International Journal on
"Technical and Physical Problems of Engineering"
(IJTPE)
Published by International Organization of IOTPE

ISSN 2077-3528
IJTPE Journal
www.iotpe.com
ijtpe@iotpe.com

September 2023 Issue 56 Volume 15 Number 3 Pages 188-196

# OPTIMIZATION FEATURE SELECTION TECHNIQUES FOR BIG DATA USING MULTI PHASE PARTICLE SWARM OPTIMIZATION ALGORITHM

M.A. Salih M.M. Hamad W.M. Jasim

*Computer Science Department, University of Anbar, Anbar, Iraq*
*mohanad.a.alhadethy@gmail.com, dr.mortadha61@gmail.com, co.wesam.jasim@uoanbar.edu.iq*

**Abstract-** Large data may not only be abundant in terms of volume, but they may also be organized into several columns, which increases the feature space's dimensionality. Many big data applications still struggle with feature selection. The suggested approach concentrates on dealing with the high dimensionality and feature selection problems. To successfully address these problems, the FS-MPSO technique is included in the suggested system. The three phases of the proposed system are preprocessing, feature selection, and evaluation. In the preprocessing stage, utilized to handle the missing data and delete the irrelevant, redundant characteristics. The primary subset of traits that can characterize Covid-19 will be selected using an efficient process. is a combined strategy that combines the filter and wrapper approaches. which, in order to give significant and accurate characteristics, chooses the most significant and instructive features for the Covid-19 patient diagnosis. Several filtering methods (IG and PC) are available to select the features that are most important for Covid-19 patient diagnosis. The results of the filter methods will then be utilized as the initial population of MPSO, which will then be used as a wrapper technique to precisely choose the best subset of features. The primary objective of feature selection is to decrease classification time by addressing big data difficulties including data velocity, data variety, and lower dimensions. The proposed system evaluation by ML algorithm is the last step (SVM, C4.5, NB). On a Covid-19. The outcomes show how the FS-MPSO strategy is superior in terms of improving accuracy and lowering the number of selected characteristics. The maximum accuracy was attained by FS-MPSO at 99.6%. Nonetheless, the outcome demonstrates that three to four characteristics were generally sufficient for COVID-19 diagnosis. ID, age, and nationality were the most often chosen features.

**Keywords:** Feature Selection, Covid-19, Big Data, High Dimensionality, Multi-Phase Particle Swarm Algorithm.

## 1. INTRODUCTION

Prior to using the data mining techniques, it is more likely for truly huge data sets that an intermediary, additional step data reduction should be carried out. Although there is a chance that huge data sets will produce better mining results than small data sets, this is not a given. Dimension reduction is the main focus of data simplification in this step [1]. Feature reduction is a technique for identifying small, crucial characteristics from high-dimensional datasets. Feature extraction and feature selection are methods for reducing number features. The term "feature selection" refers to the process of picking out key characteristics from a large set of features [2][3]. While feature extraction is the process to transforming features from one dimension to another in order to minimize the size of the feature collection [4][5]. There are various feature reduction strategies for data mining, however they have several issues with massive data. This is growing area and requires additional research. Several feature selection techniques have been introduced for large data up to this point; the majority of these techniques fall into four categories that are reviewed below and are useful for future research [5,6]. Filter, wrapper, embedding, and hybrid feature selection approaches can be generally categorized [3,7]. These methods typically take a long time to get desirable results, necessitating some type of optimization. In this case, we have opted to optimize using PSO. The PSO algorithm that was created by observing the behavior of various creatures, such as ant swarms and bird flocks [8]. Last but not least, the feature selection approach is utilized to provide the ideal quantity of features for a certain task, like categorization.

## 2. RELATED WORKS

S. Meera, et al. (2017) [9], the Accelerated Bee Colony and Neural Network is proposed in the proposed system. Modules from this research include feature selection, preprocessing, and classification. The k-Nearest Neighbor technique used in preprocessing to effectively handle the noise input. Then, using the preprocessed data, the key and pertinent features are chosen using the AABC optimization algorithm. Following that, classification is carried out using an artificial neural network, which produces classification outcomes that are more accurate given the size of the dataset. The experimental findings demonstrate that the suggested system performs better than other systems in terms of higher accuracy (90%) and recall (0.9) and precision (0.9) values.

Sharma, Shivi, et al. (2018) [10], in this study combines support vector machine and grey wolf optimization techniques are presented. It employs a combinational technique to choose the SVM's ideal parameter choices in order to increase classification accuracy, recall, precision, and f-measure. In this classification, this study created SVM GWO to determine the best SVM parameters after extracting the feature vector with the least amount of error and convergence. With a classification accuracy of 77.09%, the results suggested approach is superior to the standard SVM classification algorithm.

Chen, Hongwei, et al. (2019) [11], supposes a feature selection based on Whale optimization Algorithm (WOA). The classification accuracy of this method is increased by 4-10%. Which means that WOA has better classification accuracy and better optimization efficiency in feature selection then, the proposed improves the optimization efficiency in big data set by 42-48%.

El Hasnony, M. Ibrahim, et al. (2020) [12], this paper, a novel feature selection using a mix of grey wolf and particle swarm algorithm. To identify the best answers, the suggested model combines Euclidean separation matrices with the K-nearest neighbor. Overall accuracy is 90 percent. All datasets were processed in a total of 184.3 seconds, with GWO and PSO taking 272 and 245.6 seconds, respectively.

## 3. MULTI PHASE PARTICLE SWARM OPTIMIZATION

The PSO method is expanded by the multi-phase PSO (MPSO) algorithm. The basic purpose of MPSO is to allow groups of particles to cooperate while pursuing various transient search objectives that alter as the algorithm progresses. [13]. Depending on its short-term goals, a particle may be propelled either toward or away from its own or the world's current best location. As a result, particles are kept out of local optima through a multi-objective search technique. The equations utilized to update each particle's position and velocity are different in MPSO versus PSO. For updating velocity, use the Equation (1).

$$v_{i,n}(t+1) = C_v V_{i,n}(t) + C_g G_i(t) + C_x X_{i,n}(t) \quad (1)$$

Prior to running the MPSO algorithm, the number of phases and groups as well as the short-term search objectives for each group and phase must be determined. Particles go through phases that have various short-term objectives. The frequency of the phase change controls the shift in temporary search objectives within each group (PCF) [14]. Only adjustments to particle locations that improve things are allowed by MPSO. The configuration of the coefficients $C_v$, $C_g$, and $C_x$ determines how the updating process works. These coefficients' chosen values depend on the group and phase that each particle is in at any given time. The signs of the coefficients $C_g$ and $C_x$ ought to be different. They release signals that show a particle's direction of motion with respect to its global peak. There is a higher chance of finding new and improved candidate solutions to the optimization problem hen several objectives are being pursued with various particles. There are some instances, this necessitates leaving the area containing the finest candidate solutions thus far discovered rather than staying there [15].

The MPSO algorithm determines each particle's position similarly to the classic PSO algorithm as Equation (1), but the particle stays put until it performs better (becomes fitter), ensuring that its current position is the best one it has ever known. Several evolutionary algorithms employ restart mechanisms to prevent the objective function from becoming stuck in local optimal states. The velocity change variable (VC) of MPSO causes the initialization of particle velocities at random after a predefined number of iterations [16].

## 4. FEATURES REDUCTION

High-dimensional data are typical of most real-world data mining applications, where not all attributes are significant [17]. Modern data mining techniques are unable to overcome the abundance of redundant and weakly relevant characteristics [18]. many techniques become computationally unfeasible. As a result, numerous features can be eliminated without the mining process performing worse. The end outcome of the feature reduction method should be:

- Fewer data, allowing the data-mining system to learn more quickly.
- Improved generalization of the model from the data due to a data-mining process's increased accuracy,
- Clear data mining outcomes that are simpler to comprehend and apply.

With fewer features, allowing for the removal of unnecessary or unused features during the subsequent round of data collecting can save. Techniques for dimensionality reduction work by either picking a subset of the current features or by changing the existing features into a new, smaller collection of features. As a result, two common jobs are involved in creating a smaller collection of features, and they are categorized as follows [2].

### 4.1. Feature Selection

Referred to as variable selection. Three goals are pursued by feature selection: enhancing the performance of the model of data mining; enabling a quicker and more efficient learning process; and fostering a deeper comprehension of the underlying mechanism that produces the data. The filter, wrapper, embedding methods, and hybrid methods are generally the four conceptual frameworks in which feature selection methods are used [1].

Without directly attempting to improve the effectiveness of any particular DM approach, the filter model selects characteristics as a preprocessing step. This is often accomplished by choosing a subset of attributes that optimizes the function of a (ad hoc) evaluation function using a search strategy. They evaluate feature subsets based on how well each prospective feature subset picks up new information using the data-mining technique [18]. Embedded approaches combine the learning algorithm and feature search into a single formulation of the optimization issue. when there are a huge number of samples and dimensions [1]. In order to improve learning performance, a hybrid approach to feature selection incorporated the advantages of filter and wrapper strategies. First, a filter is used to remove features that increase size. Next, a wrapper is utilized, which operates quickly because to the tiny amount of data and provides good accuracy when additional minor features are present. In conclusion, utilizing any of the aforementioned strategies for feature selection in huge data proved to be faster and more accurate [18].

### 4.2. Feature Extraction

Feature selection techniques use wrapper techniques. Based on how well each potential feature subset utilizes the data-mining approach to assimilate new data, feature subsets are evaluated. The majority of the time, feature composition depends on application knowledge. When one wants to preserve the significance of the features and decide which of them are crucial, the choice of features over extraction or transformation is frequently preferred. Additionally, only the selected characteristics must be calculated or gathered, in contrast to transformation-based approaches where all input data are still necessary to achieve the reduced dimension [19].

## 5. PROBLEM STATEMENT AND CONTRIBUTION

Due to the high demand for medical attention in hospitals and clinics, the Covid-19 epidemic has had a significant influence on the world and has grown to be a serious issue. Every day newer coronavirus cases are being confirmed in various nations. This generates big amounts of data, which make difficult to extract important symptom features. Big data for Covid-19 may be low-quality, multidimensional and highly unstructured. The Contribution. In this work is proposed feature selection mechanism, based on two filtering criteria information gain and Pearson correlation, approved that an optimal features subset has successfully chosen minimum features.

## 6. THE PROPOSED METHODOLOGY

This effort aims to enhance the features selection process, the proposed technique consists of 5 steps as shown in the Algorithm 1, which will be explained in this paper in detail, step by step.

Algorithm 1. Optimization feature selection technique

| |
|---|
| Inputs: Patient's symptoms |
| Outputs: Diagnosis and treatment |
| Step 1: Data Collection and Understanding<br>Step 2: Data Preparation and Pre-processing using outlier detection algorithm as algorithm<br>Step 3: Feature selection using algorithms Information gain and Pearson correlation as Algorithms 2 and Algorithms 3<br>Step 4: Optimizing features selection using artificial intelligence algorithms as modified particle swarm optimization algorithm (MPSO) (4)<br>Step 5: Technique evaluation using confusion matrix |

### 6.1. Data Collection Covid-19

The goal from this study to create a suggested method for covid-19 treatment and diagnosis (DTC-19) for patients using an actual data set in order to produce accurate and significant findings that can aid hospital decision-makers. It needs to be an efficient method that can collect all of the necessary data. consequently, an online questionnaire is created. It covers the entirety of the world's nations, and promotion of it will take place on the internet via social networking sites and Twitter. patients at numerous hospitals Suffers are made up of a set of features that could influence and forecast the target Class. The linked features for the patient status influence the choice of the requested features for the training dataset. These characteristics are limited to several laboratory test factors, such as governorate, age, sex, and other individual traits (temperature, cough, headache, smell sense, taste sense, etc.). In order to forecast whether the patient status (the target class) would be favorable or unfavorable, these features are used. 5600 patients from around the world who ranged in age and gender filled out the questionnaire, which was used to identify key patterns about the patients. The following link will bring you to the page on the CET SEARCH website where the database was uploaded. https://github.com/mohanadalhadethy/covid-19database

### 6.2. Data Pre-Processing for Covid-19

After the questionnaires were given out, the data was compiled, and the process of analyzing and interpreting feedback was complete, and it was time to begin the process of preparing the Covid-19 data. The raw Covid-19 data contained some instances that did not apply, so this process was necessary. Excel sheets were used to import it. Data preprocessing and cleaning are processes that, values that are missing from the dataset are eliminated, as are values that are deemed to be outliers.

It is for this reason that the generalization of data is considered to be one of the systems for the reduction of data. in the present work After the data Covid-19 have been cleaned, prepared and processed. The goal of the outlier detection process is to identify instances of unusual or rare behavior that stand in contrast to the majority of the data points contained within a dataset. Recently, the detection of outliers has emerged as an important issue in a wide variety of applications, including the medical field. During the process of data pre-processing and cleaning, the missing data and any values that are deemed to be abnormal are removed from the dataset.

### 6.3. Feature selection of Covid-19 Data

By utilizing various feature selection algorithms like information gain (IG) and Pearson correlation, this work focuses on finding the most crucial features that could affect the precision of the model for prediction (PC), as shown in Figure 1. This could have an effect on how well the model predicts.

In this work, Feature subset generation depends on a filter approach. The proposed model calculates two statistical measurements consisting of IG and PC of each feature with the target class. Then, their feature scores rank according to the selected metrics. correlation-based filter approach produces correlation matrix [attributes, attributes] amongst features in datasets to measure feature redundancy (intra-feature correlation) utilizing Pearson correlation that is computed using Algorithm 2. then computes IG for specific attributes using Algorithm 3.

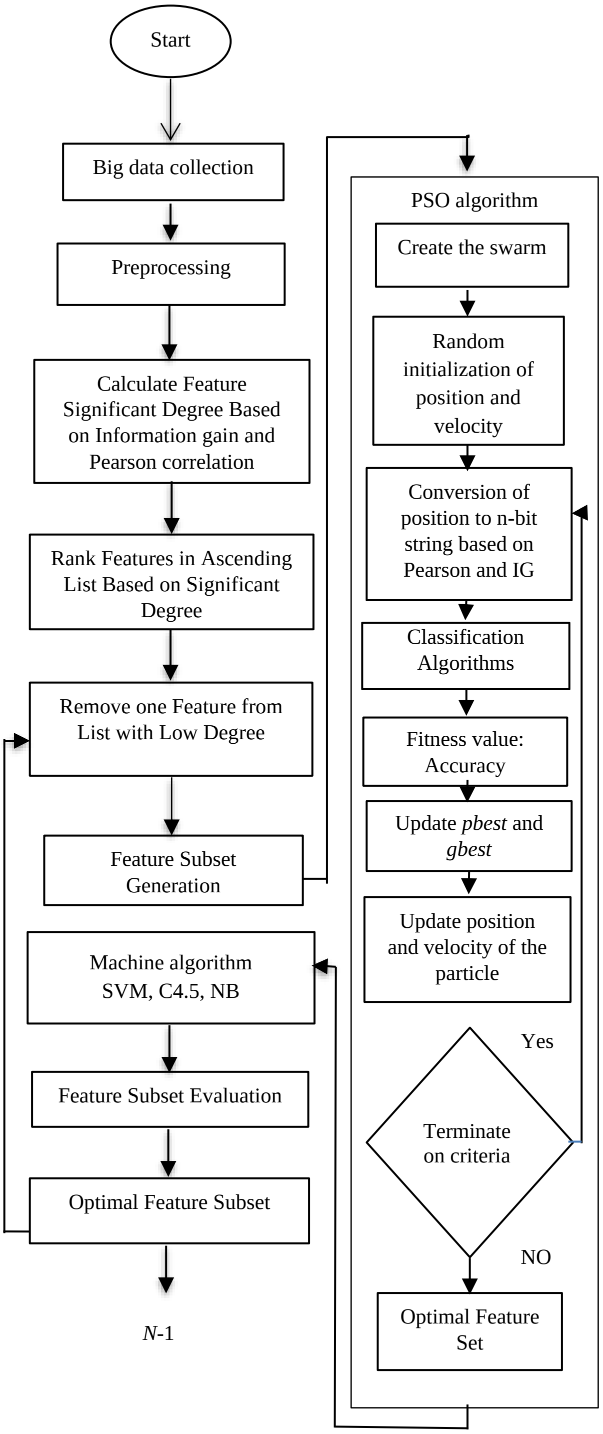
Start
Big data collection
Preprocessing
Calculate Feature Significant Degree Based on Information gain and Pearson correlation
Rank Features in Ascending List Based on Significant Degree
Remove one Feature from List with Low Degree
Feature Subset Generation
Machine algorithm SVM, C4.5, NB
Feature Subset Evaluation
Optimal Feature Subset
N-1
PSO algorithm
Create the swarm
Random initialization of position and velocity
Conversion of position to n-bit string based on Pearson and IG
Classification Algorithms
Fitness value: Accuracy
Update pbest and gbest
Update position and velocity of the particle
Terminate on criteria
Yes
NO
Optimal Feature Set

Figure 1. Feature selection for prediction Covid-19

Pearson correlation can be used to find the relationship between two variables (*I*, *j*) as shown in Algorithm 2 by calculating the Pearson correlation between them as shown in step 3. The Pearson correlation can be used to determine whether there is a relationship between the symptoms that determine the symptoms that are common in the diagnosis of Covid-19 disease, and through which the strength and direction of that relationship is determined. It also assumes that the relationship between features is linear and that the variables are normally distributed.

Algorithm 2. Pearson correlation

```
Input: dataset
Output: Pearson correlation
Step 1: Pearson correlation = []
Step 2: get all columns in dataset
Step 3: for first i in columns:
            for second j in columns:
                $x$= Values in the first i set of datasets
                $y$= Values in the second j set of datasets
                $R = \frac{\left[N(\sum xy) - (\sum x \sum y)\right]}{\left[N\sum x^2 - (\sum x)^2\right]\left[N\sum y^2 - (\sum y)^2\right]}$
            Pearson correlation [first $i$, second $j$] = $R$
Step 4: return Pearson correlation
```

In Algorithm 3 the Information Gain is computed right now. The first step involves computing the entropy of the starting state. Next, the values, which are one type for each column, are located. After that, two subsets (right and left) of the data are created based on the value that was calculated in the step before it, and finally, the value of the Information Gain is located.

Algorithms 3. Calculate information gain

```
Input: data set, target
Output: information gain
Step 1: Do the math for the original entropy: original entropy = entropy (data [target name])
Step 2: Find the values in the column that are unique.
        values = data [split name]. unique ()
Step 3: Separate the data into two subsets according to the special values.
            left split = data [data [split name] = values [0]]
        right split = data [data [split name] = values [1]]
Step 4: Calculate subset entropies by iterating through splits.
subtract = 0.
relating to the subset in [left split, right split]:
            prob = (subset data / data)
            subtract = prob * Calculate entropy (subset data [target name])
Step 5: Information gain is equal original entropy minus.
Step 6: return information gain
```

### 6.4. Optimizing Features Selection

The proposed (FS-MPSO) focuses on hybrid feature selection techniques to obtain advantages of both and to overcome the disability of them. Flowchart (1) illustrates the workflow of the proposed feature selection mechanism. where two datasets are analyzed using both the Pearson correlation and information gain. The examined features are ranked in a list according to their merits in ascending order. At each iteration, the proposed technique deletes one least significant feature from the list and observes the Machine algorithm (SVM, NB, C4.5) performance in terms of its accuracy. The selected feature subset that results in high accuracy with a smaller number of features can be considered as an optimal feature subset for other classification systems. The proposed technique prunes one feature step by step with less merit from the feature set and evaluates the rest to observe which sub-set feature results in high-performance accuracy. Elimination of features continues for N-1 iteration when the top two features remained in the sorted feature set, where N is number of features.

In order to lessen the drawbacks of wrapper systems, particularly their higher time complexities, the model is built around MPSO. In the initial step, the model assumes that the entire feature set is an n-bit binary string and that the string exclusively contains zeroes. The swarm then emerges. The location and velocity vector values are then randomly initialized while taking the PSO attribute into account. The threshold for selecting a particular feature is now determined by the mean of all the feature-class mutual information. A feature is chosen when its ranking in gbest is higher than this criterion. The accuracy is calculated after the classification algorithm has been applied, and the result is our "fitness value." Then, the particle positions and velocities, as well as the pbest and gbest values, are updated appropriately, as shown in Algorithm 4. The criteria for when the process should end are either the maximum number of iterations possible or the point at which the accuracy values converge. If the criteria are not satisfied, the algorithm will return to the fifth step as depicted in Figure 1, but if they are, it will move on to the next step and generate the best possible feature set. The proposed model (FS-MPSO) employs the machine learning algorithms (SVM, NV, and C4.5) as a classification algorithm to carry out the evaluation of the feature subset.

Algorithms 4. MPSO algorithm

```
Input: SWARM_SIZE,          MAX_ITERATION
       PROBLEM_DIMENSION
       C1, C2, ALFA, BETA
Output: best feature select
Step1: initial parameters
       pBest[SWARM_SIZE];
       fitnessValue [SWARM_SIZE]
Step2: initialize Swarm
       For each Swarm do
           Randomize problem dimension in the particle
           randomize location inside a space defined in Problem Set
           get fitnessValueList
       end for
Step3: select First fitness and  pBestLocation
       For each Swarm do
           first fitness value of each item is made as pBest
           first location of each particle is also used as pBestLocation
           VELOCITY  =  [PROBLEM_DIMENSION]
       end for
Step4 :time=0
      while MAX_ITERATION  do
  Step4-1:find the index of the particle that has the best fitness value
of all particles
        bestParticleIndex = getMaxPos(this.fitnessValue)
  Step4-2:update pBest find the particle with the best fitness value
          For each Swarm do
             If fitnessValue[i] > pBest[i]  then
                 pBest[i] =  fitnessValue t[i];
                 pBestLocation[i] =  swarm[i].Location
           endif
Step4-3:update gBest bestParticleIndex
          if (time = 0 or   fitnessValue [bestParticleIndex] >  gBest)
then
               gBest =fitnessValue [bestParticleIndex]
              gBestLocation =  swarm[bestParticleIndex].Location
           endif
Step4-4:calction wigth of swarm
         w = 0.5 + (random / 2.0)
Step4-5: Loop function to update velocity and location of each
particle
         For each Swarm do
               update velocity
               for index each PROBLEM_DIMENSION do
                   VELOCITY[i][ index] =
                   (w * Velocity.Vel[index]) + (r1 * C1) *
                              (pBestLocation[i].Loc[index] –
                               p.Location.Loc[index]) + (r2 * C2) * (
gBestLocation.Loc[index] -Location.Loc[index])
Step4-6: update location to select new attributefor
         Select value random
                for index each PROBLEM_DIMENSION do
           sv = 1 / (1 + Math.Exp(-1 * VELOCITY [index]));
                if  value < sv  then
                    Sele_Loc[index] = 1
                Else
                    Sele_Loc [index] = 0
                 Endif
Step 4-7: get all data from dataset base on Sele_Loc when the index is
1
Step4-8:  do classfy and find max accuracy
Step4-9: modify glob best
           GBEST[t] = gBest;
         Enddo
            End for
Step 5: get max accuracy and best feature select
```

## 7. EXPERIMENTS AND RESULTS DISCUSSION

The (DTC-19) system is divided into several stages. Therefore, a number of interfaces that clearly explain each system stage are included in the implementation of the system stages. The following sections will include examples of these interfaces.

### 7.1. Data Collection and Understanding

In order to construct the proposed system (DTC-19), the database that was collected through the questionnaire is called, which contains more than 5000 cases of corona disease, and each case consists of 22 features. They are the main symptoms for this disease as it was described by the doctors of the specialty.

### 7.2. Data Preparation and Pre-Processing

The first phase of the proposed system is the processing, which is an important phase to prepare data and the Discovery abnormal data which can affect negatively to the accuracy of the system. This is done Through the process of data mining and the detection of anomalies, using algorithm of the (Outlier), which divides each class (Column) into four layers, where the second and third layers are taken as good cases. The first and last layers are considered abnormal cases and are neglect it.

The identification of outliers in knowledge bases with rules that contain information on Covid-19 cases. A decision support system's knowledge base, which is typically its foundation, can be made more complete with the help of knowledge engineers or domain experts in knowledge extraction. Filtering outlier items and irrelevant features from the patient's data is the primary goal of the pre-processing phase. outlier rejection method is to locate and exclude data that have been hurriedly acquired and show significantly exceptional behavior in relation to other data. To ensure that the feature ranking phase goes smoothly, irrelevant features from the patient's laboratory results should be removed in order to choose only the best subset of features.

### 7.3. Feature Subset Selection Using Filter Model

The second phase of the system is the selection of important features. This phase is considered one of the most important phases to increase accuracy and reduce implementation time, in addition to its importance in reducing dimensions of representing data, which helps in speeding up the work of the proposed system. To prove this, the effectiveness of the system was tested on data without using the algorithms for choosing the feature Selection, where three algorithms of machine learning used to evaluate effectiveness of the system.

A) Decision Tree Algorithm (C4.5): The results showed that the accuracy was high in the decision tree algorithm at the expense of execution time in the case large data. the high execution time comes because we used all 22 features. In the data testing phase, the results show the accuracy obtained using the (C4.5) algorithm is (99 %), as shown in Table 1.

Table 1. Data testing phase using C4.5 algorithm

| Class fled | value | Class 0 | Class1 | Row sum | Precision |
|---|---|---|---|---|---|
| Class 1 | 0 | 0 | 1 | 917 | 0.9989 |
| Class 2 | 1 | 916 | 1 | 217 | 1 |
| Column Sum | | 0 | 217 | 1134 | |
| recall | | 916 | 218 | | |
| Accuracy | 0.9954 | | | | |

B) Kernel Support Vector Algorithm (SVM): The effectiveness of the system was tested using the (SVM) algorithm. In the data testing phase, the results showed that the obtained accuracy by using Gaussian Kernel in Sport Vector algorithm is equal to (80%) as shown in Table 2.

Table 2. Data Testing Phase Using SVM Algorithm

| Class fled | value | Class 1 | Class2 | Row sum | Precision |
|---|---|---|---|---|---|
| Class 1 | 0 | 916 | 218 | 1134 | 0.8078 |
| Column Sum | | 916 | 218 | 1134 | |
| recall | | 1 | 0 | | |
| Accuracy | 0.8078 | | | | |

C) Naive Bayes Algorithm (NB): The effectiveness of the system was tested using the (NB) algorithm. in the data training phase, In the data testing phase, the results show the obtained accuracy through the Nave Bayes algorithm is (72%), as shown in the Table 3.

Table 3. Data testing phase using NB algorithm

| Class fled | value | Class 2 | Class3 | Row sum | Precision |
|---|---|---|---|---|---|
| Class 2 | 1 | 729 | 187 | 916 | 0.7956 |
| Class 3 | 2 | 124 | 94 | 218 | 0.4312 |
| Column Sum | | 853 | 218 | 1134 | |
| recall | | 0.8546 | 0.3345 | | |
| Accuracy | 0.7257 | | | | |

We note from the results that appeared above that the test of the effectiveness of the system without use Feature selection algorithm was middle, where the accuracy that we obtained through the C4 algorithm was (0.99), the sport factor machine (0.80), and the Nave Bays (0.72). Figure 2 shows the accuracy for each algorithm.

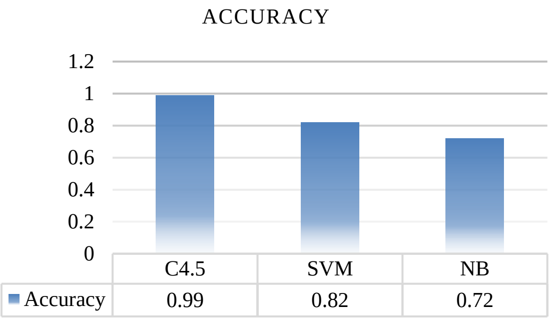

Figure 2. Data testing phase using (NB) algorithm

### 7.4. Feature Subset Selection Using Information Gain

To increase the accuracy and reduce the dimensions, filter model algorithms were used, which includes both the Pearson Correlation algorithm and the Information gain algorithm, which is one of the features selection methods. The feature selection method is demanding because features that do not seem relevant individually may be very important when taken together with other features. In addition, irrelevant features may be so useless that deleting some of them can remove unnecessary complexity. Pearson Correlation and Information gain were used to analyze the data. The Figure 7 shows the IG calculated for each feature, and arrange the features in descending order based on the IG value, which the system will choose the features with the best values. IG is calculated by Algorithm 3.

### 7.5. Feature Subset Selection Using Pearson Correlation

To choose the best features, in this work we also applied the Correlation Matrix technique, which gives attribute values confined between (-1 To 1), then the system works to find the best correlation between the attributes according to algorithm that we mentioned in Algorithms 2. We note that we got the best correlation between the features, which is (1).

After applying the Correlation Matrix algorithm, the attributes selected based on the threshold that we chose, which is (0.65), while the number zero represents the attributes that have been neglected and its value is less than the value of the threshold.

Several filter methods (IG and PC) will be deployed individually on the same Covid-19 dataset in order to swiftly pick a different subset of features in accordance with each approach, in order to verify the performance of the system after features selection utilizing filter model algorithms, we note that the accuracy obtained using each of the C4.5, SVM and NB algorithm are (0.99), (0.80), and (0.79) increased slightly in addition to reducing the dimension and improving the response time for decision makers and this is very useful Especially when dealing with big data. Table 4 shows the accuracy of the system through feature selection filter model algorithms.

Table 4. Accuracy feature selection using Filter model

| | C4.5 | SVM | NB |
|---|---|---|---|
| Precision class 0 | 0.9956 | 0.8076 | 0.8148 |
| Precision class 1 | .0000 | 0.0000 | 0.3810 |
| Recall class 0 | 1.0000 | 1.0000 | 0.9716 |
| Recall class 1 | 0.9817 | 0.0000 | 0.0734 |
| Accuracy | 0.9965 | 0.8076 | 0.7988 |

### 7.6. Optimize Features Selection Using Wrapper Module

When diagnosing Covid-19 patients, after employs the filter technique to quickly identify the most crucial features and eliminate the unnecessary and ineffective ones. The computational cost in the wrapper phase, the next step in the feature selection process, is minimized by this behavior.in addition to giving us a less representation of the data, which enables the use of the program as IOT. In this research, the algorithm (MPSO) was used to obtain the best features from the features that were selected in the previous phase during the filter model phase as shown in Algorithm 4. Figure 3 shows the speed of training of the MPSO Velocity algorithm. Where each of these lines represent a full cycle of the PSO algorithm cycles. This diagram changes depending on the parameters that were specified on the left of the image, which are (Swarm size, number of cycles, dimension, alpha, and beta).

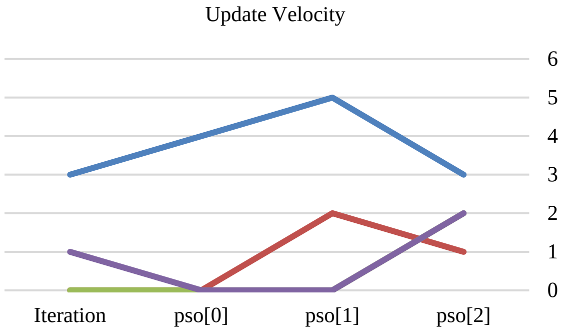

Figure 3. Training the MPSO (Velocity)

To test the effectiveness of the system, we applied SVM algorithm, the accuracy of the results we obtained after applying MPSO was (0.99) to choose the best features. Where each cycle of the PSO algorithm is represented as shown in the Figure 4, and each color in the image indicates the cycle number as shown on the right of the image. We note the importance of MPSO is also in updating the accuracy of the MPSO depending on the best accuracy we got from these cycles. The table which appears at the end of figure shows the best accuracy by each iteration.

## 8. PERFORMANCE EVALUATION

Accuracy, precision, recall, and chi-square are four evaluations that will be computed during the subsequent experiments to clarify the application results. The values of these metrics are determined using a confusion matrix. as shown in the Figure 5.

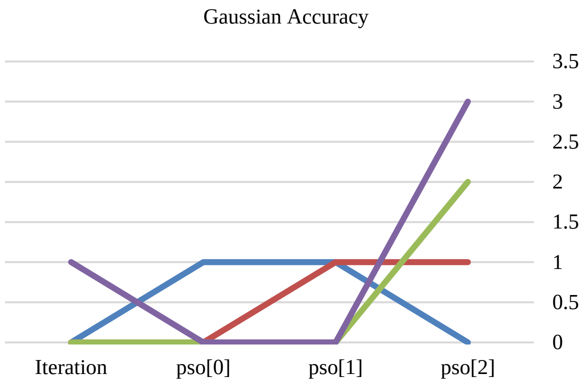

Figure 4. Accuracy MPSO to choose the best features

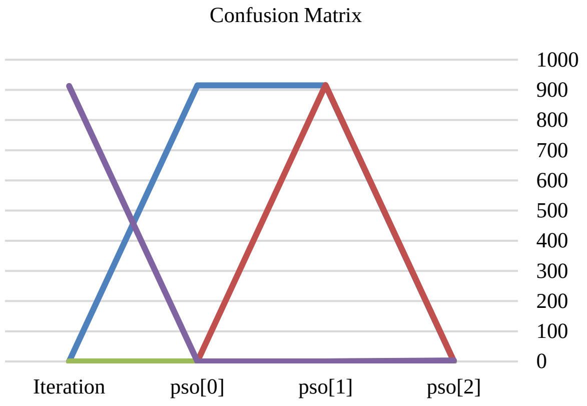

Figure 5. Calculating the Confusion Matrix

The Table 5 shows the results that were obtained, which are the number of features that were selected in MPSO, and what are those features, in addition to the accuracy that we obtained by every cycle by SVM Gaussian kernels.

Table 5. Features selection using MPSO algorithm

| Index feature | Number feature | NB Gaussian |
|---|---|---|
| 14,17,7,18,1,0,3,4… | 19 | 0.9956 |
| 9 | 1 | 0.8078 |
| 15,16,17,10,4,8,11,9… | 19 | 0.9956 |
| 16,0,5,15,8,3,2,14 … | 11 | 0.9303 |
| 7,8,5,2,6,11,16,18 … | 19 | 0.9956 |
| 14,0,2,8,5,7,17,7,1,… | 14 | 0.9700 |
| 16,0,10,15,5,10,… | 16 | 0.9832 |

## 9. RESULTS DISCUSSION

The performance of the system can be evaluated using SVM Gaussian kernels in three different phases: the phase without feature selection, the phase of feature selection by filter methods, and the phase of selecting features that combines the filter and wrapper methods based the MPSO algorithm, as shown in Figure 6. We note that the accuracy has been improved to reach (99%) while reducing dimension and improving response time.

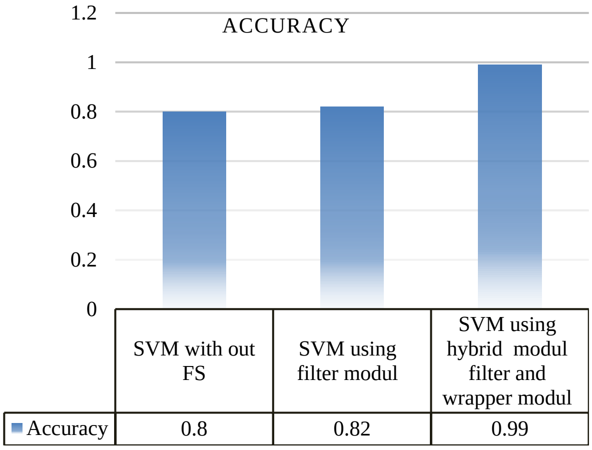

Figure 6. Comparison the results between three phase

## 10. CONCLUSIONS

Many big data applications still struggle with feature selection. The proposed system concentrates on dealing with the high dimensionality and feature selection problems. To effectively address these problems, the FS-MPSO approach is included in the suggested system. The three phases of the proposed system are preprocessing, feature selection, and evaluation. Preprocessing is used to handle missing values and remove redundant and irrelevant attributes. Effective technique will be utilized to choose the key subset of characteristics for Covid-19. is a combined strategy that combines the filter and wrapper approaches. which, in order to provide significant and accurate features, chooses the most significant and instructive features for the Covid-19 patient diagnosis. In comparison to the current research methods, there are many filtering techniques (IG and PC) to select the most important features for Covid-19 patient diagnosis results.

## REFERENCES

[1] M. Sais N. Rafalia J, Abouchabaka, "Enhancements and Intelligent Approach to Optimize Big Data Storage and Management: Random Enhanced HDFS (REHDFS) and DNA Storage", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 50, Vol. 14, No. 1, pp. 196-203, March 2022.

[2] Sh. Akbarpour, "A Review on Content-Based Image Retrieval in Medical Diagnosis", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 15, Vol. 5, No. 2, pp. 148-156, June 2013.

[3] M. Rong, D. Gong, X. Gao, "Feature Selection and its Use in Big Data: Challenges, Methods, and Trends", IEEE Access, Vol. 7, pp. 19709-19725, January 2019.

[4] M. Pourhomayoun, M. Shakibi, "Predicting Mortality Risk in Patients with Covid-19 Using Machine Learning to Help Medical Decision-Making", Smart Health 20, Vol. 5, pp. 100-178, January 2021.

[5] H. Chen, et al., "A Spark-Based Distributed Whale Optimization Algorithm for Feature Selection", The 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS) IEEE, Vol. 1, pp. 70-74, December 2019.

[6] N. Falguni, H. Shah, et al., "Big Data Feature Selection Methods: A Survey", International Journal of Research and Analytical Reviews (IJRAR), Vol. 5, Issue 2, June 2018.

[7] L. Jundong, H. Liu, "Challenges of Feature Selection for Big Data Analytics", IEEE, Issue 2,Vol. 32, pp. 9-15, April 2017.

[8] S. Wong, R. Vasilakos, et al., "Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data", IEEE, Issue 1, Vol. 9, pp. 33-45, January 2016.

[9] K. Linghe, et al., "Distributed Feature Selection for Big Data Using Fuzzy Rough Sets", IEEE, Issue 5, Vol. 28, pp. 846-857, May 2020.

[10] S. Jeetha, et al., "Acceleration Artificial Bee Colony Optimization-Artificial Neural Network for Optimal Feature Selection Over Big Data", IEEE International Conference on Power Control, Signals and Instrumentation Engineering (ICPCSI), IEEE, pp. 1698-1706, September 2017.

[11] S. Sharma, G. Rathee, H. Saini, "Big Data Analytics for Crop Prediction Mode Using Optimization Technique", The Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), IEEE, Vol. 9, pp. 760-764, December 2018.

[12] M. Al Jame, et al., "Apache Spark Implementation of Whale Optimization Algorithm", Cluster Computing, Vol. 23, pp. 1-19 July 2020.

[13] M.E. Hasnony, M. Ibrahim, et al., "Improved Feature Selection Model for Big Data Analytics", IEEE, Vol. 8, pp. 1-17, April 2020.

[14] H. Shayeghi, H. Shayanfar, G. Azimi, "A Hybrid Particle Swarm Optimization Back Propagation Algorithm for Short Term Load Forecasting", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 4, Vol. 2, No. 3, pp. 12-22, September 2010.

[15] M. Zile, et al., "Design of Power Transformers Using Heuristic Algorithms", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 38, Vol. 11, No. 1, pp. 42-47, March 2019.

[16] B. Al Kazemi, C. Mohan, "Multi-Phase Discrete Particle Swarm Optimization", Proc. The Fourth International Workshop on Frontiers in Evolutionary Algorithms (FEA 2002), Vol. 1, pp. 489-494, May 2002.

[17] M. Tabrez, et al., "Power Conversion Techniques Using Multi-Phase Transformer: Configurations, Applications, Issues and Recommendations", Machines, Issue 1, Vol. 10, pp. 1-23, December 2022.

[18] H. Chen, et al., "Co-Evolutionary Competitive Swarm Optimizer with Three-Phase for Large-Scale Complex Optimization Problem", Information Sciences, Vol. 619, pp. 2-18, January 2023.

[19] L. Wogi, et al., "Particle Swarm Optimization Based Optimal Design of Six-Phase Induction Motor for Electric Propulsion of Submarines", Energies, Issue 9, Vol. 15, pp. 1-21, April 2022.

[20] M. Lei, et al., "A Novel Wrapper Approach for Feature Selection in Object-Based Image Classi_cation Using Polygon-Based Cross-Validation", IEEE, Issue 3, Vol. 14, pp. 409-413, March 2017.

[21] R. Sathya, et al., "Economically Efficient data Feature Selection Using Big data Analysis", International Journal

of Innovative Technology and Exploring Engineering, Issue 7, Vol. 8, pp. 983-987, May 2019.

[22] A. Raweh, et al., "A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation", IEEE Access, Vol. 6, pp. 15212-15223, March 2018.

## BIOGRAPHIES

Name: **Mohanad**
Middle Name: **Ahmed**
Surname: **Salih**
Birthday: 09.11.1985
Birth Place: Baghdad, Iraq
Bachelor: Computer Science, Computer Science, College of Computer Science and Information Technology, Anbar, Iraq , 2008
Master: Computer Science, Computer Science, College of Computer Science and Information Technology, Anbar, Iraq, 2017
Doctorate: Student, Computer Science, Computer Science, College of Computer science and Information Technology, Anbar, Iraq, Since 2021
Research Interests: Data Mining , Artificail Intelligence
Scientific Publications: 7 Papers, 2 Projects, 2 Theses

Name: **Murtadha**
Middle Name: **Mohammad**
Surname: **Hamad**
Birthday: 05.04.1961
Birth Place: Hadetha, Iraq
Bachelor: Computer Sciences, College of Sciences, University of Mosul, Mosul, Iraq, 1983
Master: Computer Science, College of Sciences, University of Baghdad, Baghdad, Iraq, 1991
Doctorate: Computer Science, College of Sciences, University of Technology, Baghdad, Iraq, 2004
The Last Scientific Position: Prof., Computer Science, College of Computer Science and Information Technology, University of Anbar, Anbar, Iraq, 2010
Research Interests: Data Warehouse, Data Mining, Big Data
Scientific Publications: 60 Papers, 34 Theses, 10 Projects
Scientific Memberships: Member of the Ministerial Committee for the Development of Information Systems Curricula Iraqi Ministry of Education, Member of the Science and Technology Journal Committee, University of Wasit, Iraq, Chairman of Promotion Committee, College of Computer Science and Information Technology, University of Anbar, Iraq

Name: **Wesam**
Middle Name: **Mohammed**
Surname: **Jasim**
Birthday: 20.05.1973
Birth Place: Anbar , Iraq
Bachelor: Control and Automation, Electrical Engineering, University of Technology, Baghdad, Iraq, 1996
Master: Control and Automation, Electrical Engineering, University of Technology, Baghdad, Iraq, 2002
Doctorate: Control and Automation, School of Computer Science and Electronic Engineering, University of Essex, Essex, UK, 2016
The Last Scientific Position: Prof., Computer Science, College of Computer Science and Information Technology, University of Anbar, Anbar , Iraq, 2023 - Chairman of Computer Science, College of Computer Science and Information Technology, University of Anbar, Anbar, Iraq
Research Interests: Control, Artificial Intelligence, Robotics, Deep Learning
Scientific Publications: 30 Papers, 2 Books, 10 Theses