Leveraging Chicken Swarm Algorithm for Feature Selection Optimization Targeting Efficient Patients Backlog Elimination

Emna Bouazizi
University of Jeddah, College of Computing and Information Technology at Khulais, Department of Information Systems, Jeddah, Saudi Arabia
aaltaher@uj.edu.sa

Ayman E. Khedr
University of Jeddah, College of Computing and Information Technology at Khulais, Department of Information Systems, Jeddah, Saudi Arabia
aeelsayed@uj.edu.sa

Amira M. Idrees
Faculty of Computers and Information Technology, Future University in Egypt
Cairo, Egypt
amira.mohamed@fue.edu.eg

Abstract: *The exponential growth of data sources brings the challenge of maintaining the processing performance and reducing computation complexity. One of the vital solutions is the success in preserving the significant attributes. Consequently, this research focuses on proposing an effective method for feature selection which is based on the adaptation of the chicken swarm optimization algorithm. The research focuses on adapting the algorithm strategy in the process of searching the data space from random-based strategy to a more systematic method which ensures raising the algorithm performance. The study proposes a novel search strategy based on applying an effective clustering technique to effectively identify the main algorithm players which consequently enhance the algorithm performance. On the other hand, focusing on business objectives, this research proposes a novel framework that focuses on the patients' backlogs. The study applies the proposed enhancement to eliminate the patients' backlog while maintaining the prioritization. The proposed framework is generic and could be applied to the concept of backlogs in any domain. The experiment succeeded in confirming the applicability of the proposed adaptation for the chicken swarm optimization algorithm and reaching the business goal with a minimum accuracy* percentage *equal to 95.3% for random forest and a maximum of 98.9 for naive bayes*.

Keywords: *Optimization, chicken swarm optimization algorithm, feature selection, logistics, supply chain management, backlogs*.

Received March 10, 2024; accepted November 05, 2024
https://doi.org/10.34028/iajit/22/2/12

1. Introduction

The immense amount of the sources of data generated in a timely manner has directly led to the strong need for efficient presence of highly performance methods [2]. There have been different proposed complex methods seeking for effective data manipulation. However, the challenge of the continuous dimensionality increase has kept the road for more research open seeking higher performance with a timely manner computation. This research focuses on the pre-processing stage as one of the vital aspects for raising the techniques' performance. Efficient feature selection ensures the balance of higher performance with the least computation cost while maintaining the data quality and preserving the hidden knowledge [4]. Feature selection in its genuine definition is the process of removing redundant, misleading, and irrelevant features. Feature selection is visualized as one of the optimization methods in machine learning techniques [22]. Mathematically speaking, having a count of f features representing the dataset, the possible generated count of feature subsets reaches 2 with the power of f. According to the literature, different methods have been proposed for feature selection. Analysis of variance, correlations, and entropy are statistical methods examples; however, this direction depends on the one-to-one direct relationship between features while ignoring the hybrid relationships of a set of features as well as the impact level of these features [3].

More robust intelligent directions have been also tackled in different research [11]. In this research, the heuristics swarm intelligence direction has been on focus. This research focuses on one of the most effective swarm intelligence algorithms in feature selection namely the Chicken Swarm Optimization algorithm (CSO). CSO has proven its effectiveness in many research in identifying the most influencing features. This confidence in CSO has emerged due to its ability for fast convergence by utilizing the least possible factors during the search task [14].

However, this leverage of CSO is heavily affected by the increase in dimensionality. Therefore, leveraging this situation with a higher effective approach to be able to preserve the CSO effectiveness with no negative effect on the data dimensions' granularity has become challenging which is the scope of this research. Although CSO is efficient in feature selection tasks, however, as the dimensions' increase, the convergence

performance starts degrading due to the randomness process in the search space and the low ability for global identification [13].

The proposed enhancement of CSO provides a more systematic method in the search task ensuring reducing the time complexity. The following points represent the main research contributions. This study proposes a novel method for identifying the main players represented in the significant features. The proposed method minimizes the computation effort by moving the most influencing features to prior examination by applying a weighting collaborative methodology which will be described in detail. Moreover, one of the most important aspects of the algorithm is the fitness value which is also adapted. The collaborative methodology is also utilized in fitness value determination with other contributors. A set of machine learning algorithms is planned to contribute to a homogenous collaboration approach with a novel approach for the algorithms' results evaluation which is based on the invariance level targeting the highest accurate fitness value. Moreover, the uneven distribution of the data population is manipulated by setting an initialization setting with a more systematic distribution which enhances the ergodicity as well as the algorithm solidity. Additionally, the proposed weighting methodology is considered a key point to the dynamic adjustment of the players' position which consequently affects the algorithm execution time. The proposed location update methodology is based on a self-adapting machine learning algorithms evaluation method which is one of the key points that enhances the algorithm's accuracy and effectiveness in the exploration process.

The technical novelty of the proposed enhancement could be summarized in the following points:

1) The research proposes an adaptation for the search strategy of CSO based on a more systematic approach rather than the traditional random search approach of CSO.
2) The research proposes an adapted clustering technique for the search space segmentation targeting a parallel approach that was able to minimize the time complexity in identifying the significant algorithm players.
3) The proposed adapted algorithm is able to successfully explore the most significant dataset features while maintaining the performance enhancement and reducing the time complexity due to reducing the players' search time.

On the other hand, from the business aspect, this research focuses on enhancing the healthcare field targeting to eliminate the patients' backlog. According to Hafez *et al.* [7], the waiting list has reached 7.2 million persons in 2022 and has been increasing ever since. The waiting list has been observed since 2010, however, it has started to be monitored as a risk since the pandemic (see Figure 1) According to Hafez *et al.* [7], the waiting time reached sixty-two days for cancer patients while it reached eighteen weeks for other patients. This waiting time increased exponentially since that time.

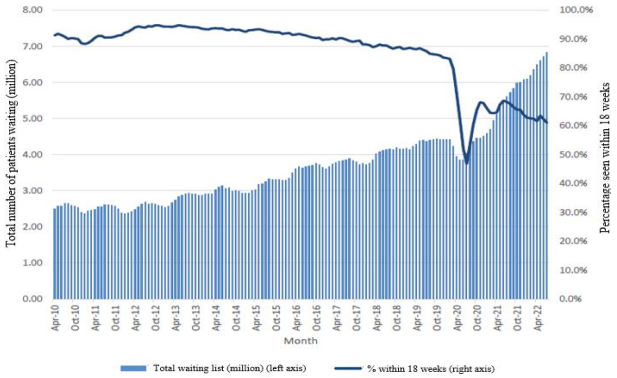

Figure 1. Waiting list distribution in eighteen weeks.

This research aims for efficient management of the waiting list targeting minimizing the backlog time and patients' waiting list. The research applies the proposed CSO enhancement by discussing a complete experiment over different patients' datasets. The applied experiment proves the proposed approach's applicability, efficiency, and reliability in reaching the required business target. The main business contributions can be summarized in the following points:

1) Efficient ordered prediction of the patients' delayed schedule by applying the adapted optimizing algorithm. The study aims to apply the proposed adapted algorithm to confirm its applicability in the business target during the applied experiment and discuss the results.
2) The study also aims to highlight the main significant factors that affect backlog management. Building a recommendation model will be performed aiming to provide efficient recommendations targeting efficient time fulfillment and management.
3) The proposed approach is applied to different datasets to confirm its applicability for any environment.

The organization of the study presents the literature in the following section, the algorithm enhancement is presented in detail in section 3, then the details of the experiment stages are discussed in section 4, and finally, the conclusion and the recommended additional research are presented in the final section.

2. Related Work

CSO owns unique advancements over other optimization algorithms. One of these advancements is its division of the search space into a set of players with different powers in which the less powerful player follows the higher power. This distribution leads to better utilization of the problem population which provides a good balance between exploration on the one

hand and exploitation on the other hand. To summarize, an efficient algorithm should be able to apply an efficient global and local search strategy with the least possible calculations and higher speed in reaching the solution. Moreover, the algorithm should be able for efficient convergence with high accuracy for all dimensions' sizes.

According to the literature, CSO suffers from some bottlenecks. One of these bottlenecks is the possibility of reaching the local optima as a result of poor tuning to the algorithm's main parameters. As these parameters are dependent, then if one of these parameters such as the roosters' parameter for example, other dependent parameters also follow the same direction. Moreover, CSO also suffers from low convergence speed. Some enhancements of CSO have been earlier proposed, this section discusses some of the recent enhancements. In Zhang *et al*. [23], the formula for searching roosters and chicken is updated to apply an inertia weighting formula which was proposed in the research to avoid the local optima. Another enhancement was proposed Idrees and Alsherif [12] which proposed a united algorithm between CSO and TLO optimization algorithms for the same target. Moreover, a formula update was introduced in Idrees *et al*. [10] using a load balance method targeting minimizing the operating time and cost of the algorithm. The balancing perspective was also tackled in Osamy *et al*. [19] for the same target by engaging the population aggregation degree in the balancing model. The chaos theory was utilized in Qaffas *et al*. [20] to avoid the local optima by improving the optimization accuracy in the wind power problem. A convergence accuracy problem solution was introduced in Liang *et al*. [15] targeting the prediction of the correct robot path. The low accuracy issue was highlighted in Hassouna *et al.* [9] by proposing the mutation strategy for evaluating the individuals. His accuracy was also highlighted in Wang *et al*. [21] by integrating CSO with the hunting optimization algorithm.

Focusing on utilizing CSO in the feature selection problem, the research in Wang *et al*. [22], utilized CSO targeting to minimize the features set and succeeded in minimizing the feature by the range of 45% to 50% of the features. Although the research Khedr *et al*. [13] proposed an enhancement for CSO for feature selection and presented more accurate results compared with other optimization algorithms, however, the algorithm only considered each feature individually and neglected the feature relations. The researchers Abdelgwad *et al*. [1] utilized the original CSO for feature selection, however, the model suffered from performance degradation for the dimensions' dataset scale. According to the literature, the field of enhancing CSO is still open for innovations. This research proposes a novel enhancement for CSO which provides an advancement in feature selection. The proposed adapted algorithm has been developed and applied to patients' datasets targeting to eliminate the patients' backlog which is a major challenge in the healthcare field.

3. Chicken Swarm Algorithm Description

The chicken swarm algorithm is based on dividing the whole space into three players, they are roosters, hens, and chickens. Each player has its own role. The roosters are the senior players who are responsible for food, hens are the second-level players who follow the roosters, and finally, the chickens follow their mothers. The rules for searching could be stated as follows:

1. The whole flock representing the search space is divided into sets. Each set has a senior rooster player, a set of hens, and a set of chickens.
2. The players' distribution is based on the fitness value. The players with the highest fitness value are elected as the senior players, as the fitness value gets lower, then the players are elected as members in the chicken set while the players with the lowest fitness value are the chickens.
3. During foraging, the senior players (roosters) select the following chickens randomly and consequently the hens are also selected randomly.
4. The dominant relationship between the roosters and their chicken remains with no change for a set of generations according to the fitness value. The parent-child relationship also follows the same rule.

Initialize *x*, *y*, and *z* as the sets of the players, roosters, chickens, and hens respectively.

Given *N* as the size of the whole flock, $f_{i,j}(m)$ is the position of the *i* member during the m iteration in the *j* search space.

Equation (1) which updates the rooster position in the workspace, is as follows:

$$fi,j\ (m+1)\ =\ fi,j\ (m)\ *\ nd(0,\sigma) \tag{1}$$

σ2=*ex*(*fr*-*fn*)/(|*fn*|+*c*) if *fn*>=*fr* and 1 otherwise, where: *fr* is the fitness value of one of the roosters, *fn* is the fitness value of any member, and *c* is a constant.

Equation (2) which updates the hen position in the workspace, is as follows:

$$\begin{aligned} fi,j(m+1)\ =\ & fi,j\ (m)\ +\ c1\ *\ var\ *\ (fr1,j\ (m) \\ & -\ fi,j\ (m))\ +\ c2\ *\ var\ *\ (fr2,j\ (m) \\ & -\ fi,j\ (m)) \end{aligned} \tag{2}$$

Where *var* is a random number, and *fr*1,*j*(*m*), *fr*2,*j*(*m*) are the fitness of two random members in the flock. *c*1 is defined in Equation (3) and *c*2 is define in Equation (4).

$$c1\ =\ ex((fr1-fi)/|fi|\ +\ c)) \tag{3}$$

$$c2\ =\ ex((fr2-fi) \tag{4}$$

Equation (5) updates the chicken's position in the workspace is as follows:

$$fi,j\ (m+1)=\ fi,j\ (m)+\ c3\ *\ var\ *\ (fm,j\ (m)-\ fi,j\ (m)) \tag{5}$$

Where *c*3 < = 2 and m is the position of the parent.

4. Leveraging Feature Selection CSO Algorithm (LCSO)

According to Abdelgwad *et al*. [1], CSO is defined to be one of the effective optimization algorithms for the task of feature selection. As discussed in the previous section, the first rule is dividing the chicken swarm into a set of groups. The suitable number of groups is one of the vital parameters that play a significant role in reaching the optimized solution with high accuracy and efficient time management. The proposed adaptation aims to perform with high performance under the supply chain constraints such as the exponential growth in both contributors and time.

4.1. Contributing Parameters and Constraints

The setting parameters rely on the problem at hand. The number of roosters is determined to be equal to the number of clusters with the condition of having the maximum fitness value of all cluster members, while the number of chickens is set to be the number of members with a fitness value above the threshold, and the remaining members are the hens. The number of dimensions is determined as the number of significant features that contribute to exploring the players. Constraints rely on identifying the contributing attributes to be more than one with identifying the significance threshold.

4.2. Adaptive Clustering for Data Distribution Representing Chicken Groups

This stage develops the required chicken swarm clustering using the proposed k-means algorithm adaptation. The proposed adaptation tackles the two main issues of k-means, configuring the optimal clusters' count that should be pre-identified and accurate configuration for the initialization clustering point. The following sub-sections explain the proposed adaptation in detail while Table 1 illustrates the set of notations used in the provided explanation.

Table 1. Notations summary.

Notation	Description
CD	Cluster density
H^k	Hypersphere radius dimension
cm	Contributing measure
TD	Training dataset
S_j	Roosters set
Θ	Represents the search space
wm	Weighting measure
f	feature
φ	deviation

4.2.1. Identify The Optimal K Clusters

Identifying the optimal clusters' count follows the hypersphere density-based algorithm [8]. The algorithm is planned to run n times; each time the clusters are set to be identified as an integer ranging from 1 to n. In each run, the silhouette score is observed. The cluster density is then calculated following Equation, where Γ represents the Leonhard gamma integral, K represents the dimensions and H represents the hypersphere radius. R is calculated as the largest distance between the cluster centroid and any of the cluster elements. The clusters' density CD is then calculated following Equation (2), where x is the number of the cluster elements and i is the cluster index. The mean density is represented and the recommended optimal count for the clusters is determined by the elbow region in Equation (6).

$$CD = \frac{\pi^{\frac{k}{2}}}{\Gamma(\frac{k}{2}+1)} H^k \tag{6}$$

Where CD is a single cluster density, Γ represents the Leonhard gamma integral, K represents the dimensions and H represents the hypersphere radius. The density is represented in Equation (7).

$$Density = \frac{CD_i}{x} \tag{7}$$

Where x is the number of the cluster elements and i is the cluster index.

Optimizing the number of iterations is one of the main challenges that are tackled in this research. Identifying the iteration number parameter is considered a key aspect of this phase.

4.2.2. Explore the Clustering Initialization Based on Hybrid-Measurements

Exploring the initialization point follows the hybrid measurement approach. The set of contributing measures is identified and calculated based on the derived eigenvalue for each measure. The set of contributing measures is identified as a set of measures in Equation (8).

$$Measures = \{cm1, cm2, \dots cmc | \, c \in \mathbb{N}\} \tag{8}$$

Where cm is the contributing measure.

The initialization point is identified given the training dataset, TDset (i), and the attributes dataset AttSet (i), then the set of contributing measures values in defined in Equation (9).

$$CMVal = \{cmval1, cmval2, \dots, cmvalc\} \; where \; c \in \mathbb{N}, cmvalc \in \mathbb{N}. \tag{9}$$

Calculating the initialization point follows Equation (10) where z_v is a member in the roosters set S_j set and $|S_j|$ is the total count of S_j members in Equation (10).

$$CP_i = \frac{1}{|S_j|} \sum_{z_v \in S_j} z_v \tag{10}$$

Where z_v is a member in the roosters set S_j set and $|S_j|$ is the total count of S_j members.

The next step is to identify the centroids of the clusters whose count is previously set. The centroids are set to be the points surrounding the initialization point having equal distance and equal sine angles in a sphere perception with

varying directions for sphere angles. The distance is identified to be the farthest distance between the initialization and the point elected to be a centroid with respect to the identified angle while the angles' values are determined according to the identified number of clusters. The distance between the initialization and each data point is determined following Equation (11).

$$\varphi(p_x, ct_y) = \sum_{z=1}^{n_t} (w_z^t \, (d_{xz}^t - c_{yz}^t))^2 + \sum_{z=1}^{n_c} (\Omega \, (d_{xz}^c - c_{yz}^c))^2 \quad (11)$$

This equation calculates the deviation between the p_x and ct_y which are the data point and initialization respectively.

The data points that are elected as centroids are determined by φ calculating the initial centroids starts with identifying the sphere latitude from 90° south to 90° north representing the radians from -π/2 to π/2 and the longitude from 180° south to 180° north representing the radians from -π to π.

On the other hand, determining the training set that contributes to the initialization phase could be described starting with identifying the space as Θ. Θ is visualized as the possibilities of the attributes' values. The count of members in Θ is then determined to be multiplying all possible values of the members of the attributes' set. Mathematically speaking.

Consider Θ includes a set of y attributes. The attributes A has a count of p possible values and attribute B has a count of q possible values, $|\Theta| = p \times q$. Θ is described in Equation (12). to be the Cartesian product of all the possible values of all the contributing attributes.

$$\begin{gathered} |\Theta| = \times_{att_p} |\mathrm{Val(p)}| \in \Theta \\ \forall \theta \quad \theta, dset\,(\theta) \subseteq DSet \\ |dseti| = |\,dset\,(\theta)| * 8/10 \; and \; i \;\; \{1,2,3,\ldots\, 2j\} \\ dseti \subseteq dset\,(\theta) \end{gathered} \quad (12)$$

In each iteration i, the training set is determined as $\bigcup_{k=1}^{2^j} dset_k$ Where each element t contributes to only n-1 iterations in the dset.

4.3. Identifying Search Space Players

In this phase, the clusters of players represented in features are determined. Each cluster is considered an individual search space. The players are identified according to the features' weights which follows the strategy in [16]. The collaborative approach is extended to this phase in which each feature is labeled as strong, weak, or not significant. The rooster and chickens belong to the strong significant features as a first priority, and weak significance as a second priority, while the chicks are the weak significant, moreover, insignificant features are eliminated. Identifying the feature labels is performed by the following steps.

The features set *fset* of a cluster i and the weighting measures set are represented in Equation (7); while the weighting of each feature in a defined cluster is identified in Equations (8), (9), and (10) as the set of the feature j weights in cluster i by the contributing weighting measures *wm*. Identifying the significance according to the weighting threshold.

$$F(ci) = \{f1i, f2i, f3i, \ldots, fxi \mid x \in \mathbb{N}\} \quad (13)$$

Where F is the set of features and f_{xi} is the feature x for cluster i.

$$WM = \{wm1, wm2, wm3, \ldots, wmy\} \quad (14)$$

Where *WM* is the set of weighting methods and *WMy* is the weighting method y.

$$WM\,(fj\,(ci)) = \{wmji1, wmji2, wmji3, \ldots, wmjiy \mid y \in \mathbb{N}\} \quad (15)$$

Where *WM* (*fj* (*ci*)) is the value set after determining the weights for the feature *fj* with respect to the cluster ci by applying the weighting methods that are members in the weighting measures set *WM* Equation (16).

$$W(fj) = \bigcup_{k=1}^{x} f\,(WM) \quad (16)$$

Where $WM\,(f_j\,(c_i))$ is the value set after determining the weights for the feature f_j for all sets of clusters by applying the weighting methods that are members in the weighting measures set *WM* Equation (17).

$$W(fij) = \bigcup_{k=1}^{i} \bigcup_{k=1}^{x} f\,(WM) \quad (17)$$

The features are considered significant by the following steps Equation set (18).

$$\begin{gathered} \forall fj \in F, \;\; \exists wmji \in fj\,(WM) \rightarrow S\_fj \\ = count\,(fj\,(WM)) \mid wmjy > WMTh) \\ \text{If S\_fj} = |\,\mathrm{fx\,(WM)}| \rightarrow \mathrm{SF} = \mathrm{SF} \cup \mathrm{fj} \\ Else\;If\;S\_fj > |\,fj\,(WM)| * 60/100 \rightarrow WF = WF \cup fj \\ Else\;RF = RF \cup fj \end{gathered} \quad (18)$$

4.4. Initialize the Chicken Swarm

In this step, exploring the main players is performed. This step relies on the determined weights of the earlier step. The feature is a member in either one of three sets, it is either nominated, neutral, or rejected. The feature is nominated if it is considered one of the strongly significant features, it is neutral if it is a weakly significant feature, and rejected otherwise. The rejected features are considered members of the chicks' group with no further investigation. This decision proved to minimize the algorithm cost effectively as will be demonstrated in the next sections. Following the backward approach, the contributing features in this step are arranged in an ascending order based on their average gained weight. The ordered set is utilized by eliminating one feature each round and the data are examined by applying classification task as will be discussed in the next sections. The evaluation results determine whether the feature should be excluded from

the final set. In this step, the contributing features are represented in Equation (19).

$$\begin{aligned} &Temp\_F = \{F1, F2, \dots, Fg\} \; where \; W(fx) > W(Fy) \\ &Ex\_F = \emptyset, In\_F = \emptyset \\ &Contributing \; set \; of \; features = ft : ft \in Temp_F, \\ &ft \notin Ex_F, ft \notin In_F, avg_{Weight}(ft) > avg_{Weight}(fu), \\ &fu \in Strong\_Features \\ &If \; success \; (ft) \; In\_F = In\_F + ft \\ &Else \; Ex\_F = Ex\_F + ft \end{aligned} \tag{19}$$

Where F is the feature, $W(f)$ is the weight.

4.4.1. Determine Fitness Value

In this step, the selection assessment is performed according to the evaluation of the classification task based on the final set of features. The fitness value is determined following a novel approach which is based on the collaboration of a set of evaluation measures in evaluating a set of classification algorithms. This approach is utilized to ensure avoiding the bias of using a single method of classification while ensuring in-depth evaluation. In order to accept the results of the feature selection method, the fitness value should be higher than eighty percent of each algorithm individually and the overall average as well. In other words, the selected set of features should provide at least eighty percent positive performance for at least eight out of ten contributing algorithms.

In this step, determining the contributing machine learning algorithms and the evaluation parameters are performed. Two main sets are identified in Equation (20) set:

$$\begin{aligned} Fit\_S &= \{fit_g, \dots, fit_t, \dots, fit_u\} \\ ML &= \{ ML_1, ML_2, \dots ML_i \mid i \in \mathbb{N}, i>0\} \\ Ev\_P &= \{Ev\_P_1, Ev\_P_2, \dots Ev\_P_j \mid j \in \mathbb{N}, j>0\} \end{aligned} \tag{20}$$

4.4.1.1. Evaluation Significance Level Exploration

Exploring the evaluation significance permits a clear vision for completing the task efficiently. For example, the accuracy identifies the level of success for the algorithm to be able to identify the correct class for the data item. On the other hand, the sensitivity highlights the success of the algorithm in identifying the positive clusters' members. Looking from another perspective, these measures should not be monitored as perfect. It is a fact that the measures have their own error margin. For example, if we considered that the sensitivity has 70% accuracy in its task, then three out of ten members could be calculated as correctly classified in the positive cluster while in fact, they do not belong to this cluster. So, if the sensitivity is determined to be 95%, then it is actually 72%. This perspective provides more clearer perspective for the algorithm's performance and could lead to more accurate decisions in this stage. Accordingly, raising the flag for the necessity of multiple contributing evaluation parameters has been performed. Following the approach that is discussed in [17] solves the situation of exploring the most effective measure to consider. In this research, the invariance level is set to be the main determinant for the measures which identifies the level of accuracy for these measures. According to Mourad [17], the higher level of invariance provides higher measuring accuracy. Eight invariance parameters contribute to this step which could be represented as follows:

- *Invariance measure* 1 the measure is invariant if the replacement of the TP and TN has no effect on the results.
- *Invariance measure* 2 the measure is invariant if the change of the TN has no effect on the results.
- *Invariance measure* 3 the measure is invariant if the change of the TP has no effect on the results.
- *Invariance measure* 4 the measure is invariant if the change of the FN has no effect on the results
- *Invariance measure* 5 the measure is invariant if the change of the FP has no effect on the results.
- *Invariance measure* 6 the measure is invariant if the replacement of the positive parameters and negative parameters has no effect on the results.
- *Invariance measure* 7 the measure is invariant if the vertical replacement of the positive parameters and negative parameters has no effect on the results.
- *Invariance measure* 8 the measure is invariant if the horizontal replacement of the positive parameters and negative parameters has no effect on the results.

4.4.1.2. Determination of Fitness Value by Machine Learning' Cross Validation

Determining the fitness value is performed by the contribution of a set of machine learning algorithms. Each algorithm is examined against a set of adapted evaluation measures. The minimum evaluation value is considered the fitness value to ensure maximum performance. Only the feature with the evaluation higher than the threshold is considered in the final features set. Equation (21) set represent the process formally.

$$\begin{aligned} &ML\_Eval \; (ML\_Algc) = \{FEvalc1, \dots., FEvalcs\} \\ &Where \; c, s \in N, c, s > 0, ML\_Algc \in ML, FEvalcs \in Eval\_P \\ &Avg\_FEval \; (ML\_Algc) = \sum_{d=c1}^{cs} FEval_{cs} \\ &EVAL\_ML \\ &= \{ Avg\_FEval \; (ML\_Alg1), \dots Avg\_FEval \; (ML\_Algc) \} \\ &Fitness\_Value \; (Features) = Min \; (EVAL\_ML) \end{aligned} \tag{21}$$

4.4.2. Determine Training Dataset

Determining the training dataset is one of the vital parameters in successfully exploring the fitness value. The research aims to follow the five-fold strategy, which means that the training dataset will represent eighty percent of the whole population. The following strategy is followed targeting to maintain a balanced dataset distribution. An iterative process is performed which divides each contributing feature into two clusters with

a determinant of the mean value of the feature. This process is performed for all contributing features, then the final subset is the integrated subset. Mathematically representing the process, considering the contributing features count is x, then there are $2x$ subsets. The training dataset is represented to be eighty percent of this integrated set. This process is inspired by the decision tree algorithm which proved to provide a balance in its branches until reaching the leaves with the minimum processing effort [6]. However, the current research succeeds in avoiding the main decision tree limitation as it does not suffer from the instability in the training dataset and it identifies the mean value as a milestone for identifying the training dataset. Following this milestone, the need to update the training dataset is minimized unless there is a major change in the dataset.

Identifying the environment space as Θ represents the possible key values of all contributing attributes. Consequently, the cardinality of Θ equal to multiplying the values of the members of the set of all attributes' values as follows:

Given that a, b are attribute members in Θ and v, w are the attributes' values of a and b respectively. Then $|\Theta|$ equal to $v \times w$. As a general representation, Θ is defined as the Cartesian product of the values of all members residing in the attributes' set.

$$|\theta| = \times_{att_j} |att\_Vals_j| \in \theta \quad (22)$$

$\forall \theta, \theta \in \Theta, \exists s$, s is subset (θ) $\subseteq$ population Fragment$_i$ $\subseteq$ subset (θ), |Fragment$_i$|=|subset (θ)|*8/10 and $i \in \{1,\ldots 2^j\}$

The contributing dataset representing the training data of an iteration is represented in Equation (23):

$$Tr\_set\ (t) = \cup_{i=1}^{2^j} Subset_i\ ,\ Sub_i = \{f | f \in population\}, \quad (23)$$

Occurrences ϵ representing an element in *Tr_set* is less than or equal 4 which represents that, for a total of five iterations, the element f will contribute in at most four iterations in the training phase.

5. Experimental Study

The contributing data in the experiment is a public data which resides in Kaggle website [18]. The dataset includes one thousand patients' records and is characterized by twenty-three attributes that describe the lifestyle of lung cancer patients. Al attributes are normalized to be of numerical type. One of the attributes represents the cancer severity level. The attributes are divided into five categories. The first category is the main data including the patient's age and gender. The second category represents the surrounding environment including the level of pollution in the air and the hazards of the workplace. The third category represents the main medical status of the patient including the allergic level to dust and the possibility level of having the disease on a genetic basis. The fourth category represents the lifestyle of the patient including if the patient is an alcoholic, current or previous smoker, the diet style, and obesity level. The fifth category represents the medical case of the patient including if there is a chronic lung disease if the patient feels continuous fatigue, has a continuous cough with blood or the cough is dry, loses weight regularly, has difficulty in swallowing, feels pain in chest and hard to breeze, snoring, continuously feeling cold and nails are clubbing. The disease severity has three levels, low, medium, and high. The dataset is relatively balanced as three hundred and two records follow the low severity level, three hundred and thirty-three follows the low severity level, and thirty thousand and sixty-five records follow the low severity.

5.1. Dataset Preparation

The data [18] is observed to have no missing values. However, the age attribute required to be transformed into a discrete attribute rather than its current type as a continuous attribute. The minimum age is seventeen while the maximum age is seventy. In order to minimize data generalization, the age is transformed into ranges of five years. Consequently, the age attribute has been transformed into a discrete value attribute of eleven values. Statistically based, Table 2 provides a brief description of the contributing attributes. Moreover, Tables 3 and 4 present the dataset distribution by grouping the data by age and gender respectively.

Table 2. Contributing attributes description.

Attribute	Description
Age	Patient age
Gender	Patient gender
Air pollution	Level of air pollution exposure
Alcohol use	Patient use of alcohol
Dust allergy	Level of allergy to dust
Occupational hazards	Patient occupational hazard
Genetic risk	The level of genetic risk
Chronic lung disease	Level of lung disease
Balanced diet	Level of the patient diet balance
Obesity	Whether the patient is obese
Smoking	Whether the patient is smoking
Passive smoker	Whether the patient was smoking
Chest pain	Whether the patient has pain in chest
Coughing of blood	Whether the patient has a cough with blood
Fatigue	Whether the patient feels tired most of the time
Weight loss	Whether the patient is losing weight
Shortness of breath	Whether the patient usually has shortness of breath
Wheezing	Whether the patient has wheezing
Swallowing difficulty	Whether the patient has swallowing Difficulty
Clubbing of finger nails	Whether the patient has
Frequent cold	Whether the patient has frequently caught cold
Dry cough	Whether the patient has a dry cough
Snoring	Whether the patient is snoring
Level	The level of cancer

Table 3. Dataset distribution grouped by age.

Age category	Age range	Count of records
1	17-27	67
2	28-38	353
3	39-49	446
4	50-60	114
5	61-70	20

Table 4. Dataset distribution grouped by gender.

Gender	Count of records
1- Male	598
2-Female	402

The second task is providing a primary understanding of the dataset. This is accomplished by visualization. Some examples are demonstrated in Figures 2 to 4. The sample demonstrates an attribute for each category. Figure 2 demonstrates the age distribution which highlights the normal distribution, while the gender distribution demonstrates the larger set of men patients. Moreover, one-third of the dataset records have high risk due to genetic properties.

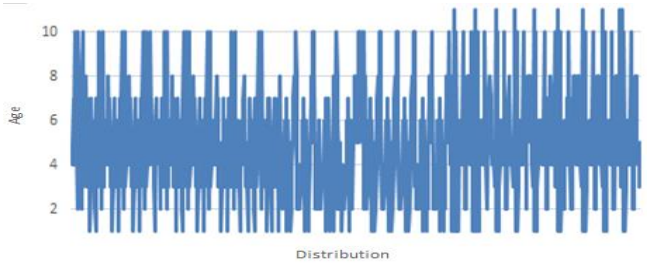

Figure 2. Age attribute.

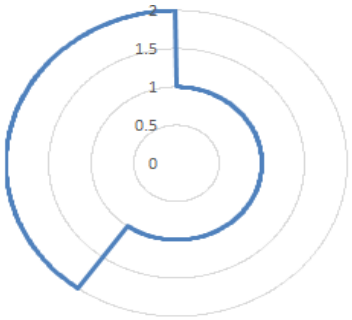

Figure 3. Gender attribute distribution.

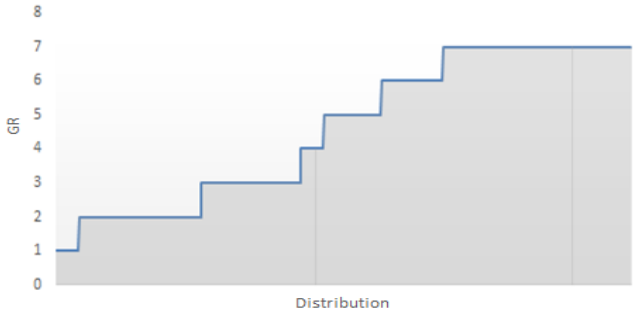

Figure 4. Genetic risk attribute.

5.2. Applying the Proposed Collaborative Based Chicken Swarm Optimization Algorithm for Feature Selection Stage (CCSO)

Following the proposed adapted algorithm, the following represents the details of applying an iteration that clarifies how the chicken groups are initiated and how the roosters and the chicken election are performed. The first step is applying the clustering algorithm in order to distribute the dataset into a set of clusters, each cluster will be considered as one of the chicken groups. As previously discussed in the proposed algorithm, the adopted k-means algorithm is applied.

5.2.1. Apply Adopted K-Means for Data Distribution Representing Chicken Groups

In this stage, the chicken groups are performed. The proposed adopted k-means is applied, and the following sections demonstrate how the two challenges of the k-means are tackled.

5.2.1.1. Identify the Optimal Clusters Count

The first step is identifying the optimal clusters count. In this step, the silhouette distribution index is applied to the dataset with an index from two to thirteen. The experiment halted at the index equal thirteen as it is observed that the evaluation measures moved towards the almost same value starting at index eight. According to the applied folds, the most suitable clusters number was five. Therefore, it is decided to set the required number of clusters to five. Figure 5 demonstrates the Silhouette index elbow for the applied folds starting from index two to thirteen. The elbow shows that the most suitable number of clusters is five. This decision solved the first challenge in k-means. Then the experiment moved to solve the second challenge which is identifying the initial point.

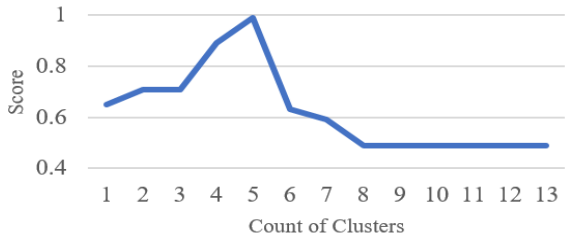

Figure 5. Silhouette index score elbow.

To confirm the accuracy in identifying the number of clusters, the results of the clustering are evaluated using the Accuracy clustering AC Equation (24), average accuracy clustering ($\overline{AC}$) Equation (25),, and Standard Deviation (SD) Equation (26) measures. The three measures are determined by Equations (24), (25), and (26) respectively. The results reveal that distributing the data into five clusters reached the highest accuracy (see Table 5).

$$AC = \frac{\sum_{k=1}^{q} dp_k}{c} \quad (24)$$

Where dp_k is the total count of the data points that joined their correct class in the total of c classes.

$$\overline{AC} = \frac{\sum_{i=1}^{u} AC_i}{u} \quad (25)$$

Where U is the total count of the rounds and AC represents the percentage of accuracy in correctly clustering the data points in the i^{th} round.

$$SD_{AC} = \sqrt{\frac{\sum_{i=1}^{T}(AC_i - \overline{AC})^2}{T}} \quad (26)$$

Table 5. Evaluation results.

No of clusters	$\overline{AC}$	SD_{AC}
2	0.69	0.346
3	0.73	0.084
4	0.73	0.049
5	0.97	0.003
6	0.82	0.014
7	0.63	0.049
8	0.59	0.077
9	0.61	0.063
10	0.53	0.261
11	0.53	0.190
12	0.53	0.120
13	0.53	0.155

5.2.1.2. Apply the Multi-Model Measuring Approach for Identifying the Optimal Initial Starting Point

The number of clusters that are previously determined plays a vital role in the centroids' initiation. The five centroids are determined as described earlier. The following steps are applied to determine the centroids of each attribute:

1. Determine the minimum and maximum value for the attribute.
2. Determine the attribute range AR (maximum value-minimum value).
3. Determine the value range VR for each cluster (AR/5).
4. Determine the centroids accordingly (C_1=minimum, then C_i=C_{i-1}+VR where i ranges from 2 to 5).

For example, the genetic risk attribute has minimum value equal to 1 and maximum value equal to 7. Therefore, following the five possible centroids, they are {1, 2.5, 4, 5.5, 7}. The approach is determined for each attribute, then the values are merged. Table 2 presents the initial centroids of the five clusters. After identifying the five initial centroids, then k-means algorithm is applied. The results of applying k-means have been evaluated to ensure the enhancement applicability and move to the next step (see Table 6). Moreover, Figure 6 presents the attribute distribution for the five clusters to confirm the uniform distribution for the centroids. The order of the attributes conforms with the order in Table 6.

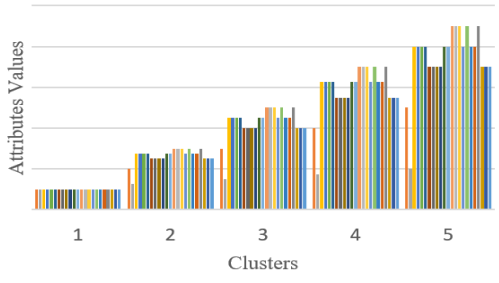

Figure 6. Evaluation of adopted k-means for the first iteration.

Following the proposed adaptation of k-means to two different datasets for confirmation, the results of the original k-means and the proposed enhancement are presented in Table 7.

Table 6. Initial centroids.

1	Age	1	2	3	4	5
2	Gender	1	1.25	1.5	1.75	2
3	Air pollution	1	2.75	4.5	6.25	8
4	Alcohol use	1	2.75	4.5	6.25	8
5	Dust allergy	1	2.75	4.5	6.25	8
6	Occupational hazards	1	2.75	4.5	6.25	8
7	Genetic risk	1	2.5	4	5.5	7
8	chronic lung disease	1	2.5	4	5.5	7
9	Balanced diet	1	2.5	4	5.5	7
10	Obesity	1	2.5	4	5.5	7
11	Smoking	1	2.75	4.5	6.25	8
12	Passive smoker	1	2.75	4.5	6.25	8
13	Chest pain	1	3	5	7	9
14	Coughing of blood	1	3	5	7	9
15	Fatigue	1	3	5	7	9
16	Weight loss	1	2.75	4.5	6.25	8
17	Shortness of breath	1	3	5	7	9
18	Wheezing	1	2.75	4.5	6.25	8
19	Swallowing difficulty	1	2.75	4.5	6.25	8
20	Clubbing of Finger Nails	1	3	5	7	9
21	Frequent Cold	1	2.5	4	5.5	7
22	Dry Cough	1	2.5	4	5.5	7
23	Snoring	1	2.5	4	5.5	7

Table 7 illustrates the values of three evaluation measures, they are Avg $\overline{AC}$ (average accuracy clustering for all clustering iterations), Avg. SD_{AC} (average standard deviation for all clustering iterations), and execution time for each algorithm (k-means and the proposed adaptation). The results highlight the advancement that the enhancement performed over the original algorithm by raising the accuracy percentage and reducing the execution time.

Table 7. Evaluation comparison.

Algorithm	Avg$\overline{AC}$.	Avg. SD_{AC}	Execution time
k-means	0.82	0.397	500
Proposed Adaptation	0.97	0.003	300

5.2.2. Identifying Search Space Players

For the whole flock, the search for the rooster and the two chickens are identified in each cluster by identifying the significant attributes. The roaster is identified to be the highest significant attributes which are defined to be significant are elected as roasters. While the chickens are weak significant attributes with having a weight less than the threshold by at least one of the weighting measures with no more than 50% of the measures. The attributes below the threshold by more than 50% of the weighting measures are defined to be insignificant and elected as the hens which will not be included in the final features set. These steps are iteratively applied until there is no change in the features' subsets. The weighting measures are determined and applied to the dataset as demonstrated in Table 8. The threshold has been defined to be 50%. The results of one iteration are illustrated in Table 9.

Table 8. Results of the contributing weighting methods.

Feature	Information gain	Information gain ratio	Correlation	Chai square	Deviation	l
1	0.89	0.75	0.46	0.81	0.96	0.47
2	1.30	0.58	0.47	0.76	1.01	0.55
3	1.10	0.59	0.52	1.00	0.93	0.49
4	0.39	0.56	0.47	0.88	0.39	0.33
5	0.79	0.46	0.53	0.69	0.79	0.35
6	1.20	0.59	0.57	1.10	0.92	0.57
7	0.99	0.54	0.53	0.41	1.24	0.52
8	1.00	0.50	0.58	0.99	1.25	0.40
9	0.94	0.45	0.23	0.00	0.94	0.46
10	0.98	0.51	0.27	0.67	1.12	0.51
11	1.10	0.67	0.54	0.94	0.69	0.53
12	0.94	0.65	0.79	0.00	0.78	0.52
13	0.98	0.61	0.76	1.00	0.87	0.68
14	1.12	0.59	0.53	0.00	1.01	0.63
15	0.45	0.36	0.37	0.00	0.71	0.10
16	1.00	0.57	0.54	0.57	0.90	0.50
17	0.89	0.66	0.68	0.51	0.89	0.47
18	0.96	0.63	0.62	0.39	0.82	0.39
19	0.84	0.56	0.61	0.20	0.75	0.28
20	1.10	0.68	0.62	1.00	0.74	0.41
21	1.20	0.59	0.57	1.10	0.92	0.57
22	0.94	0.65	0.79	0.00	0.78	0.52
23	1.30	0.58	0.47	0.76	1.01	0.55

Table 9. Attributes' statistics.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Above Th	4	5	5	2	4	6	5	5	2	5	6	5
Below Th	2	1	1	4	2	0	1	1	4	1	0	1
Significance	W	W	W	N	W	S	W	W	N	W	S	W
Weight %	72.3	77.8	77.2	-	60.2	82.5	70.5	78.7	-	67.7	74.5	61.3
	13	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	
Above Th	6	5	1	6	5	4	4	5	6	2	5	
Below Th	0	1	5	0	1	2	2	1	0	4	1	
Significance	S	W	N	S	W	W	W	W	S	N	W	
Weight %	81.7	64.7	-	68.0	68.3	63.5	54.0	75.8	82.5	-	77.8	

The next step is determining the fitness value to confirm the feature selection. The features that are elected as chicken will contribute to this step. This step follows the greedy approach. The Iterative Dichotomiser 3 (ID3) algorithm; which is one of the well-known classification algorithms; is determined to contribute to this step. In each iteration, one of the chickens is eliminated and the results are evaluated. A feature is elected as one of the chickens when its absence affects the classification results.

Table 10. Measures descriptions.

Measure	Definition	Formula
Confusion matrix - false positive	Number of incorrect records that are predicted to belong to the positive class	fp
Confusion matrix -false negative	Number of incorrect records that are predicted to belong to the negative class	fn
Confusion matrix -true positive	Number of correct records that are predicted that are predicted to belong to the positive class	tp
Confusion matrix -true negative	Number of correct records that are predicted to belong to the negative class	tn
F1-measure	A relational balancing percentage between precision and recall	2*Precision*recall/(precision+recall)
recall (Sensitivity)	The percentage of the correctly classified records in the whole test set	tp/(tp+fn)
Precision	The percentage of the correctly classified records in a certain class	tp / (tp + fp)
Accuracy	The percentage of the correctly classified records	(tp+tn)/total dataset
Error Rate	The percentage of incorrect examples' classification	(fp+fn) / (tp+tn+fp+fn)
Youden	A relational balancing percentage between the sensitivity and the specificity	Sensitivity + specificity -1
Specificity	The percentage that the classification succeeded to determine the negative examples	tn/(fp+tn)

Table 11. The evaluation measure invariance (×denotes non-invariance, √ denotes invarianvec).

	Inv1	**Inv2**	**Inv3**	**Inv4**	**Inv5**	**Inv6**	**Inv7**	**Inv8**
F measure	N	Y	N	N	N	Y	N	N
Recall	N	Y	N	Y	N	Y	N	Y
Precision	N	Y	N	N	Y	Y	Y	N
Accuracy	Y	N	N	N	N	Y	N	N
Error Rate	Y	N	N	N	N	Y	N	N
Youden	N	N	N	N	N	Y	N	N
Specificity	N	N	Y	N	N	Y	N	Y

In this step, identifying the evaluation measures is vital for the classification algorithms evaluation. As discussed in the literature section [17], the most beneficial evaluation measures are determined. Table 10

presents the contributing evaluation measure and a brief description. The evaluation measures have been examined by the invariance parameters and the examination results are illustrated in Table 11. In binary classification, the measures are more accurate if its invariance is minimized. The number of non-invariance is computed to reflect the accuracy weight for the evaluation measure (Table 12).

Table 12. Weight of the evaluation metrics.

	DT	KNN	NB	LMT	RF	SVM	ID3
Accuracy	95.94	90.16	98.17	92.94	97.99	96.58	98.72
Precision	91.13	85.85	95.45	94.32	99.7	98.9	98.9
Recall	65.82	52.52	99.7	79.83	96.50	90.9	93.8
F_Measure	76.43	65.17	97.53	86.47	98.07	94.73	96.28
Specificity	98.7	99.9	98.73	96.9	100	100	100
Youden	48	50	96	51	98	88	96
Error Rate	4.06	9.84	1.83	7.06	2.01	3.42	1.28

The evaluation results of the contributing algorithms are determined and presented in Tables 13 and 14 before and after the invariance weighting. The best algorithm after considering the invariance is revealed to be the random forest algorithm (Figure 7).

Table 13. Classification task evaluation for contributing algorithms.

	DT	KNN	NB	LMT	RF	SVM	ID3
Accuracy	95.94	90.16	98.17	92.94	97.99	96.58	98.72
Precision	91.13	85.85	95.45	94.32	99.7	98.9	98.9
Recall	65.82	52.52	99.7	79.83	96.50	90.9	93.8
F_measure	76.43	65.17	97.53	86.47	98.07	94.73	96.28
Specificity	98.7	99.9	98.73	96.9	100	100	100
YOUDEN	48	50	96	51	98	88	96
Error Rate	4.06	9.84	1.83	7.06	2.01	3.42	1.28

Table 14. Final classification task evaluation for contributing algorithms wrt weighting measures confidence.

Measure (weight)	DT	KNN	NB	LMT	RF	SVM	ID3
Accuracy (0.75)	71.96	67.62	73.63	69.71	73.49	72.44	74.04
Precision (0.5)	45.57	42.93	47.73	47.16	49.85	49.45	49.45
Recall (0.5)	32.91	26.26	49.85	39.92	48.25	45.45	46.90
F_measure (0.75)	57.33	48.88	73.15	64.85	73.56	71.05	72.21
Specificity (0.625)	61.69	62.44	61.71	60.56	62.50	62.50	62.50
YOUDEN (0.875)	42.00	43.75	84.00	44.63	85.75	77.00	84.00
Error Rate (0.75)	3.05	7.38	1.37	5.30	1.51	2.57	0.96
Average	44.93	42.75	55.92	47.45	56.42	54.35	55.72

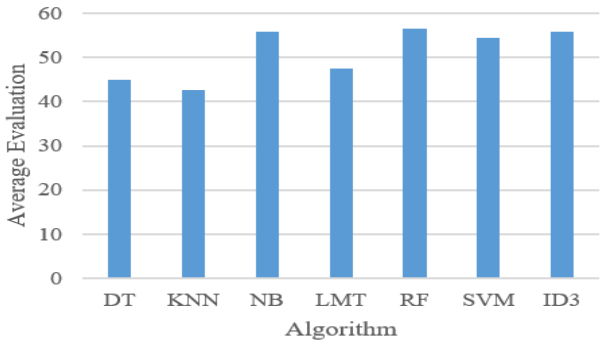

Figure 7. Avg evaluation for the contributing algorithms.

The proposed CCSO mechanism resulted in the election of eleven features as final feature set members. Moreover, further investigations for the elected features as well as the non-elected features are applied. As an example, the "balanced diet" had a p-value of 0.96 (*x*-squared=3.38, *df*=10, *p*=96%). These measures confirmed the feature's insignificance as it does not affect the logs' delay. Another example is "smoking" which had a p-value f 2.2e16 (*x*-squared=1694, *df*=3, p≈0.0%) These measures confirmed the feature significance.

6. Patients' Backlogs Prediction

The main task is the prediction of the patients' backlog. Identifying the patients' backlogs provides the highlights targeting elimination. Tables 15 and 16 illustrate the results following the same approach of determining the fitness value. According to the results, it is confirmed that CCSO has successfully led to the most effective features that could affect exploring the patients' backlogs as cleared in the results of the classification algorithms evaluation.

Table 15. Classification task evaluation for contributing algorithms.

	DT	RF	NB	LMT	KNN	SVM	ID3
Accuracy	97.3	95.3	98.9	96.5	98.7	96.1	98.1
Precision	93.4	91.1	95.3	93.3	99.9	99.1	99.1
Recall	78.5	86.1	94.9	89.8	97.9	94.1	94.5
F_measure	85.30	88.53	95.10	91.52	98.89	96.54	96.75
Specificity	97.9	98.2	99.3	99.7	99.6	99.1	99.5
YOUDEN	84	86	96	85	97	96	94
Error Rate	2.7	4.7	1.1	3.5	1.3	3.9	1.9

Table 16. Final classification task evaluation for the contributing algorithms wrt weighting measures confidence.

Measure (weight)	DT	RF	NB	LMT	KNN	SVM	ID3
Accuracy (0.75)	72.98	71.48	74.18	72.38	74.03	72.08	73.58
Precision (0.5)	46.70	45.55	47.65	46.65	49.95	49.55	49.55
Recall (0.5)	39.25	43.05	47.45	44.90	48.95	47.05	47.25
F_measure (0.75)	113.74	118.04	126.80	122.02	131.85	128.71	128.99
Specificity (0.625)	61.19	61.38	62.06	62.31	62.25	61.94	62.19
YOUDEN (0.875)	73.50	75.25	84.00	74.38	84.88	84.00	82.25
Error Rate (0.75)	2.03	3.53	0.82	2.63	0.97	2.93	1.43
Average	58.48	59.75	63.28	60.75	64.70	63.75	63.60

7. Conclusions

This research proposed a novel adaptation for the chicken swarm optimization algorithm targeting the enhancement of the algorithm performance. As the challenge of CSO performance lies in its random strategy in the process of searching the data space, therefore, this study proposed an enhancement that was based on two main pillars. The first pillar is reducing the search space and the second is reducing the computation time. The study presented a novel model to identify the space players based on the features' contribution and then these highly influencing features move for early

examination targeting their stability with no further examination in the next stages. The fitness value of the algorithm was adapted by a collaborative approach while the evaluation included a set of evaluation measures and accuracy was also examined according to their invariance level. The research confirmed that the algorithm adaptation outperformed other machine learning techniques by the experimental results comparison.

The business aspect is also tackled in this study. Successful identification of the patients' schedules' significant factors in addition to identifying the significance polarity was accomplished. The proposed model was able to predict the bottlenecks in the patient supply chains by the early prediction of the patients' delay and eliminate backlogs. Recommendations for avoiding the backlog bottlenecks were endorsed according to the features' significance and effectiveness in the search space which confirms the efficient fulfillment on time and attain the customers' satisfaction.

The proposed framework is generic and could be applied to any type of supply chain. The experiment succeeded in confirming the applicability of the proposed adaptation for CSO and reaching the business goal with a minimum accuracy percentage equal to 95.3% for random forest and a maximum of 98.9 for naïve bayes.

Future research could focus on other swarm algorithms to confirm the proposed approach's applicability and generality. Moreover, applying the proposed adaptation to other datasets from different fields. Additionally, considering different business tasks to confirm the effectiveness of the proposed adaptation in other tasks. Moreover, including other diversity of parameters could confirm the generality.

Acknowledgment

This work was funded by the University of Jeddah, Jeddah, Saudi Arabia, under grant no. (UJ-23-DR-95). Therefore, the authors thank the University of Jeddah for its technical and financial support

References

[1] Abdelgwad M., Abed A., and Bahloul M., "Authenticated Diagnosing of COVID-19 Using Deep Learning-Based CT Image Encryption Approach," *Future Computing and Informatics Journal*, vol. 8, no. 2, pp. 31-58, 2022. https://digitalcommons.aaru.edu.jo/fcij/vol8/iss2/4

[2] Abogabal F., Ouf S., and Idrees A., "Proposed Framework for Applying Data Mining Techniques to Detect Key Performance Indicators for Food Deterioration," *Future Computing and Informatics Journal*, vol. 7, no. 2, pp. 33, 2022. DOI:10.54623/fue.fcij.7.2.4

[3] Aggarwal A., Han L., Sullivan R., Haire K., Sangar V., and Der Meulen J., "Managing the Cancer Backlog: A National Population-Based Study of Patient Mobility, Waiting Times and 'Spare Capacity' for Cancer Surgery," *The Lancet Regional Health Europe*, vol. 30, pp. 100642, 2023. DOI:10.1016/j.lanepe.2023.100642

[4] Almazroi A., Idrees A., and Khedr A., "A Proposed Customer Relationship Framework Based on Information Retrieval for Effective Firms' Competitiveness," *Expert Systems with Applications*, vol. 176, pp. 114882, 2021. https://doi.org/10.1016/j.eswa.2021.114882

[5] Bourahouat G., Abourezq M., and Daoudi N., "Word Embedding as a Semantic Feature Extraction Technique in Arabic Natural Language Processing: An Overview," *The International Arab Journal of Information Technology*, vol. 21, no. 2, pp. 313-325, 2024. https://doi.org/10.34028/iajit/21/2/13

[6] Deb S., Gao X., Tammi K., Kalita K., and Mahanta P., "A Novel Chicken Swarm and Teaching Learning Based Algorithm for Electric Vehicle Charging Station Placement Problem," *Energy*, vol. 220, pp. 119645, 2021. https://doi.org/10.1016/j.energy.2020.119645

[7] Hafez A., Zawbaa H., Emary E., Mahmoud H., and Hassanien A., "An Innovative Approach for Feature Selection Based on Chicken Swarm Optimization," *in Proceedings of the 7th International Conference of Soft Computing and Pattern Recognition*, Fukuoka, pp. 19-24, 2015. DOI:10.1109/SOCPAR.2015.7492775

[8] Han B. and Liu S., "An Improved Binary Chicken Swarm Optimization Algorithm for Solving 0-1 Knapsack Problem," *in Proceedings of the 13th International Conference on Computational Intelligence and Security*, Hong Kong, pp. 207-210, 2017. DOI:10.1109/CIS.2017.00052

[9] Hassouna D., Khedr A., Idrees A., and Elseddawy A., "Intelligent Personalized System for Enhancing the Quality of Learning," *Journal of Theoretical and Applied Information Technology*, vol. 98, no. 13, pp. 2199-2213, 2020. https://www.jatit.org/volumes/Vol98No13/1Vol98No13.pdf

[10] Idrees A., Alhusseny N., and Ouf S., "Credit Card Fraud Detection Model-Based Machine Learning Algorithms," *IAENG International Journal of Computer Science*, vol. 51, no. 1, pp. 1649-1662, 2024. https://www.iaeng.org/IJCS/issues_v51/issue_10/IJCS_51_10_22.pdf

[11] Idrees A., Khedr E., and Almazroi A., "Utilizing Data Mining Techniques for Attributes' Intra-Relationship Detection in a Higher Collaborative Environment," *International Journal of Human-Computer Interaction*, vol. 40, no. 2, pp. 190-202,

2022. https://doi.org/10.1080/10447318.2022.2112029
[12] Idrees M. and Alsherif F., "A Collaborative Evaluation Metrics Approach for Classification Algorithms," *Journal of Southwest Jiaotong University*, vol. 55, no. 1, pp. 1-14, 2020. DOI:10.35741/issn.0258-2724.55.1.1
[13] Khedr E., Alsahafi Y., and Idrees A., "A Proposed Multi-Level Predictive WKM_ID3 Algorithm, Towards Enhancing Supply Chain Management in Healthcare Field," *IEEE Access*, vol. 11, pp. 125897-125908, 2023. DOI:10.1109/ACCESS.2023.3330691
[14] Khedr E., Idrees A., and El Seddawy A., "Enhancing Iterative Dichotomiser 3 Algorithm for Classification Decision Tree," *WIREs Data Mining Knowledge Discovery*, vol. 6, pp. 70-79, 2016. https://doi.org/10.1002/widm.1177
[15] Liang X., Kou D., and Wen L., "An Improved Chicken Swarm Optimization Algorithm and its Application in Robot Path Planning," *IEEE Access*, vol. 8, pp. 9543-49550, 2020. DOI:10.1109/ACCESS.2020.2974498
[16] Mousa M., Khedr A., and Idrees A., "Hierarchical Method for Automated Text Documents Classification," *The International Arab Journal of Information Technology*, vol. 22, no. 1, pp. 11-19, 2025. https://doi.org/10.34028/iajit/22/1/2
[17] Mourad R., Cancer Dataset, https://www.kaggle.com/code/rawanmourad/cancer-dataset/notebook, Last Visited, 2024.
[18] Nanjundan S., Sankaran S., Arjun C., and Anand G., "Identifying the Number of Clusters for K-Means: A Hypersphere Density Based Approach," *arXiv Preprint*, vol. arXiv:1912.00643, pp. 1-5, 2019. https://doi.org/10.48550/arXiv.1912.00643
[19] Osamy W., El-Sawy A., and Salim A., "CSOCA: Chicken Swarm Optimization Based Clustering Algorithm for Wireless Sensor Networks," *IEEE Access*, vol. 8, pp. 60676-60688, 2020. DOI:10.1109/ACCESS.2020.2983483
[20] Qaffas A., Alharbi I., Idrees A., and Kholeif S., "A Proposed Framework for Student's Skills-Driven Personalization of Cloud-Based Course Content," *International Journal of Software Engineering and Knowledge Engineering*, vol. 33, no. 4, pp. 603-613, 2023. https://doi.org/10.1142/S0218194023500092
[21] Wang H., Chen Z., and Liu G., "An Improved Chicken Swarm Optimization Algorithm for Feature Selection," *in Proceedings of the International Conference on Wireless Communications, Networking, and Applications*, Wuhan, pp. 177-186, 2021. https://doi.org/10.1007/978-981-19-2456-9_19
[22] Wang J., Zhang F., Liu H., Ding J., and Gao C., "Interruptible Load Scheduling Model Based on an Improved Chicken Swarm Optimization Algorithm," *CSEE Journal of Power and Energy Systems*, vol. 7, no. 2, pp. 232-240, 2021. DOI:10.17775/CSEEJPES.2020.01150
[23] Zhang T., Yu M., Li B., and Liu Z., "Capacity Prediction of Lithium-Ion Batteries Based on Wavelet Noise Reduction and Support Vector Machine," *Transactions of China Electrotechnical Society*, vol. 35, no. 14, pp. 3126-3136, 2020. DOI:10.19595/j.cnki.1000-6753.tces.190620

Emna Bouazizi is currently working as an Assistant Professor in the Department of Information Systems at Khulais College. Her research interests include (Scientific) Data and Model Management, Data Science, Big Data, IoT, E-Learning, Data Mining, Bioinformatics, and Cloud Computing.

Ayman E. Khedr currently a Professor at the University of Jeddah. He have been the Vice Dean of post-graduation and research and the Head of the Information Systems Department in the Faculty of Computers and Information Technology, at Future University in Egypt. He is a Professor in the Faculty of Computers and Information, at Helwan University in Egypt. He has previously worked as the general manager of the Helwan E-Learning Center. his research is focused on the Themes (scientific) Data and Model Management, Data Science, Big Data, IoT, E-learning, Data Mining, Bioinformatics, and Cloud Computing.

Amira M. Idrees a Professor in Information Systems. She has been the Head of Scientific Departments and the Vice Dean of Community Services and Environmental Development, at the Faculty of Computers and Information, at Fayoum University. She is a Professor in the Faculty of Computers and Information Technology at Future University, she is the Head of IS Department, and the Head of the University Requirements Unit. Her research interests include Knowledge Discovery, Text Mining, Opinion Mining, Cloud Computing, E-Learning, Software Engineering, Data Science, and Data Warehousing.