Cluster Analysis of Novel Corona Virus Through Nature Inspired Algorithms

Enrollment No.s- 17104039, 17104041, 17104068

Name of the students- Sakshi Gupta, Manvi Chawla, Adhar Agrawal

Name of the Supervisor- Dr. Parul Agarwal

December 2020

Submitted in partial fulfillment of the Degree of

Bachelor of Technology

In

Computer Science Engineering

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING &

INFORMATION TECHNOLOGY

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

ACKNOWLEDGEMENT

We would like to express our special thanks of gratitude to our mentor Dr. Parul Agrawal for teaching us, helping us with the project as well as guiding us with the relevant resources, and ultimately providing us with an opportunity to make and present this report. We would like to express our gratitude towards our parents & member of Jaypee Institute Of Information Technology for their kind co-operation and encouragement which help me in completion of this project.

Date- 4th December, 2020

DECLARATION

We hereby declare that the project report is based on our own work carried out during the course of our study under the supervision of our mentor Dr. Parul Agrawal.

I. We assert the statements made and conclusions drawn are an outcome of our research work.

II. I further certify that the work contained in the report is original. This work has never been submitted to any other Institution or any other University of India or abroad.

III. We have prepared the report based on the guidelines provided by our institute.

IV. Whenever we have used data and text from other resources, we have given the respective citations for them in the reference section.

Sakshi Gupta(17104039)
Manvi Chawla(17104041)
Adhar Agrawal(17104068)

Table of contents

List Of Figures

List Of Acroynms Used

WHO	Wealth Health Organization
SARS-COV	Severe Acute Respiratory Syndrome Coronavirus
PSO	Particle Swarm optimization
GA	Genetic Algorithm
GWO	Grey Wolf Optimization
HHO	Harris Hawk Optimization
DBC	Distance between clusters
DWC	Distsnce within clusters
FF	Fitness Function

Chapter - 1

Introduction

On 31st December 2019 in China, a disease capable of spreading at a very high rate, identified as Corona virus disease (COVID-19), was reported by the Municipal Health Commission, Wuhan. Further, on 23 January 2020, Wuhan was locked down on the basis of the reports that it is spreading due to community transmission in the cities. The COVID-19 has been spread to the most of the parts of the world and is spreading at a very fast rate which is very difficult to control.

Till date, almost the whole world is affected with COVID-19 and is suffering a lot economically. Till now , there is no specific treatment or vaccine or drug for the disease caused by the COVID-19. The outbreak of COVID-19 is forcing many countries to take harsh steps and policies for improving medical facilities such as ventilators, testing kits, masks, sanitizers etc. for protection of people. It has resulted more than 10 lakhs deaths of people of different age groups. A highly effective care should be taken of the patients suffering from COVID-19 should be taken. The COVID-19 pandemic was declared as a global health emergency by the World Health Organization(WHO) and it encouraged the nation to start improving their health policies and facilites and start preparing for the health emergencies that the nation would be facing due to COVID-19.

Data Science and related technologies plays a very important role in fighting against any pandemic like 2003 Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV), COVID-19 to help governments and health managements to figure out the best preparation they can do to make the world safe from the widespread of COVID-19. Data mining, machine learning and several other technologies can be used to analyze data quickly and effectively to track and control the spread of COVID-19 around the world.

The research commuinty is putting in lots of efforts to explore the medical, and sociological impact of the COVID-19 pandemic so that immediate actions can be taken and the world is made free from COVID-19 as soon as possible. The present study presents a novel analysis which results to clustering countries with respect to active cases and the number of deaths. The cluster analysis presented in the project would be highly beneficial to the managers of the health sectors, finance experts and sociologists to take an immediate action for the highly affected areas and make them safe from the spread of COVID-19.

Chapter – 2

Literature Review

S.NO	Paper	Method/Algorithm	Year	Drawback
1.	Genetic Algorithm with New Fitness Function for Clustering [1]	Genetic algorithm is used for clustering and data mining is used to infer useful information from the dataset.	**2020**	The specified method failed at the outliers.
2.	Genetic cluster analysis of SARS-CoV-2 and the identification of those responsible for the major outbreaks in various countries. [2]	The phylodynamic of 247 high quality genomic sequences were tested to trace the evolution route of the COVID-19 virus. Among these, 4 clusters are chosen as super-spreaders that are considered to be the major cause of the outbreak spread. This process is repeated to get the total number of cities that are highly contributing the widespread of the COVID-19 pandemic.	**2020**	**NA**

3.	Clustering method for spread pattern analysis of corona-virus (COVID-19) infection in Iran.[3]	This research work helps in clustering analysis for time series modelling and provides the pattern of spread of the disease in Iran.	**2020**	This research includes work in which the data of 5 different cities in Iran has been classified. To improve the results a large number of cities should have been considered and the perfor-mance metrices should also have been included in the reuslts.
4.	COVID-19 Optimizer Algorithm, Modeling and Controlling of Coronavirus Distribution Process [4]	An algorithm has been proposed to provide the COVID-19 optimizer and then three scenarios are proposed to solve the optimization of the COVID-19 in these regions.	**2020**	The proposed algorithm fails in achieving high accuracy and the efficiency was also low.

5.	Clustering of Country-Based Data in COVID-19 Infections By Coronavirus outbreak features [5]	This study investigated connections between the infection cycles of states round the world. Utilizing factors like the Day of most Infections, the overall Infections and therefore the Day of most Infections, and Deaths and Recoveries per Million. additionally, countries that have completed the infection cycle were compared to know similarities and variations amongst the same factors.	**2020**	The results obtained were in the form of binary i.e. whether the patient has the disease or not. This study can be extended to further classify the patient on the base of the severity.

6.	Applying Data Clustering Feature to Speed Up Ant Colony Optimization. [6]	The data is firstly divided into local clusters using an improved ACO algorithm and further small TSP routes were calculated which were then esembled to form a large TSP. The running speed of the ACO was increased by a factor of 200.	**2014**	The running time of the ACO is very large which is a very big drawback of this algorithm.
7.	Improved Ant Colony Clustering Algorithm and Its Performance Study [7]	This work proposes a new clustering algorithm that improves the efficiency and the accuracy of the Ant colony clustering algorithm. This algorithms is used to cluster the benchmark problems.	**2015**	The major limitation of ant colony clustering algorithms is that the number of parameters for the two ant colony algorithms are large which makes the running of the algorithm a little difficult task.

8.	Visual tracking using improved flower pollination algorithm [8]	An improved flower pollination algorithm tracking architecture is presented in this study and the comparison of the accuracy is made to verify the tracking ability with PSO is also presented.	**2018**	A lot of work is needed to improve the efficiency of the model as it has been implemented on a very small dataset.
9.	Data clustering using particle swarm optimization [9]	This paper presents how PSO can be used in finding out the centroids of a user specified number of clusters and further K-means clustering is used to feed the intial clusters. Secondly, PSO is used to refine the clusters that are intitally formed using the K-means, and the results shows that both the PSO clustering methods have a high potential in clustering and can be used in the future.	**2013**	This study requires more extensive study on the higher dimensional problems and is to be tested on a larger dataset to prove the model accurate.

10.	Visual tracking using CNN and Genetic Algorithm [10]	This paper first uses CNN which fails due to limited dataset. Further transfer learning is used which uses trained models for feature extraction of network layers. At the last CNN with the pre-trained model provides a good result and helps in providing a good accuracy in clustering the data.	**2017**	This model was tested on a very small dataset and hence work is required in increasing the dataset so that the results obtained are accurate.

11.	Cluster of coronavirus disease 2019 (Covid-19) in the French Alps, 2020 [11]	The connections between the infection cycle from all over the world is considered in this study. The factors such as Maximum infections per day and deaths and recoveries per day is considered. Moreover, the countries where is the infection cycle is over is also compared with the countries where the infection cycle is going on as it would provide the future results and would help in preparing for the coming future.	2020	NA
12.	Monitoring Novel Corona Virus (COVID-19) Infections in India by Cluster Analysis [12]	A data mining techniques known as clustering analysis is applied in this model to detect the spread of the COVID-19 infection in the differnet parts of India.	2020	More work is required t prove the accuray of th model as the dataset used i very less. And the dat collected was not proper a well.

13.	Grey Wolf Optimizer Based on Powell Local Optimization Method for Clustering Analysis. [13]	An extended version of Grey Wolf Optimization is implemented in this paper that is known as powell local optimisation method called as PGWO. This algorithm helps in solving large scale clustering problems likewise we need in the clustering of the COVID-19 data.	2015	Work is needed to dynamically determine the optimal number of clusters. And some optimization is also required to increase the dataset capacity so that higher dimensional problems can be solved.
14.	Improved Binary Grey Wolf Optimizer and Its application for feature selection. [14]	This paper presents the application of binary grey wolf optimization algorithm that helps in large scale clustering of data. This algorithm gives the result in binary form that whether the disease is present or not.	2020	Some more feature selection methods are required to make sure that the error occured in the data is less and more appropriate classification is required.

15.	Target specific mining of COVID-19 scholarly articles using one-class approach [15]	The clustering of the COVID-19 dataset present in India is done in this research article using parallel one class support vector machines. The previous dataset of the SARS and the MERS is also considered and the trend that it followed is considered in this article that provides a better future prediction that the COVID-19 would do to the world using the pre-trained data.	2020	This dataset is not feasible for very large dataset as K-means is followed by the parallel one-class support vector machine algorithm which makes the working of the large datasets a little difficult.
16.	A Novel Hybrid Harris Hawks Optimization for Color Image Multilevel Thresholding Segmentation[16[	This study presents a hybrid algorithm for segmenting colour image. This helps in extracting the features from a high performance of the two existing algotithm. Two techiques, Ostu's method and kapur's method are used as fitness function that helps in feature extraction and eventually determines the segmentation threshold values.	2019	This model has to be made for general use that means all the types of images can be segmented by this algorithm in future. Moreover, work is required in increasing the efficiency of the algorithm.

17.	A chaotic sequence-guided Harris hawks optimizer for data clustering. [17]	In this study, the Harris Hawk optimization is implemented for clustering the data. Further, after the clustering of the data, it is compared with 12 other benchmark algorithms and it is proved that the Harris Hawk optimization algorithm has a higher effectiveness and accuracy than the other 12 benchmark algorithms.	2020	The proposed algorithm as such do not have any drawback, but it is suggested that the model can be implemented on real-world applications and a multi-objective version of the proposed algorithm can also be formed.
18.	Harris hawks optimization: Algorithm and applications [18]	Harris Hawk Optimization is implemented on the dataset to which clustering is performed on. Harris hawks algorithm reveals enumerous chasing patterns based on the dynamic nature of how the prey escapes from the hunter. It gives very good results when compared with other nature inspired algorithms that were 29 benchmark algorithms.	2019	Efficient work is required in making the algorithm more accurate and it has to be made capable of working on a larger dataset.

19.	A novel Harris hawks' optimization and k-fold cross-validation predicting slope stability [19]	This study introduces a novel metaheuristic optimization technology known as Harris Hawk Optimization that enhances the accuracy of the multilayer perceptron technique that helps in predicting the safety in the vicinity of the rigid foundations. This model implments Harris Hawk optimization model that will help in the clustering of the data and would be beneficial in clustering the Coivd-19 dataset that we have.	2019	K-fold cross validation reduces the stabilty of the model and the dataset is also very large that is disturbed by this algorithm. Hence another algorithm is required that will be efficient in handling such a large dataset.
20.	Modified Harris Hawks Optimization Algorithm for Global Optimization Problems. [20]	In this study, the Harris Hawk optimization is implemented for clustering the data. Further, after the clustering of the data, it is compared with 14 other benchmark algorithms and it is proved that the Harris Hawk optimization algorithm has a higher effectiveness and accuracy than the other 14 benchmark algorithms.	2020	The proposed algorithm as such do not have any drawback, but it is suggested that the model can be implemented on real-world applications and a multi-objective version of the proposed algorithm can also be formed.

21.	Grey Wolf Optimizer (GWO) Algorithm to Solve the Partitional Clustering Problem [21]	In this paper, grey wolf optimization (GWO) algorithm which is modelled according to the social behaviour of grey wolves is applied to partition the data samples by searching the optimal center of the clusters. The clustering performance of the GWO is compared with the performances of the three clustering algorithms: k-means, k-medoids and fuzzy c-means algorithms. The experiments show that the GWO algorithm has generally better results than the other clustering algorithms and can be alternatively applied on the clustering problem.	2019	GWO outperforms all the other algorithms applied for clustering that were used as a tester in this model but what happens with GWO is that it works efficiently on a small dataset. So work is required to make it work
22.	Grey wolf optimization based clustering algorithm for vehicular ad-hoc networks [22]	In this article, grey wolf optimization based clustering algorithm for VANETs is proposed, that replicates the social behaviour and hunting mechanism of grey wolfs for creating efficient clusters. The proposed method is compared with well- known meta-heuristics from literature and results show that it provides optimal outcomes that lead to a robust routing protocol for clustering of VANETs, which is appropriate for highways and can accomplish quality communication, confirming reliable delivery of information to each vehicle.	2018	The process of clustering in VANET can be further investigated by implementing different bio-inspired algorithms like Moth-Flame Optimizer, Salp Swarm Algorithm, Dragon-Fly Optimizer Algorithm, Ant Lion Optimizer and Whale Optimization Algorithm. Moreover,the proposed work can be further enhanced by customizing the objective function as per user requirements, and it can be used forth emulti-objective functions as well.

23.	Clustering analysis using a novel locality-informed grey wolf-inspired clustering approach. [23]	This algorithm benefits from stochastic operators, but it is still prone to stagnation in local optima and premature convergence when solving problems with a large number of variables (e.g., clustering problems). To alleviate this shortcoming, the GWO algorithm is hybridized with the well-known tabu search (TS). To investigate the performance of the proposed hybrid GWO and TS (GWOTS), it is compared with well-regarded metaheuristics on various clustering datasets. The comprehensive experiments and analysis verify that the proposed GWOTS shows an improved performance compared to GWO and can be utilized for clustering applications.	2020	There are many spatial applications in the field of location-based services (LBS) that the proposed GWOTS-based clustering approach can also be evaluated. In future works, we will investigate the performance of swarm-based and evolutionary clustering methodssuch as GWOTS on synthetic datasets with different sizes and arbitrary shapes. We will alsoutilize parallel computing to reduce the run time of the proposed GWOTS method.

24.	Hybridizing Grey Wolf Optimization (GWO) with Grasshopper Optimization Algorithm (GOA) for text feature selection and clustering. [24]	The proposed technique is that it produces a mature convergence rate and requires minimal computational time and is trapped in local minima in a low dimensional space. The text data is fed as the input and pre-processing steps are performed in the document. Next, the text feature selection is processed by selecting the local optima from the text document and then selecting the best global optima from local optimum using hybrid GWO-GOA. Furthermore, the selected optima are clustered using the Fuzzy c-means (FCM) clustering algorithm. This algorithm improves the reliability and minimizes the computational time cost. Eight datasets are used in the proposed algorithm and the performance is envisaged efficaciously. The evaluation metrics used for performing text feature selection and text clustering are accuracy, precision, recall, F-measure, sensitivity, specificity and show better quality when comparing with various other algorithms. When comparing with GWO, GOA and the proposed hybrid GWO-GOA algorithm, the proposed methodology reveals 87.6% of efficiency.	2020	Improvement in this work might involve the use of other functional selection algorithms and different fitness functions that are expected to strengthen the success rates.

25.	Genetic algorithms applied to clustering problem and data mining [25]	In this paper the authors investigate the use of Genetic Algorithms to determine the best initialization of clusters, as well as the optimization of the initial parameters. The experimental results show the great potential of the Genetic Algorithms for the improvement of the clusters, since they do not only optimize the clusters, but resolve the problem of the number K cluster, which had been giving it form a priori. The techniques of clustering are most used in the analysis of information or Data Mining, this method was applied to Data Set at mining.	2017	Work is required to make this model efficient of large dataset.
26.	Genetic Algorithm-Based Clustering Technique [26]	A genetic algorithm-based clustering technique, called GA-clustering, is proposed in this article. The searching capability of genetic algorithms is exploited in order to search for appropriate cluster centres in the feature space such that a similarity metric of the resulting clusters is optimized. The superiority of the GA-clustering algorithm over the commonly used K-means algorithm is extensively demonstrated for four artifcial and three real-life data sets.	2015	The implementation of this model was not successful for all the general models. It has it flaws of small dataset capacity and less efficiency.

27.	MGKA: A genetic algorithm-based clustering technique for genomic data [27]	In this paper, a genetic algorithm-based unsupervised clustering method that searchs forthe optimal centers of clusters based on the concept of k-means is proposed. The genetic algorithm reduces k-means sensitivity to randominitialized centers and reduces the probability of converging tolocal minima. Two clustering validity indexes are introduced tothe selection process to automatically determine the appropriatenumber of clusters. The proposed algorithm is applied to 16 disease datasets and four single cell datasets to demonstrate its performance. Results show that our approach outperforms current state of the art algoritms on a majority of the datasets.	2019	This model has to be implemented for multi-objective model and work is required to increase its efficiency.

28.	Particle swarm optimization algorithm and its application to clustering analysis [28]	Clustering analysis is applied generally to Pattern Recognition, Color Quantization and Image Classification. It can help the user to distinguish the structure of data and simplify the complexity of data from mass information. The user can understand the implied information behind extracting these data. In real case, the distribution of information can be any size and shape. A particle swarm optimization algorithm-based technique, called PSO-clustering, is proposed in this article. We adopt the particle swarm optimization to search the cluster center in the arbitrary data set automatically. PSO can search the best solution from the probability option of the Social-only model and Cognition-only model[1, 2, 3J. This method is quite simple and valid and it can avoid the minimum local value. Finally, the effectiveness of the PSO-clustering is demonstrated on four artificial data sets.	2012	NA

29.	Clustering Big Datasets Using Orthogonal Gray Wolf Optimizer [29]	This study proposes a new variant of GWO called as Orthogonal Grey Wolf Optimization (OGWO). It is different from the original GWO in a sense that the position of wolves are not merely updated by averaging the movement towards three global leaders. Instead a combination termed as orthogonal methodology is used to determine the effective update position of the leader wolves. Here the methodology objective is to obtain the best possible combination of positions from the three global leaders. The simulation analysis on standard benchmark function reveals that the results obtained from the proposed algorithm are more optimal and have lesser standard deviation than the previous approach. In addition to this, the proposed algorithm is also successfully used on cluster analysis and very competent results are obtained when compared to other nature-inspired algorithms like original GWO, Particle Swarm Optimization (PSO), Orthogonal PSO (OPSO), Orthogonal Genetic Algorithm with Quantization (OGA).	2019	This paper has a high efficiency for smaller datasets but fails in most of the cases where very large variable dataset is present. It can work on a continous large dataset.

30.	Grey wolf optimization based clustering algorithm for vehicular ad-hoc networks. [30]	In this article, grey wolf optimization based clustering algorithm for VANETs is proposed, that replicates the social behaviour and hunting mechanism of grey wolfs for creating efficient clusters. The proposed method is compared with well- known meta-heuristics from literature and results show that it provides optimal outcomes that lead to a robust routing protocol for clustering of VANETs, which is appropriate for highways and can accomplish quality communication, confirming reliable delivery of information to each vehicle.	2018	The process of clustering in VANET can be further investigated by implementing different bio-inspired algorithms like Moth-Flame Optimizer, Salp Swarm Algorithm, Dragon-Fly Optimizer Algorithm, Ant Lion Optimizer and Whale Optimization Algorithm. Moreover,the proposed work can be further enhanced by customizing the objective function as per user requirements, and it can be used forth emulti-objective functions as well.

Chapter – 3

Algorithms Used

3.1 Clustering through Genetic Algorithm

Crossover

Suppose chromosomes C_1-C_{100} belong to Cluster 1. and C_{101}-C_{200} belong to Cluster 2. In crossover, whole clusters are swapped with each other. All the chromosomes of Cluster 1 are swapped with Cluster 2 and the same is followed for every cluster. The cluster numbers to be swapped are chosen randomly in every generation [1].

Mutation

Suppose chromosomes C_1-C_{100} belong to Cluster 1. Then in mutation process, randomly some chromosomes from one cluster are selected and mutated with same number of chromosomes from another cluster. We have used Roulette wheel to select the fittest chromosomes.

Fitness function

The fitness function is calculated as:
We used distance between clusters (DBC) and distance within clusters (DWC) and silhoutte width as our parameters. DBC is defined as the distance between centroids of each cluster whereas, DWC is defined as the distance between chromosomes within each cluster [2].

DBC is calculated as: …………….(1)

$$Dbc_{m,n} = \sqrt{\sum_{j=1}^{r}(X_i - D_j)^2}$$

where m and n are respective clusters and r is the number of chromosomes in cluster n.

$$WC_{a,a} = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{p}(Xi - Xj)^2}{p * p}}$$

DWC is calculated as:

………..(2)

Where a is a particular cluster and p are the number of chromosomes.

Then the sum of DBC and DWC are calculated. The Silhouette value tells us how dense the cluster is, where ai is the average distance between chromosome i and other chromosomes within the cluster whereas bi is the average distance between i and the chromosomes in the nearest cluster.

The SW is calculated as:

$$SW_i = \frac{b_i - a_i}{\max(b_i, a_i)}$$

…………….(3)

where s is the sample size.

$$FF = \frac{Sum(DBC)}{SUM(DWC)} + SW$$

The fitness function is calculated as:

…………….(4)

The DBC and SW need to be maximized whereas, DWC needs to be minimised.

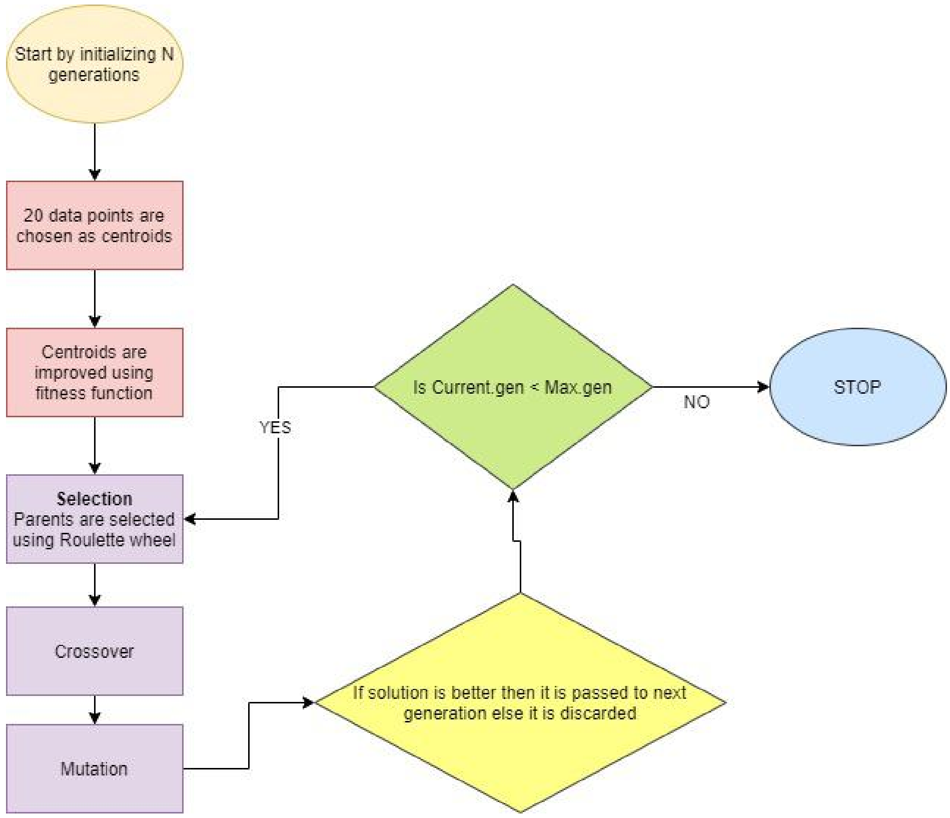


Fig 1. Flowchart of Genetic Algorithm Clustering

3.2 Particle Swarm Optimization Clustering

Particle Swarm Optimization (PSO) is a method that is used for optimization of the continuous non-linear function that helps in simulation of the social behaviors. This optimization technology is reated with the birds flocking, fish schooling and swarming theory basically. It is very effective for the models where this a wide range of functions. A global swarm algorithm uses mutliple individual particles that helps in exploring the search space for the optimal solution. PSO algorithm utilizes the overall best solution and a particles inertia to determine how to move Each particle about the search space [28].

A solution as a set of n-coordinates is defined in which each one corresponds to a cluster centroid of c-dimension. In the problem of PSO Clustering it follows that we can have more than one possible solution, in which every n solution consists of c-dimensional cluster positions, i.e., cluster centroids. In PSO clustering there are chances that one or more solution is present but it is important to notice that the algorithm can be used in any dimensional space irrespective of the dimension space that is taken as the input. The aim of the proposed algorithm is then to find the best evaluation of a given fitness function or, in our case, the best spatial configuration of centroids. Since each particle represents a position in the Nd space, the aim is then to adjust its position according to

- the particle's best position found so far, and
- the best position in the neighborhood of that particle.

To fulfill the previous statements, each particle stores these values:

- xi , its current position
- vi , its current velocity
- yi , its best position, found so far.

Using the above notation, a particle's position is adjusted according to:

$$\boxed{v_{i,k}(t+1) = wv_{i,k}(t) + c_1r_{1,k}(t)(y_{i,k}(t) - x_{i,k}(t)) + c_2r_{2,k}(t)(y(t) - x_{i,k}(t)}$$

………………………..(5)

$$\boxed{x_i(t+1) = x_i(t) + v_i(t+1)a}$$

…………………..……(6)

In Equation (5) w is called the inertia weight, c1 and c2 are the acceleration constants, and both r1,j (t) and r2,j (t) are sampled from an uniform distribution U(0, 1). The velocity of the particle is then calculated using the contributions of

(1) the previous velocity,

(2) a cognitive component related to its best-achieved distance, and

(3) the social component which takes into account the best achieved distance over all the particles in the swarm.

The best position of a particle is calculated using the trivial Equation, which simply updates the best position if the fitness value in the current i-timestep is less than the previous fitness value of the particle.

Before closing this section we need to introduce how to evaluate the PSO performance at each time step, i.e., a descriptive measure of the fitness of the whole particle set. Equation (7) implements this measure, where |Ci,j | is the number of data vectors belonging to cluster Cij , z(p) is the vector of the input data belonging the Cij cluster, m(j) is the j-th centroid of the i-th particle in cluster Cij , Nc is the number of clusters, and it can be described as follows.

$$\boxed{J_c = \frac{\sum_{j=1}^{N_C}[\sum_{\forall Z \in Ci_j} d(z_p, m_j) / |C_{i,j}|]}{N_c}}$$

………………….(7)

The particles parameter represents how many parallel swarms should be executed at the same time. Recall that each swarm, called also particle, represents a complete solution of the problem, i.e., in the case of two centroids within a two-dimensional space, a couple two coordinates that localize the centroids. The dataset subset parameter allows to resize the original 5-dimensional Covid dataset to the specified value, allowing a 2D or 3D visualization.

w = 0.72; %INERTIA
c1 = 1.49; %COGNITIVE
c2 = 1.49; %SOCIAL

The local fitness is then defined as the mean of all the distances between the points belonging to each centroid.

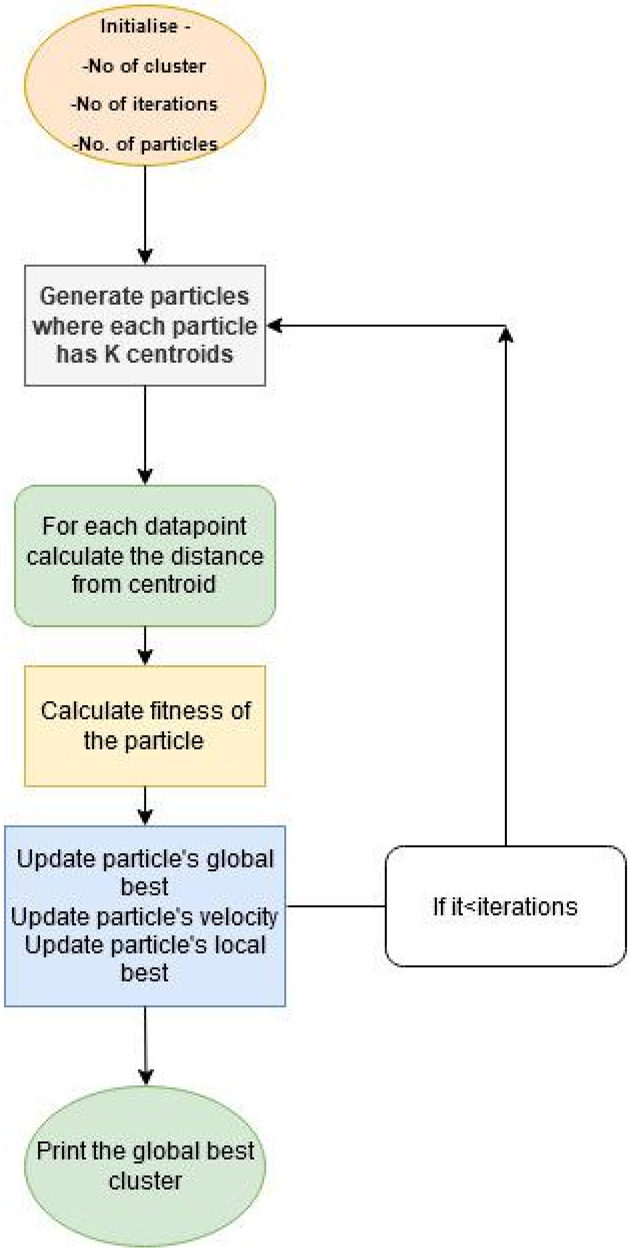


Fig 2. Flowchart of PSO Clustering

3.3 Harris Hawk optimization clustering

The behavior characteristic of Harris hawks is that they trace, encircle, approach, and finally attack the potential prey (rabbits in most cases) by means of good teamwork. A skillful manoeuvring called "surprise pounce" will be effectively carried out in hunting escaping preys. The concrete implementing process of surprise pounce is: team members make active attacks from different directions respectively and then converge to the intended rabbit. Similar to other meta-heuristic algorithms, the HHO algorithm also contains the phases of exploration and exploitation. The hawks will perch randomly on some locations for monitoring various regions, so as to track and detect the rabbit during the exploration stage. Whereas the hawks will conduct surprise pounce or team rapid dives to exploit the neighborhood of intended prey in the stage of exploitation. The positions of hawks are considered as candidate solutions. The best position of is defined as the intended rabbit location.

The 20 hawks are chosen randomly as centroids and then their fitness value is calculated according to the fitness function [19], [20].

A. Exploration Phase

In this phase, Harris hawks update their position through two tactics, and both have the equal probability to be chosen. Which can be described in detail as follows: A random value of $p<0.5$ means the hawks perch on some locations according to the position of other team members, and the position of each hawk is updated by eqn no 8 . In this way, all members can ensure to be close enough when attacking the intended prey. On the other hand, a random value of $p\geq 0.5$ indicates that the hawks perch on giant trees randomly to explore the desert site, and the population using the eqn no 9 to update positions.

$$X(t+1)=(X_{prey}(t)-X_m(t))-r_3(LB+r_4\Delta B)\ldots\ldots\ldots\ldots\ldots\ldots(8)$$
$$X(t+1)=X_{rand}(t)-r_1|X_{rand}(t)-2r_2X(t)|\ldots\ldots\ldots\ldots\ldots\ldots..(9)$$

where $X(t+1)$ is the position vector of hawks in the next iteration. $X_{prey}(t)$ represents the position of intended rabbit. $X_{rand}(t)$ is the position of a hawk which is chosen randomly from current team. r_1,r_2,r_3 and r_4 are random numbers, which can provide more diversification trends and make sure the hawks explore different regions of the search space. $\Delta B=UB-LB$, UB and LB are the upper and lower bounds of search space. t is the current iteration counter. And $X_m(t)$ is the average position of the current population of hawks which is calculated by the following equation.

$$X_m(t)=\sum_{i=1}^{N} X_i(t)\ X_i(t)\ldots\ldots\ldots(10)$$

B. Transition From Exploration to Exploitation

As the intended prey try to run away from the attack, the retained energy of prey constantly decreases, which can be modeled as follows:

$$E=2E_0(1-tT)\ldots\ldots\ldots\ldots\ldots(11)$$

where E_0 ranged from -1 to 1 denotes the energy of initial state. Note that, the intended prey is physically flagging in the case of $E_0\in[-1,0]$, whilst when the value of $E_0\in[0,1]$, it means that the intended prey is strengthening. t is the current iteration counter. And T represents the max iteration.

Different values of E establish the basis for a transition from exploration to exploitation smoothly, and determine the unique exploitative behaviors in the process of chasing intended prey. The hawks search the promising region in the case of $|E|\geq 1$, which is also known as exploration stage. On the contrary, when the value of escaping energy $|E|<1$, the hawks are in the step of exploitation.

C. Exploitation Phase

When the hawks carry out "surprise pounce" strategy, the intended rabbit will try to rush to the safety instinctively. Hence, the exploitation phase is consisted of four models with respect to the escaping behaviors and chasing tactics of the hawks. Assume that r is a random number ranged from 0 to 1, where if $r<0.5$ then the rabbit successfully escapes from dangerous situations; otherwise, the result is failure of escape. And the retained escaping energy $|E|$ is utilized to determine that the besiege is soft or hard.

1) Soft Besiege

Although the rabbit has enough energy, but it does not succeed in escaping from attack due to some random misleading jumps in the case of $r\geq 0.5$ and $|E|\geq 0.5$. Moreover the hawks encircle the rabbit from different directions softly to make it more exhausted, and then conduct the surprise pounce. The behavior of hawks is modeled as follows:

$$X(t+1)=\Delta X(t)=\Delta X(t)-E|JX_{prey}(t)-X(t)|\ldots\ldots\ldots\ldots(12)$$

$$\Delta X(t)=X_{prey}(t)-X(t)\ldots\ldots\ldots\ldots(13)$$

where $\Delta X(t)$ defines the gap between the location of intended rabbit and the position of current hawk in iteration t . r_5 is a random number ranged from [0, 1], and $J=2(1-r_5)$, denotes the random jump strength of intended rabbit in the process of escaping, which can mimic the natural motions of rabbit by virtue of random change.

2) Hard Besiege

The rabbit is very exhausted, as well as has a low escaping energy in the case of $r\geq 0.5$ and $|E|<0.5$. Therefore, the hawks pay almost no effort to encircle intended rabbit before the surprise pounce performed. Each hawk updates its current position using the following equation.

$$X(t+1)=X_{prey}(t)-E|\Delta X(t)|\ldots\ldots\ldots\ldots(14)$$

3) Soft Besiege With Progressive Rapid Dives

The intended rabbit has enough energy to escape from attack, and the hawks still construct a soft besiege in the case of $r<0.5$ and $|E|\geq 0.5$. In addition, the levy flight, an optimal searching tactic for predators in non-destructive foraging conditions, is utilized to model the escaping patterns of rabbit and leapfrog movements of hawks mathematically and accurately in this situation. According to the real behaviors of hawks, assume that hawks can progressively select the best possible dive toward the intended prey. In another word, the hawks compare the possible result of each move to detect that will is be a good dive or not, and then implement the following rules correspondingly. To be more specific, the position of hawk is updated by eqn (15) if next position is better than the current. Otherwise, the hawks will perform team rapid dives based on levy flight which can enhance exploitation capacity using eqn (16).

$$Y=Z=X_{prey}(t)-E|JX_{prey}(t)-X(t)|\ldots\ldots\ldots\ldots(15)$$
$$Z=Y+S\times LF(D)\ldots\ldots\ldots\ldots(16)$$

where D denotes the dimension of search space. S represents random selected vector which are sized at $1\times D$.

4) Hard Besiege With Progressive Rapid Dives

The intended rabbit has too low energy to escape in the case of $r<0.5$ and $|E|<0.5$, and the hawks perform a hard besiege at the same time. The strategy for updating the positions of hawks is similar to that in soft besiege with progressive rapid dives. Note that, the team members try to shrink the distance between their average location and the location of intended rabbit in this situation.

$$Y=X_{prey}(t)-E|JX_{prey}(t)-X_m(t)|\ldots\ldots\ldots\ldots(17)$$

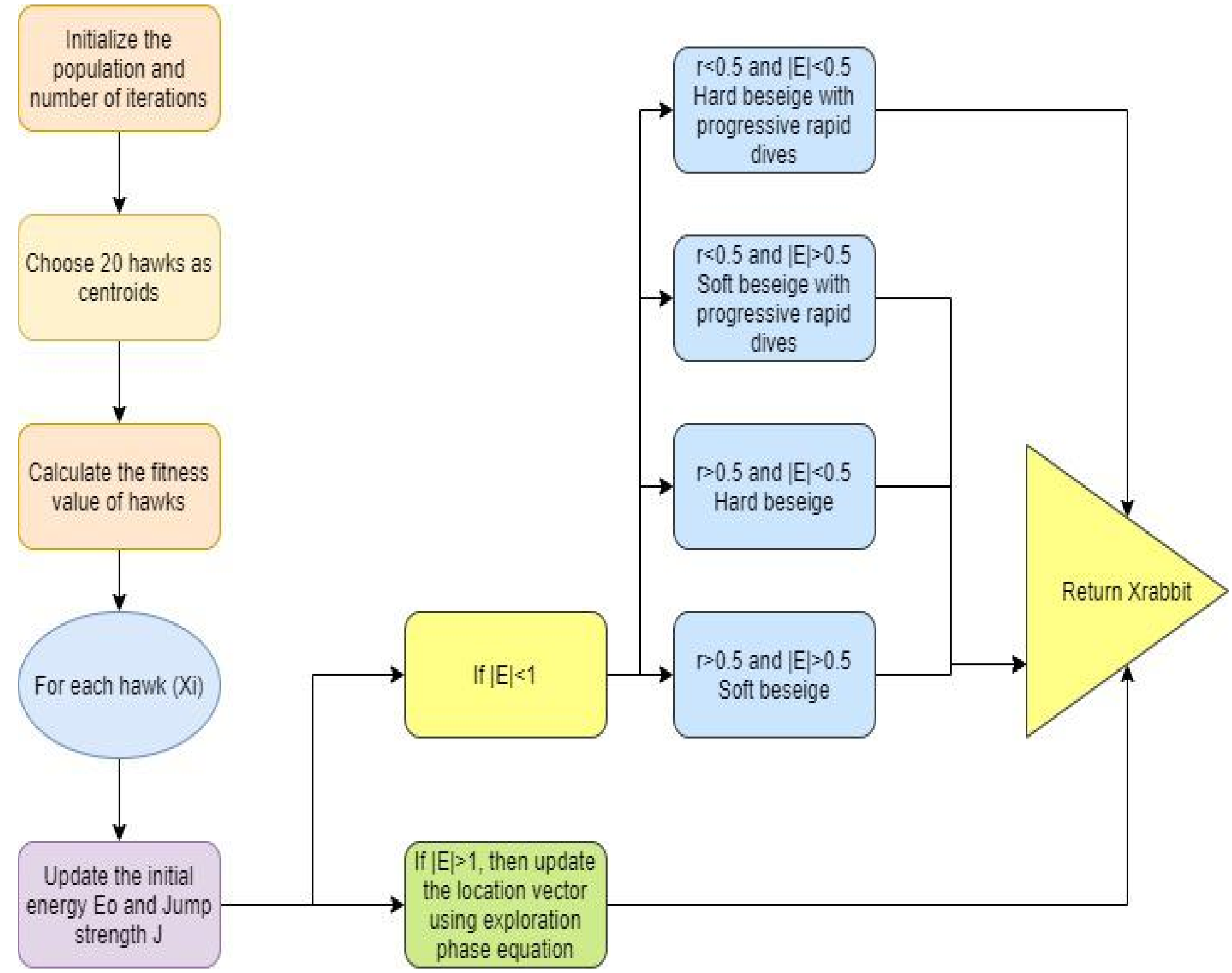


Fig 3. Harris Hawk clustering flowchart

3.3 GREY WOLF OPTIMIZATION ALGORITHM

The Grey Wolf Optimizer (GWO) algorithm is a new bio inspired metaheuristic that has been introduced in 2014 . It is inspired by both the social hierarchy of wolves, as well as their hunting behaviour. In GWO, the search starts by a population of randomly generated wolves (solutions). During hunting (optimization), these wolves estimate the prey's (optimum) location through an iterative procedure. In order to formulate the hierarchy of wolves, the fittest solution is referred to as the alpha . While the second and third best solutions are called beta and delta , respectively. The wolves alpha,beta , and delta are responsible for guiding the search (hunting), while other wolves follow [21]. The hunting behaviour is mainly divided into three steps: tracking, encircling and attacking the prey. The encircling behaviour is mathematically modelled according to the following equations:

$$\vec{D} = |\vec{C}.\vec{X}_P(t) - \vec{X}(t)| \quad \ldots\ldots\ldots\ldots\ldots(18)$$

$$\vec{X}(t+1) = |\vec{X}_P(t) - \vec{A}.\vec{D}| \quad \ldots\ldots\ldots\ldots\ldots(19)$$

Where D represents the distance between the position vector of both the prey X_p and a wolf X, and t represents the current iteration number. A and C are coefficient vectors, and are calculated as follows:

$$\vec{A} = 2\vec{a}.\vec{r}_1 - \vec{a} \quad \ldots\ldots\ldots\ldots\ldots(20)$$

$$\vec{C} = 2\vec{r}_2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(21)$$

Where r_1 and r_2 are random vectors in [0,1] and the value of a is linearly decreased from 2 to 0 as iterations proceed.
According to (18) and (19), a wolf can randomly update its position in the area around the prey.

In order to model how the three best solutions, and guide the other search agents, the following formulaes are used:

$$\vec{D}_\alpha = |\vec{C}_1.\vec{X}_\alpha - \vec{X}|.\vec{D}_\beta = |\vec{C}_2.\vec{X}_\beta - \vec{X}|.\vec{D}_\delta = |\vec{C}_3.\vec{X}_\delta - \vec{X}| \quad \ldots\ldots\ldots\ldots(22)$$

$$\vec{X}_1 = \vec{X}_\alpha - \vec{A}_\alpha.(\vec{D}_\alpha), \vec{X}_2 = \vec{X}_\beta - \vec{A}_\beta.(\vec{D}_\beta), \vec{X}_3 = \vec{X}_\delta - \vec{A}_\delta.(\vec{D}_\delta) \quad \ldots\ldots\ldots(23)$$

$$\vec{X}(t+1) = \frac{\vec{X}1+\vec{X}2+\vec{X}3}{3} \quad \ldots\ldots\ldots\ldots\ldots(24)$$

From (22), (23), (24) it can be seen that the position of X of any wolf is determined by the position of the three best solutions.
The hunting ends by attacking (approaching) the prey, this step represents the exploitation phase. It is performed by decreasing the value of a, linearly from 2 to 0. This in turn decreases the value of A. To avoid local stagnation, random values of A greater than 1 are employed to force the wolf away from the prey. This step represents the exploration phase. Using adaptive values of a and consequently A, which is depicted by linearly decreasing the values of afrom 2 to 0, guarantees a balance between exploration and exploitation. This balance is achieved since half of the iterations are dedicated to exploration when |A| • 1, while the rest of the iterations is dedicated for exploitation when |A| >1. This balance is one of the GWO's strengths. To sum up, values of |A| <1 oblige the search agents to move towards the prey while values of |A| >1 oblige them to diverge from it. The value of C represents another component that influences exploration, where C belongs to [0,2]. From (18) we can conclude that C represents the weight of the prey in defining the distance; values of C>1 emphasize its effect while C < 1 reduce it.

The K-means clustering algorithm assigns points to k clusters based on their proximity to the cluster's centroid. Initially centroids are selected at random and then iteratively reallocated until a predetermined stopping criterion is met. The K-means algorithm may be divided into an initialization phase, a cluster assignment and centroids update phase, and finally an exploration and evaluation phase. K-means algorithm suffers from two main drawbacks, namely: the dependence on initial centroid values, and its tendency to fall into local optima. In our proposed K-GWO algorithm, wolves represent solutions. Each wolf holds a set of K centroids that correspond to K clusters. Each centroid is a D dimensional vector. Consequently, each wolf is represented by a (K X D) vector. A population of N wolves collaboratively hunts for the best possible configuration of the clusters (prey). The best configuration of clusters is reflected by the optimal positions of the centroids. These centroids are subsequently used to grow clusters following the basic K-means principle: points are assigned to the cluster with nearest centroid. The algorithm aims at minimizing an objective function defined as:

$$F = \sum_{j=1}^{k} \sum_{i=1}^{N} \| x_{i,j} - cen_j \|^2 \quad \text{……………..(25)}$$

Where, k is the number of clusters, N is the number of points to be clustered, $x_{i,j}$ is the i^{th} data point belonging to the j^{th} cluster, and cen_j is the centroid of cluster j . As mentioned earlier, each search agent represents a set of k centroids (cen_1, $cen_{2....}cen_n$) which, when used in equation (20), gives an indication on how fit this agent is. The fittest search agent is the one associated with the minimum value of [22].

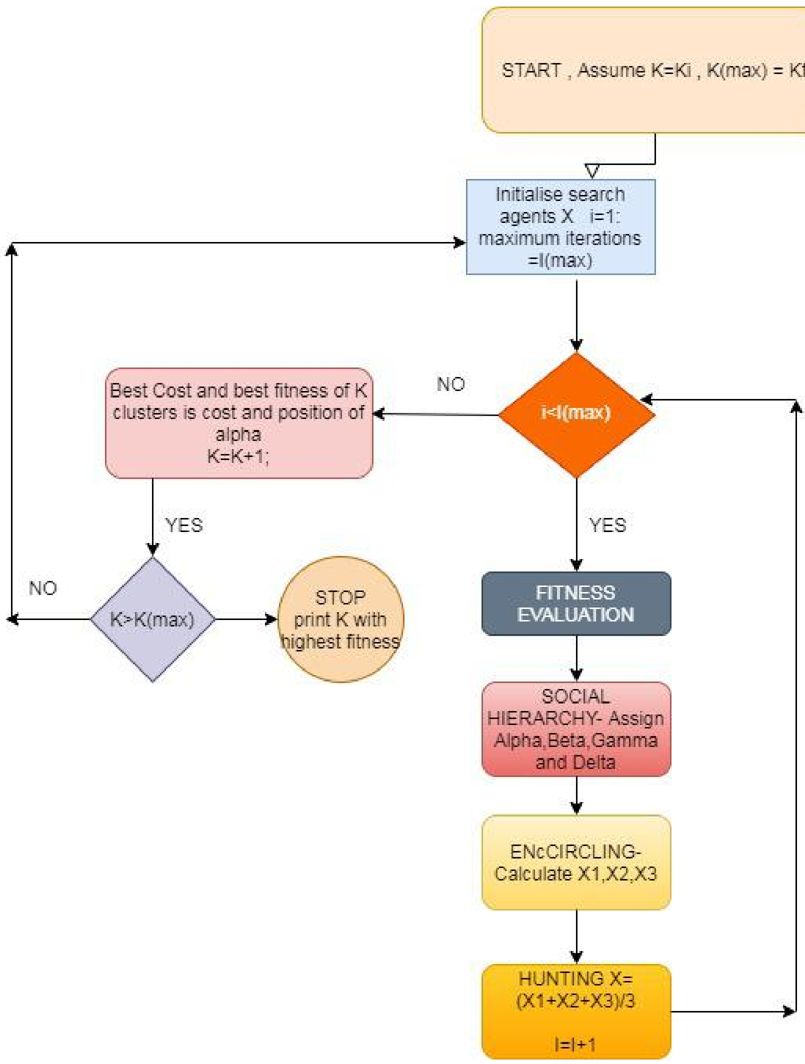


Fig 4. GWO Flowchart

Chapter – 4 Dataset and Libraries used

Dataset used is the original dataset of covid-19 of different countries is mentioned in [31]:

Libraries Used:

Seaborn, Pandas,

Numpy,

Matplotlib

Chapter- 5

Testing

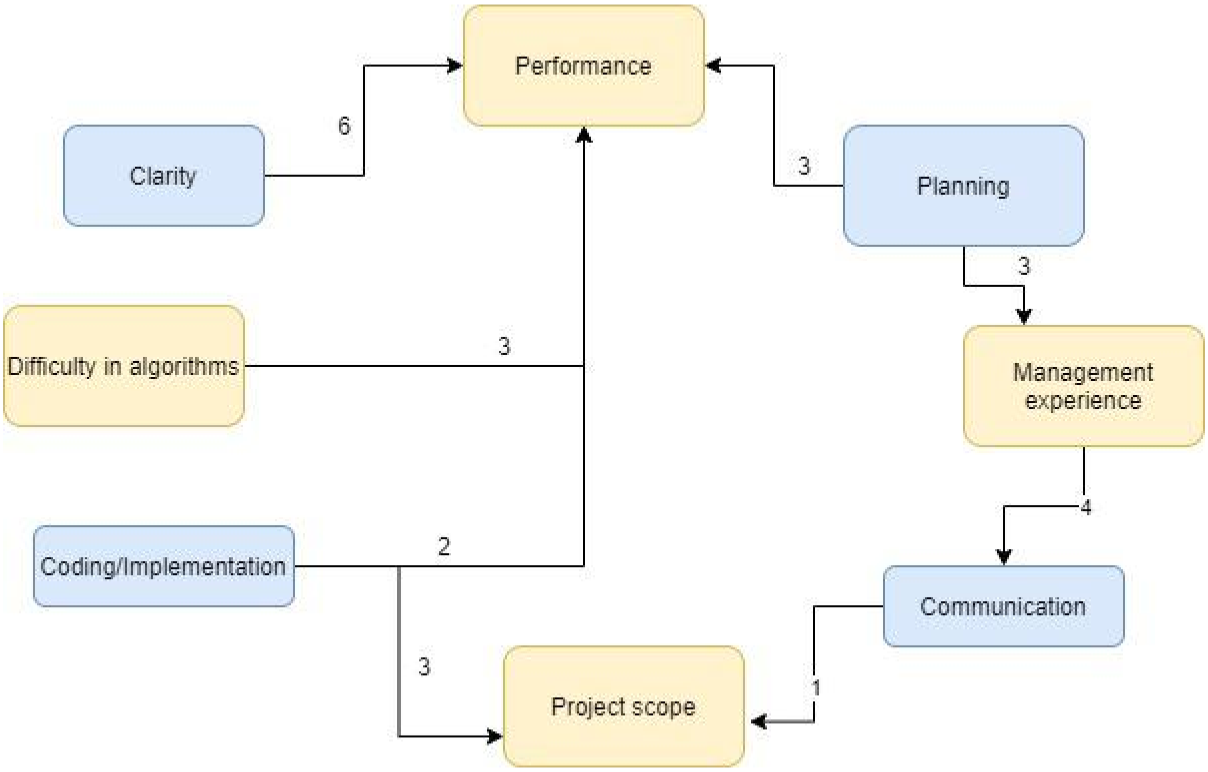
5.1 Testing plan

S no.	List of various components that require testing	Types of testing required	Techniques for writing test cases
1	Genetic algorithm clustering	Performance	White box testing
2	Particle swarm optimization clustering	Performance	White box testing
3	Harris hawk clustering	Performance	White box testing
4	Grey wolf optimization clustering	Performance	White box testing

5.2 Test cases in described format

Test case id	input	Expected output	result
1	43,39,4	A cluster is assigned	pass
2	56,46,11	A cluster is assigned	pass
3	23,0,0	A cluster is assigned	pass
4	12,6,0	a cluster is assigned	fail

5.3 Error and exception handling



Risk Area	Weights	Total weights	Priority
Performance	6+3+2+3	14	1
Clarity	6	6	2
Planning	3	3	3
Management experience	3	3	4
Coding/Implementation	2	2	5
Communication	4	4	6
Difficulty in algorithm	3	3	7
Project scope	3+1	4	8

Chapter - 6

Results and Analysis

In this study, we used clustering analysis to classify countries on the basis of confirmed, dead and recovered cases. Genetic, PSO, Harris Hawk optimization and Grey wolf optimization clustering was used to determine the severity of countries in context to state of corona cases. The cluster analysis grouped into 20 clusters depending on the state of severity in various states. The severe country cluster needs more medical facilities (ventilators, testing kits, masks etc), treatment etc to reduce number of deceased persons. The clusters we have obtained are based on where the COVID-19 condition
is more worse. This analysis can further be applied to one country also, as to maximize the distribution of resources in the most efficient way.

Result analysis

The centroids are plotted on a graph and different clusters are represented with different colors. In each clustering, 20 clusters have been formed. Each cluster is based on the severity of corona cases. The countries with maximum number of deaths have been clustered in one cluster and one with low number of deaths in another. One of the figures represent Genetic clustering and the other one represents PSO clustering. By this information of clustering, we can analyze which areas need more medical facilities and we can handle the situation in a a better way.

1)The Silhouette Coefficient (sklearn.metrics.silhouette_score) is an example of such an evaluation, where a higher Silhouette Coefficient score relates to a model with better defined clusters. The Silhouette Coefficient is defined for each sample and is composed of two scores:

- **a**: The mean distance between a sample and all other points in the same class.
- **b**: The mean distance between a sample and all other points in the *next nearest cluster*.

The Silhouette Coefficient *s* for a single sample is then given as:

$$S= \frac{b-a}{\max(b-a)} \text{……………..(26)}$$

2) Davies-Bouldin index
A lower relates to a model with better separation between the clusters.

This index signifies the average ‘similarity’ between clusters, where the similarity is a measure that compares the distance between clusters with the size of the clusters themselves.

The index is defined as the average similarity between each cluster C_i for i=1,...,k and its most similar one C_j. In the context of this index, similarity is defined as a measure R_{ij} that trades off:

- s_i, the average distance between each point of cluster i and the centroid of that cluster – also know as cluster diameter.
- d_{ij}, the distance between cluster centroids i and j.

$$R_{ij}=s_i+s_j/d_{ij} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(27)$$

$$DB=1/k \sum_{I=1}^{K} \max\,(Rij) \quad \ldots\ldots\ldots\ldots\ldots..(28)$$

Algorithm	Silhouette coefficient	Davis-Bouldin index
1)Genetic algorithm	0.782	0.479
2) Particle swarm optimization	0.786	0.478
3) Harris hawk	**0.813**	**0.345**
4) Grey wolf optimization	0.798	0.453

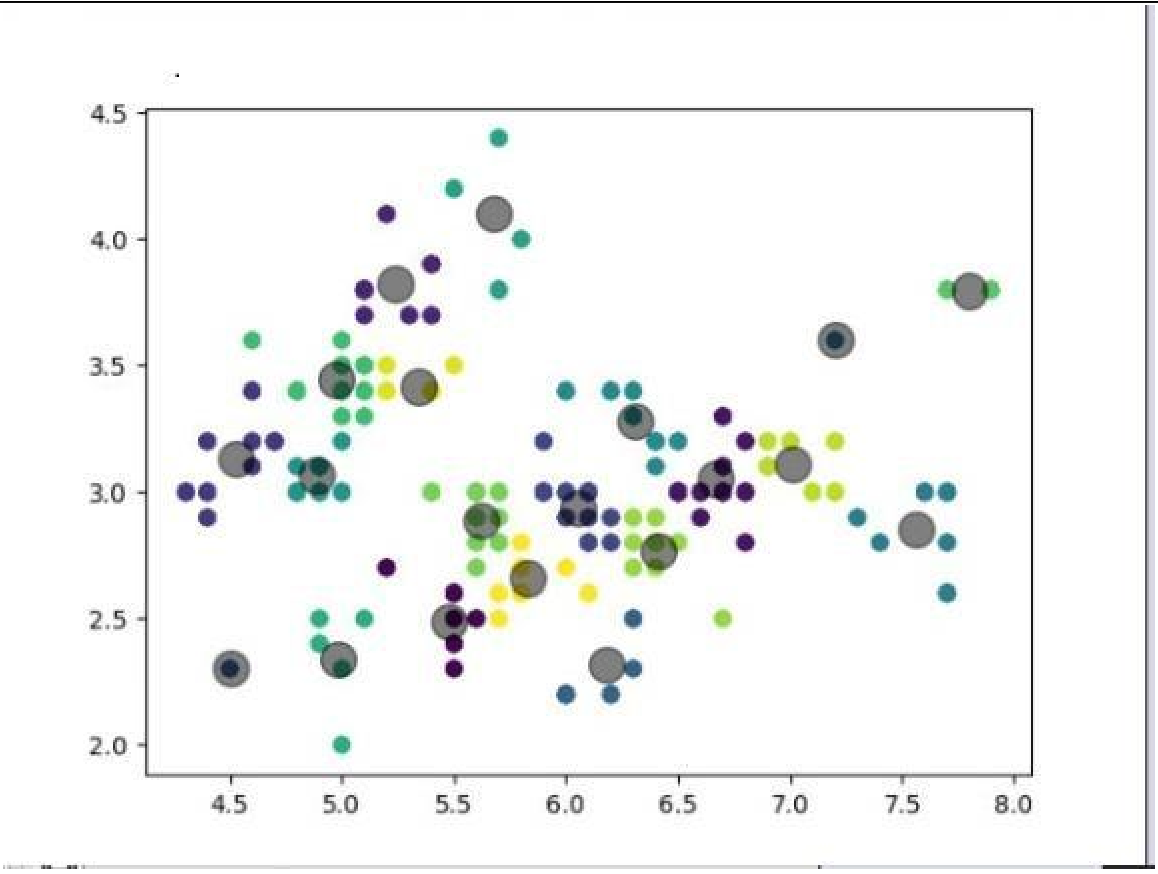


Fig. 5 PSO CLUSTERING

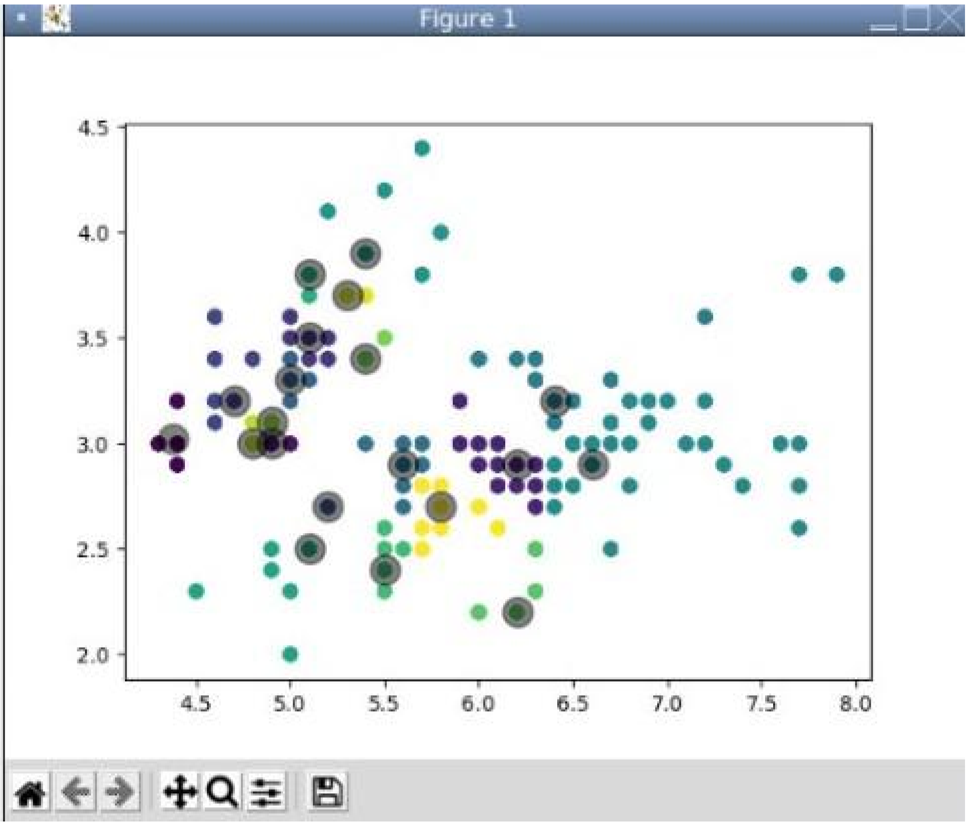


Fig.6 GA CLUSTERING

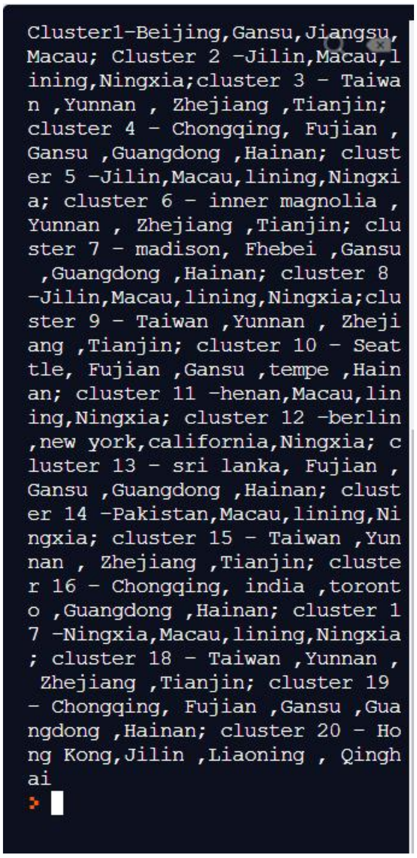


Fig. 8 Best Clusters Formed using GA clustering

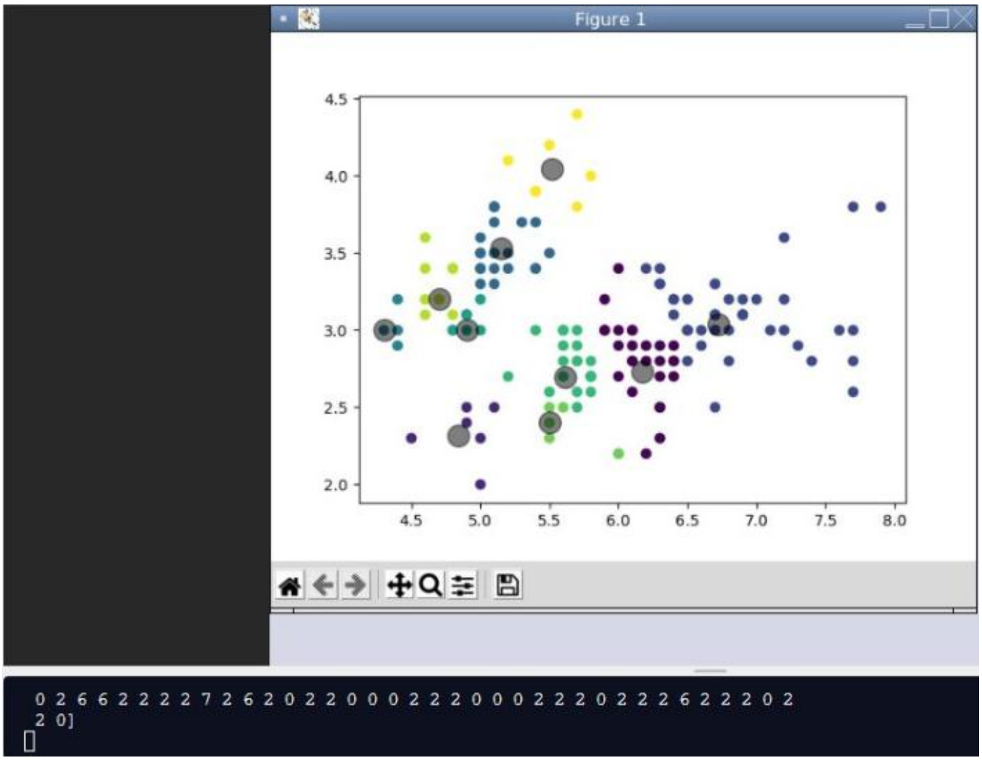


Fig. 9 GWO Clustering

```
Cluster1 Tibet,Washington,Gansu,Ningxia Cluster2-Macau,Sichuan,S
handong Cluster3-Beijing,Gansu,Jiangsu,Hainan,Chongqing,Shandong
 Cluster4-Zhejiang,Ningxia,Taiwan Cluster5-Beijing,Chongqing,Gui
zhou Cluster6-Zhejiang,Qinghai,Jilin Cluster7-Hong Kong,Inner Mo
ngolia,Jilin Cluster8-Sichuan,Taiwan,Chongqing Cluster9-Illinois
,Chicago,Beijing,Chongqing Cluster10-Hainan,Heilongjiang Cluster
11-Tianjin,Chongqing,Shaanxi Cluster12-Illinois,Anhui,Inner Mong
olia Cluster13-Taiwan,Shandong,Gansu Cluster14-Shaanxi,Chongqing
,Heilongjiang,Macau,Tibet Cluster15-Zhejiang,Hong Kong,Anhui,Bei
jing, Cluster16-Shaanxi,Heilongjiang,Sichuan Cluster17-Jilin,Inn
er Mongolia, Cluster18-Macau,Washington,Jiangsu Cluster19-Anhui,
Chongqing,Jiangsu Cluster20-Hainan,Chongqing,Shandong,Macau
>
```

Fig 10. Best Clusters formed using GWO Clustering

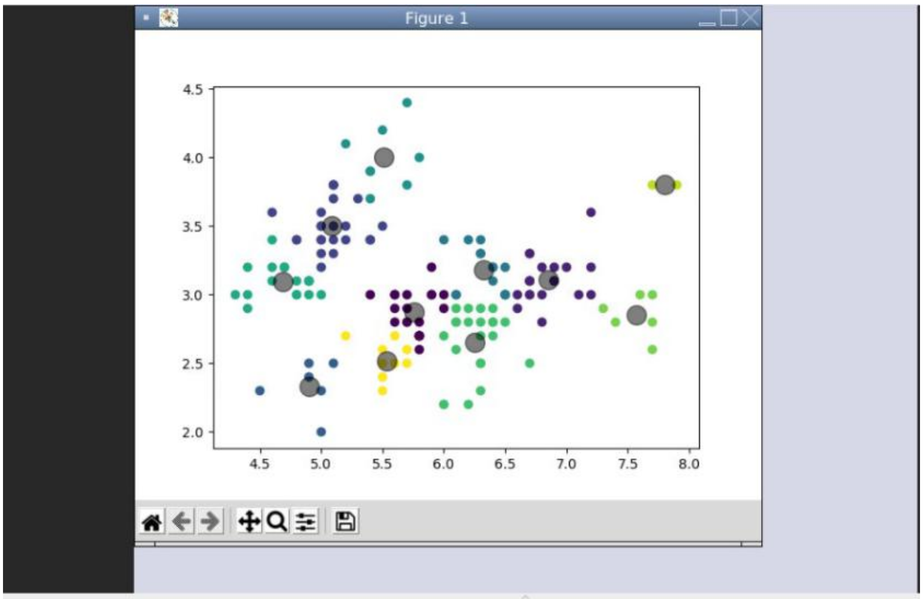


Fig. 11 Harris Hawk Optimization Clustering

```
Cluster1-Zhejiang,Hong Kong,Anhui,Beijing,Ningxia Cluster2-Macau
,Sichuan,Shandong Cluster3-Hainan,Chongqing,Shandong Cluster4-Zh
ejiang,Ningxia,Taiwan Cluster5-Beijing,Chongqing,Guizhou Cluster
6-Zhejiang,Qinghai,Jilin Cluster7-Hong Kong,Inner Mongolia,Jilin
 Cluster8-Sichuan,Taiwan,Chongqing Cluster9-Illinois,Chicago,Bei
jing,Chongqing Cluster10-Hainan,Heilongjiang Cluster11-Tianjin,C
hongqing,Shaanxi Cluster12-Illinois,Anhui,Inner Mongolia Cluster
13-Taiwan,Shandong,Gansu Cluster14-Shaanxi,Chongqing,Heilongjian
g,Macau,Tibet Cluster15-Tibet,Washington,Gansu Cluster16-Shaanxi
,Heilongjiang,Sichuan Cluster17-Jilin,Inner Mongolia, Cluster18-
Macau,Washington,Jiangsu Cluster19-Anhui,Chongqing,Jiangsu Clust
er20-Beijing,Gansu,Jiangsu,Macau
>
```

Fig. 12 Best Clusters formed using Harris Hawk Clustering

Chapter 7

Conclusion

In this study, we used clustering analysis to classify countries on the basis of dead and recovered cases. Genetic, PSO, Harris Hawk and Grey-wolf optimization clustering was used to determine the severity of countries in context to state of corona cases. The cluster analysis grouped into 20 clusters depending on the state of severity in various states. The severe country cluster needs more medical facilities (ventilators, testing kits, masks etc), treatment etc to reduce number of deceased persons. The clusters we have obtained are based on where the covid condition is more worse.

References

1. Akay, Özlem & Tekeli, Erkut & Yuksel, Guzin. (2020). Genetic Algorithm with New Fitness Function for Clustering. Iranian journal of science and technology. transaction a, science. 1-10. 10.1007/s40995-020-00890-8.
2. Yang, Xuemei & Dong, Ning & Chan, Edward & Chen, Sheng. (2020). Genetic cluster analysis of SARS-CoV-2 and the identification of those responsible for the major outbreaks in various countries. Emerging Microbes & Infections. 9. 1287-1299. 10.1080/22221751.2020.1773745.
3. Azarafza, Mehdi & Azarafza, Mohammad & Akgun, Haluk. (2020). Clustering method for spread pattern analysis of corona-virus (COVID-19) infection in Iran. 10.1101/2020.05.22.20109942.
4. E. Hosseini, K. Z. Ghafoor, A. S. Sadiq, M. Guizani and A. Emrouznejad, "COVID-19 Optimizer Algorithm, Modeling and Controlling of Coronavirus Distribution Process," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 10, pp. 2765-2775, Oct. 2020, doi: 10.1109/JBHI.2020.3012487.
5. Doha, Akad & Markman, Ofer & Jared, Lefkort. (2020). Clustering of Country-Based Data in COVID-19 Infections By Coronavirus outbreak features -First wave (data up to date 28/5). 10.13140/RG.2.2.28070.34886.
6. Pang, Chao-Yang & Hu, Ben-Qiong & Zhang, Jie & Hu, Wei & Shan, Zheng-Chao. (2014). Applying Data Clustering Feature to Speed Up Ant Colony Optimization. Abstract and Applied Analysis. 2014. 1-8. 10.1155/2014/545391.
7. Gao, Wei. (2016). Improved Ant Colony Clustering Algorithm and Its Performance Study. Computational Intelligence and Neuroscience. 2016. 1-14. 10.1155/2016/4835932.
8. Gao, Mingliang & Shen, Jin & Jiang, Jun. (2017). Visual tracking using improved flower pollination algorithm. Optik - International Journal for Light and Electron Optics. 156. 10.1016/j.ijleo.2017.11.155.
9. Merwe, D. & Engelbrecht, Andries. (2003). Data clustering using particle swarm optimization[C]. Proc of 2003 Congress on Evolutionary Computation (CEC'03). 1. 215-220. 10.1109/CEC.2003.1299577.
10. Gao, Mingliang & Shen, Jin & Jiang, Jun. (2017). Visual tracking using improved flower pollination algorithm. Optik - International Journal for Light and Electron Optics. 156. 10.1016/j.ijleo.2017.11.155.
11. Alexandra & Berthelot, Philippe & Saura, Christine. (2020). Cluster of coronavirus disease 2019 (Covid-19) in the French Alps, 2020. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 71. 10.1093/cid/ciaa424.
12. Kumar S. (2020). Monitoring Novel Corona Virus (COVID-19) Infections in India by Cluster Analysis. *Annals of Data Science*, 1–9. Advance online publication. https://doi.org/10.1007/s40745-020-00289-7.
13. Zhang, Sen & Zhou, Yong-Quan. (2015). Grey Wolf Optimizer Based on Powell Local Optimization Method for Clustering Analysis. Discrete Dynamics in Nature and Society. 2015. 1-17. 10.1155/2015/481360.
14. Hu, Pei & Pan, Jeng-Shyang & Chu, Shu-Chuan. (2020). Improved Binary Grey Wolf Optimizer and Its application for feature selection. Knowledge-Based Systems. 195. 105746. 10.1016/j.knosys.2020.105746.

15. Sonbhadra, Sanjay & Agarwal, Sonali & Nagabhushan, P.. (2020). Target specific mining of COVID-19 scholarly articles using one-class approach. Chaos, Solitons & Fractals. 140. 110155. 10.1016/j.chaos.2020.110155.
16. X. Bao, H. Jia and C. Lang, "A Novel Hybrid Harris Hawks Optimization for Color Image Multilevel Thresholding Segmentation," in IEEE Access, vol. 7, pp. 76529-76546, 2019, doi: 10.1109/ACCESS.2019.2921545.
17. Singh, Tribhuvan. (2020). A chaotic sequence-guided Harris hawks optimizer for data clustering. Neural Computing and Applications. 32. 10.1007/s00521-020-04951-2.
18. Heidari, Ali Asghar & Mirjalili, Seyedali & Faris, Hossam & Aljarah, Ibrahim & Mafarja, Majdi & Chen, Huiling. (2019). Harris hawks optimization: Algorithm and applications. Future Generation Computer Systems. 97 aliasgharheidari.com. 849-872. 10.1016/j.future.2019.02.028.
19. Moayedi, Hossein & Osouli, Abdolreza & Nguyen, Hoang & A Rashid, Ahmad Safuan. (2019). A novel Harris hawks' optimization and k-fold cross-validation predicting slope stability. Engineering With Computers. 35. 1-11. 10.1007/s00366-019-00828-8.
20. Zhang, Yang & Zhou, Xizhao & Shih, Po-Chou. (2020). Modified Harris Hawks Optimization Algorithm for Global Optimization Problems. Arabian Journal for Science and Engineering. 45. 1-26. 10.1007/s13369-020-04896-7.
21. Inan, Onur & Karakoyun, Murat & Akto, ihtisam. (2019). Grey Wolf Optimizer (GWO) Algorithm to Solve the Partitional Clustering Problem. International Journal of Intelligent Systems and Applications in Engineering. 7. 201-206. 10.18201/ijisae.2019457231.
22. Khan, Muhammad & Aadil, Farhan & Rehman, Zahoor & Khan, Salabat & Shah, Dr. Peer Azmat & Muhammad, Khan & Lloret, Jaime & Wang, Haoxiang & Lee, Jong & Mehmood, Irfan. (2018). Grey wolf optimization based clustering algorithm for vehicular ad-hoc networks. Computers & Electrical Engineering. 70. 10.1016/j.compeleceng.2018.01.002.
23. Aljarah, Ibrahim & Mafarja, Majdi & Heidari, Ali Asghar & Faris, Hossam & Mirjalili, Seyedali. (2020). Clustering analysis using a novel locality-informed grey wolf-inspired clustering approach. Knowledge and Information Systems. 62. 10.1007/s10115-019-01358-x.
24. Purushothaman, R. & Rajagopalan, S.P. & Dhandapani, Gopinath. (2020). Hybridizing Grey Wolf Optimization (GWO) with Grasshopper Optimization Algorithm (GOA) for text feature selection and clustering. Applied Soft Computing. 96. 106651. 10.1016/j.asoc.2020.106651.
25. Jimenez, Julio & Cuevas, Francisco & Carpio Valadez, Juan. (2007). Genetic algorithms applied to clustering problem and data mining.
26. Maulik, Ujjwal & Bandyopadhyay, Sanghamitra. (2000). Genetic Algorithm-Based Clustering Technique. Pattern Recognition. 33. 1455-1465. 10.1016/S0031-3203(99)00137-5.
27. Nguyen, Hung & Louis, Sushil & Nguyen, Tin. (2019). MGKA: A genetic algorithm-based clustering technique for genomic data. 103-110. 10.1109/CEC.2019.8790225.
28. C. Chen and F. Ye, "Particle swarm optimization algorithm and its application to clustering analysis," 2012 Proceedings of 17th Conference on Electrical Power Distribution, Tehran, 2012, pp. 789-794.
29. S. J. Nanda, M. Sharma and A. Panda, "Clustering Big Datasets Using Orthogonal Gray Wolf Optimizer," 2019 International Conference on Information Technology (ICIT), Bhubaneswar, India, 2019, pp. 353-358, doi: 10.1109/ICIT48102.2019.00069.
30. Jaime & Wang, Haoxiang & Lee, Jong & Mehmood, Irfan. (2018). Grey wolf optimization based clustering algorithm for vehicular ad-hoc networks. Computers & Electrical Engineering. 70. 10.1016/j.compeleceng.2018.01.002.

31. https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=covid_19_data.csv