Nature inspired optimization model for classification and severity prediction in COVID-19 clinical dataset

Suma L. S. · Anand H. S. · Vinod chandra S. S.

Received: 11 January 2021 / Accepted: 6 July 2021 / Published online: 31 July 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
The spread rate of COVID-19 is expected to be high in the wake of the virus’s mutated strain found recently in a few countries. Fast diagnosis of the disease and knowing its severity are the two significant concerns of all physicians. Even though positive or negative diagnosis can be obtained through the RT-PCR test, an automatic model that predicts severity and the diagnosis will help medical practitioners to a great extent for affirming medication. Machine learning is an efficient tool that can process vast volume of data deposited in various formats, including clinical symptoms. In this work, we have developed machine learning models for analysing a clinical data set comprising 65000 records of patients, consisting of 26 features. An optimum set of features was derived from this data set by the proposed variant of artificial bee colony optimization algorithm. By making use of these features, a binary classifier is modelled with support vector machine for the screening of COVID-19 patients. Different models were tested for this purpose and the support vector machine has showcased the highest accuracy of 96%. Successively, severity prediction in COVID positive patients was also performed successfully by the logistic regression model. The model managed to predict three severity status viz mild, moderate, and severe. The confusion matrix and the precision-recall values (0.96 and 0.97) of the binary classifier indicate the classifier’s efficiency in predicting positive cases correctly. The receiver operating curve generated for the severity predicting model shows the highest accuracy, 96.0% for class 1 and 85.0% for class 2 patients. Doctors can infer these results to finalize the type of treatment/care/facilities that need to be given to the patients from time to time.

Keywords Artificial bee colony optimization · Logistic regression · Severity prediction · Support vector machine · COVID-19

1 Introduction
COVID-19 caused by the virus SARS-COV 2 has become a pandemic that keeps on spreading worldwide, taking people’s lives. Since the outbreak at Wuhan in China in December 2019, millions of people got infected due to the virus’s highly transmissive nature. Though this disease is not a serious one, it may lead to high infections and even cause death in particular cases. Despite various measures taken by Governments to control the spread, more and more people are getting affected. Moreover, a new variant of the disease caused by the mutating virus has also been reported in a few countries. Isolating people with specific symptoms would help to reduce the intensity of spreading. Most countries have started a system wherein positive people who belong to the moderate/mild category stay home. In this context, early diagnosis and severity prediction of patients have become necessary to prevent disease spread.

The diagnosis process still relies on Reverse Transcription Polymerase Chain Reaction (RT-PCR) to detect the virus. Although this method is reliable and widely used, many factors affect results Tahamtan and Abdollah 2020. Moreover, the false-negative and false-positive results are
A vote-based deep learning model was proposed by incorporating neural network’s EfficientNet family to detect COVID-19 patients (Silva et al. 2020). A Convolution Neural Network (CNN) model was also developed to classify COVID-19 positive or negative cases out of chest CT images. A multi-objective differential evolution algorithm supported the model performance for fine-tuning the hyperparameters (Singh et al. 2020). As the DE algorithm considers only the exploitation aspect, it may get stuck in local optima. Also, the performance was compared with only ANN model. A survey was conducted for image-based deep learning approaches for COVID-19 prediction, describing and evaluating various implementation and performance aspects (Bhattacharya et al. 2021). A deep learning-based VB-Net system was applied for lung infection quantification through segmenting COVID infected regions of chest CT images Shan et al. 2020. The method results were not compared with other similar segmentation techniques other than the dice similarity coefficient measure. A classification system for COVID-19 was developed using chest X-ray radiographs, experimenting with five different pre-trained CNN Res-Net models (Narin et al. 2020). Although the system has shown higher accuracy for one particular data set, the results were not compared with other deep learning techniques. On overall analysis of image-based deep learning techniques, one possible limitation is the high volume of data samples required to generate accurate features. Practically, such image data curation is found to be a tedious task.

Apart from the image analysis, clinical data associated with COVID-19 patients was also curated at hospitals, treatment centers, and other agencies. By analyzing this data, we can extract inherent features that could help physicians diagnose and make appropriate decisions. This text data could be a valuable resource for machine learning models to classify the positive and negative cases and predict severity status. In the RT-PCR results’ false-negative cases, such computational tools capable of analyzing symptoms data would help physicians decide whether to go for a re-test. A few works were published in 2020 for predicting COVID status based on text data. Test results of COVID-19 and influenza patients were studied and based on which the machine learning-based diagnosis tool was developed. The appropriate clinical features were used to cluster the patients, and the XGBoost model was designed to classify patients into COVID positive and negative (Li et al. 2020). The feature elimination was done in this work by computing the correlation coefficient for every pair of variables. For variables of different nature, different statistical methods were used, which increases the computational effort. Analyzing the classification model, the authors have used RIDGE, Random Forest, XGBoost approaches and suggest XGBoost as the best one with an AUC of 98.6%. However, the sample size used in the study is too less, which could be attributed to the higher accuracy of the model.

Another machine learning approach applied multiple algorithms in the WEKA tool to predict RT-PCR results using necessary clinical data. The framework has achieved an accuracy of 91.4%, with a sensitivity of 94.1% and specificity of 88.7% (Thomas et al. 2020). The study involves a patient set consisting only of 199 patients. Another downside of this method is that among the various models tried out for the classification, the best model achieved an accuracy of only 91.4% compared to other techniques. Another classifier that relied only on clinical text data was able to detect COVID-19 cases. The limitations of this approach include insufficient variables used for prediction and the bias of self-reported data (Zoabi et al. 2020).
Predicting the mortality risk of COVID-19 patients was handled by an online prognostic tool developed on machine learning models (Das et al. 2010). Here, the best model identified has produced an area under the curve measure of 0.83. However, the important properties such as clinical information on symptoms, risk factors and details of earlier illness were not considered in the study. Other than these conventional techniques, the Nature Inspired approach was also suggested for predicting the risk of covid among people with various clinical conditions (Singh et al. 2020). A hybrid particle swarm optimization and Sine Cosine algorithm, along with individual algorithms, were implemented for the analysis. However, the obtained accuracy measures were found to be less for the applied algorithms.

Similar to the prediction and analysis of the covid related data, the security of personal information used for various studies is also very important. An approach based on blockchain technology was proposed for maintaining covid data security in a recent work (MK et al. 2020). The regularization process used in the classification model certainly affects the performance. A ridge-adaline stochastic gradient descent technique was proposed in an early disease prediction system (Saritha and Vinod Chandra 2016). While searching through literature, it has been noticed that only a few ML techniques were proposed for predicting severity of COVID-19 using the clinical data. In the onset, we outline this work as a classification problem for screening COVID-19 patients out of the clinical data set, followed by the severity class prediction of the identified positive cases.

While using a machine learning model for classification, selecting the most relevant clinical information to predict disease becomes crucial for improving model performance. Reducing the number of features would help minimize computing costs as well. Thus, any machine learning model designed for COVID-19 prediction is provided with an optimum set of features to expect maximum result accuracy. In this work, the feature selection process has been framed as combinatorial optimization and a nature-inspired algorithm is proposed for solving it. Particle swarm optimization (PSO) algorithm and its variants were used to solve many optimization problems effectively (Saritha and Vinod Chandra 2016). Parameter tuning is a major task in machine learning algorithms that requires efficient techniques. An ant lion optimization algorithm was implemented on a deep network model to derive optimal parameter values in reduced time (Reddy et al. 2020). Artificial Bee Colony optimization (ABC) algorithm is a powerful technique for finding global optimum solutions by utilizing a few parameters (Karaboga and Basturk 2007). The swarm-based metaheuristic was successfully used for many applications in various fields (Suma and Vinod Chandra 2017). An efficient optimization algorithm must include the provisions for local search and a global search to escape from local minima (Saritha and Vinod Chandra 2018). A major aspect of ABC algorithm is that it handles exploitation through neighbor finding and exploration through introducing scout bees, ensuring global optimum solutions. Therefore, we chose to apply ABC for the feature selection process. However, the neighbor selection process in the employee bee phase in the basic algorithm does not guarantee the optimum solution in some instances. Thus, we proposed a variant of ABC algorithm with a modified employee bee phase.

We would pass the optimum feature set to the machine learning model to predict COVID-19 infection. A binary classifier SVM is proposed to classify the patients as COVID positive cases and negative cases. After screening out the patients with COVID-19, knowing the present severity of patients is essential in deciding future clinical requirements and treatment decisions. For a COVID-19 patient with clinical details, any system capable of predicting which severity class he/she belongs to would help physicians make necessary decisions. Accordingly, we have framed a multiclass classification problem to predict the severity class for a particular patient and proposed a Logistic Regression (LR) model for the same. We have applied different machine learning models for both the classification tasks and finally chosen the highest accuracy ones. Accordingly, SVM model was selected for the initial classification tasks and finally chosen the highest accuracy ones.

| Feature Name | Type/Description |
|--------------|------------------|
| Fever        | Symptom          |
| Tiredness    | Symptom          |
| Dry cough    | Symptom          |
| Breathing difficulty | Symptom    |
| Sore throat  | Symptom          |
| Pains        | Symptom          |
| Nasal congestion | Symptom     |
| Runny Nose   | Symptom          |
| Diarrhea     | Symptom          |
| Age group (0–9) | Personal     |
| Age group (10–19) | Personal |
| Age group (20–24) | Personal |
| Age group (25–59) | Personal |
| Age group (60+) | Personal     |
| Female       | Personal         |
| Male         | Personal         |
| Transgender  | Personal         |
| Mild         | Severity level of symptoms |
| Moderate     | Severity level of symptoms |
| Severe       | Severity level of symptoms |
| Non_Contact  | Non_Symptom      |
| Contact      | Non_Symptom      |
binary classification with optimum features. Similarly, based on the performance evaluations of considered algorithms, LR seemed to be good for severity prediction.

Contributions: In this work, we proposed a new variant of the ABC optimization algorithm with an improved neighbor finding criteria. This algorithm was applied for the feature selection process to improve the classification process’s performance. This methodology helps to reduce the complexity of the feature selection process required in other methods. Moreover, an efficient early disease detection system has been developed out of minimum clinical data, which will help physicians to make appropriate decisions.

2 Materials and methods

2.1 Dataset

We have used the data set comprising 26 different features, including disease symptoms and other related personal information (COVID-19 dataset 2020). These data will help identify whether any person has coronavirus disease or not based on some predefined standard symptoms. These symptoms are based on the World Health Organization (WHO) guidelines and the Ministry of Health and Family Welfare. Existing machine learning models for prediction take information from both CT images and text data. However, we make use of symptom data which is in text form. The dataset contains six major variables that are supposed to impact whether someone is affected or not. The description of each variable is given below.

1. Country: List of countries person visited.
2. Age: Classification of the age group for each person, based on WHO Age Group Standard
3. Symptoms: According to WHO, 5 are major symptoms of COVID-19, Fever, Tiredness, Difficulty in breathing (Smell loss), Dry cough, and sore throat (Taste lost).
4. Experience any other symptoms: Pains, Nasal Congestion, Runny Nose, Diarrhoea and Other.
5. Severity: The level of severity, Mild, Moderate, Severe
6. Contact: Has the person contacted some other COVID-19 Patient?

The ranking of parameters is done to confirm the impact contributed to prediction. Each label in the variable is created with all these categorical variables and in total 316800 combinations are checked. The features are represented as binary values indicating presence or absence. Out of 26 features in the raw data, four were eliminated because they did not seem relevant in this analysis. The features taken for our study are
shown in Table 1. This feature set is passed through the feature selection phase, and the obtained optimum features are then used to predict whether the patient is infected or not.

2.2 Methodology

The aim of this work is to detect COVID-19 cases and to predict the severity by analysing the symptoms data. For any ML model to produce accurate results, the feature set selected should be the most relevant. Adding more features into the feature vector may not provide the expected accuracy. Obviously, selection of features becomes so crucial in traditional ML approaches. Accordingly, the methodology starts with a feature selection process. Although a number of feature selection techniques are available in literature, we propose an efficient optimization approach for this task. A variant of ABC optimization algorithm is designed for this purpose. The input data set is passed initially through the feature selection phase to produce an optimum feature set.

The next phase in the pipeline is the training of this generated optimum feature set with an ML model. Here different models including SVM are applied for training and testing process to choose the best performing one. Towards this, the input data set is to be divided into positive cases as well as negative cases. Once the model returns sufficient classification accuracy, then the next process of severity prediction is initiated.

To proceed with this task, we need to prepare the input data so as to enable the model to perform multi class classification. Here, the severity deciding feature is given as three attributes Mild, Moderate and Severe. As there exists three levels of severity, the corresponding binary values for positive cases are replaced with 1, 2 and 3 respectively. After getting the final data set, the LR model is used to perform severity level prediction. This step is implemented as a multi class classification. The overall workflow of the paper is depicted as Fig. 1.

2.3 Feature selection through optimization

2.3.1 ABC algorithm solution encoding

We use the ABC algorithm for choosing the optimum features from the available feature set. While mapping the problem into an optimization task, the first thing needed is a suitable representation of solutions. Here, each solution in solution space corresponds to a combination of features chosen from the feature set. We apply the same binary encoding for the solution space to make further computations easier. The feature vector with \( D \) number of features in binary representation is \([1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 1 \ 1]\).

The first 1 indicates the presence of \( F_1 \), and the next 0 indicates the absence of \( F_2 \), and so on. As the number of bits in the solution vector is 22, dimension \( D \) of solution space takes the value 22.

Initially, the vectors will hold random real values between 0 and 1. Based on a threshold value comparison, each entry is assigned with 0 or 1. Accordingly, the solution space is mathematically represented as a matrix \( S \) of size \( N \times D \), where \( N \) denotes an important algorithm-specific parameter that corresponds to the number of solutions in \( S \).

2.3.2 Objective function

The most important aspect of the optimization algorithm is defining a suitable objective function capable of evaluating the chosen solution’s quality. As each solution consists of different features, the objective function is described in terms of feature values. Here we perform match operation between feature in solution vector and the original feature data in the data set. If the values in the corresponding positions match, then that feature is given a score. It enables quantification of the prominence of each feature that is correctly chosen. Thus, a particular feature that is frequently occurring would be assigned a higher score. The objective function \( f \) defined in terms of feature components is depicted as

\[
f = \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{\sum_{j=1}^{D} \text{score}(\text{match}(S_i(j), V_k(j)))}{D} \right\}
\]

Here, \( N \) is the total number of vectors given in the extracted COVID data set, \( D \) is the number of features in each vector. In our algorithm, each solution vector is analyzed to find its quality. Thus, \( S_i(j) \) corresponds to the \( j \)th position in the \( i \)th vector of the solution space, which is compared with the corresponding feature component in the data set’s \( k \)th vector. In the case of a match, the corresponding score value for the vector is incremented. The cumulative score is measured after making comparisons for all the \( D \) feature components. This computation would give the match score of a single vector in the data set. We need to check the significance of our chosen features in \( S_i \) against the entire data set, corresponding to all the patients. Accordingly, the objective function computes the final feature score of \( S_i \) by comparing \( N \) records in the data set. As far as any optimization problem is concerned, the algorithm either maximizes or minimizes the objective value. As we are looking for a solution with a maximum score, the objective function defined here is expected to produce the highest value, and the problem became a maximization problem.

3 ABC optimization algorithm

Nature-inspired algorithms have emerged as powerful techniques for solving optimization problems in various fields. This class of algorithms, from Genetic Algorithm to recent plant-based algorithms has found its place in
combinatorial optimization Saritha and Vinod Chandra 2016. Among these algorithms, the ABC optimization algorithm is well known for finding global optimum solution with a perfect blending of exploitation and exploration. The algorithm works following the food foraging behavior of bees. Based on category of bees - employee bee, onlooker bee, and scout bee tasks are assigned. The algorithm’s ultimate aim is to arrive at the best possible food source with the maximum nectar amount. Employee bees are initially located at specific food sources and are responsible for finding higher quality food sources. The quality analysis is performed using the amount of nectar/fitness function within the source. Once all the employee bees finish this process, the information about the quality sources is passed on to the next bee group, through “waggle dance”. The onlooker bees then employs a probabilistic selection and repeat the neighbour search process. Finally, the best solution with highest fitness value is returned by the algorithm. The basic steps in the ABC algorithm are depicted as Procedure ABC.

### Algorithm 1: Procedure ABC

**Initialise**: Define solution space of size \((N, D)\) and initialize with random values  
**Input**: Input number of iterations, lower bound and upper bound  
**Repeat**  
Find the new solution for each initial solution assigned to employee bee  
Evaluate solution quality and choose the best solution  
Generate probability vector to be used by onlooker bees  
Onlooker bees choose the solution based on probability value  
Update chosen solution by onlooker bees  
Evaluate solution quality and choose the best one  
Introduce scout bee if there is an abandoned solution  
Save the best solution and increment iteration count  
Until iteration count > maximum value  
Return the best solution saved

In the feature selection process, the solution in space \(S\) corresponds to the initial feature set. The contents of \(S\) will be initialized randomly. After generating intermediate solutions by different bees, the algorithm will memorize the optimum solution. Detailed steps of basic ABC optimization algorithm are given in Algorithm 2.

### Table 2 Parameters chosen for artificial bee colony optimization

| Parameter | Description                   | Type   | Value |
|-----------|-------------------------------|--------|-------|
| N         | No of solutions in S          | Hyper  | 20    |
| UB, LB    | Upper and lower limit values  | Specific| (1,0) |
| Limit     | Iteration count               | Hyper  | 10    |
| D         | Dimension                     | Problem| 22    |

### Algorithm 2: Basic ABC Optimization Algorithm

Step 1: Define solution space with dimension \((N, D)\)  
Initialise the lower and upper bound values of feature variables, \(b, u_b\) and \(limit\)

Step 2: Assign random initial values for solution by \[ x(i) = l_b + r(u_b - l_b) \] where \(r\) is a random number between \((0, 1)\)

Step 3: **while** \(iter > number of iterations\) go to step 14

Step 4: Employee bees find the neighbour of current solution \(i\) through \[ x'(i,j) = x(i,j) + r \times (x(k,j) - x(i,j)) \] where \(k \neq i, r \in [0, 1]\)

Step 5: Evaluate the new solution by fitness value measure (4) to select best one  
**fit**\((x) = 1 + f(x)\), \(f(x) > 0\)

\[ f(x) = \frac{1}{(1 + f(x))} \text{, } f(x) \leq 0 \text{, where } f(x) \text{ is the objective function} \]

Step 6: If the quality of \(x(i)\) is not improved, update the depletion counter

Step 7: Onlooker bees compute probability value by \[ P = \frac{f_i}{\sum f_i} \] where \(f_i\) is the objective function

Step 8: Choose the solution with probability value greater than random number

Step 9: Update the solution by processing steps 4 and 5

Step 10: Memorize the best food source

Step 11: Check for any solution which is depleted as per \(limit\)

Step 12: Scout bee is introduced to replace the solution by using equation (2)

Step 13: Increment \(iter\) value

Step 14: Return the optimum solution
If a solution cannot be improved even after the given iteration count, then the scout bee comes in and replace that solution with a new solution with random values. This feature of the ABC algorithm provides the exploration ability to find the global optimum solution. It ensures that the algorithm does not get stuck in any local optima point.

While implementing the algorithm, we need to put values to different parameters. The various parameters required for the proposed algorithm include problem dependent parameters as well as hyper-parameters. Table 2 gives parameters along with the details and values taken for implementing the algorithm.

To constitute the solution space, a number of solutions \( N \) are to be decided initially. This is the same as the number of working bees also. If \( N \) value is a higher number, then the algorithm will process more solutions and may return a better quality solution as the optimum. However, if the value is too high, then the processing time will increase. Thus, while choosing \( N \), the tradeoff between the solution quality as well as time is to be considered.

Here, we have run the algorithm by putting different values 20,30,40 and 50 and finally chosen the value as 20 by considering the time of execution and value of the objective function. Another hyper parameter limit is also chosen based on the optimum value for the objective function. In addition, two other problem dependent parameters Dimension and \((lb, ub)\) were also required. Dimension indicates the size of each solution, and here it is taken as the number of features. \((lb,ub)\) values indicate the minimum and maximum values required to initialize the solution space. As the solution vector encodes the data as presence or absence of features, binary values 0 and 1 are taken as the bound values.

As the problem-dependent parameters dimension and \((lb, ub)\) values are used for algorithm implementation, they do not directly impact the produced solutions. However, the \( N \) and limit values may influence the quality of the generated solution. If these values are not correct, then the final feature set may not be the optimum set. This is why we fine tune the algorithm by running multiple times and finalizing the parameters based on the objective value.

Bees measure food source quality with a fitness function defined using objective function value as per Eq. (1).

As our problem is framed as a maximization task, the fitness computation is slightly modified to increase the quality measure as the objective value increases.

### 3.1 Improving ABC optimization algorithm

The two major aspects of the ABC algorithm are exploitation and exploration ability for achieving the global optimum solution. During the employed bee phase, the algorithm finds the next neighbor solution by choosing a random intermediate solution. However, the random nature of the selected intermediate solution may affect the quality of the solution, which will affect the efficiency of the exploitation process. To overcome this limitation, a number of ABC variants incorporating global best-guided search were proposed (Zhu and Kwong 2010). In one such g-best guided search approach, the intermediate solution is taken through finding the best from neighbors of the given solution (Peng et al. 2019). Here, the limitation is that the varying neighbor block size taken by the algorithm may affect the quality of the solution. More runs are needed to derive the best value for this attribute.

We propose a multilevel quality checking strategy to find the best neighbor of the current solution in this work. An additional problem-dependent function is introduced to enable the subset of neighbors to compete among each other. This quality grading function is defined as the ranks assigned to features and represented as, \( g = \frac{1}{m} \sum_{j=1}^{D} R_j \times \mu \) where \( m \) is the number of active features in the solution and \( \mu \) is the rank proportionality constant and it takes the value 0.5. The solution space consisting of feature vectors is divided into \( Ng \) number of groups which depends on the size of the solution space. When a particular solution in a group is to be updated, instead of choosing the random neighbour, the remaining solutions will be analyzed with the new function \( g \). The solution with highest \( g \) value is taken as the winner and it becomes the neighbour solution to take part in update process. The new update equation becomes

\[
x'(i,j) = x(i,j) + r \times [x(w,j) - x(i,j)]
\]

(2)

The schematic representation of neighbour selection by employee bee in the new ABC is given in Fig. 2.

The new procedure is depicted using the detailed steps in the following Algorithm 3.
4 Support vector machine and logistic regression classifiers

The derived optimum feature set is passed to the next phase of our workflow, predicting patients' COVID status. For this, we have used a supervised machine learning approach, SVM. It is a powerful model among machine learning models used to solve various classification and regression problems. One upside of SVM is that it can solve both linear and non-linear problems by using different kernels. SVM implements a classification task by generating the optimized or maximally separated hyperplane among the multiple planes generated to partition the data points. As we used the data in binary format, no pre-processing or data transformation is needed.

We have applied different machine learning models, including Naive Bayes, Random Forest, and Multilayer Perceptron, for this binary classification. However, the performance of these models was not seemed to be comparable with our chosen SVM model. After procuring the classified list of positive cases, features were extracted for the subsequent analysis. This data was passed to the next phase for the severity prediction of patients. For data processing, the target labeling was done based on severity classes. The prognosis was then implemented as a multiclass classification task and solved by an efficient logistic regression model. Although we have applied other machine learning models for this classification, none of them showed better performance.

5 Results and discussions

This work aims to classify the patient conditions into COVID positive & negative and predicting severity status using the machine learning models. Towards this, we have used a patient text data set comprising 65,000 records.

Implementation of the proposed ABC optimization algorithm for feature selection requires selecting final values for hyper-parameters and problem-specific parameters. The parameter values for the modified ABC algorithm were chosen in the same manner as basic ABC algorithm. To analyse the performance of the proposed algorithm, we have compared the convergence of the proposed algorithm with basic ABC and PSO algorithms. As shown in Fig. 3, the proposed algorithm converges after 100 iterations with the highest objective value 0.70, and the optimum solution is the solution vector corresponding to this maximum objective value. We have chosen the final set of features as those in the data set corresponding to the 1's in the solution vector 11100111000000011001. Thus, the ABC algorithm has succeeded in deriving the optimum feature set.

Table 3 shows the optimum feature set derived by the ABC optimization algorithm. Among the 22 features submitted, basic ABC algorithm outputs 11 features as optimum and proposed IABC algorithm outputs 6 features as optimum. The ranking of features that is done based on frequency criteria proves the relevance of chosen features. We used these features to predict the COVID-19 positivity in patients with different symptoms. Compared with the input feature set (Table 1) comprising 22 features, 11 were eliminated. The majority of the eliminated features belong to personal information like age group, sex, etc. All the age
groups ranging from 0-9 to 60+ were included in the data set. However, the result optimum feature set excludes such data. This indicates that various symptoms are to be given more weightage while performing computational analyses of COVID-19 data irrespective of age groups. Moreover, the severity of patient conditions in three grades was also present in the input feature set. The algorithm-generated feature set has two of them, excluding the mild cases.

Feature ranking: Apart from this feature selection process, we have also ranked the features based on frequency. Based on the frequency of each feature in the given data set, a score value was assigned. Accordingly, ranked all features from the highest scored feature to the least scored feature. The highest rank was shared by nasal congestion and runny nose with a score of 0.5502. This means that 55.0% of the patients with COVID positive had nasal congestion and runny nose. The third column of Table 3 details the score procured by each feature, along with the rank assigned. Although we have ranked all the input features, the rank of only the final features is shown in Table 3.

The derived optimum feature set was analyzed with a novel classification model, SVM. Apart from the features shown in Table 1, we have used the variable “Nonexperiencing” as the target. We used Python Scikitlearn package for implementing the training and testing processes of the model. Out of 65535 records in our data set, all rows with target label 0 were considered as COVID positive cases. Also, 70% of the data was taken for training, and 30% was taken for testing. As part of analyzing the SVM model’s power in predicting COVID-19, we have computed various metrics, including confusion matrix and performance accuracy. The confusion matrix generated out of the predicted variables is shown in Fig. 4.

The four quadrants of the matrix (Q1, Q2, Q3, Q4) are placed in a clockwise direction. We have considered the test data comprising a total of 19661 records for this analysis. It is seen that out of 17808 actual COVID positive cases, 17128 were predicted correctly and denoted as True Positives (TP in Q1).

However, 680 cases were wrongly predicted as negative, indicating False Negatives (FN in Q4). Similarly, out of the actual negative cases, 1200 cases were predicted as True Negative (TN in Q3), and the remaining 669 cases were wrongly predicted as False Positive (FP in Q2). The overall accuracy of the model performance is obtained to be 94% using the equation.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3}
\]

Based on the theory behind the relationship between different performance measures, we have also used precision - recall to evaluate and visualize results. We chose this measure due to an imbalance between positive and negative observations in the data set. It is also reported that precision-recall is more informative than the Receiver Operating characteristics (ROC) curve when imbalanced data is analysed for binary classification problems (Saito and Marc 2015). Before plotting the precision-recall curve, these two measures were computed using the equations. Precision looks for the proportion of correct positive predictions with respect to the total number of positive predictions.

\[
\text{Precision (True Positive Rate)} = \frac{TP}{FP + TP} \tag{4}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{5}
\]

On analyzing the given data with SVM, the precision value is found to be 0.96. And recall is computed as 0.97. The
The real power of the SVM model lies in the kernel that is used for the classification process. Here, we have tried different kernels and obtained good performance with the sigmoid. Another parameter of the SVM classifier, randomness, was assigned a 0 to get the maximum accuracy. Moreover, the plot is also generated concerning the accuracy measures and given in Fig. 5.

After analyzing the performance of SVM with the feature set selected by basic ABC, we have also run the model with the feature set selected by the newly proposed IABC algorithm. This new feature set consists of 7 features and the SVM model has produced a classification accuracy of 96%. Apart from these ABC algorithms, PSO algorithm was also executed to derive features and a classification accuracy of 90% was obtained with the generated features. The various feature sets derived and the corresponding accuracy of classification algorithm are shown in Table 4.

### 6 Predicting severity

On analyzing the input data set, we have observed three features indicating severity class severity mild, severity moderate, and severity severe. Thus, the same data set is found to be suitable for predicting the severity status in patients. As the data is binary, labeling was done by encoding these features into integer values 1, 2, and 3. Feature values of all positive cases were extracted from the dataset and was submitted to the machine learning algorithms for performing multiclass classification. The data set was split into training and testing parts by the models. As in the COVID-19 screening problem, we have used different machine learning models for the prediction as well, and finally the logistic regression model has produced better results. The resulting accuracy values got plotted in the ROC curve and shown in Fig. 6. The ROC curve depicts the relationship between sensitivity and specificity measures. As shown in Fig. 6, the curve corresponding to class1 prediction has better performance.

ROC curve usually evaluates the performance of binary classifiers. However, here we have transformed the performance criterion of a three-class classifier into three separate binary classifications. Each classification corresponds to a given class and the total of the other two classes. On evaluating the performance with the measure Area Under the Curve (AUC), the prediction of class 1 severity achieved the best performance with the highest value of 0.966. It is

![Performance of Artificial Bee Colony optimization algorithm through convergence.](image)

![Confusion Matrix - P and N represent positive and negative COVID cases.](image)
also observed that the prediction of class 3 has got the least accuracy with the smallest AUC value of 0.779.

7 Comparing performance

As part of evaluating the model performance, we have compared the accuracy measures of various methods and consolidated them in Table 5. The integrated SVM-DL approach has produced 95.4% accuracy with the support of the automatic feature generation technique. Also, this method has utilized X-ray image data of less size as input. The successive DL approaches also succeeded in gaining higher accuracy values with the automatic feature extraction facility. The actual comparison of our proposed method happens with the results of text data-based approaches. Here, out of two such methods, the Weka tool acquired a lower accuracy of 91.4%. The second Gradient Boosting technique have used clinical data analysis of around 50,000 patients and procured area under PR curve score 0.66. The classification model developed with the optimum feature set derived by IABC has obtained an average accuracy of 96%, and its precision-recall curve is shown in Fig. 7. This is an improvement over the average accuracy of 94% obtained for SVM with ABC derived features.

8 Conclusion

COVID-19 pandemic clinical scenario has entered into another phase with the introduction of the vaccine. Nevertheless, the threat of pandemic does not seem to end in the wake of new variants identified. As per reports, these...
variants would be capable of multiplicating the spreading rate. Thus, speedier detection of positive cases by analyzing clinical data remains a significant problem. In this paper, we have developed an optimization-based machine learning approach to screen COVID-19 patients. The data set we analyzed consisted of clinical and personal information of 65,500 patients from many countries. The proposed novel framework predicted COVID-19 relying on textual data and succeeded in predicting the severity status. A preliminary pre-processing was made for the feature selection based on the ABC optimization algorithm. Being a powerful problem-solving technique, it has succeeded in deriving optimum features. We relied on ABC because of its ability to produce global optimum solutions. The generated features consisted only of clinical symptoms, and the relevance was also proved by ranking. Subsequently SVM classifier was built with sigmoid kernel. When analyzed the performance of this model, we have observed an overall accuracy of 96%. With the other two performance metrics- precision and recall, the model has shown a higher accuracy of 0.96 and 0.97, respectively. Thus, SVM returns above 90% of accuracy with all the considered performance metrics in this binary classification problem. This indicates the efficacy of the proposed approach in predicting the COVID status out of text data. It is also made the severity prediction of positive cases with good accuracy. The machine learning approach interleaved with the optimization strategy has proved its efficacy in various aspects of data analysis in COVID-19 prediction. This would help physicians put more care for needy people to make necessary arrangements for further treatment.

The downside of the proposed ML approach is that more clinical data, specifically the records of previous diseases, were not included in the available data set. The accuracy of the model could have been improved with such additional features. Secondly, we have implemented other ML models such as Random Forest and Perceptron for the classification, but they have not produced comparable results. Therefore, our future work includes developing and testing other competent models that can showcase higher accuracy for both classification and severity predictions using complete data set.

**Funding** We also confirm that no Funding/Research Grants has been obtained for this work. The authors did not receive support from any organization for the submitted work. No funding was received to assist with the preparation of this manuscript. No funding was received for conducting this study. No funds, grants, or other support was received. The authors have no relevant financial or non-financial interests to disclose. The authors have no conflicts of interest to declare that are relevant to the content of this article. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no financial or proprietary interests in any material discussed in this article.

**Declarations**

**Conflict of interest** Authors hereby confirm that there is no conflict of interest or competing interest for the article and work submitted in this journal. All Ethical norms are being followed during each and every phase of the work from the inception to the submission of the manuscript.

**Ethical approval** This article does not contain any studies with human participants or animals performed by the author. In addition to this, Non-financial interests like personal relationships or personal beliefs, position on editorial board, advisory board or board of directors or other type of management relationships, writing and/or consulting for educational purposes, expert witness, mentoring relations has not been utilised.

**References**

Ali T et al. (2020) Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology

Apostolopoulos ID, Bessiana T (2020) COVID-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci Med 43:635–640

Bhattacharya S, Maddikunta PKR, Pham QV, Gadekallu TR, Chowdhary CL, Alazab M, Piran MJ (2021) Deep learning and medical image processing for coronavirus (COVID-19) pandemic: a survey. Sustain City Soc vol 65
COVID-19 Open Research Dataset (2020) (CORD-19), Version dated 2020-12-12, Retrieved from COVID-19 Open Research Dataset (CORD-19), 2020
Karaboga D, Basturk B (2007) A powerful and efficient algorithm for numerical function optimization: artificial bee colony algorithm. J Global Optim 39:459–471
Li WT, Jiayan M, Neil S, Grant C, Jaideep C, Joseph CT, Lauren A et al. (2020) Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. BMC Med Inf Decis Making 20:1–13
MK M, Srivastava G, Somayaji SRK, Gadekallu TR, Maddikunta PKR, Bhattacharya S (2020) An incentive based approach for COVID-19 using blockchain technology
Narin A, Kaya C, Pamuk Z (2020) Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks
Peng H, Deng C, Wu Z (2019) Best neighbor-guided artificial bee colony algorithm for continuous optimization problems. Soft Comput 23(18):8723–8740
Reddy T, Bhattacharya S, Maddikunta PKR, Hakak S, Khan WZ, Bashir AK, Tariq U (2020) Antion re-sampling based deep neural network model for classification of imbalanced multimodal stroke dataset. Multimed Tools and Appl pp 1–25
Saito T, Marc R (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced data. PLoS One vol 10
Saritha R, Vinod Chandra SS (2016) An approach using particle swarm optimization and rational kernel for variable length data sequence optimization. Lect Notes Comput Sci Spring 9712:401–409
Saritha R, Vinod Chandra SS (2018) Multi modal foraging by honey bees toward optimizing profits at multiple colonies. IEEE Intell Syst 34:14–22
Sethy PK, Behera SK, Ratha PK, Biswas P (2020) Detection of coronavirus disease (COVID-19) based on deep features and support vector machine. Int J Math Eng Manag Sci 5(4):643–651
Shan F, Yaozong G, Jun W, Weiya S, Nannan S, Miaofei H, Zhong X, Dinggang S, Yuxin S (2020) Lung infection quantification of COVID-19 in CT images with deep learning
Shis Kumar Das, Shiba Mishra, Saji Saraswathy Gopalan (2020) Predicting COVID-19 community mortality risk using machine learning and development of an online prognostic tool. PeerJ, vol 8
Silva P, Luz E, Silva G, Moreira G, Silva R, Lucio D, Menotti D (2020) COVID-19 detection in CT images with deep learning: a voting-based scheme and cross-datasets analysis. Inf Med Unlock vol 20
Singh D, Vijay K, Manjik K (2020) Classification of COVID-19 patients from chest CT images using multi-objective differential evolution based convolutional neural networks. Eur J Clin Microbiol Infect Dis pp 1–11
Singh N, Singh SB, Houssein EH, Ahmad M (2020) COVID-19: risk prediction through nature inspired algorithm. World J Eng
Suma LS, Vinod Chandra SS (2017) Identification of common structural motifs in RNA sequences using artificial bee colony algorithm for optimization. Lecture Notes in Computer Science. Springer, Berlin
Tahamtan A, Abdollah A (2020) Real-time RT-PCR in COVID-19 detection: issues affecting the results. pp 453–454
Thomas L, Martina F, Riccardo G, Gabriele B et al. (2020) Development of machine learning models to predict RT-PCR results for severe acute respiratory syndrome coronavirus2 (SARS-COV 2) in patients with influenza-like symptoms using only basic clinical data. Scand J Trauma Resuscit Emerg Med 28:113
Vinod Chandra SS (2015) Smell detection agent based optimization algorithm. J Inst Eng Ser B 97:431–436
Vinod Chandra SS, Anand HS (2020) Machine learning: a practitioners approach. PHI Learning, New Delhi
Vinod Chandra SS, Anand HS (2020) Artificial intelligence: principles and applications. PHI Learning, New Delhi
Vinod Chandra SS, Saju Sankar S, Anand HS (2020) Multi-objective particle swarm optimisation for cargo packaging. Lect Notes Comput Sci Spring 12145:415–422
Wang D, Hu B, Hu C et al (2020) Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus infected pneumonia in Wuhan. JAMA 323(11):1061–1069
Wang Y, Kang H, Liu X et al (2020) Combination of RT-qPCR testing and clinical features for diagnosis of COVID-19 facilitates management of SARS-CoV-2 outbreak. J Med Virol
Yang Y, Lu Q, Liu M et al (2020) Epidemiological and clinical features of the 2019 novel coronavirus outbreak in China
Zhu G, Kwong S (2010) Gbest guided artificial bee colony algorithm for numerical function optimization. Appl Math Comput 217(7):3166–3173
Zoabi Y, Dera-Rozov S, Shomron N (2021) Machine learning-based prediction of COVID-19 diagnosis based on symptoms. Digit Med 4(1):1–5
Zu ZY et al (2020) Coronavirus disease 2019 (COVID-19): a perspective from China. Radiology

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.