int. J. Comput. Trends Technol., 48 (3), 128–138, (2017).

© 2023 NSP
Natural Sciences Publishing Cor.

[12] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," Comput. Biol. Med., 136, p. 104672, (2021).
[13] E. A. Zanaty, "Support Vector Machines (SVMs) versus Multilayer Perception (MLP) in data classification," Egypt. Informatics J., 13 (3), 177–183, (2012).
[14] J. M. Ferreira, I. M. Pires, G. Marques, N. M. Garcia, E. Zdravevski, P. Lameski, F. Flórez-Revuelta,and S. Spinsante., "Identification of Daily Activites and Environments Based on the AdaBoost Method Using Mobile Device Data: A Systematic Review," Electronics, 9 (1), 192, (2020).
[15] J. Hu, "Automated detection of driver fatigue based on AdaBoost classifier with EEG signals," Front. Comput. Neurosci., 11, 1–10, (2017).
[16] A. Kadiyala and A. Kumar, "Applications of python to evaluate the performance of decision tree-based boosting algorithms," Environ. Prog. Sustain. Energy, 37 (2), 618–623, (2018).
[17] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-Based Feature Selection: Introduction and Review Ryan," J. Biomed. Inform., 85, 189–203, (2018).
[18] J. Miao and L. Niu, "A Survey on Feature Selection," Procedia Comput. Sci., 91, 919–926, (2016).
[19] R. Salgotra and U. Singh, "The naked mole-rat algorithm," Neural Comput. Appl., 31 (12), 8837–8857, (2019).
[20] W. Zhao, L. Wang, and Z. Zhang, "Supply-Demand-Based Optimization: A Novel Economics-Inspired Algorithm for Global OptImization," 1–16, (2019).
[21] M. A. Al-Betar, M. A. Awadallah, I. Abu Doush, E. Alsukhni, and H. ALkhraisat, "A Non-convex Economic Dispatch Problem with Valve Loading Effect Using a New Modified β-Hill Climbing Local Search Algorithm," Arab. J. Sci. Eng., 43 (12), 7439–7456, (2018).
[22] M. A. Al-Betar, β-Hill climbing: an exploratory local search, Neural Comput. Appl., 28, 153–168, (2017).
[23] P. K. Sethy, S. K. Behera, P. K. Ratha, and P. Biswas, "Detection of coronavirus Disease (COVID-19) based on Deep Features and Support Vector Machine," Int. J. Math. Eng. Manag. Sci., 5 (4), 643–651, (2020).
[24] D. C. R. Novitasari, R. Hendradi, R. E. Caraka, Y. Rachmawati, N. Z. Fanani, A. Syarifudin, T. Toharudin, and R. C. Chen., "Detection of COVID-19 chest x-ray using support vector machine and convolutional neural network," Commun. Math. Biol. Neurosci., 2020, 1–19, (2020).
[25] T. Tuncer, S. Dogan, and F. Ozyurt, "An automated Residual Exemplar Local Binary Pattern and iterative ReliefF based corona detection method using lung X-ray image," Chemom. Intell. Lab. Syst., 203, 104054, (2020).
[26] M. Turkoglu, "COVIDetectioNet: COVID-19 diagnosis system based on X-ray images using features selected from pre-learned deep features ensemble," Appl. Intell., 51 (3), 1213–1226, (2021).
[27] M. S. Iraji, M. R. Feizi-Derakhshi, and J. Tanha, "COVID-19 Detection Using Deep Convolutional Neural Networks and Binary Differential Algorithm-Based Feature Selection from X-Ray Images," Complexity, 2021, (2021).
[28] A. Narin, "Accurate detection of COVID-19 using deep features based on X-Ray images and feature selection methods," Comput. Biol. Med., 137, 104771, (2021).
[29] M. Canayaz, "Biomedical Signal Processing and Control MH-COVIDNet: Diagnosis of COVID-19 using deep neural networks and meta-heuristic-based feature selection on X-ray images," Biomed. Signal Process. Control, 64 (June 2020), 102257, (2021).
[30] P. Bhowal, S. Sen, and R. Sarkar, "A two-tier feature selection method using Coalition game and Nystrom sampling for screening COVID-19 from chest X-Ray images," J. Ambient Intell. Humaniz. Comput., (2021).
[31] S. Sen, S. Saha, S. Chatterjee, S. Mirjalili, and R. Sarkar, "A bi-stage feature selection approach for COVID-19 prediction using chest CT images," Appl. Intell., (2021).
[32] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. A. Abo-Elsoud, "Anew COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier," Knowledge-Based Syst., 205, 106270, (2020).
[33] 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, 8, (2020).
[34] S. M. Rezaeijo, R. Abedi-Firouzjah, M. Ghorvei, and S. Sarnameh, "Screening of COVID-19 based on the extracted radiomics features from chest CT images," J. Xray. Sci. Technol., 29 (2), 229–243, (2021).
[35] 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," Appl. Soft Comput., 111, 107698, (2021).
[36] A. Dey, S. Chattopadhyay, P. K. Singh, A. Ahmadian, M. Ferrara, N. Senu, and R. Sarkar., "MRFGRO: a hybrid meta-heuristic feature selection method for screening COVID-19 using deep features," Sci. Rep., 11 (1), 1–15, (2021).
[37] E. H. Houssein, M. M. Emam, and A. A. Ali, "Improved manta ray foraging optimization for multi-level thresholding using COVID-19 CT images," Neural Comput. Appl., 33 (24), 16899–16919, (2021).
[38] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. A. Abo-elsoud, "Accurate detection of COVID-19 patients based on distance biased Naïve Bayes (DBNB) classification strategy," January, (2020).
[39] M. abdulameer kadhim alsaeedi and S. Kurnaz, "Feature selection for diagnose coronavirus (COVID-19) disease by neural network and Caledonian crow learning algorithm," Appl. Nanosci., 0123456789, (2022).
[40] H. Chauhan, K. Modi, and S. Shrivastava, "Development of a classifierwith analysis of feature selection methods for COVID-19 diagnosis," World J. Eng., December, (2021).
[41] K. H. Abdulkareem, M. A. Mohammed, A. Salim, M. Arif, O. Geman, D. Gupta, and A. Khanna., "Realizing an Effective COVID-19 Diagnosis System Based on Machine Learning and IOT in Smart Hospital Environment," IEEE Internet Things J., 1–8, (2021).

© 2023 NSP
Natural Sciences Publishing Cor.

[42] 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 Dyn., December 2019, (2021).
[43] https://www.kaggle.com/code/midouazerty/symptoms-covid-19-using-7-machine-learning-98/data.
[44] B. Kovács, F. Tinya, C. Németh, and P. Ódor, "SMOTE: Synthetic Minority Over-sampling Technique," Ecol. Appl., 30 (2), 321–357, (2020).
[45] K. K. Ghosh, R. Guha, S. K. Bera, N. Kumar, and R. Sarkar, "S-shaped versus V-shaped transfer functions for binary Manta ray foraging optimization in feature selection problem," Neural Comput. Appl., 33 (17), 11027–11041, (2021).
[46] Y. Wang and T. Du, "An improved squirrel search algorithm for global function optimization," Algorithms, 12 (4), (2019).
[47] G. Nagarajan and L. D. Dhinesh Babu, "A hybrid feature selection model based on improved squirrel search algorithm and rank aggregation using fuzzy techniques for biomedical data classification," Netw. Model. Anal. Heal. Informatics Bioinforma., 10 (1), 1–29, (2021).
[48] D. Karaboga, "An idea based on honey bee swarm for numerical optimization. Technical Report TR06, Erciyes University," Erciyes Univ., 46(2), 1–10, (2005).
[49] Mudita Juneja and S. K. Nagar, "Particle swarm optimization algorithm and its parameters: A review," no. Iccccm, (2016).
[50] I. Fister, X. S. Yang, S. Fong, and Y. Zhuang, "Bat algorithm: Recent advances," CINTI 2014 - 15th IEEE Int. Symp. Comput. Intell. Informatics, Proc., 163–167, (2014).
[51] H. Faris, I. Aljarah, M. A. Al-Betar, and S. Mirjalili, "Grey wolf optimizer: a review of recent variants and applications," Neural Comput. Appl., 30 (2), 413–435, (2018).
[52] M. Sharawi, H. M. Zawbaa, E. Emary, and E. Hossamzawbaagmailcom, "Feature Selection Approach Based on Whale Optimization Algorithm".
[53] D. Wang, H. Chen, T. Li, J. Wan, and Y. Huang, "A novel quantum grasshopper optimization algorithm for feature selection," Int. J. Approx. Reason., 127, 33–53, (2020).
[54] Y. Zhang and Y. Mo, "Dynamic optimization of chemical processes based on modified sailfish optimizer combined with an equal division method," Processes, 9 (10), 2021.
[55] Y. Zhang, R. Liu, X. Wang, H. Chen, and C. Li, "Boosted binary Harris hawks optimizer and feature selection," Eng. Comput., 37 (4), 3741–3770, (2021).
[56] J. Too and A. Rahim Abdullah, "Binary atom search optimisation approaches for feature selection," Conn. Sci., 32 (4), 406–430, (2020).
[57] N. Neggaz, E. H. Houssein, and K. Hussain, "An efficient henry gas solubility optimization for feature selection," Expert Syst. Appl., 152, 113364, (2020).
[58] K. Mishra and S. K. Majhi, "A binary Bird Swarm Optimization based load balancing algorithm for cloud computing environment," Open Comput. Sci., 11 (1), 146–160, (2021).
[59] https://covid19.who.int/.

© 2023 NSP
Natural Sciences Publishing Cor.